UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA IN
Filologia, Letteratura italiana, Linguistica

CICLO XXX

COORDINATORE Prof.ssa Paola Manni

Proposta di un metodo di valutazione automatica della leggibilità di pagine web in lingua italiana

Settore Scientifico Disciplinare: L-FIL-LET/12

Dottoranda
Dott. Lucia Francalanci

Tutore
Prof. Marco Biffi

Coordinatore
Prof.ssa Paola Manni

Anni 2014/2019

Indice

Introduzione

Scopo del presente lavoro è la costruzione di un metodo per la valutazione della leggibilità dei testi presenti nei siti web in lingua italiana.

La nascita della ricerca moderna sulla leggibilità e lo sviluppo di strumenti per misurarla risalgono agli anni Venti del secolo scorso. Eppure la valutazione della leggibilità continua ad essere un settore di ricerca attivo e di grande interesse. In molti paesi è ormai una pratica comune fare riferimento a standard di leggibilità per la produzione di testi destinati a un vasto pubblico. Negli Stati Uniti, ad esempio, le formule di leggibilità sono ampiamente utilizzate nell'editoria scolastica, nell'industria, nelle strutture amministrative e governative.

I tradizionali indici di leggibilità sono stati applicati con successo per molti anni e in numerosi campi, grazie anche alla loro facilità di applicazione. Nonostante questo, sono molte le obiezioni che sono state rivolte a tali tecniche. Le critiche principali riguardano il fatto che le formule non tengono conto di diversi fattori che influenzano il processo di comprensione come il vocabolario impiegato, la correttezza ortografica, grammaticale e sintattica del testo, la struttura logica, l'impaginazione, la dimensione e il tipo di carattere impiegati, la presenza di tabelle, immagini, grafici o di accorgimenti volti a facilitare la decodifica, come titoli, sottotitoli, sottolineature, grassetti, ecc. Non tengono inoltre conto delle caratteristiche che riguardano il lettore, come il suo livello culturale, la sua preparazione, la sua motivazione, il suo interesse, ecc. È inoltre possibile che alcuni fattori, nonostante rappresentino dei buoni indicatori di difficoltà, vengano lasciati fuori dalle formule perché troppo complessi da misurare.

Inoltre, la maggior parte delle formule è stata creata prima della diffusione del web: essendo progettati esclusivamente per l'analisi dei testi scritti, gli indici non prendono in considerazione le caratteristiche tipiche dei contenuti web. Uno dei problemi riguarda la dimensione e la varietà del campione di testi presenti in rete: è possibili trovare testi abbastanza lunghi ma anche brani di poche parole, affiancati da immagini e video o corredati di tabelle ed elenchi. Le classiche formule di leggibilità sono state sviluppate per valutare generalmente brani o porzioni di testo di almeno 100 parole e risultano invece inattendibili nel caso di testi più brevi. Le pagine web possono inoltre presentare una struttura sintattica diversa da quella dei documenti tradizionali; l'individuazione stessa dei confini della frase diventa problematica, in quanto la presenza di numerosi collegamenti ipertestuali potrebbe confondere gli algoritmi che conteggiano le frasi.

La natura estremamente varia e non tradizionale dei contenuti web, dai commenti dei blog, ai post e ai tweet dei social, alle pagine dei risultati dei motori di ricerca fino alla pubblicità online, porta a nuove sfide per la previsione della leggibilità (Collins-Thompson 2014).

A partire dalla prima metà degli anni 2000, i ricercatori hanno dimostrato un rinnovato interesse per la leggibilità. Il desiderio di superare le limitazioni degli indici tradizionali, insieme ai progressi compiuti nel campo del *machine learning* e lo sviluppo di efficienti tecniche di *Natural Language Processing* (NLP), hanno contribuito alla nascita di nuovi approcci alla valutazione della leggibilità.

Da una parte, l'opportunità di sfruttare metodi computazionali sempre più sofisticati e una crescente disponibilità di nuove fonti di dati hanno consentito ai ricercatori di esplorare una

più ampia varietà di caratteristiche linguistiche e sperimentare variabili più complesse; dall'altra, l'uso di modelli di previsione avanzati basati sull'apprendimento automatico ha permesso di costruire nuovi strumenti e algoritmi per la misurazione della leggibilità.
Si è quindi registrato un passaggio dalle misure tradizionali a favore dei nuovi approcci alla valutazione della leggibilità; tali metodi sono rivolti alla costruzione di un modello che permetta di classificare in modo automatico un insieme di documenti testuali in base al loro livello di difficoltà.
Uno dei vantaggi di questi modelli è che sono dinamici e possono essere riadattati facilmente in base a nuovi dati e a diverse applicazioni: possono imparare ad evolversi automaticamente via via che muta il vocabolario e variano i tipi di testo da analizzare. Ciò diviene particolarmente importante nel contesto del web.
All'interno di una società dell'informazione, in cui gli utenti hanno a disposizione un'enorme quantità di dati, le attività di recupero e organizzazione dei contenuti divengono sempre più fondamentali. In quest'ottica, la valutazione automatica della leggibilità gioca un ruolo chiave, in particolar modo in quei domini applicativi in cui l'accesso alle risorse è particolarmente importante. La necessità di fornire contenuti chiari e accessibili a un gruppo di persone ampio ed eterogeneo coinvolge infatti diversi campi di applicazione. Si pensi, ad esempio, alle informazioni reperibili sui siti istituzionali, che dovrebbero essere accessibili a tutti i membri della società, a prescindere dal loro livello di istruzione, dalle loro limitazioni fisiche e cognitive o dal fatto che si tratti di persone che apprendono l'italiano come lingua straniera.

Attualmente, in Italia, non è stato ancora messo a punto un indice di leggibilità che sia specificamente calibrato sulla lingua dei siti web e gli strumenti esistenti sembrano inadeguati a valutare testi contemporanei scritti per il web. L'indice GULPEASE, oltre a considerare soltanto due variabili (lunghezza delle parole e delle frasi), presenta il problema di essere tarato sull'italiano scritto degli anni '80 e molto probabilmente non rispecchia i livelli di lettura/difficoltà attuali. Inoltre la formula è tarata soltanto su bambini e ragazzi in età scolare e mancano invece verifiche sistematiche ed estese su gruppi di adulti. Anche il più recente strumento READ-IT, che si basa su un approccio di apprendimento automatico e considera parametri linguistici sempre più complessi, è comunque rivolto allo studio della leggibilità di testi scritti; inoltre, READ-IT nasce come supporto al processo di semplificazione dei testi e pertanto si rivolge a un pubblico di destinatari specifico, cioè lettori caratterizzati da una bassa alfabetizzazione o da lieve deficit cognitivo.
La lingua del web rappresenta un caso particolare: collocata sull'asse diamesico in una posizione intermedia tra scritto e parlato, tale varietà linguistica viene spesso definita "italiano trasmesso scritto". Essa, però, non si sposta soltanto sull'asse diamesico, condividendo caratteristiche proprie sia dello scritto che del parlato, ma si muove nello spazio linguistico. "Un aspetto su cui è bene riflettere, e su cui ancora mi pare ci siano forti oscillazioni, è la vera natura del trasmesso scritto, soprattutto in relazione alla sua "dimensione": troppo spesso se ne parla come di una varietà monolitica, senza aggettivi, senza cioè individuarne le ovvie e naturali sfaccettature" (Biffi 2014). Sul web, invece, è possibile trovare tutta la gamma delle variazioni secondo i diversi assi: diacronico (sono pubblicati sia testi recenti che antichi), diatopico (l'estensione va dall'italiano standard al dialetto, con le gradazioni intermedie di dialetto italianizzato e italiano regionale),

diastratico (i testi sono rappresentativi dei vari strati sociali; tratti caratterizzanti possono essere l'età o il grado di istruzione degli scriventi), diafasico (l'estensione può andare dalla conversazione informale a testi con altissimi livelli di formalità).
Alla luce di queste considerazioni, proponiamo un metodo di valutazione della leggibilità dei contenuti presenti sui siti web in lingua italiana. La linea di ricerca che abbiamo scelto prevede l'abbandono dei metodi tradizionali di costruzione delle formule di leggibilità a favore di un approccio di valutazione automatica basata su tecniche di *machine learning*, che risulta, almeno per quanto riguarda la lingua inglese, ampiamente sperimentata per la classificazione di pagine web. La metodologia proposta consentirà lo sviluppo di un sistema di misurazione basato sul livello di lettura e comprensione della popolazione attuale. La taratura specifica per i contenuti web terrà inoltre conto della variabilità propria della lingua italiana in rete.
Tale strumento può costituire un supporto sia alla produzione che alla semplificazione dei testi in tutti quegli ambiti in cui la comprensione è cruciale per la comunicazione: i siti web degli enti istituzionali (pubblica amministrazione, enti sanitari), i siti delle testate giornalistiche, i portali relativi all'ambito educativo, i siti web delle aziende (per la promozione del marchio e dei prodotti), ecc. Si consideri, ad esempio, le applicazioni educative e le finalità didattiche: la produzione o la ricerca di materiale didattico destinato agli studenti non può infatti prescindere dal livello di lettura degli studenti. Un sistema che valuta la difficoltà del testo può essere utile sia a chi scrive libri di testo o produce materiale online, sia agli insegnanti che cercano risorse in rete da integrare alle lezioni.
Un altro campo di applicazione sono i motori di ricerca: un sistema di misurazione della leggibilità potrebbe essere incorporato direttamente nei sistemi di recupero delle informazioni per fare in modo che i risultati delle *query* possano essere personalizzati in base al livello di istruzione dell'utente. Oppure, potrebbe concentrarsi sul riconoscimento automatico dei livelli di lettura degli utenti in base ai termini da loro impiegati nelle interrogazioni sui motori di ricerca. Gli algoritmi di apprendimento automatico potrebbero perfino essere applicati alla costruzione dei profili di leggibilità degli utenti in base a diverse variabili, come il livello di difficoltà delle pagine lette di recente, le caratteristiche linguistiche delle *query* effettuate e altre caratteristiche legate alla cronologia dell'utente. "More generally, we foresee the need for topic-specific models of readability that reflect a user's expertise on specific topics but not others, in addition to their general reading proficiency" (Collins-Thompson 2014).
Considerando l'importanza della chiarezza dei contenuti nel rispondere ai bisogni informativi delle persone e l'importanza del web come mezzo per diffondere tali informazioni, "le implicazioni per lo sviluppo di un metodo efficace di valutazione automatica della leggibilità dei testi sono tanto diverse quanto lo sono gli usi del testo stesso" (Collins-Thompson 2014).

La prima parte della tesi si concentra sulla ricostruzione dello stato dell'arte della ricerca sulla misurazione della leggibilità. Nel primo capitolo si tenta di dare una definizione del concetto di leggibilità, delimitando il suo campo di applicazione e stabilendo il suo ruolo nel processo di comprensione della lettura, anche in riferimento al concetto di comprensibilità dei testi. Si analizzano inoltre quegli indirizzi di studio che hanno costituito la base per la

nascita della ricerca sulla leggibilità: gli studi sulle frequenze lessicali dei testi e le analisi di statistica linguistica.
Il secondo capitolo prende in considerazione quelli che vengono definiti "studi classici sulla leggibilità", cioè tutte quelle ricerche sulla leggibilità dei testi e sulle formule sviluppate per la lingua inglese a partire dagli anni Venti fino agli anni Sessanta.
Il terzo capitolo è dedicato ai "nuovi studi di leggibilità", che costituiscono una fase di consolidamento e approfondimento delle ricerche che arriverà fino agli anni Novanta. Questo periodo è caratterizzato dall'utilizzo di strumenti informatici, che consentono di analizzare una grande quantità di testi e considerare un maggior numero di variabili, dallo sviluppo di formule di leggibilità per lingue diverse dall'inglese e dall'introduzione della procedura cloze come criterio per lo sviluppo degli indici.
Il quarto capitolo è rivolto proprio allo studio di formule di leggibilità per lingue diverse dall'inglese e, in particolare, alle ricerche sulla leggibilità dei testi in lingue straniere negli Stati Uniti e agli indici sviluppati in Europa e nel resto del mondo.
Il quinto capitolo riguarda gli studi di leggibilità in Italia. In particolare, si analizzano la formula di Vacca, che costituisce il primo adattamento all'italiano dell'indice di Flesch e la formula GULPEASE, che rappresenta invece la prima formula tarata sulla lingua italiana.
Il sesto capitolo è dedicato ai nuovi approcci al tema della leggibilità.
Dopo una breve panoramica sul *machine learning* e alcuni algoritmi di apprendimento automatico, verranno presentati i metodi più recenti di valutazione della leggibilità, sia per la lingua inglese, che come sempre è la lingua da cui parte l'impulso alla ricerca, sia per le altre lingue, tra cui l'italiano. Vedremo, in particolare, che la tendenza è ormai lo sviluppo di strumenti rivolti a misurare in modo automatico testi e risorse presenti sul web. Parte di questo capitolo riguarda READ-IT, il primo, e attualmente unico, strumento italiano di valutazione automatica della leggibilità.

La seconda parte del presente lavoro è rivolta alla costruzione di un metodo di valutazione automatica della leggibilità di siti web in lingua italiana.
Nel settimo capitolo, sono illustrate in modo dettagliato le diverse fasi di realizzazione del progetto; per ciascuna, sono esposti la metodologia che abbiamo scelto di seguire e gli eventuali approcci alternativi. Cercheremo inoltre di affrontare le diverse problematiche e le varie questioni che possono emergere.
L'ottavo capitolo è dedicato alla costruzione vera e propria del corpus di addestramento su cui si baserà il modello di apprendimento. Per lo sviluppo del nostro metodo di valutazione, abbiamo deciso di concentrarci su una specifica varietà linguistica, la lingua istituzionale degli enti sanitari. In particolare, abbiamo raccolto un campione di testi informativi destinati ai cittadini dai siti web delle Aziende Sanitarie Locali (ASL) italiane. Dopo una breve introduzione sulla comunicazione sanitaria in rete e la valutazione della qualità e della leggibilità delle informazioni sanitarie dei siti web, presentiamo i criteri per la selezione dei testi e la composizione del corpus. Questa parte si conclude con un'analisi critica dei diversi siti delle ASL; in particolare, la valutazione riguarda alcuni aspetti legati alla reperibilità di specifici contenuti, all'organizzazione testuale e, più in generale, alla navigabilità del sito.
L'ultimo capitolo mostra infine i risultati dell'annotazione linguistica del corpus delle ASL: il monitoraggio delle caratteristiche lessicali, sintattiche e morfosintattiche dei testi ci

consentirà di ricostruire il profilo linguistico del corpus, punto di partenza per l'individuazione dei parametri legati alla complessità e più in generale per la valutazione automatica della leggibilità.

PARTE PRIMA

La leggibilità

Io ringrazio Dio perché parlo in lingue sconosciute più di tutti voi;
ma quando la comunità è riunita, preferisco dire cinque parole che si capiscono,
piuttosto che diecimila incomprensibili.
(1 Cor 14, 8-9)

Le parole sono fatte, prima che per essere dette, per essere capite.
(Tullio De Mauro)

1. La leggibilità dei testi

1.1. Leggibilità e comprensibilità

Nell'uso corrente il termine *leggibilità* viene impiegato con diversi significati: può riferirsi ad aspetti che influiscono sulla decifrabilità materiale del testo, come le caratteristiche calligrafiche, il corpo tipografico, il tipo di caratteri, la qualità della grafica, l'uso del colore; può riferirsi alla capacità del testo di coinvolgere e interessare il lettore; può indicare le caratteristiche formali del testo, come il lessico, la sintassi, l'organizzazione dei contenuti, la coerenza, la coesione, le scelte stilistiche; può essere associato alla facilità di lettura o difficoltà di comprensione del materiale scritto. Può fare infine riferimento agli aspetti logico-semantici del testo: in questo caso, è più appropriato parlare di *comprensibilità*.
Analogamente, in inglese sono impiegati diversi termini: *readability, legibility* (aspetti grafici del testo), *comprehensibility, ease of reading*, *ease of understanding*. *Ease of reading* indica i problemi percettivi legati alla lettura, *ease of understanding* riguarda le caratteristiche della scrittura legate alla comprensione del lettore, si riferisce cioè ai problemi di comprensibilità del testo (Klare 1984). Questi termini vengono spessi confusi, probabilmente perché appartengono allo stesso campo semantico, e le varie definizioni di leggibilità sono associate di volta in volta al concetto di comprensione (o la mancanza di essa), all'abilità della persona di leggere un dato testo ad una velocità ottimale, ai fattori motivazionali che influiscono sull'interesse del lettore.
Klare (1963) definisce la leggibilità come "the ease of understanding or comprehension due to the style of writing". Lorge (1944) evidenzia l'interazione tra leggibilità e abilità di lettura nel processo di comprensione: "What a person understands of the material he reads depends upon his general reading ability and the readability of the text he is reading. His reading ability, moreover, depends upon his intelligence, education, environment, and upon his interest and purpose in reading. The readability of a text depends upon the kind and number of ideas it expresses, the vocabulary and its style, and upon format and typography. Reading comprehension must be viewed as the interaction between reading ability and readability. Reading ability can usually be estimated by a person's success with an adequate reading test. Readability, however, must be measured in terms of the success that large numbers of persons have in comprehending the text". McLaughlin (1969), il creatore della formula SMOG, definisce invece la leggibilità come "the degree to which a given class of people find certain reading matter compelling and comprehensible". Come sottolineano Dale e Chall (1948), questi tre elementi della definizione di leggibilità non sono separati, ma interagiscono fra loro. Per spiegare queste interazioni Gilliland (1972) fornisce il seguente esempio: "in a scientific article, complex technical terms may be necessary to describe certain concepts. A knowledge of the subject will make it easier for a reader to cope with these terms and they, in turn, may help him to sort out his ideas, thus making the text more readable. This interaction between vocabulary and content will affect the extent to which some people can read the text with ease". La definizione di Dale e Chall (1949) sembra più completa: la leggibilità è "the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of

readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting".

In italiano, il termine *leggibilità* è spesso considerato sinonimo di comprensibilità e i loro significati tendono a sovrapporsi. Benché sia difficile tracciare delle linee di separazione tra i due diversi aspetti, è preferibile tenerli distinti. La leggibilità si riferisce agli aspetti superficiali del testo, come la decifrabilità materiale e le variabili linguistiche, sintattiche e lessicali. La comprensibilità riguarda invece gli aspetti profondi, logico-semantici del testo, quali l'articolazione dei contenuti, la densità delle informazioni, la maggiore o minore esplicitezza, il contesto nel quale avviene la comunicazione. La leggibilità è una caratteristica intrinseca del testo e la comprensibilità è una caratteristica relativa, che deriva dalla relazione che si stabilisce tra il lettore e il testo.

Il processo di comprensione della lettura può quindi essere analizzato da almeno due punti di vista: da un lato la lettura come decifrazione della *superficie* dei testi, dall'altro la lettura come comprensione, cioè come processo di interazione tra il testo e il lettore (Lucisano e Piemontese 1986).

La scelta di distinguere l'aspetto superficiale da quello profondo è preferibile perché consente di rendere conto sia dei diversi metodi di misurazione o valutazione delle due dimensioni, sia dei diversi problemi o livelli di difficoltà nel processo di comprensione.

Gli ostacoli di fronte ai quali possono trovarsi i lettori di un testo possono essere infatti legati alla decifrazione materiale del testo o alla sua comprensione. Nel primo caso si tratta di *ostacoli superficiali*, nel secondo di *ostacoli profondi* (cfr. Lumbelli 1989); i primi riguardano la leggibilità, i secondi la comprensibilità. La leggibilità veicola la comprensibilità ma non è detto che porti effettivamente alla comprensione del testo. Controllando la leggibilità di un testo non si può avere l'assoluta certezza di determinare i processi di comprensione dell'utente ma si possono eliminare i fattori linguistici di ostacolo all'elaborazione dell'informazione (Vedovelli 1995). Tuttavia, l'eliminazione degli ostacoli superficiali non implica il superamento automatico degli ostacoli profondi; quasi sempre, invece, è vero il contrario e cioè che non è possibile eliminare gli ostacoli profondi senza aver prima eliminato o ridotto gli ostacoli superficiali.

La leggibilità è misurabile tramite criteri quantitativi. Un testo è più o meno leggibile se le sue caratteristiche quantitative, oggettivamente misurabili e controllabili, rispettano alcuni criteri ricavati dall'uso di formule matematiche basate su leggi di statistica linguistica (Piemontese 1996). Queste formule non sono però in grado di rendere conto delle variabili sociolinguistiche che intervengono nella comprensione del testo, né delle connessioni logiche dell'impianto concettuale. La comprensibilità è infatti valutata secondo criteri qualitativi, più complessi, che richiedono l'acquisizione di una metodologia di analisi che, partendo dal testo e dalle sue caratteristiche formali, tende a ricostruire i percorsi mentali automatici, individuali e imprevedibili che portano alla comprensione (o all'incomprensione) del testo (Lumbelli 1989).

1.2. Verso la leggibilità

Il problema della leggibilità ha da sempre suscitato un vasto interesse. Klare (1984), nel suo lavoro di rassegna sulle ricerche che affrontano questo tema, sottolinea l'importanza della

leggibilità come settore di ricerca, affermando di aver trovato centinaia di lavori scientifici sull'argomento e ben oltre mille rifermenti bibliografici.

In molti paesi è ormai da anni una pratica comune fare riferimento a standard di leggibilità per la produzione di testi destinati a un vasto pubblico. Negli Stati Uniti, ad esempio, le formule di leggibilità sono ampiamente utilizzate non solo nell'editoria scolastica ma anche nell'industria e nelle strutture amministrative e di governo[1]. Roger Farr, presidente emerito dell'IRA (International Reading Association), ha stimato che il 40% dei distretti scolastici locali e statali fanno riferimento ai valori di leggibilità come uno dei criteri per la scelta dei libri di testo[2]; oltre metà degli stati ha inoltre stabilito livelli standard di riferimento per le polizze assicurative. La leggibilità è direttamente coinvolta anche in molti casi giudiziari: nonostante l'opposizione di molti avvocati, formulazioni ambigue di regolamenti, libretti di istruzioni e documenti informativi hanno causato non poche controversie legali. Per esempio, in una *class action* che ha coinvolto ricorsi sanitari nazionali, gran parte della contesa riguardava la leggibilità di un documento informativo inviato dall'ufficio di Medicare di New York. Medicare è il programma di assicurazione medica amministrato dal governo degli Stati Uniti e che riguarda i cittadini che hanno più di 65 anni. Il malinteso nasceva dal fatto che il documento non era chiaro e lasciava intendere che Medicare avrebbe pagato una percentuale elevata delle fatture ospedaliere di tali persone, quando invece ne avrebbe pagato soltanto la metà. I consumatori sostenevano quindi di non essere stati adeguatamente informati sull'assicurazione né sui loro diritti di ricorso. È stato dimostrato che il documento aveva un livello di leggibilità superiore a quello adeguato ai destinatari, ovvero che il livello di istruzione di tali consumatori non permetteva loro di comprendere del tutto il contenuto della comunicazione. Il giudice Weinstein (1984) della US District Court, il distretto orientale di New York, ha ordinato al Segretario della Salute e dei Servizi Umani di riscrivere il documento e il governo federale ha così perso il caso.

A partire dalla prima metà degli anni Settanta sono comparse negli Stati Uniti le prime *plain language laws*, che stabiliscono che alcuni tipi di documenti devono soddisfare determinati standard di leggibilità, pena l'invalidità. Nel 1978 il presidente Carter ha emanato un ordine esecutivo che impegna tutti gli uffici governativi a migliorare la chiarezza delle loro comunicazioni scritte; sulla spinta di questa raccomandazione circa trenta stati hanno firmato il *Plain Language Act* ('Legge sul parlar chiaro'), che stabilisce di redigere in *plain language* documenti amministrativi, polizze e contratti e di usare l'indice di Flesch come criterio di leggibilità. Connecticut, Delaware, Hawaii e Maine hanno poi esteso l'applicazione della legge ad altri tipi di polizze ed atti giuridici; in Oklahoma, il Dipartimento di Stato ha richiesto un'analisi della leggibilità di tutte le proposte elettorali.

Nel 2010 il presidente Obama ha firmato la *Plain Writing Act*, la legge che impone alle agenzie federali di utilizzare "clear Government communication that the public can understand and use"[3].

[1] Per un resoconto dei campi di applicazione della misurazione di leggibilità negli Stati uniti cfr. Fry 1987.

[2] Secondo un sondaggio condotto nel 1977 da Rob LaRue (Texas Department of Insurance).

[3] Plain Writing Act of 2010 - An act to enhance citizen access to Government information and services by establishing that Government documents issued to the public must be written clearly, and for other purposes (H. R. 946, Public Law 111-274), 13 October 2010.

Quasi tutte le democrazie europee hanno istituito organismi per promuovere l'uso del *parlar chiaro* nella comunicazione pubblica. Merita un cenno particolare la Svezia, dove il Ministero della Giustizia ha istituito un'apposita divisione con il compito di esaminare sistematicamente tutti i disegni di legge e, se necessario, convertirli in *plain Swedish*.

La leggibilità ha sempre avuto una posizione di rilievo nel campo della ricerca educativa; il problema della comprensione è infatti strettamente collegato ai processi di alfabetizzazione e agli studi sulla misurazione delle abilità di lettura.
Fino alla seconda metà dell'Ottocento, negli Stati Uniti, il sistema scolastico non prevedeva una divisione in classi e la maggior parte delle scuole riuniva in un'aula studenti di ogni età e provenienza. Nel 1847 ha aperto a Boston la prima scuola con un sistema di istruzione strutturato in più cicli, che prevedeva tra l'altro l'adozione di libri di testo pensati specificamente per ogni livello di lettura. Da allora, anche grazie al rapido aumento del numero di studenti iscritti, il raggruppamento per classi (*grading system*) è entrato a pieno nel sistema scolastico statunitense. Fondamentale è stata la promozione, da parte degli educatori, di test standardizzati di lettura che definivano un target per ciascun grado di istruzione. Il più importante è stato *Standard Test Lessons in Reading*, sviluppato da McCall e Crabbs (1925) al Teachers College della Columbia University; il test, che misura la comprensione di alcuni brani tramite prove a scelta multipla, è divenuto anche il criterio per lo sviluppo e la validazione delle formule di leggibilità per la lingua inglese, almeno fino agli anni Cinquanta.
Contemporaneamente i ricercatori si sono occupati di indagini sull'alfabetizzazione degli adulti, suscitando un vasto interesse sul problema della comprensione in aree diverse dall'ambito scolastico. Le prime sperimentazioni sistematiche sono condotte sulle forze armate degli Stati Uniti nel 1917 per poi passare a quelle sui civili a Chicago nel 1937[4]. Si tratta di ricerche finalizzate alla definizione di prove che misurano esclusivamente le capacità di lettura.

La ricerca empirica in senso stretto sulla leggibilità ha inizio negli anni Venti, collegata da un lato allo studio delle frequenze lessicali e dall'altro alle analisi di statistica linguistica. Per entrambi gli indirizzi di studio, le prime applicazioni hanno coinvolto l'ambito scolastico, con l'analisi della frequenza del lessico dei libri di testo e della loro comprensibilità.

1.2.1. Studi sulla frequenza

Le prime liste di frequenza di testi scritti compaiono nell'Ottocento ad opera di pedagogisti. Il primo approccio scientifico alla questione della frequenza lessicale è adottato da W. Gamble; il suo lavoro (*Two lists of selected characters containing all in the Bible and twenty-seven other books*, Shanghai, 1861) è un conteggio di frequenza di ideogrammi cinesi, concepito come un aiuto per migliorare i metodi di stampa. Il secondo e il più vasto studio di tipo quantitativo realizzato con spogli manuali si deve a F. W. Kaeding (*Häufigkeitswörterbuch der deutschen Sprache*, Berlino, 1898): si tratta di una lista di circa

[4] Guy Buswell ha condotto un'intervista a Chicago su un migliaio di adulti con vari livelli di istruzione; per misurare la comprensione della lettura ha usato prove tradizionali, ma anche inserzioni pubblicitarie di alimenti o di cinema, elenchi telefonici, ecc. I risultati sono stati pubblicati nel volume *How Adults Read*, University of Chicago Press, 1937.

11.000.000 occorrenze, compilata con lo scopo di ottimizzare i sistemi stenografici, che fornisce, oltre alla frequenza delle 5.000 forme più utilizzate, anche quelle dei singoli grafemi e delle sillabe. Il materiale lessicale deriva dallo spoglio di dibattiti parlamentari, testi amministrativi e commerciali, giornali, libri di storia, classici della letteratura tedesca e straniera (traduzioni), documenti militari. Kaeding ha dimostrato che le 15 parole più frequenti rappresentano il 25% delle occorrenze totali e le 66 forme più frequenti coprono circa il 50% dei testi. La lista di frequenza di Keading costituisce il punto di partenza per la compilazione dei dizionari di base.

Negli Stati Uniti, a partire dal 1920, si registra un incremento della popolazione scolarizzata, in particolare un aumento di studenti delle scuole secondarie, probabilmente collegato al flusso migratorio. Gli insegnanti iniziano a prendere coscienza dell'eccessiva difficoltà dei libri di testo proposti agli studenti.

Nel 1921 E. L. Thorndike pubblica il *Theacher's Word Book*, il primo lessico di frequenza della lingua inglese. L'intenzione dell'autore è fornire agli insegnanti uno strumento obiettivo per misurare le difficoltà dei testi; egli aveva infatti notato che i docenti in Germania e Russia usavano il conteggio delle parole per la classificazione dei testi. Il presupposto è che più frequentemente una parola è usata, più è familiare e dunque più facile da comprendere. Il *Theacher's Word Book* contiene le prime 10.000 parole inglesi per frequenza d'uso, tratte dalla letteratura per bambini, manuali scolastici, la Bibbia, libri di cucina e altri settori, giornali. A questo volume seguono *The Theacher's Word Book of 20.000 Words* nel 1932, e *The Theacher's Word Book of 30.000 Words* nel 1944 insieme a I. Lorge.

Fino al 1972 la maggior parte delle ricerche riguarda l'inglese, il francese, il tedesco e lo spagnolo, mentre "l'italiano è stata un po' la lingua dimenticata in questi spogli di frequenza" (Bortolini et al. 1971). Fino ad allora erano stati fatti solo due brevi saggi, quello di T. M. Knease (*An italian Word List from Literacy Sources,* 1933), basato su uno spoglio manuale di circa 4.000 parole tratte da fonti letterarie che vanno dalla seconda metà dell'Ottocento alla prima metà del Novecento, e quello di B. Migliorini (*Der grundlegende Wortschatz des Italienischen*, 1943), in cui erano elencate le 1.500 parole ritenute dall'autore fondamentali nella lingua italiana e utili per l'insegnamento dell'italiano agli stranieri[5].

Il *Lessico di frequenza della lingua italiana contemporanea* (LIF)[6], elaborato al Centro Nazionale Universitario di Calcolo elettronico di Pisa nel 1971, rappresenta il primo grande progetto di costruzione di un lessico di frequenza per la lingua italiana e il primo dizionario di frequenza realizzato con l'ausilio dei calcolatori elettronici. È il risultato dello spoglio di

[5] In realtà, un primo conteggio sulla frequenza dei vocaboli nell'italiano era stato già effettuato da M. E. Thompson (*A study in italian vocabulary frequency*, 1927); si tratta di una tesi di dottorato non pubblicata che riporta una lista di 500 occorrenze destinata a studenti che studiano l'italiano come L2. È interessante notare che tra questo studio e quello di Migliorini, cioè la prima ricerca ad opera di un italiano, trascorrono 17 anni e che dunque gli studiosi italiani riconoscono relativamente tardi l'importanza di tali lessici di frequenza.

Sgroi (1994) osserva che i primi vocabolari fondamentali e di frequenza dell'italiano sono opere di studiosi stranieri (Thompson 1927, Knease 1933, Juilland eTraversa 1973, Sciarone 1977) o di italiani operanti all'estero (Migliorini 1943, G. A. Russo 1947, J. A. Russo 1962) e solo successivamente di italiani in Italia (Bortolini et al. 1971, De Mauro et al. 1980, VELI 1989, Katerinov et al. 1991, De Mauro et al. 1993, ecc.).

[6] Bortolini et al. 1971.

un corpus di 500.000 occorrenze della lingua italiana contemporanea scritta, dal quale sono ricavati dati statistici di diverso tipo su circa 5.000 lemmi. Il corpus del LIF si basa su cinque gruppi di testi (cinema, teatro, romanzi, periodici, sussidiari delle scuole elementari), datati tra il 1947 e il 1968. Per ciascuna di queste categorie sono prese 100.000 occorrenze. Per cercare di creare un corpus rappresentativo della lingua contemporanea viene preso come limite cronologico il 1945 e cioè la fine della seconda guerra mondiale (anche se in realtà il testo più antico è edito nel 1947): "questa data, come del resto le altre fissate per i periodi storici o letterari, pur non rappresentando un confine netto è significativa soprattutto perché rappresenta la chiusura di un periodo storico che ha avuto notevoli ripercussioni sulla lingua e un diverso orientamento delle fonti di informazioni. Prima di tutto avviene un notevole rinnovamento del lessico con l'uscita dall'uso di molte voci legate a particolari istituzioni storico-politiche del passato regime, in secondo luogo viene a cessare o a diminuire moltissimo l'influsso di modelli francesi, lasciando il posto a modelli prevalentemente anglosassoni (inglesi e soprattutto americani)" (Bortolini et al. 1971, p. XIX).

Per ogni forma flessa e ogni lemma sono forniti tre valori: la frequenza, l'indice di dispersione e il valore d'uso. La dispersione è una misura che rende conto di quanto un dato temine è distribuito tra le diverse classi di testi; indica cioè il numero di testi diversi in cui la parola appare. Più le parole compaiono in diverse tipologie di testi più hanno la probabilità di essere incontrate e imparate da un maggior numero di persone: per questo motivo, tra due parole che presentano lo stesso indice di frequenza, la più conosciuta sarà probabilmente quella che ha una dispersione maggiore. Il valore d'uso deriva dalla combinazione dei valori di frequenza e di dispersione e fornisce la stima più attendibile della diffusione di una parola nella lingua. "Dalla moltiplicazione di frequenza e dispersione abbiamo ciò che i linguisti chiamano 'uso' della parola" (De Mauro 2007).

Immediatamente successiva al LIF è la pubblicazione di A. Juilland e V. Traversa, *Frequency Dictionary of Italian Words* (1973); il volume fa parte della collana *The Romance Languages and their Structures*, diretta da Juilland, che propone lo studio quantitativo del lessico, della grammatica e della fonematica delle principali lingue romanze (spagnolo, rumeno, francese, italiano, portoghese). Il dizionario italiano segue la pubblicazione delle liste per lo spagnolo (A. Juilland, E. Chang Rodriquez, *Frequency Dictionary of Spanish Words*, The Hague, 1964), il rumeno (A. Juilland, P.M.H. Edwards, I. Juilland, *Frequency Dictionary of Rumanian Words*, The Hague, 1965) e il francese (A. Juilland, D. Brodin, C. Davidovitch, *Frequency Dictionary of French Words*, The Hague, 1970). Il corpus è rappresentato da circa 500.000 occorrenze, tratte da testi pubblicati tra il 1920 e il 1940, appartenenti a cinque generi diversi: opere teatrali, romanzi e novelle, saggistica, periodici e quotidiani, letteratura tecnico-specialistica.

Nel 1977 viene elaborata una terza lista di frequenza: il *Vocabolario fondamentale della lingua italiana*, di A. G. Sciarone. Questa lista è ricavata dall'analisi di un corpus di 1.500.000 occorrenze, ottenuto dalla combinazione dei corpora del LIF e del *Frequency Dictionary of Italian Words* con un nuovo campione di 500.000 occorrenze tratte da testi che vanno fino al 1974.

Il LIF è servito anche come base per la compilazione del *Vocabolario di Base della lingua italiana* (VdB) nel 1980[7]; T. De Mauro ha individuato, all'interno del lessico italiano, un settore particolare da lui definito *vocabolario di base*, formato da circa 7.000 parole che costituiscono, appunto, la base di tutti i testi scritti e parlati della nostra lingua.

Il *Vocabolario di Base* è suddiviso in tre fasce: il *lessico fondamentale*, il *lessico di alto uso* e il *lessico di alta disponibilità*. Il lessico fondamentale comprende circa 2.000 vocaboli (i primi 2.000 lemmi del LIF) che costituiscono il 95% dei testi più semplici e l'80% di quelli più tecnici: si tratta sia di parole funzionali (preposizioni, avverbi, articoli, congiunzioni, ausiliari) sia di nomi, aggettivi e verbi più comuni e frequenti, noti praticamente a tutti coloro che parlano italiano (*mano*, *casa*, *gatto*, *pioggia*, *bello*, *forte*, *andare* ecc.). "Sono i vocaboli che chi parla una lingua ed è uscito dall'infanzia conosce, capisce e usa. Sono le parole di massima frequenza nel parlare e nello scrivere e disponibili a chiunque in ogni momento" (De Mauro 1980, p.106).

Il lessico di alto uso comprende tra le 2.500 e le 3.000 parole (i successivi lemmi del LIF), impiegate frequentemente sia nel parlato che nello scritto e note a tutti coloro che hanno almeno un livello di istruzione medio (*pregiudizio*, *privilegio*, *definire*, ecc.).

Il lessico di alta disponibilità è costituito da circa 2.300 vocaboli che ricorrono con frequenza molto bassa nei testi scritti (e dunque non risultano nella lista di frequenza nel LIF) ma che sono ben noti ad ogni parlante (*dentifricio*, *forchetta*, *matita*, *abbronzare* ecc.). "Partendo dall'esame dei dizionari dell'italiano comune, si sono isolate le parole di maggiore 'disponibilità'. Si tratta delle parole che può accaderci di non dire né tanto meno di scrivere mai o quasi mai, ma legate a oggetti, fatti, esperienze ben noti a tutte le persone adulte nella vita quotidiana. Sono le parole che diciamo o scriviamo raramente, ma che pensiamo con grande frequenza" (De Mauro 2007, p. 162).

Se usiamo parole del vocabolario di base possiamo avere buone probabilità di essere capiti da chi ha almeno la licenza media inferiore (71% della popolazione italiana); se usiamo solo le parole del vocabolario fondamentale, possiamo sperare di essere compresi da chi possiede almeno la licenza elementare (91% della popolazione italiana)[8]. Più cresce in un testo il numero di parole estranee al vocabolario di base, più si restringe il numero di persone che sono in grado di capirlo. Il vocabolario di base "è il riferimento fondamentale per il controllo del lessico di testi scritti in italiano, quando si vuole verificare la rispondenza del lessico a criteri oggettivi di comprensibilità" (Mastidoro 1992, p. 126).

Altri 45.000 lessemi appartengono al cosiddetto *vocabolario comune* e compaiono in testi più complessi, soprattutto scritti, comprensibili a chi è fornito di un'istruzione medio-alta. Il *vocabolario di base* e il *vocabolario comune* costituiscono il *vocabolario corrente*, al di fuori del quale si situano i lessemi che sono propri della lingua letteraria o dei vari linguaggi settoriali.

Nel 1999 De Mauro pubblica un'altra grande impresa lessicografica, il *Grande Dizionario Italiano dell'uso* (GRADIT), in sei volumi, che comprende circa 250.000 lemmi; si tratta del

[7] La lista di parole che compone il VdB è apparsa per la prima volta in appendice alla prima edizione di un libro di Tullio Di Mauro, *Guida all'uso delle parole* (1980), uno dei primi volumi della collana dei Libri di Base degli Editori Riuniti. Le 7.000 parole sono ordinate alfabeticamente e accompagnate dalla qualifica grammaticale.

[8] I dati sono ricavati dal sito dell'Istat su una statistica relativa al grado di istruzione della popolazione italiana nel 2011 (circa 56 milioni di individui).

primo grande dizionario sincronico dell'italiano, in cui troviamo la specificazione sistematica di tutti i lemmi del *Vocabolario di Base*. In particolare, il GRADIT indica la marca d'uso non del lemma, ma di ogni singola accezione, registrando anche voci letterarie, termini specialistici, varianti formali, regionalismi, parole straniere, voci gergali, latinismi, sigle, abbreviazioni, ecc.

Ancora dedicato all'italiano scritto è il *Vocabolario Elettronico della Lingua Italiana* (VELI 1989), realizzato dall'IBM, con la consulenza scientifica di De Mauro. Il corpus, che propone i valori d'uso dei primi 10.000 lemmi risultati dallo spoglio di un corpus di 26 milioni di occorrenze, è costituito da testi orientati prevalentemente al linguaggio giornalistico-finanziario: notizie economiche dell'ANSA e testi da *Il Mondo*, *Europeo*, *Domenica del Corriere*, apparsi nel biennio tra il settembre 1985 e il giugno 1987.

Nel 1993 viene presentato il *Lessico di frequenza dell'italiano parlato* (LIP)[9], che costituisce il primo esempio di analisi statistica in grande scala della lingua italiana parlata. È il risultato dello spoglio di registrazioni di conversazioni di diverso tipo, della durata complessiva di circa 57 ore, effettuate a Milano, Firenze, Roma e Napoli tra il 1990 e il 1992. Il corpus è costituito da circa 500.000 parole grafiche, che derivano dalle trascrizioni delle registrazioni.

1.2.2. Statistica linguistica

Accanto agli studi che hanno prodotto le prime liste di frequenza si collocano le ricerche di statistica linguistica[10]. Gli indici di frequenza si basano infatti sui procedimenti della statistica lessicale che prevede l'applicazione di metodi statistici all'esame dei fatti linguistici: "Le unità costitutive di una lingua (fonemi, parole, ecc.) soprattutto considerate sotto il profilo della frequenza con cui appaiono nei testi, costituiscono un tipico insieme di fenomeni di massa e sono perciò suscettibili di indagini statistiche per rilevare le frequenze medie del loro distribuirsi nel discorso e, nel tempo, le eventuali trasformazioni di tali frequenze." (De Mauro 1961).

La statistica linguistica mira quindi a individuare le regolarità statistiche delle diverse unità testuali, con particolare attenzione al lessico. In quest'ottica si collocano una serie di studi legati ai nomi dello statunitense George K. Zipf, del polacco Benoit Mandelbrot, dei francesi Pierre Guiraud e Charles Muller, dei praghesi e di Gustave Herdan nell'Europa orientale (Chiari 2007).

L'interesse per questo tipo di analisi si deve inizialmente a stenografi, come il già citato F. W. Kaeding, che nel 1898 coordina una ricerca sulle frequenze dei grafemi, delle sillabe e delle parole della lingua tedesca, e J. B. Estoup che, in uno studio pubblicato nel 1907, definisce la nozione di *rango* come la posizione occupata da una parola in una lista ordinata per frequenze decrescenti.

Nella statistica linguistica si definisce *frequenza* il numero di volte che un termine occorre in un testo o in un corpus di testi; generalmente in una lista di frequenza gli elementi sono

[9] De Mauro et. al. 1993

[10] Sulla statistica linguistica cfr. Guiraud 1954, 1954b e 1960; De Mauro 1961; Heilman 1961; Herdan 1964. Per una panoramica aggiornata di lavori italiani nel settore cfr. De mauro e Chiari 2005. Per le leggi statistiche cfr. Chiari 2007 e Lenci et al. 2005.

ordinati in ordine di frequenza decrescente[11]. Con *rango* si intende la posizione che una parola occupa nella lista di frequenza ordinata, per cui la parola con la frequenza più alta avrà rango 1, la successiva rango 2, e così via.
Estoup (*Gammes Sténographiques*, 1916) individua per la prima volta la relazione che lega rango e frequenza di una parola, principio che fornisce uno dei principali supposti teorici alla successiva sistematizzazione da parte di Zipf (1935, 1949). È infatti lo statunitense George K. Zipf, filologo e docente di linguistica ad Harvard, ad enunciare i termini generali di quella che verrà poi definita *Legge armonica di Zipf* (1949), o *Legge di Zipf-Estoup*, secondo la quale la relazione tra la frequenza (*f*) di un termine e il suo rango (*r*) in una data lista ordinata per frequenza decrescente è costante; in altre parole rango e frequenza sono inversamente proporzionali. Il prodotto del rango per la frequenza è costante (*c*) per ogni parola della lista:

$$f \; x \; r \; = \; c$$

Un classico esempio fatto dallo stesso Zipf è tratto dallo studio lessicale dell'*Ulysse* di Joyce:

al rango	10	la frequenza è	2653	$f \times r = 26.530$
al rango	100	la frequenza è	265	$f \times r = 25.500$
al rango	1000	la frequenza è	26	$f \times r = 26.000$
al rango	10000	la frequenza è	2	$f \times r = 29.000$

La legge prevede quindi un decremento progressivo della frequenza delle parole proporzionale all'aumentare del rango:

Per *r=1* *f= c*
Per r=2 *f= c / 2*
Per r=3 *f = c / 3*

La distribuzione delle parole in un dato testo seguirebbe approssimativamente una serie armonica, la cui espressione grafica è un'iperbole equilatera. Zipf chiama questa rappresentazione *curva canonica.* Parole che compaiono molto in basso nella lista di frequenza tendono ad avere frequenze simili; la coda della curva conterrà quindi molte parole con frequenza 1 (*hapax*)[12].

[11] La frequenza può essere assoluta o relativa: la *frequenza assoluta* è data dal numero di *occorrenze* o *repliche* (*token*) in un testo o in un corpus di testi; la *frequenza relativa* è invece data dal rapporto tra *repliche* e *tipi* (*type*). Il numero dei *tipi* è il numero di *parole diverse* presenti in un dato testo o corpus di testi.
[12] La legge è stata verificata su varie lingue: sull'inglese scritto (cfr. Miller e Newman 1958), sull'inglese americano orale (cfr. Dahl 1979), su testi letterari francesi (cfr. Guiraud 1954b), sul cinese (cfr. Rousseau e Zhang 1992).

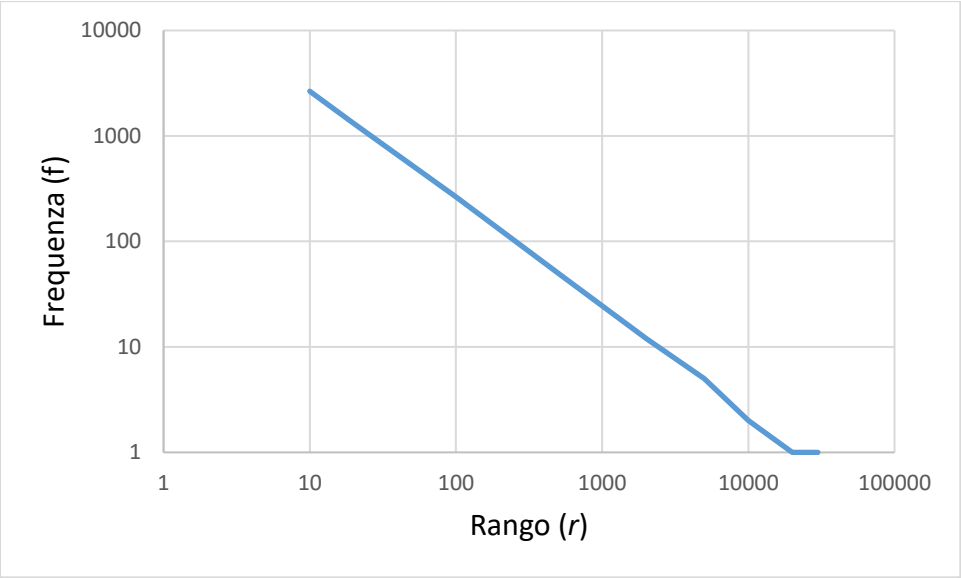


Figura 1. Curva canonica delle parole inglesi (da Zipf 1935)

Nel 1954 B. Mandelbrot pubblica *Structure formelle des textes et communication*, in cui espone la sua legge canonica sulla distribuzione della frequenza delle parole in un testo o corpus di testi: si tratta di una generalizzazione della legge armonica di Zipf, che rappresenta invece un caso particolare.
Secondo Mandelbrot, la legge di Zipf può essere formulata più rigorosamente determinando la frequenza o probabilità di un vocabolo (Pr) in funzione del rango (r), tenendo però anche conto di altri parametri, come la *varietà* (p) delle frequenze dei vocaboli e la *temperatura informazionale* (b), cioè la distribuzione del numero di parole diverse del vocabolario del testo. La legge di Zipf si verifica per p=0 e b=1 (De Mauro 1961)[13].

Oltre a questa legge, Zipf individua altre regolarità nella lingua, come le correlazioni tra la frequenza di una parola e il numero di significati e tra la frequenza e il numero dei fonemi.
Zipf spiega queste regolarità statistiche non come effetto del caso, bensì delle caratteristiche di finitezza psicobiologica dell'essere umano (Chiari 2007). Esiste un principio che governa i comportamenti umani collettivi e individuali, che egli chiama il *principio del minimo sforzo*: "any human action will be a manifestation of the *Principle of Least Effort* in operation" (Zipf 1949). La distribuzione delle parole nei testi rifletterebbe proprio questa economicità nella comunicazione. Secondo tale principio, i comportamenti umani sono dominati da due tendenze che si oppongono costantemente: la *tendenza all'unificazione*, volta ad esprimersi nel modo più economico possibile e dunque tesa a ridurre il numero di unità distintive (*speaker's* economy) e la *tendenza alla diversificazione*, volta ad esprimersi chiaramente, all'efficacia comunicativa e tesa dunque a differenziare il più possibile le unità (*auditor's economy*).
A queste tendenze si ispira la nozione di *economia linguistica* di Martinet (1955) che consiste in un equilibrio "tra le esigenze espressive che richiedono unità più numerose, più

[13] Per la legge canonica di B. Mandelbrot e il concetto di *temperatura* informazionale si veda anche Plantera 2005.

specifiche e relativamente meno frequenti, e l'inerzia naturale che spinge verso un numero più ristretto di unità più generali e di impiego più frequente". Questo equilibrio si individua ad esempio nel rapporto di proporzionalità inverso tra la complessità di un fonema e la sua frequenza relativa: quanto più un fonema è frequente tanto meno esso tende ad essere nettamente articolato (*Legge di Zipf-Martinet*).
"Tanto Zipf che Martinet si avvidero che dal continuo opporsi di queste forze proveniva la capacità della lingua di cambiare, evolversi, plasmarsi sulle esigenze comunicative degli esseri umani. Un esempio del continuo e profondo agire di queste due forze opposte lo troviamo osservando il vocabolario delle lingue storico-naturali: in esse continuamente sono registrate parole nuove, coniate appositamente per individuare nuovi referenti e parole già esistenti che via via nel tempo acquistano nuovi significati e coprono aree sempre più vaste di senso" (Carloni 2005).

La legge che individua la relazione tra la frequenza di una parola e il numero di significati viene enunciata da Zipf nel 1945, qualche anno prima della sua legge più importante. È facile constatare che le parole più frequenti possiedono una certa generalità semantica.
Secondo questa formulazione la frequenza e il numero di accezioni di ciascuna parola sono direttamente proporzionali: il numero dei significati cresce secondo la radice quadrata della frequenza.

$$m = \sqrt{f}$$

dove *m* è il numero delle accezioni e *f* è la frequenza della parola[14].
La funzione che descrive questa relazione è detta *retta di regressione*; di questa funzione il dato importante è il coefficiente angolare: se è di segno positivo indica che tra le variabili esiste una relazione diretta, se invece è di segno negativo indica che esiste una relazione di tipo inverso. Inoltre è il valore del coefficiente angolare a determinare l'ordine di grandezza di questa relazione.

Un'altra tendenza individuata da Zipf è il rapporto tra la frequenza di una parola e il numero di fonemi che la compongono. Basandosi su uno studio empirico effettuato sull'inglese, il latino e il cinese, lo studioso nota che esiste una relazione tra la lunghezza fonematica delle parole e la loro frequenza nei testi[15]: le parole più brevi tendono ad avere una frequenza maggiore delle parole più lunghe.
In inglese le parole più frequenti sono monosillabi (Zipf 1935). Anche il computo eseguito da Kaeding su circa 11 milioni di parole della lingua tedesca scritta mostra che il 50% di queste sono monosillabi; molte altre lingue danno risultati analoghi. Questo è dovuto

[14] La legge è stata verificata sul francese da Guiraud (1954b), il quale l'ha riformulata come segue: $m / \sqrt{f} = k$
dove *m* e rappresenta il numero di accezioni, *f* la frequenza e *k* una costante di proporzionalità. La costante sembra essere un numero quasi sempre approssimabile a 1; nella formulazione di Zipf la costante viene omessa perché egli la pone uguale a 1. Per l'applicazione della legge sulla lingua italiana cfr. Carloni 2005.
[15] Kaeding (1898) aveva già notato per il tedesco la relazione di proporzionalità inversa tra il numero di sillabe delle parole e la loro frequenza; il fatto la lunghezza delle parole sia valutata in sillabe dipende dagli interessi stenografici dell'autore.

anche al fatto che quando in una lingua una parola inizia ad essere usata frequentemente dalla comunità linguistica, si tende ad abbreviarla: Zipf propone come esempio i termini *movies*, *talkies*, *gas* che sono abbreviazioni per troncamento di *moving picture*, *talking picture* e *gasoline*. Parole come *constitutionality*, *quintessentially*, *idiosyncrasy* non vengono invece troncate perché utilizzate con bassa frequenza. L'abbreviazione si verifica anche per sostituzione di parole lunghe con parole più corte, come nel caso di *car* per *automobile*. Queste sostituzioni possono essere temporanee, come nel caso dei pronomi o degli avverbi, o permanenti, come nel caso dello *slang*. Le sostituzioni permanenti riflettono un aumento generale della frequenza media relativa di una parola all'interno dell'intera comunità linguistica, quelle temporanee riflettono invece un aumento temporaneo della frequenza. Zipf chiama questa tendenza alla riduzione dipendente dalla frequenza *Legge dell'Abbreviazione*: "in view of the evidence of the stream of speech we may say that the length of a word tends to bear an inverse relationship to its relative frequency; and in view of the influence of high frequency on the shortenings from truncation and from durable and temporary abbreviatory substitution, it seems a plausible deduction that, as the relative frequency of a word increases, it tends to diminish in magnitude. This tendency of a decreasing magnitude to result from an increase in relative frequency, may be tentatively named the Law of Abbreviation" (Zipf 1935, p. 38).

Anche questa legge sembra riflettere il principio di economia linguistica: "the law of abbreviation seems to reflect on the one hand an impulse in language toward the maintenance of an equilibrium between length and frequency, and on the other hand an underlying law of economy as the *causa causans* of this impulse toward equilibrium. That the maintenance of equilibrium is involved is clear from the very nature of the statistics. That economy, or the saving of time and effort, is probably the underlying cause of the maintenance of equilibrium is apparent from the fact that the purpose of the all truncations and transitory contextual substitutions is almost admittedly the saving of time and effort" (id.).

Le leggi di Zipf non vengono ben accolte, anzi in alcuni casi sono aspramente criticate e contestate da diversi linguisti[16]. Sono studiosi di altri campi a riconoscerne l'importanza e ad estenderne il campo di applicazione: la legge armonica è stata generalizzata ad altri fenomeni naturali e sociali, dalla distribuzione del diametro dei crateri lunari alla distribuzione della popolazione nei centri abitati, alla distribuzione del reddito, alla frequenza di accesso alle pagine dei siti web, ai terremoti, ai sistemi di catalogazione bibliotecaria[17].

Come sostiene Chiari (2007), le implicazioni filosofico-biologiche del lavoro di Zipf e alcune sue ingenuità linguistiche e matematiche hanno fortemente limitato la diffusione del suo lavoro, almeno fino alla ripresa di alcuni aspetti da parte di Pierre Guiraud.

[16] Mandelbrot (1974) definisce Zipf "autore di numerosi libri che combinano strettamente e in modo inconsueto verità e follia" (p. 313). Herdan (1971) discute l'uso che è stato fatto delle leggi di Zipf e Mandelbrot: "a parte il fatto che esse siano state citate parecchie volte, spesso da autori che non erano in grado di giudicare il loro valore matematico (…) il massimo difetto sta nella acritica, facile accettazione di certi rapporti funzionali (parlando con il linguaggio matematico) nel linguaggio, e trascurando il fatto che, di regola, qui possiamo sperare di stabilire soltanto leggi statistiche" (p. 94).

[17] Per approfondimenti sulle applicazioni della legge di Zipf alla distribuzione degli utenti nei siti web si veda Adamic e Huberman 2002 e Breslau et al. 1999.

Guiraud prende in esame la relazione tra frequenza delle parole e lunghezza fonematica e individua quella che viene definita *Legge di Zipf-Guiraud*. Secondo tale legge, il numero di fonemi (*k*) di una data parola è direttamente proporzionale al suo rango (*r*) e cioè diminuisce al crescere della frequenza della parola:

$$\frac{k}{\log r} = costante$$

Uno dei contributi più significativi di Guiraud (*Les caractères statistiques du vocabulaire*, 1954) è l'aver notato come le parole si distribuiscono statisticamente nei testi. Vi sono pochissime parole molto frequenti (in gran parte parole appartenenti a *classi chiuse*), che coprono più della metà delle occorrenze di qualsiasi testo, mentre vi sono un grandissimo numero di parole a bassa frequenza o rare (*hapax*): "un très petit nombre de mots convenablement choisis couvrent la plus grande partie de n'importe quel texte, et il est possible d'établir une liste de mots telle que: les 100 premiers mots couvrent 60% de n'importe quel texte, les 1000 premiers mots couvrent 85% de n'importe quel texte, les 4000 premiers mots couvrent 97,5% de n'importe quel texte, le reste (40 à 50.000 mots) couvre 2,5% de n'importe quel texte"[18]. Il *Vocabolario di base della lingua italiana* ad esempio contiene nella sua fascia più interna il vocabolario fondamentale, 2.000 parole che coprono circa il 90% delle occorrenze di un qualunque testo (Chiari 2008).
Klare (1968), nella sua revisione della ricerca sulla frequenza delle parole, afferma: "Not only do humans tend to use some words much more often than others, they recognize more frequent words more rapidly than less frequent, prefer them, and understand and learn them more readily. It is not surprising, therefore, that this variable has such a central role in the measurement of readability" (p. 20).
Guiraud identifica inoltre una relazione costante tra frequenza relativa ed estensione del vocabolario. La frequenza relativa è data dal rapporto tra il numero di parole diverse e il numero totale delle occorrenze di un dato testo; questa proporzione, nota anche come *type/token ratio* (TTR), consente di valutare la *ricchezza lessicale* di un testo, ossia l'ampiezza del vocabolario utilizzato.
Indichiamo con V il numero di parole diverse presenti in un dato testo, cioè l'insieme delle parole che compongono il vocabolario di quel testo e con N il numero di parole totali presenti nel medesimo testo; l'indice di ricchezza lessicale è dato dunque dal rapporto V/N.
Questa misura permette di confrontare due testi di uguale dimensione; per limitare la dipendenza dalle dimensioni del corpus, è consigliabile utilizzare un indice più complesso, proposto da Guiraud:

$$G = \frac{V}{\sqrt{N}}$$

La formula originariamente proposta da Guiraud considerava V come numero dei diversi lemmi ma viene sostituita da questa formulazione più semplice: "la ragione del successo

[18] "Un numero molto piccolo di parole opportunamente scelte coprono la maggior parte di un testo qualsiasi; è possibile stabilire una lista di parole in cui: le prime 100 parole coprono il 60% di un testo qualsiasi, le prime 1000 parole coprono l'85%, le prime 4000 il 97,5% e il resto (da 40.000 a 50.000 parole) copre il 2,5% di un qualsiasi testo".

della formula semplice e intuitiva riportata poco sopra, nonostante una certa povertà da un punto di vista linguistico più profondo, è la sua immediata praticità e la possibilità di ottenerla in modo automatico senza particolare trattamento per lingue diverse, anche a noi sconosciute". (Chiari 2007, p. 50).
Secondo Guiraud, nel confronto tra due testi di uguale dimensione, è possibile valutare la ricchezza lessicale utilizzando anche il semplice rapporto V_1/V, cioè la proporzione di *hapax* sul totale delle parole diverse: quanto il valore di questo rapporto è maggiore, tanto più ampia è la varietà del vocabolario del testo.

In conclusione, prima la compilazione delle liste di frequenza del vocabolario delle lingue e poi l'osservazione, sulla base di queste, delle regolarità statistiche che hanno portato alla definizione di leggi, hanno aperto la strada a nuove direzioni di ricerca sul tema della comprensione della lettura, portando infine allo sviluppo di strumenti oggettivi di misurazione della leggibilità e comprensibilità dei testi.

2. Studi classici sulla leggibilità

A partire dagli anni Venti nascono negli Stati Uniti gli studi sulla leggibilità dei testi e iniziano ad essere sviluppate le prime formule matematiche in grado di misurarla, i cosiddetti *indici di leggibilità*.
Le prime ricerche provengono dall'ambito scolastico; da una parte si concentrano sul controllo del vocabolario dei libri di testo, dall'altra sulla comprensione dei materiali di lettura proposti agli studenti. Gli studiosi cercano di ideare dei metodi oggettivi di misurazione della difficoltà dei materiali destinati all'apprendimento.
Il metodo seguito dai ricercatori prevede che siano effettuate delle misurazioni sui lettori e successivamente delle misurazioni sui testi.
Vengono messi a punto dei test di comprensione della lettura per determinare il grado di facilità o di difficoltà di testi scritti. Se i lettori rispondono in modo rapido e corretto alle domande formulate sul testo in esame, il testo è considerato facile; se invece essi commettono errori, il testo è considerato più o meno difficile. In questo modo si costruisce una scala di leggibilità: i vari campioni di testi vengono ordinati su una scala che va dal più difficile (quello che ha fatto registrare il maggior numero di errori) al più facile (quello che ha fatto registrare il minor numero di errori). Successivamente, si analizzano statisticamente i brani, confrontando le caratteristiche linguistiche dei campioni per individuare se esiste una variabile statistica strettamente correlata con la difficoltà.
Sono molti gli indici statistici che possono essere misurati: la percentuale delle parole più frequenti, il numero delle parole difficili, il numero delle parole diverse, il numero dei pronomi, il numero delle frasi complesse, il numero delle figure retoriche, ecc.
Una volta compilata una lista degli indici statistici e stabiliti i punteggi dei vari brani per ciascun indice, è possibile correlare i punteggi con quelli relativi alla leggibilità. Se il coefficiente di correlazione è prossimo a 1 o -1, l'indice statistico misura qualcosa di correlato alla leggibilità. Se la correlazione è prossima allo zero, l'indice statistico non ha rapporto con la leggibilità (Miller 1972, pp. 188-190). In base alle diverse variabili statistiche considerate, i ricercatori costruiscono le varie formule di leggibilità. Tale procedimento è chiamato *analisi di correlazione multipla.*
Questo capitolo prende in considerazione quelli che vengono definiti "studi classici sulla leggibilità", cioè tutte quelle ricerche sulla leggibilità dei testi e sulle formule sviluppate per la lingua inglese a partire dagli anni Venti fino agli anni Sessanta[19].

2.1. Lively e Pressey: la prima formula di leggibilità

Nel 1923 B. Lively e S. L. Pressey pubblicano il loro studio *A method for measuring the "Vocabulary Burden" of textbooks*, a seguito di un'indagine sulla quantità dei tecnicismi presenti nei libri di testo di scienze di una scuola media. Gli insegnanti si erano infatti lamentati della notevole quantità dei termini tecnici nei materiali, tanto che erano costretti a spendere più tempo nello studio del lessico scientifico che nello studio del contenuto stesso.

[19] Una sintesi dei principali studi è proposta da Klare 1974-75, 1984 e Dubay 2004, 2006, 2007.

L'articolo di Lively e Pressey descrive la prima formula di leggibilità e misura la difficoltà dei testi considerando il numero di parole diverse su un campione di 1.000 parole e il numero di parole che non sono presenti sulla lista di Thorndike (*Theacher's Word Book*, 1921). Il loro metodo produce un coefficiente di correlazione di 0,80, tuttavia viene criticato in quanto assegna punteggi in base alla frequenza delle parole e non alla misurazione diretta della difficoltà dei testi (Gray 1947).
Il merito di Lively e Pressey è quello di aver mostrato l'efficacia di un approccio statistico alla previsione della difficoltà dei testi; l'uso dell'elenco di parole di Thorndike ha avuto inoltre una notevole influenza su molte delle formule successive. Come affermano gli autori: "the fundamental value of Thorndike's contribution is obvious; the *Word Book* has opened up a whole new field for investigation".

2.2. Vogel e Washburne: la Formula Winnetka

M. Vogel e C. Washburne (1928) di Winnetka, in Illinois, sono i primi a creare una formula di leggibilità che correla la difficoltà dei materiali scritti a livelli di lettura specifici.
Gli studiosi tentano di classificare i libri per gradi di istruzione appropriati, in base alla misurazione delle abilità di lettura dei bambini. Per farlo, chiedono a 36.750 bambini di compilare una lista sui libri letti nel corso dell'anno; di questi, circa 700 libri sono indicati da almeno 25 soggetti come letti e apprezzati. L'abilità di lettura dei bambini è valutata tramite lo *Stanford Achievement Test*.
I risultati di questi giudizi sono riuniti nella *Winnetka Graded Book List*[20]. La lista è classificata secondo un punteggio di difficoltà assegnato ai libri: il grado di posizionamento di un libro rappresenta l'abilità media di lettura dei bambini che hanno letto quel libro.
Assieme ad una ventina di insegnanti, gli autori esaminano i libri della lista al fine di individuare i possibili fattori che influenzano la difficoltà dei testi e ne scelgono dieci che sembrano maggiormente correlati al punteggio medio di lettura dei bambini. Tuttavia, poiché molti di questi elementi sono correlati fortemente tra loro, la scelta si riduce a soli quattro fattori. La formula di Vogel e Washburne (detta *Formula Winnetka*) si basa quindi su queste quattro variabili[21]:

- Numero di parole diverse su 100 parole
- Numero di parole non comuni su 100 parole[22]
- Numero di frasi semplici in 75 frasi successive
- Numero di preposizioni su 100 parole

Nel 1938 Washburne e Morphett rivedono la formula, eliminando il conteggio delle frasi preposizionali e considerando il numero di parole comuni in base alle 1.500 parole usate più frequentemente nella lista di Thorndike. Il numero di parole differenti sembra essere il miglior indicatore della difficoltà di un brano in quanto è maggiormente correlato con i punteggi dei test di lettura.

[20] Vogel e Washburne 1926.
[21] La correlazione multipla di questi indici con i punteggi dei test di lettura è di 0,845. Nella revisione alla formula (1938), la correlazione multipla è corretta in 0,869.
[22] Numero di parole non presenti tra le prime 1.000 della lista di Thorndike (*Theacher's Word Book*, 1921)

La formula Winnetka, oltre ad essere la prima a misurare i testi in base al livello di lettura dei bambini, è anche la prima a considerare l'influenza della struttura della frase sulla difficoltà di lettura; gli studi precedenti si basavano infatti sul solo vocabolario dei testi.
Il modello della formula di Vogel e Washburne è seguito da Lewerenz 1929, Ojemann 1933[23], Dale e Tyler 1934, Gray e Leary 1935, Lorge 1939, Flesch 1943 e Dale e Chall 1948.

2.3. Dale e Tyler: la prima formula per gli adulti

Nel 1934 E. Dale e R. Tyler pubblicano la prima formula di leggibilità destinata agli adulti.
Il contributo del loro studio è stato l'uso di materiali specificamente progettati per adulti con capacità di lettura limitate. Essi considerano infatti inadeguati, e spesso troppo difficili, i materiali proposti nelle biblioteche.
Dale e Tyler partono dal presupposto che è impossibile determinare i fattori che intervengono nel processo di comprensione dei testi a meno che non si tengano separati quelli relativi ai materiali stessi da quelli esterni, legati al lettore: "the reader's interest in the topic treated in the reading matter, his ability to read, the kind of comprehension appropriate to the purposes of the reading matter, and the difficulty of the ideas developed in the reading matter are all factors which greatly affect his comprehension of the material read but are distinct from the characteristics involved in the materials themselves which may be changed so as to make these ideas understandable to adults of limited reading ability" (1934, p. 384).
I due studiosi compiono quindi un'indagine esplorativa dei fattori "interni" che influenzano la difficoltà dei testi, mantenendo costanti il gruppo di lettori, l'argomento trattato e lo scopo della lettura. Gli autori scelgono come criterio una selezione di 74 brani tratti da riviste, quotidiani, libri di testo, ecc. che hanno come argomento quello della salute personale[24]. La difficoltà di ogni specifico brano è valutata tramite test a scelta multipla.
Tra i 25 diversi fattori trovati significativi per la comprensione, Dale e Tyler ne selezionano 3, particolarmente predittivi della difficoltà:

- numero di termini tecnici differenti
- numero di parole difficili non tecniche
- numero di frasi indeterminate

Combinando questi parametri in una formula, risulta una correlazione multipla di 0,511 con i testi criterio.

2.4. Gray e Leary: *What Makes A Book Readable*

Nel 1935 W. S. Gray e B. Leary pubblicano il loro libro *What Makes a Book Readable*, divenuto un punto di riferimento per la metodologia usata. Si tratta di una ricerca esplorativa sui fattori che in qualche modo possono contribuire alla leggibilità, condotta

[23] Alfred S. Lewerenz (1929, 1935, 1939) ha prodotto diverse formule di leggibilità per il distretto scolastico di Los Angeles. Ralph Ojemann (1934) non ha sviluppato una formula, ma un metodo di valutazione della difficoltà di materiali destinati all'educazione degli adulti.
[24] D. Waples e R. Tyler hanno pubblicato uno studio sugli interessi di lettura degli adulti (*What People Want To Read About,* 1931). Intervistando 107 gruppi diversi hanno dimostrato che il tema della salute personale era un argomento di grande interesse per tutti.

attraverso interviste ad insegnanti, editori, autori e redattori. Gli autori identificano ben 289 elementi, che possono essere raggruppati in 4 categorie:

- contenuto
- stile e presentazione
- formato
- organizzazione

Il contenuto, con un leggero margine sullo stile, viene considerato il più importante. Non potendo però misurare questo criterio in modo preciso, Gray e Leary si concentrano su 64 variabili dello stile, legate alla struttura sintattica del testo o a scelte lessicali.
Gli autori calcolano quindi le correlazioni tra le misure di queste variabili e i punteggi delle prove effettuate sui testi criterio. Il set di testi include 48 brani di circa 100 parole ciascuno, per la maggior parte appartenenti al genere della narrativa, tratti da libri, riviste, quotidiani. La difficoltà dei brani è valutata con test di comprensione *(l'Adult Literacy Test* degli stessi autori) effettuati su circa 800 soggetti.
Solo 20 di questi indici presentano correlazioni significative; tra questi, cinque sono presi per la formula definitiva[25]:

- lunghezza media della frase, misurata in parole
- numero di parole difficili (parole non presenti nella lista di Dale di 769 parole ad alta frequenza[26])
- numero di pronomi personali
- percentuale di parole diverse
- numero di frasi preposizionali

La formula di Gray e Leary raggiunge una correlazione di 0.645 con i testi criterio.
Fino almeno agli anni Ottanta gli sviluppatori delle formule di leggibilità seguono lo stesso modello, escludendo gli aspetti di contenuto, organizzazione e formato, "più per la difficoltà che comportava l'introduzione di queste variabili che per la scarsa considerazione del loro peso. […] Fu presto evidente che la capacità di predizione di formule complesse non giustificava lo sforzo necessario per applicarle; la ricerca si indirizzò quindi verso formule di più facile uso e tuttavia capaci di una buona potenzialità predittiva" (Lucisano 1992, p. 30).
I ricercatori iniziano quindi a concentrarsi su due sole componenti, considerate come maggiormente predittive della difficoltà testuale: la variabile semantica, come misura della difficoltà del vocabolario e la variabile sintattica.

2.5. Irving Lorge e la ricerca di un criterio per lo sviluppo della formula

Irving Lorge, autore di *The Semantic Count of the 570 Commonest English Words* (1938) e co-autore insieme a Thorndike del *The Teacher's Word Book of 30,000 Words* (1944),

[25] I 20 fattori presentano una correlazione di almeno 0,35 ma nessuna supera lo 0,52. La correlazione più alta si ha con lunghezza media della frase (- 0,52). Le altre correlazioni sono: numero di parole difficili (- 0,50), numero di pronomi personali (0,48), numero di parole diverse (- 0,38), numero di frasi preposizionali (- 0,35).
[26] E. Dale, *A Comparison of Two Word Lists*, in Educational Research Bulletin, X (December 9, 1931), p. 484. L'elenco è costruito sulla base di altre due liste: l'*International Kindergarten Union List* (1928) e le prime 100 parole del *Teacher's Word Book* di Thorndike.

collabora al Readabilty Laboratory del Teachers College della Columbia University come assistente di Lyman Bryson[27]. Lorge si occupa del problema della misurazione della leggibilità e della comprensibilità: "if readability of a passage could be evaluated adequately, the estimate would have two major values – one, placing the book on some scale of comprehensibility, the other, indicating to writers of books for specified populations, the nature of the difficulty of their product. […] One criterion for readability is its comprehensibility, or negatively, its difficulty on a scale of comprehension" (Lorge 1939, p. 229).

Nel 1939 Lorge pubblica un articolo, *Predicting Reading Difficulty of Selections for Children*, con l'intento di riportare l'attenzione sulla valutazione della difficoltà dei testi, nella ricerca di un criterio soddisfacente e definito con sufficiente rigore. Fino ad allora il criterio impiegato per la valutazione della difficoltà è quello di Vogel e Washburne (1928), ovvero l'abilità di lettura dei bambini (misurata tramite lo *Stanford Achievement Test*) e i loro giudizi sui libri letti; per gli adulti, è invece impiegato il criterio di Dale e Tyler (1934) e Gray e Leary (1935), ovvero il punteggio di comprensione della lettura misurato tramite test ideati espressamente per gli adulti.

L'intento di Lorge è quello di applicare il metodo di Grey e Leary alla comprensione della lettura da parte di bambini, così da ottenere una semplice formula che misura la difficoltà di lettura in base al livello di istruzione (anni di scuola che servono per comprendere un dato testo). Come testi criterio, lo studioso sceglie 376 brani tratti da *Standard Test Lessons in Reading* di McCall e Crabbs (1926). I brani sono divisi in livelli di difficoltà sulla base del numero di risposte corrette alle domande poste alla fine di ciascun brano, secondo il punteggio della *Reading Scale* di Thorndike e McCall.

Utilizzando gli stessi 5 indici di Grey e Leary, Lorge valuta le intercorrelazioni tra le diverse variabili con il criterio, con l'aggiunta di un ulteriore elemento, detto *indice ponderato delle parole difficili*, ottenuto dando a ciascuna parola un peso in base alla frequenza di occorrenza secondo la lista di Thorndike (*The Theacher's Word Book of 20.000 Words*) e dividendo per il numero di parole totali.

Le correlazioni multiple ottenute risultano essere maggiori di quelle di Grey e Leary[28]. Le variabili con il più alto valore predittivo sono quelle legate al vocabolario impiegato: "vocabulary load is the most important concomitant of difficulty" (Lorge 1939, p. 229).

La formula di Lorge è descritta in un articolo del 1944, *Predicting Readability*; l'indice si basa su tre delle variabili impiegate da Grey e Leary: lunghezza media della frase (*sl*), numero di frasi preposizionali su 100 parole (*pp*), numero di parole difficili su 100 parole (*wd*):

---

[27] L. Bryson si è occupato della formazione degli adulti a New York; nel 1936 ha istituito il Readability Laboratory presso la Columbia University con lo scopo di raccogliere tutto il materiale fino ad allora conosciuto sulla leggibilità e di mettere queste conoscenze al servizio di una produzione di testi leggibili destinati agli adulti. Il suo più grande contributo, probabilmente, è stata l'influenza sui suoi due allievi, Irving Lorge e Rudolph Flesch.

[28] Usando l'indice ponderato, il numero di frasi preposizionali, la percentuale di parole diverse, la lunghezza media della frase, il numero di parole difficili e il numero dei pronomi personali, la correlazione multipla è 0,7722. Eliminando i pronomi personali, la correlazione diventa 0,7721; togliendo la percentuale di parole diverse diventa 0,7711; lasciando fuori anche l'indice di Thorndike la correlazione diventa 0.7669. Considerando solo le frasi preposizionali e le parole diverse, la correlazione multipla è 0,7456; solo la lunghezza media della frase e il numero di parole diverse è 0,7406; solo le frasi preposizionali e la lunghezza media della frase è 0,6949.

$$Grade = 0{,}07sl + 0{,}1073wd + 0{,}1301pp + 1{,}6126$$

Nel 1948 escono due articoli, uno dello stesso Lorge e uno di Dale e Chall, con alcune correzioni alla formula[29]. Gli autori hanno infatti trovato un errore computazionale che riguarda il coefficiente di correlazione tra la lunghezza della frase e il criterio, originariamente riportato come 0,6174. Il coefficiente viene corretto in 0,4681 da Dale e Chall e 0,467 da Lorge.
Anche se ideata per la stima della difficolta di lettura dei bambini, la formula di Lorge è ampiamente usata anche per gli adulti ed è una delle prime formule di leggibilità di facile applicazione. L'uso di *Standard Test Lessons in Reading* di McCall e Crabbs come criterio semplifica notevolmente il problema di abbinare i lettori ai testi (Klare 1985) e rimane lo standard per le formule di leggibilità fino agli studi di J. Bormuth nel 1969.

2.6. Rudolf Flesch: *The art of Plain Talk*

È soprattutto R. Flesch a divulgare attraverso i suoi libri[30] e i suoi articoli il concetto di leggibilità e a pubblicizzare l'esigenza del *plain talk*.
Flesch nasce in Austria e si laurea in giurisprudenza presso l'Università di Vienna nel 1933; pratica legge fino al 1938, quando giunge negli Stati Uniti come rifugiato politico. Dal momento che la sua laurea in legge non viene riconosciuta, si dedica ad altri studi: nel 1939 riceve una borsa di studio per rifugiati presso la Columbia University e nel 1940 si laurea con lode in biblioteconomia. Nello stesso anno, diventa assistente di Lyman Bryson al Readabilty Lab del Teachers College della Columbia University, assieme a Lorge.
Nel 1942 consegue un master in Educazione degli adulti e l'anno successivo un dottorato di ricerca in Ricerca educativa. Nella sua tesi, *Marks of a Readable Style* (1943), Flesch pubblica la sua prima formula di leggibilità; la formula viene poi ripresentata nel 1946 in *The art of Plain Talk*, scritto intenzionalmente da Flesch in "modo leggibile". Nell'introduzione si legge: "about two years ago, I published my Ph.D. dissertation "Marks of a Readable Style", which contained a statistical formula for measuring readability. The dissertation was quite a success, as dissertations go, and the formula is now being used in many organizations and government agencies. This has been gratifying, but also somewhat embarrassing to me: for "Marks of a Readable Style", being a Ph.D. dissertation, was not a very readable book. I tried to rewrite it in simple language, but when I was through, a natural thing had happened ad I had written a new book. This is the book".
Flesh è convinto che le formule fino ad allora elaborate non abbiano individuato quei caratteri che incidono maggiormente sulla leggibilità; in particolare, esse forniscono indicazioni esatte per quanto riguarda la valutazione della difficoltà di lettura dei bambini ma non riescono a registrare quelle degli adulti.

[29] I. Lorge, *The Lorge and Flesch Readability Formulae: A Correction*, School and Society, Vol. 67, pp. 141-142, 21 febbraio 1948. - E. Dale, J. S. Chall, A *Formula for Predicting Readability* in Educational Research Bulletin, Vol. 27, No. 1 (Jan. 21, 1948), pp. 11-20+28.
[30] Tra i principali: *The Art of Plain Talk* (1946), *The Art of Readable Writing* (1949), *The Art of Clear Thinking* (1951), *Why Johnny Can't Read - And What You Can Do About It* (1955), *The ABC of Style: A Guide to Plain English* (1964), *How to Write in Plain English: A Book for Lawyers and Consumers* (1979).

La formula elaborata da Flesch si basa sul conteggio di tre elementi: lunghezza media della frase, numero di affissi e numero di riferimenti personali. La complessità delle frasi è un ottimo indice di difficoltà sia per i bambini che per gli adulti. Un altro buon indice di difficoltà è la quantità di affissi (prefissi, infissi, suffissi): gli affissi contraddistinguono in genere parole astratte, la cui comprensione richiede passaggi logici più complessi. L'altro fattore da considerare è l'interesse che uno scritto suscita nel lettore; per misurare questo elemento si contano i riferimenti personali: nomi propri, pronomi personali, nomi di persona, ecc.

Dopo la sua pubblicazione, la formula trova largo impiego in molti campi: quotidiani e riviste, pubblicità, pubblicazioni del governo, materiale per l'educazione degli adulti, libri di testo, libri per bambini, corsi di scrittura creativa, ecc. Il suo uso ne mostra la validità, ma al tempo stesso ne evidenzia anche i difetti. Uno di questi è la difficoltà di applicazione, ad esempio nel conteggio del numero di affissi; altre persone trovano complesso usare il sistema dei punteggi, che generalmente va da 0 (molto facile) a 7 (molto difficile); inoltre, il conteggio dei riferimenti personali viene considerato arbitrario.

Uno dei limiti maggiori è il troppo tempo che richiede la sua applicazione. Il tempo medio necessario per testare un campione di 100 parole è 6 minuti; questo rende l'applicazione della formula più veloce rispetto a formule più semplici, che richiedono un riferimento a liste di parole (ad esempio Gray e Leary o Lorge), ma ancora troppo lunga per l'uso pratico (Flesch 1948).

Per superare queste carenze e renderla più pratica, nel 1948 lo studioso pubblica una seconda formula. Flesch seleziona, come criterio, brani tratti da *Standard Test Lessons in Reading* di McCall e Crabbs (1926). I coefficienti di correlazione con i punteggi di difficoltà dei testi si basano in parte sui risultati statistici stabiliti dallo studio di Lorge; i livelli di difficoltà sono ottenuti da prove di comprensione effettuate sui bambini di scuole elementari. Si tratta di dati non proprio ottimali per misurare la facilità e l'interesse con cui leggono gli adulti ma sono gli unici disponibili al momento dello sviluppo e della revisione della formula.

Gli elementi considerati sono quattro:

- lunghezza media della frase misurata in parole;
- lunghezza media della parola misurata in sillabe (sostituisce il conteggio degli affissi; i risultati ottenuti sono simili ma risulta più semplice da misurare);
- percentuale di parole personali (elemento già impiegato nella precedente formula, viene ripreso con una nuova definizione: tutti i nomi con genere naturale, tutti i pronomi eccetto quelli neutri, la parola *people* 'persone' usata col verbo al plurale e la parola *folks* 'gente');
- percentuale di frasi personali (discorsi diretti, domande, richieste e altre frasi indirizzate direttamente al lettore, esclamazioni, frasi incomplete il cui significato è dedotto dal contesto; questo elemento è progettato per correggere il difetto strutturale della formula precedente, che non sempre riusciva a mostrare l'alta leggibilità dei periodi contenenti il discorso diretto).

La correlazione multipla dei quattro elementi con il criterio non mostra però un aumento significativo rispetto al valore predittivo della formula precedente. Flesch decide allora di calcolare due correlazioni multiple: una impiegando i primi due elementi, l'atra usando gli

ultimi due. Su queste intercorrelazioni costruisce quindi una formula in due parti: una misura l'indice di difficoltà e l'altra l'interesse che il testo suscita nel lettore.
Nella prima parte, *Reading Ease*, le variabili considerate sono il numero di sillabe per parola (cioè la lunghezza media delle parole) e il numero di frasi ogni 100 parole (cioè la lunghezza media delle frasi). Il numero di sillabe per parola viene assunto come indice di difficoltà semantica, mentre il numero di parole per frase viene assunto come indice di complessità sintattica.
La relazione tra la lunghezza delle frasi e la difficoltà è intuitivamente comprensibile: più una frase è lunga, più è sintatticamente complicata. La relazione tra la lunghezza delle parole e la difficoltà si basa sui risultati della statistica linguistica che stabilisce un rapporto tra la frequenza delle parole e la loro lunghezza: le parole più frequenti sono in genere le più brevi; esse sono generalmente parole concrete, con pochi affissi; i concetti astratti sono invece espressi da termini lunghi e composti. "Le parole frequenti sono quelle familiari, e le parole frequenti e familiari sono brevi, e le parole frequenti, familiari e brevi hanno pochi affissi, e lo scrittore che usi parole frequenti familiari brevi senza affissi presenta un basso rapporto tipo-replica" (Miller 1972, pp. 190-191).
La formula *Reading Ease* è la seguente:

$$\text{Reading Ease Score} = 206{,}835 - 0{,}846\text{W} - 1{,}015\text{S}$$

dove
W = numero medio di sillabe per parola (*Word*), ottenuto dividendo il numero di sillabe per il numero di parole;
S = numero medio di parole per frase *(Sentence)*, ottenuto dividendo il numero di parole per il numero di frasi;
206,835 è un coefficiente numerico scelto per fare in modo che i valori oscillino da 0 a 100.

Questo indice ha una correlazione di 0,70 con i brani dello *Standard Test Lessons in Reading* di McCall e Crabbs del 1926 e di 0,64 con la versione dello stesso test del 1950.
La formula, di facile applicazione anche per i non addetti ai lavori, predice la facilità di lettura su una scala da 1 a 100; un punteggio di 100 significa che un testo è molto semplice e che un bambino che ha completato la quarta elementare (5° grado) risponderà correttamente a ¾ di domande del test; un testo con un punteggio inferiore a 30 è considerato molto difficile e può essere capito da chi ha almeno una laurea.

Risultato	Descrizione dello stile	Parole per frase	Sillabe su 100 parole	Tipo di rivista	Livello di istruzione
0 - 30	Molto difficile	≥ 29	≥ 192	Scientifica	College
30 - 50	Difficile	25	167	Accademica	13° - 16° grado
50 - 60	Abbastanza difficile	21	155	Di qualità	10° - 12° grado
60 - 70	Normale	17	147	*Digests*	8° - 9° grado
70 - 80	Abbastanza facile	14	139	Narrativa leggera	7° grado
80 - 90	Facile	11	131	Narrativa scadente	6° grado
90 - 100	Molto facile	≤8	≤ 123	Fumetti	5° grado

Tabella 1. Interpretazione degli indici di facilità di lettura (Flesh 1948).

In realtà la formula, così come le altre finora analizzate, ha un buon valore predittivo solo fino al 7° grado; oltre questo, sottovaluta il livello di istruzione. Questo dipende probabilmente dal fatto che la maggior parte di queste formule sono basate su materiali destinati ai bambini (Vogel e Washburne, Lorge, Flesch) o destinati ad adulti con limitate capacità di lettura (Dale e Tyler, Gray e Leary).
L'indice ha comunque molto successo e trova una lunga serie di applicazioni: viene impiegata per valutare articoli di giornale, romanzi, testi pubblicitari, documenti governativi, contratti di assicurazione, testi scolastici; viene anche insegnata in vari corsi presso diverse università. Sono stati fatti anche molti tentativi di calcolo automatico della formula tramite programmi informatici.
La seconda parte della formula misura l'interesse umano (*Human interest*) e si ottiene contando il numero di vocaboli personali (pronomi personali, nomi propri, nomi di persona) e di frasi personali (citazioni, esclamazioni, discorsi diretti, frasi incomplete, frasi che contengono una domanda, un ordine, una richiesta diretta al lettore). La formula è la seguente:

$$\text{Human Interest } = 3{,}64\text{p} + 0{,}31\text{f}$$

dove
p = percentuale di parole personali su 100 parole;
f = percentuale di frasi personali su 100 frasi.

La formula non contiene costanti statistiche. Il coefficiente di correlazione multipla è 0,43.
Se il valore si avvicina a 0 significa che ci sono pochi riferimenti personali; al contrario, se si avvicina a 100 indica che il testo è ricco di riferimenti personali ed è considerato molto semplice. Un buon romanzo ad esempio ottiene un punteggio di interesse umano compreso tra 60 e 100, una rivista di cultura (*il New Yorker*) tra 40 e 60, le riviste commerciali tra 10 e 20 e le opere scientifiche tra 0 e 10.

Risultato	Descrizione dello stile	Percentuali di parole personali	Percentuali di frasi personali	Carattere della rivista
0 – 10	Noioso	≤ 2	0	Scientifica
10 - 20	Abbastanza interessante	4	5	Commerciale
20 - 40	Interessante	7	15	*Digests*
40 - 60	Molto interessante	11	32	*New Yorker*
60 - 100	Drammatico	≥ 17	≥ 58	Narrativa

Tabella 2. Interpretazione degli indici di interesse umano (Flesh 1948).

Da questo momento Flesch produce "una vera e propria girandola di formule" (Biagioli et al. 1984). Nel 1950 pubblica la formula per misurare il *livello di astrazione*, che impiega come variabili il conteggio delle *definits words* e la lunghezza delle parole in sillabe. Le *definits words* sono una lista di parole che danno concretezza al testo, cioè nomi propri, nomi comuni con un significato specifico, aggettivi possessivi, pronomi personali, relativi, riflessivi, negazioni. Seguono la formula che misura il *realismo* e la *vivacità* (1954), quella che misura il *formalismo*/*colloquialità* (1958)[31]. Tutte queste formule incontrano subito critiche e obiezioni da parte degli studiosi del settore.

2.7. La formula di Dale e Chall

Quando la prima formula di Flesch viene rilasciata (1943), Edgar Dale (professore di pedagogia presso l'Ohio State University) e Jeanne Chall (fondatrice e direttrice del Reading Laboratory di Harvard) stanno lavorando alla valutazione dei materiali didattici pubblicati dalla National Tuberculosis Association, con il compito di analizzare gli opuscoli già pubblicati e riscriverli in modo da renderli comprensibili per un adulto medio. I due studiosi usano la formula di Flesch per misurare la difficoltà di questi materiali ed individuano alcune carenze: "The most serious shortcoming was the count of affixes, which we found to be rather arbitrary, in the sense that two people making a count on the same sample would usually come out with a different number of affixes. […] The second shortcoming of the Flesch formula was the count of personal references. In our numerous analyses we found that the personal-reference count was not a reliable index of difficulty" (Dale e Chall 1948, p. 12-14).

Per correggere queste lacune, Dale e Chall sviluppano una propria formula (1948). Come testi campione gli studiosi utilizzano *Standard Test Lessons in Reading* di McCall e Crabbs (1926), servendosi delle schede tecniche relative ai brani compilate da Lorge, le quali includono il conteggio degli affissi e dei riferimenti personali di Flesch. In ognuno dei brani contano inoltre il numero di parole che non appartengono alla lista di Dale delle 3.000 parole più frequenti, la quale include a sua volta quella di Thorndike. Si tratta di vocaboli ritenuti di uso comune e di facile comprensione; l'80% dei vocaboli della lista è infatti conosciuto da bambini che frequentano la quarta elementare.

Vengono calcolate le correlazioni multiple di quattro elementi: lunghezza media della frase, numero di affissi e di riferimenti personali di Flesch, numero di parole difficili di Lorge

[31] Cfr. Klare 1963.

(parole non presenti nella lista di Dale di 769 vocaboli ad alta frequenza) e numero di parole che non appartengono alla lista di Dale delle 3.000 più frequenti.
Vi è un'alta correlazione tra il criterio e il numero di parole fuori dalla lista di Dale (0,6833). I punteggi successivi sono il numero di parole difficili di Lorge e il numero di affissi di Flesch. L'intercorrelazione tra questi 3 fattori è alta (Dale-Flesch: 0,7932, Flesch-Lorge: 0,7441, Dale-Lorge: 0,7988). Questo avvalora la scoperta di Lorge, cioè che il vocabolario impiegato è il fattore più determinante per la difficoltà di lettura e che tutte le misure del vocabolario sono strettamente correlate. La seconda variabile con il più alto valore predittivo è la lunghezza media della frase (la correlazione con il criterio è 0,4681).
In base a queste correlazioni, Dale e Chall costruiscono una formula a due variabili, che utilizza la lunghezza delle frasi come misura della difficoltà sintattica e il numero di parole non comuni come difficoltà semantica. La formula è la seguente:

$$reading\ grade\ score\ = 0{,}1579x_1 + 0{,}0496x_2\ + 3{,}6365$$

dove
x_1 indica il numero di parole fuori dalla lista di Dale
x_2 indica la lunghezza media della frase
3,6365 è una costante

Il coefficiente di correlazione multipla dei due fattori con il criterio è 0,70. Per gli adulti, il punteggio del grado di istruzione va interpretato come numero di anni di scuola richiesti per capire il materiale.

Formula Score	**Corrected Grade Levels**
4.9 and below	Grade 4 and below
5.0 to 5.9	Grades 5-6
6.0 to 6.9	Grades 7-8
7.0 to 7.9	Grades 9-10
8.0 to 8.9	Grades 11-12
9.0 to 9.9	Grades 13-15 (college)
10 and above	Grades 16 and above (college graduate)

Tabella 3. Interpretazione dei punteggi per la formula di Dale e Chall.

La formula riscuote un grande successo negli Stati Uniti, soprattutto tra gli insegnanti.
Dale e Chall (1948), tuttavia, invitano ad essere cauti nell'applicazione meccanica delle formule di leggibilità: "we must be cautious about "writing for a readability formula." We must remember at all times that a formula is a statistical device. It means that, on the whole, longer sentences make comprehension more difficult. This does not mean that all long sentences are hard to read and understand. There are some very short sentences that

may be harder to comprehend than longer ones. The same holds true for the use of familiar words. On the whole, the more unfamiliar the words used, the harder the material will be to understand. But sometimes familiar words are used in a symbolic or metaphoric sense. "To be or not to be" is not an easy idea although the sentence is short and the separate words used would usually be called simple and familiar ones. Readability formulas are not sensitive to such subtle variations in meaning" (p. 20).

2.8. Farr, Jenkins e Paterson: modifiche alla formula di Flesch

Nel tentativo di semplificare la formula di Flesch, J. N. Farr, J. J. Jenkins e D. G. Paterson (1951) propongono una nuova versione dell'indice *Reading Ease*. Gli autori ritengono che sostituire il conteggio delle sillabe con il conteggio dei monosillabi possa ridurre il tempo di applicazione dell'indice ed eliminare la necessità per l'analista di conoscere le regole di sillabazione; il valore predittivo della formula rimarrebbe invece invariato (la correlazione tra le due variabili – conteggio delle sillabe e monosillabi - è di 0,91).
La formula modificata considera quindi la lunghezza media della frase misurata in parole e il numero di parole monosillabiche su 100 parole:

$$\text{New Reading Ease Index} = 1{,}599\text{os} - 1{,}015\text{s} - 31{,}517$$

dove:
os = numero di monosillabi (*one-syllable*) ogni 100 parole
s = numero medio di parole per frase *(sentence)*

I punteggi della nuova formula su due serie di brani campione hanno una correlazione di 0,93 e 0,95 con i punteggi della vecchia formula sugli stessi brani.

2.9. Robert Gunning e il Fog Index

Robert Gunning è tra i primi studiosi ad applicare le ricerche di leggibilità al mondo del lavoro. Nel 1935 entra nel settore dell'editoria scolastica; gli educatori sono molto preoccupati per il fatto che molti studenti si diplomano ma non sono in grado di leggere e comprendere neanche i quotidiani. Gunning si rende conto che gran parte dei problemi legati alla "lettura" dipendono da problemi legati alla "scrittura": i testi dei giornali e i documenti che riguardano il lavoro quotidiano sono infatti pieni di nebbia (*fog*) e di inutili complessità.
Nel 1944 fonda la Robert Gunning Associates, la prima società di consulenza specializzata sulla leggibilità. Gunning si rivolge a scrittori che vogliono migliorare la propria scrittura, a giornalisti e soprattutto al personale di aziende che si trova a dover scrivere per lavoro senza essere un professionista del settore (commercio e industria, governo, forze armate). La consulenza è particolarmente necessaria durante il periodo della Seconda Guerra Mondiale, quando una migliore scrittura e una maggiore facilità di lettura diventano strumenti fondamentali per la comunicazione di massa.

In *The Technique of Clear Writing* (1952)[32], Gunning descrive come applicare la propria formula di leggibilità destinata agli adulti, il *Fog Index*. Invece di contare il numero di sillabe (Flesch) o le parole monosillabiche (Farr, Jenkins, e Paterson), lo studioso propone un conteggio delle parole polisillabiche. La formula utilizza due variabili, la lunghezza media delle frasi e il numero di parole difficili (numero di parole con più di due sillabe per ogni 100 parole):

$$Reading\ Grade\ Level = 0{,}4\ (lunghezza\ media\ delle\ frasi\ + percentuale\ di\ parole\ difficili)$$

L'indice correla perfettamente con i livelli di lettura divisi per grado di istruzione definiti dallo *Standard Test Lessons in Reading* di McCall eCrabbs.

Punteggi	Estimated Reading Grades
17	College graduate
16	College senior
15	College junior
14	College sophomore
13	College freshman
12	High school senior
11	High school junior
10	High school sophomore
9	High school freshman
8	Eight grade
7	Seventh grade
6	Sixth grade

Tabella 4. Interpretazione dei punteggi dell'Indice Fog.

Il punteggio ideale per una buona leggibilità è 7 o 8; qualunque testo al di sopra di 12 è troppo complesso per la maggior parte dei lettori. Per fare un esempio, la Bibbia, Shakespeare e Mark Twain hanno un Indice Fog di circa 6; le principali riviste come il *Time*, il *Newsweek* e il *Wall Street Journal* hanno una media di 11.
La validazione dell'Indice Fog non viene mai pubblicata. Secondo i calcoli dell'autore, tuttavia, la formula ha una correlazione di 0,93 con i testi di Chall et al. 1996[33].
La formula di Gunning, divenuta popolare grazie alla sua facilità d'uso, viene scelta dall'esercito, dalla marina e dall'aeronautica per valutare i loro manuali di scrittura.
Gunning sottolinea più volte che si tratta di un semplice sistema di allarme, di uno strumento che va impiegato come controllo dopo la "scrittura" e non di un modello per scrivere o di una regola per "scrivere bene": "although we have often given permission for reprinting the Fog Index, our means of measuring reading difficulty, we have sometimes

[32] Gunning 1952, l'edizione a cui si fa riferimento in questo capitolo è quella aggiornata del 1968.
[33] Chall, J. S., G. L. Bissex, S. S. Conard, and S. Harris-Sharples. 1996. *Qualitative assessment of text difficult: A practical guide for teachers and writers.* Cambridge, MA: Brookline Books.

cringed at the use made of it. In our work we emphasize that the Fog Index is a tool, not a rule. It is a warning system, not a formula for writing. Testing without the support of experienced analysis can be detrimental" (Gunning 1969, p. 10). E ancora: "Besides, we never looked upon the Fog Index as a rule for writing or a "scientific" key to the mysteries of readability. Anyone who loves the language and has used it effectively knows that to expect a yardstick to do such a job is folly. […] The Fog Index serves as a simple warning system. No formula will guarantee that you write well" (p. 12). In ogni caso, conclude Gunning, "I don't know of a better warning system presently at hand" (p. 13).

2.10. Powers, Sumner e Kearl: nuove versioni di formule classiche

Powers, Sumner e Kearl (1958) ricalcolano le formule di Flesch (*Reading Ease*), Dale e Chall, Farr, Jenkins e Paterson, e l'indice Fog in base alla nuova revisione dei brani di McCall e Crabbs *Graded Test Lessons in Reading* (1950), stabilendo una comparazione diretta tra i vari indici.
La loro versione della formula *Reading Ease* è la seguente:

$$Reading\ Ease = -2{,}2029 + 0{,}0455\mathrm{W} + 0{,}0778\mathrm{S}$$

La correlazione con i brani dello *Standard Test Lessons in Reading* di McCall e Crabbs del 1926 era di 0,70 mentre con questa versione è di 0,64.
La formula di Dale e Chall ricalcolata è invece:

$$reading\ grade\ score\ = 0{,}1155x_1 + 0{,}0596x_2\ + 3{,}2672$$

Il coefficiente di correlazione multipla dei due fattori con i brani è 0,71, praticamente lo stesso valore trovato nella versione precedente (0,70). Questo suggerisce che la formula di Dale e Chall è l'indice più preciso a disposizione almeno fino al 1960 (Klare 1974).
La nuova versione della formula Farr, Jenkins e Paterson ha un punteggio di correlazione di 0,58 con il nuovo criterio:

$$\text{New Reading Ease Index} = 0{,}0648\mathrm{os} + 0{,}0923\mathrm{s} + 8{,}4335$$

L'indice Fog ricalcolato è il seguente (la correlazione con i testi è 0,59):

$$Grade\ level = 3{,}0680 + 0{,}0877\ (lunghezza\ media\ della\ frase) + 0{,}0984\ (percentuale\ di\ monosillabi)$$

2.11. La formula Spache

Nel 1953 George D. Spache sviluppa una formula di leggibilità per valutare i materiali destinati ai bambini dalla prima alla terza elementare. Nessuna delle formule precedenti infatti si applica a livelli di istruzione inferiori al 4° grado (Flesch ad esempio parte dal 5° grado, Gunning dal 6°, cioè dalla prima media). La formula è la seguente:

$$Grade\ level = 0{,}141\ \mathrm{x}_1\ +\ 0{,}086\ \mathrm{x}_2\ +\ 0{,}839$$

dove:
x_1 indica la lunghezza media delle frasi misurata in parole;
x_2 indica il numero di parole difficili (parole non presenti nella lista delle 769 parole di Dale)

Spache convalida la sua formula testando 152 libri destinati ai gradi 1-3 e trova un coefficiente di correlazione multipla di 0,818.
Clarence R. Stone (1956) ritiene che la precisione della formula di Spache potrebbe essere aumentata sostituendo 173 parole della lista di Dale con un numero uguale di vocaboli tratti da *The Author's Word List* di L. L. Krantz e *A Graded Vocabulary for Primary Reading* dello stesso Stone[34]. Utilizzando l'elenco rivisto come nuovo parametro della formula, i materiali ottengono un grado di difficoltà leggermente inferiore rispetto alla valutazione di Spache.
Nelle sue pubblicazioni successive (1966), Spache segue questa procedura, utilizzando la *Revised Word List* di Stone. Nel 1974 rivede la sua formula, utilizzando il *Basic Elementary Reading Vocabularies* di Harris e Jacobson (1972) e ottenendo un punteggio di correlazione multipla di 0,95.

2.12. Il cloze test

Nel 1953 W. L. Taylor presenta il cloze test come nuova tecnica di misurazione della comprensione dei testi[35]. La procedura consiste nel cancellare determinate parole da un brano, per farle poi reintegrare dallo studente con l'aiuto degli indizi forniti dal testo. Le parole vengono eliminate in modo casuale, senza tenere conto della loro funzione o del loro significato specifico. Di solito si toglie una parola ogni cinque; al loro posto vengono inseriti degli spazi bianchi di misura standard, in modo da non influenzare la ricerca della parola mancante.
Il punteggio viene calcolato sulla base della percentuale di buchi riempiti, ovvero di parole esatte mancanti inserite. Se lo studente individua facilmente e correttamente le parole cancellate, il testo è più leggibile di un altro testo che presenta difficoltà di completamento. Il fatto che la scelta delle parole da eliminare avvenga in modo casuale consente di superare tutta una serie di obiezioni che erano state mosse per i test a scelta multipla. Taylor (1953), che applica principalmente il cloze come strumento per valutare la leggibilità, lo descrive così: "Si può pensare alla procedura cloze come al mettere in pentola tutte le variabili che possono influenzare la leggibilità, lasciarle interagire per poi scremare i risultati".
Come osserva Lucisano (1992), alla base del lavoro di Taylor ci sono la teoria dell'informazione e la psicologia della Gestalt. Dalla teoria dell'informazione ricava l'idea di misurare la quantità di ridondanza di un testo, cioè quella parte di testo che può essere cancellata senza perdita di informazione; dalla teoria della Gestalt riprende il concetto di pattern di completamento. La mente umana tende a percepire un insieme di elementi

[34] L. L. Krantz, *The Author's Word List*, Curriculum Research Co, Minneapolis, 1945 - C. R. Stone, *A Graded Vocabulary for Primary Reading*, Webster Publishing Co, St. Louis, Missouri, 1936
[35] Cfr. Taylor 1953. Sulla procedura cloze si veda inoltre Taylor 1956, Bormuth 1965, 1967, 1968, 1969 e in italiano Marello 1984, 1989, 1991, De Grafenstein e Pierdonati 1985, Lucisano 1989, 1992, Nuccorini 2001, Chiari 2002.

come un'unità formale, piuttosto che come parti distinte, per cui tende a completare certe strutture familiari anche quando elementi della struttura stessa sono stati cancellati. Questa tendenza porta ad esempio chiudere inconsciamente le figure aperte, creando le informazioni mancanti per completare il pattern. "Al cuore di questa procedura c'è una unità di misura funzionale che a titolo sperimentale ho chiamato cloze. Si pronuncia come il verbo dose e deriva da closure[36]. Quest'ultimo termine è usato dalla psicologia della Gestalt in riferimento alla tendenza umana a completare le strutture familiari ma incomplete, a vedere un cerchio interrotto come un intero, per esempio, completando mentalmente le parti mancanti" (Taylor 1953).
Si riporta di seguito un esempio di un test cloze classico:[37]

Dimenticavo di dire che (1)_____ signora Teresa ha avuto (2)_____ bella idea di presentarmi (3)_____ suoi parenti, facendomi passare (4)_____ un suo nipote "ospite (5)_____ di lei per un (6)_____ periodo di convalescenza," e (7)_____, colto di sorpresa, non (8)_____ la prontezza di contraddirla, (9)_____ dato il via a (10)_____ reazione a catena di (11)_____.
Ben presto mi ritrovo (12)_____ in un intrico crescente (13)_____ parentele. Incontro persone che (14)_____ di avermi visto nascere. (15)_____ cugini ormai non si (16)_____ più; il cortile è pieno (17)_____ zii al sole, la (18)_____ invasa dalle zie, e (19)_____ gabinetto sempre occupato dai "(20)_____ nipotini."
Come se ciò (21)_____ bastasse, ecco che, potenza (22)_____ parte, mi ritrovo ad (23)_____ a braccia tese un (24)_____ venuto, a far da (25)_____ a qualche smarrita famigliola, (26)_____ a consolare qualche sconosciuta (27)_____ in lacrime.

Risposte:
1) la 2) la 3) ai 4) per 5) presso 6) lungo 7) io 8) trovando 9) ho 10) una 11) equivoci 12) impigliato 13) di 14) giurano 15) I 16) contano 17) di 18) cucina 19) il 20) cari 21) non 22) della 23) accogliere 24) nuovo 25) guida 26) o 27) signora

Come si può osservare, la procedura garantisce una generale varietà nei tipi di parole cancellati.
Oltre al cloze classico, esistono tutta una serie di varianti: dall'originaria cancellazione di parole, si è passati alla cancellazione di lettere, sillabe, morfemi, sintagmi, frasi, segni d'interpunzione. Una variante recente è la cancellazione di classi di parole: si possono eliminare le parole "vuote", come preposizioni e articoli o si possono occultare le parole "piene", come verbi, aggettivi, nomi, avverbi.
Nonostante il cloze test sia sviluppato da Taylor nel 1953, non viene utilizzato nelle formule di leggibilità fino al 1965[38]. La sua facilità e praticità di utilizzo lo rendono ben presto un

[36] Taylor non inventa solo la tecnica, ma anche il termine *cloze*. Egli ritiene che reintegrare nel testo le parole cancellate sia un esempio di *closure* e quindi decide di coniare il tecnicismo *cloze* per riferirsi alla sua procedura.
[37] L'esempio è tratto da Chiari 2002. Il brano è scelto da *L'ombra e la meridiana* di Paolo Maurensig, 1998.
[38] La procedura cloze è stata impiegata nello sviluppo e nella validazione delle formule di leggibilità prima da Coleman (1965 e 1967) poi da Bormuth (1966 e 1969).

criterio molto importante per determinare il grado di leggibilità dei testi. Il cloze produce infatti coefficienti di correlazione superiori rispetto allo *Standard Test Lessons in Reading* di McCall e Crabbs nel confronto con le stesse formule.
Oltre a questa destinazione iniziale, la procedura viene ampiamente utilizzata con varie finalità: controllo della leggibilità di giornali, libri di testo, rapporti aziendali, ecc. La tecnica risulta molto valida anche nella misurazione delle competenze testuali, in particolare per valutare la comprensione di testi in L2. Uno dei vantaggi del cloze è infatti la rapidità di preparazione e correzione delle prove e il fatto che anche un programma informatico possa cancellare in modo automatico le parole ad intervalli prestabili, verificare e calcolare le risposte corrette e anche il tempo impiegato per terminare il test.

La rassegna presentata finora mostra che gli studi sulla leggibilità si sono concentrati su due aspetti: da una parte la ricerca di un criterio adeguato a determinare la difficoltà di brani specifici, dall'altra l'identificazione di quegli elementi testuali che influenzano tale difficoltà e il modo migliore per misurarli.
Per quanto riguarda gli indici statistici correlati con la difficoltà, gli studiosi hanno preso in considerazione quasi esclusivamente variabili dello stile, legate alla struttura sintattica del testo o a scelte lessicali. Il vocabolario impiegato risulta avere il più alto valore predittivo. In particolare, nelle formule vengono misurati la diversità (il numero di parole diverse), il numero di parole non comuni, il numero di parole difficili, il numero di tecnicismi, l'interesse umano (ad esempio il numero di pronomi personali), la lunghezza delle parole (in realtà Flesch è l'unico ad impiegarla come variabile semantica).
La struttura sintattica viene misurata principalmente tramite la lunghezza della frase (il parametro è introdotto da Grey e Leary nel 1935 ed è mantenuto in tutte le formule successive) ma sono usati anche il numero di frasi semplici e la percentuale di frasi preposizionali.
Lo *Standard Test Lessons in Reading* di McCall e Crabbs (1926, 1950) sembra essere il miglior criterio per determinare la difficoltà di lettura dei brani. Viene scelto per la prima volta da Lorge (1939) e rimane lo standard per le formule di leggibilità fino agli anni Sessanta. Ogni brano ha un punteggio di difficoltà (già stabilito dagli autori) che può essere usato come parametro di confronto; il test misura la comprensione della lettura per ogni specifico brano. Questo è il suo punto di forza ma anche la sua debolezza di fondo.
Le valutazioni sono effettuate tramite un questionario e le domande sono pensate in base ai brani considerati. Ma le domande variano sia per quanto riguarda il linguaggio impiegato, sia a livello dei concetti considerati; se viene usato un linguaggio troppo difficile il lettore potrebbe non rispondere correttamente anche se ha compreso bene il brano. Il numero di risposte corrette, quindi, dipende anche dal tipo di domande sottoposte al lettore. La misurazione della difficoltà di un testo è inevitabilmente legata alla qualità delle domande che lo valutano. Questa indeterminatezza determina l'ambiguità del criterio (Lorge 1949).
L'introduzione del cloze test come nuovo metodo di validazione degli indici apre la strada a misurazioni più precise delle variabili testuali e della comprensione della lettura.

3. Nuove formule di leggibilità

La pubblicazione delle formule di Flesch, Dale e Chall e Gunning segna la fine del primo periodo di studi sulla leggibilità; gli autori hanno il merito di aver portato la questione della leggibilità all'attenzione del pubblico, stimolando l'esigenza di produrre (e leggere) testi in un linguaggio semplice e comprensibile. A questo primo momento segue un periodo di consolidamento e approfondimento delle ricerche, che arriverà fino agli anni Novanta[39].
Gli studiosi si sforzano di migliorare le formule attuali e renderle sempre più di facile applicazione. Un forte impulso alla ricerca è dovuto alla disponibilità di strumenti informatici che consentono di analizzare una grande quantità di testi e considerare un maggior numero di variabili senza però perdere il vantaggio della facilità d'uso.
Questo periodo è inoltre caratterizzato dallo sviluppo di formule di leggibilità per lingue diverse dall'inglese e dall'introduzione della procedura cloze come criterio per lo sviluppo degli indici.

3.1. La formula Devereaux

Nel 1961 Edgar A. Smith pubblica la prima versione della sua formula, detta *formula Devereaux*, dal nome della fondazione in cui lavora. Il suo metodo si differenzia da quelli precedenti in quanto utilizza il conteggio di caratteri-spazi (lettere, numeri, segni di punteggiatura) per la stima della difficoltà delle parole. Si tratta di un calcolo più veloce e più semplice rispetto alla misurazione di altre variabili, come ad esempio il numero delle sillabe; inoltre si presta bene al computo automatico.
La versione originale della formula, progettata per coprire i gradi da 4 a 12, è la seguente:

$$Grade\ Placement\ = 1{,}56\ \mathrm{W_L}\ + 0{,}19\mathrm{S_L} - 6{,}49$$

dove
W = lunghezza della parola misurata in caratteri-spazi
S = lunghezza della frase misurata in parole

Il coefficiente di correlazione multiplo è di 0,74. Per semplificare l'equazione, Smith propone una seconda versione della formula, che però non prevede livelli di istruzione.

$$Readability\ Index\ = 8\ \mathrm{W_L}\ + \mathrm{S_L}$$

Smith ha verificato la validità della sua formula sui testi della collana *Reading for Meaning* (Guiler e Coleman 1955) e dello *California Achievement Test* (Tiegs e Clark 1957).

[39] Uno dei principali teorici della leggibilità è George R. Klare, Illustre Professore Emerito di Psicologia e decano all'Università dell'Ohio. Il suo contributo principale consiste in una sintesi periodica dello stato di avanzamento della ricerca (1952, 1963, 1974-75, 1984, 1988): Klare ha effettuato recensioni critiche ai vari studi e ha diretto e partecipato a ricerche sulla validazione delle formule, non solo per la lingua inglese.

3.2. La formula di Rogers per la comprensione orale

La maggior parte delle formule sviluppate fino a questo momento è dedicata alla valutazione della comprensione di materiali scritti. Nel 1962 John Rogers pubblica una formula per la predizione della difficoltà di materiale orale. Come criterio, lo studioso utilizza 480 campioni di parlato spontaneo improvvisato di studenti di ogni grado scolastico (elementari, medie e superiori). La formula è la seguente:

$$G = 0{,}669I + 0{,}4981LD - 2{,}0625$$

dove:
G = livello di istruzione (grade level)
I = lunghezza media di *idea unit*
LD = numero medio di parole che non compaiono sulla lista di Dale (3.000 parole più frequenti) su un campione di 100 parole.

Idea unit è un termine coniato dallo stesso Rogers per indicare una misura di complessità linguistica che si riferisce all'unità informativa di una proposizione indipendente. La lunghezza dell'*idea unit* è determinata dividendo il numero di parole del testo trascritto per il numero di proposizioni indipendenti. La formula ha una correlazione multipla di 0,727 con il grado di istruzione dei campioni di testi.

3.3. Danielson e Bryan e le prime formule automatizzate

A. W. Danielson e S. D. Bryan (1963) sviluppano i primi due programmi informatici per l'applicazione delle formule di leggibilità. La loro è la prima formula creata specificatamente per l'uso automatico. Per facilitare il procedimento, usano la stessa variabile impiegata da Smith (1961), ovvero il conteggio dei caratteri, per misurare sia la lunghezza della frase che quella delle parole. Rispetto al computo delle sillabe, il numero di caratteri può essere infatti calcolato più facilmente e con maggiore precisione da un programma per il computer.
Gli studiosi presentano due versioni del loro indice: la prima formula è un'equazione pura di regressione e la seconda calcola i punteggi su una scala da 0 a 100 in base alle formule di Flesch e Farr, Jenkins e Paterson.

$$DB\#1 = 1{,}0364CPSp + 0{,}0194CPSt - 0{,}6059$$

$$DB\#2 = 131{,}059 - 10{,}364CPSp - 0{,}194CPSt$$

dove
$CPSp$ = caratteri per spazio
$CPSt$ = caratteri per frase

Gli autori usano lo *Standard Test Lessons in Reading* di McCall e Crabbs come criterio e ottengono una correlazione di 0,575.

3.4. Il grafico di Fry

Nel 1961 Edward Fry lavora come borsista Fulbright[40] al Makerere College in Uganda, affiancando un gruppo di insegnanti africani in un progetto formativo dell'Unesco per l'insegnamento dell'inglese come seconda lingua. Durante la sua ricerca sviluppa il suo *Readability Graph* per la previsione della leggibilità, raccomandandolo come un modo per risparmiare tempo e fatica. Il grafico viene pubblicato per la prima volta in Inghilterra nel 1965 e in seguito negli Stati Uniti (1968, 1969a, 1969b).

Il grafico consente una stima diretta del livello di istruzione necessario per capire un dato materiale, in base a due variabili: la lunghezza delle parole misurata in sillabe e la lunghezza della frase misurata in parole. Per convenienza Fry chiama il suo metodo "formula" ma in realtà non presenta nessuna equazione; il punteggio della leggibilità è ottenuto confrontando direttamente i punteggi delle variabili nel grafico. Il grafico originale determina la leggibilità fino alla scuola superiore (dal 4° al 7° grado). Nella versione del 1969 la valutazione è estesa alle elementari e in quella del 1977 agli anni del college (Figura 2).

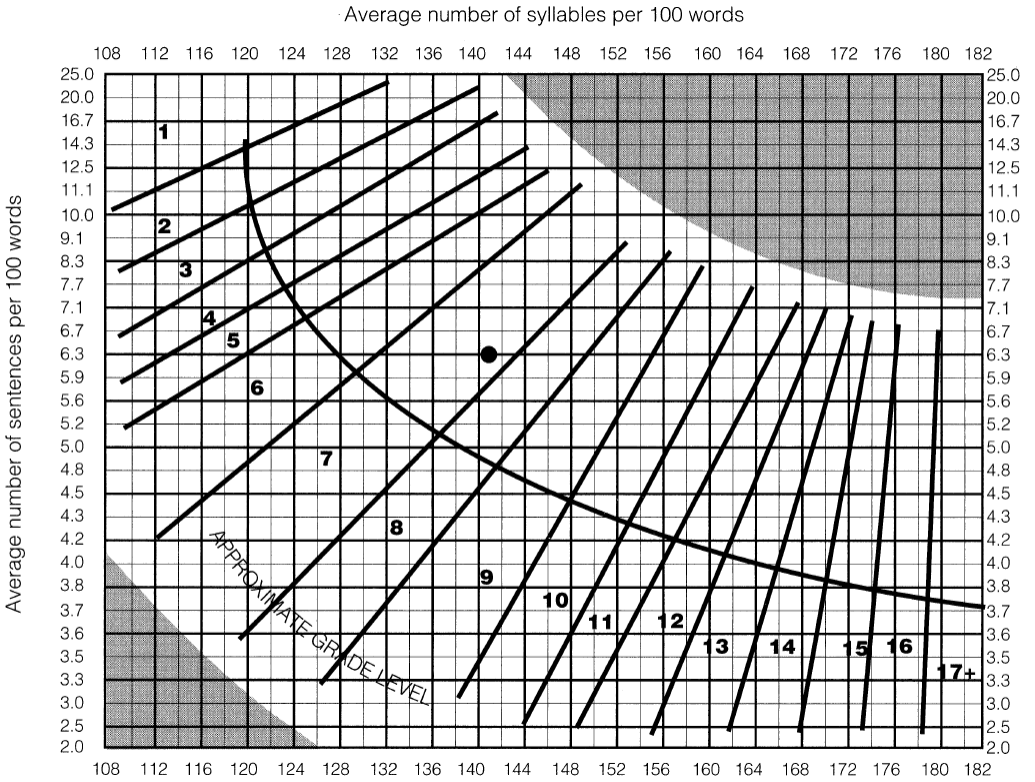


Figura 2. La versione del *Readability Graph* del 1977 (da Fry 2002).

[40] Il Programma Fulbright è un progetto di scambio internazionale molto competitivo destinato a studiosi, artisti e scienziati che partecipano a progetti di ricerca di particolare rilevanza. Il programma nasce negli Stati Uniti nel 1946, con la legge proposta dal Senatore J. William Fulbright dell'Arkansas. La legge, approvata dal Congresso statunitense, prevede il finanziamento di borse di studio per lo studio, la ricerca, l'insegnamento in modo da favorire il processo di pace attraverso lo scambio di idee e di cultura tra gli Stati Uniti e le altre nazioni nel mondo.

Fry propone alcune indicazioni circa l'uso del grafico di leggibilità:

1. Selezionare tre campioni di 100 parole ciascuno da un libro o un articolo; i brani devono essere presi all'inizio, a metà e alla fine del libro. Eliminare tutti i nomi propri e i numeri.
2. Contare il numero totale di frasi per ciascun campione e calcolare la media.
3. Contare il numero totale di sillabe in ciascun campione e calcolare la media.
4. Tracciare sul grafico i valori corrispondenti al numero medio di frasi e al numero medio di sillabe. L'area in cui le coordinate si incontrano mostra il punteggio relativo al grado di istruzione. I punteggi che si collocano nelle aree scure sono da considerarsi non validi.

Il grafico è altamente correlato con le formule di Dale e Chall (0,94), di Flesch (0.96) e di Spache per quanto riguarda le classi elementari (0,90). Nella sua versione più recente (1977) il grafico di Fry è uno dei metodi più utilizzati per la valutazione della leggibilità (Klare 1988).

Al 25° incontro annuale della *National Reading Conference* (1975), Fry presenta la sua *Kernel Distance Theory* ('Teoria della distanza del nucleo') e la ripropone anche nell'articolo del 1977.
La teoria cerca di spiegare perché due frasi che hanno la stessa lunghezza e lo stesso vocabolario (parole che hanno la medesima lunghezza o frequenza) possono presentare diversi gradi di difficoltà. Secondo lo studioso, quanto più il nucleo è vicino all'inizio della frase, tanto più semplice sarà la frase; anche una minore distanza tra gli elementi del nucleo incide sulla semplicità dell'enunciato. La distanza tra soggetto e predicato incide maggiormente sulla difficoltà rispetto alla distanza tra predicato e oggetto. Nella formulazione di Fry, il nucleo è composto dal soggetto, dal predicato e talvolta dall'oggetto; la distanza è misurata in numero di parole.
La ricerca si pone come uno strumento utile per gli scrittori, tuttavia questi accorgimenti non influiscono particolarmente sul punteggio di leggibilità.

3.5. Le formule di Coleman

Nel 1965, in un progetto di ricerca sponsorizzato dalla National Science Foundation, Edmund B. Coleman pubblica quattro formule di leggibilità per uso generale[41]; è il primo studio in cui si utilizza il cloze come criterio al posto dei più convenzionali test di lettura a scelta multipla o classificazioni da parte di esperti.
Le quattro formule utilizzano diverse variabili:

$$C\% = 1{,}29w - 38{,}45$$

$$C\% = 1{,}16w + 1{,}48s - 37{,}95$$

$$C\% = 1{,}07w + 1{,}18s + 0{,}76p - 34{,}02$$

[41] E. B. Coleman, *On understanding prose: some determiners of its complexity*, NSF Final Report GB-2604, Washington, D.C.: National Science Foundation, 1965. Dal momento che è difficile rintracciare la relazione, per ulteriori informazioni è possibile consultare l'articolo di Szalay (1965) sulla validazione delle formule di Coleman.

$$\mathrm{C\%} = 1{,}04\mathrm{w} + 1{,}06\mathrm{s} + 0{,}56p - 0{,}36prep - 26{,}01$$

dove
C% = percentuale di completamenti cloze corretti;
w = numero di monosillabi su 100 parole
s = numero di frasi su 100 parole
p = numero di pronomi su 100 parole
prep = numero di preposizioni su 100 parole

I coefficienti di correlazione multipla tra le formule e i punteggi ottenuti con la procedura cloze sono rispettivamente 0,85, 0,89, 0,90 e 0,91. L'uso del cloze come criterio produce dunque coefficienti di correlazione superiori rispetto allo *Standard Test Lessons in Reading* di McCall e Crabbs.
Szalay (1965) conduce uno studio di validazione incrociata con un nuovo set di 7 brani di 150 parole ciascuno; i coefficienti così ottenuti sono 0,83, 0,88, 0,87e 0,89.

3.6. Easy Listening Formula (ELF)

"*Listenability* is not necessarily "*readability*": così Irving E. Fang presenta la sua *Easy Listening Formula*, una formula di leggibilità destinata alla valutazione di materiali orali. "A readability formula should be simple, or it will not be used when it should be used, and it will not be used by writers, the people who should use it. A "listenability" formula should meet the same requirements" (Fang 1966-67, p. 64).
Fang sviluppa quindi una formula di facile applicazione, basata sullo studio comparato di testi provenienti da telegiornali o programmi televisivi[42]:

$$ELF = numero\ di\ sillabe\ (>1)\ per\ parola\ in\ una\ frase$$

Un punteggio medio per frase inferiore a 12 è considerato auspicabile per l'ascoltabilità di massa. L'autore individua un coefficiente di correlazione di 0,96 tra la sua formula e quella *Reading Ease* di Flesch per le 36 sceneggiature televisive e i 36 campioni di giornali testati.

3.7. Gli studi di Bormuth

Anche John Bormuth inizia a sperimentare la procedura cloze come nuovo criterio. Il suo primo studio (1966) fornisce una panoramica di tutte quelle variabili, oltre al vocabolario e la lunghezza delle frasi, che possono influire sulla comprensione; il cloze gli permette di valutare gli effetti di questi fattori non solo sulla difficoltà di interi brani ma anche su singole parole o frasi.
I dati della ricerca sono ottenuti dai risultati della valutazione della difficoltà tramite il cloze test di 20 testi campione, i quali hanno livelli di leggibilità dal 4° all'8° grado in base alla formula di Dale e Chall. I brani hanno tra le 275 e le 300 parole e sono tratti da materiali

[42] Le reti ABC, CBS e NBC; telegiornali locali sulle stazioni della rete di Los Angeles KABC-TV, KNXT e KNBC; Il *New York Times*, il *Wall Street Journal*, il *Christian Science Monitor*, il *Los Angeles Times*, il *Chicago Tribune* e il St. Louis Post-Dispatch.

didattici di vario genere (letteratura, storia, geografia, biologia, scienze fisiche). I soggetti esaminati includono l'intero corpus di studenti dal 4° all'8° grado (675 studenti) di una scuola elementare e media californiana, con livelli di lettura che vanno dal 2° al 12° grado (dalla seconda elementare all'ultimo anno di scuola superiore).
Nel corso dei due lavori successivi (1967 e 1968), Bormuth cerca di fornire un quadro di riferimento per l'interpretazione dei risultati delle prove cloze; il suo obiettivo è determinare quali punteggi percentuali prendere convenzionalmente come riferimento per la valutazione della comprensione della lettura. In base allo studio pubblicato da Thorndike (1916), il 75% di risposte corrette nei test a scelta multipla è da considerarsi il criterio della difficoltà ottimale per l'apprendimento assistito in classe e il 90% per la lettura indipendente. Bormuth stabilisce allora una corrispondenza tra i punteggi ottenuti nelle prove di comprensione della lettura a scelta multipla e quelli ricavati con il cloze sugli stessi testi: percentuali di risposte corrette di 50%, 75% e 90% nei test a scelta multipla equivalgono a percentuali di 35%, 45% e 55% di completamenti cloze corretti. Questi valori vengono presi come punteggi di riferimento per indicare il livello di frustrazione (35% di riempimenti corretti in un cloze), il livello di lettura scolastica (45%) ed il livello di lettura indipendente (55%).
Nel 1969 Bormuth conduce la più ampia analisi di leggibilità che sia stata fatta, fornendo una nuova base empirica per le formule successive. Lo studioso misura la comprensione di 330 brani di circa 100 parole ciascuno, tratti da libri di testo di dieci materie (biologia, chimica, educazione civica, notizie di attualità, economia, geografia, storia, letteratura, matematica e fisica); i livelli di difficoltà dei testi vanno dalla prima elementare all'università. I soggetti testati sono 2600 alunni di un distretto scolastico di Minneapolis, dalla quarta elementare all'ultimo anno delle superiori; gli studenti sono divisi in 50 gruppi in base ai punteggi ottenuti con il *California Reading Achievement Test* (edizione del 1963).
Come criterio di misurazione viene usata la procedura cloze, sia nella sua versione originale che in altre cinque varianti, per un totale di più di 2 milioni di risposte da analizzare. Oltre a questo, i soggetti sono sottoposti anche a test a scelta multipla.
Lo studioso controlla il potenziale valore predittivo di 169 variabili (molte delle quali mai analizzate prima), sviluppando ben 24 formule di leggibilità.
Le formule sono raggruppate in 4 serie: formule "senza restrizioni", che includono ogni variabile che abbia una correlazione anche parziale ma significativa con la difficoltà e che contengono quindi dai 14 ai 20 indici statistici; la forma breve delle formule senza restrizioni, ottenute utilizzando soltanto 10 variabili; formule destinate all'uso manuale e formule per il computer. La difficoltà è misurata sia a livello dell'intero brano (vedi formule seguenti) sia a livello di singole frasi e parole.
Ogni serie prevede inoltre quattro forme diverse: tre basate sui valori criterio (35%, 45% e 55%) e una generalizzata, basata su una media (formula *Mean Cloze*), calcolata in modo tale che l'utente possa scegliere un qualsiasi punteggio.
Le formule indicano il livello minimo di lettura necessario per ottenere un dato punteggio cloze in un test che misura la leggibilità di un determinato brano.
Formule per il calcolo manuale:

$$Cloze\ Mean\ =\ 1{,}051674 - 0{,}099691\ (LET/W) - 0{,}004236\ (LET/MPU) + 0{,}000015\ (LET/MPU)^2$$

$$GP\ (35) = 0{,}861207 + 1{,}279050\ (LET/W) + 0{,}050548\ (LET/MPU) — 0{,}000172\ (LET/MPU)^2$$

$$GP\ (45) = 1{,}849494 + 1{,}307968\ (LET/W) + 0{,}053930\ (LET/MPU) — 0{,}000191\ (LET/MPU)^2$$

$$GP\ (55) = 1{,}231834 - 2{,}764035\ (LET/W) — 0{,}023845\ (LET/W)^3 + 0{,}051591\ (LET/MPU) — 0{,}000186\ (LET/MPU)^2$$

dove
GP = grado di posizionamento (*grade placement*) per percentuali di completamenti corretti cloze (35), (45), (55).
LET/W = lettere per parola
LET/MPU = lettere per unità minima di punteggiatura (*minimal punctuation units*)

Formule per il calcolo automatico:

$$Cloze\ Mean = 0{,}886593 - 0{,}083640\ (LET/W) + 0{,}161911\ (DLL/W)^3 - 0{,}021401\ (W/SEN) + 0{,}000577\ (W/SEN)^2 - 0{,}000005\ (W/SEN)^3$$

$$GP\ (35) = 3{,}761864 + 1{,}053153\ (LET/W) - 2{,}138595\ (DLL/W)^3 + 0{,}152832\ (W/SEN) - 0{,}002077\ (W/SEN)^2$$

$$GP\ (45) = 3{,}398999 + 1{,}107014\ (LET/W) + 0{,}155327\ (W/SEN) - 0{,}002184\ (W/SEN)^2 + 6{,}672669\ (DLL/W)^2 - 7{,}523689\ (DLL/W)^3 - 5{,}266225(modal\ v)$$

$$GP\ (55) = 3{,}450806 + 1{,}094841\ (LET/W) + 0{,}153830\ (W/SEN) - 0{,}002242\ (W/SEN)^2 + 11{,}478313\ (DLL/W)^2 - 11{,}224816\ (DLL/W)^3 - 5{,}427013(modal\ v)$$

dove
LET/W = lettere per parola
DLL/W = numero di parole appartenenti alla lista di Dale delle 3.000 parole più frequenti sulle parole totali (*Dale Long List Words*)
W/SEN = parole per frase
Modal v = verbi modali (*modal verbs*)

Le formule sono state validate usando un nuovo set di 20 brani; per quelle che misurano la difficoltà a livello di brano sono riportate anche le validazioni incrociate.
Le formule senza restrizioni presentano coefficienti di correlazione con i 330 brani criterio che vanno da 0,86 a 0,89 e coefficienti che vanno da 0,67 a 0,80 nella convalida incrociata con i nuovi 20 brani. Per le forme brevi, i valori di correlazione con il criterio sono più bassi (0,83 - 0,87) ma più alti con i nuovi testi (0,88 - 0,93).
Le formule destinate al calcolo manuale presentano coefficienti di correlazione rispettivamente di 0,81 - 0,79 - 0,80 - 0,79; i coefficienti che derivano dalla convalida

incrociata sono 0,83 - 0,83 - 0,84 - 0,82. Infine, le formule per il computer mostrano correlazioni che vanno da 0,81 a 0,83 con i 330 brani criterio e che vanno da 0,92 a 0,93 con i 20 nuovi brani.

Nel 1975 Bormuth propone una nuova formula di facile applicazione, progettata per l'uso manuale da parte degli insegnanti. Come variabili sceglie la lunghezza della parola e la lunghezza della frase, "facili da analizzare, senza farsi coinvolgere in liste di parole o complesse analisi grammaticale" (p. 85). La formula è la seguente:

$$d = 1{,}069 - (0{,}106\ \frac{1}{w}) - (0{,}0036\ \frac{1}{s}) + \left[0{,}0000002\ \left(\frac{1}{s}\right)^2\right]$$

dove:
d = difficoltà del testo in termini di completamenti cloze
1/w = numero medio di lettere per parola in un campione scelto a caso di 250 parole.
1/s = numero medio di lettere per frase in un campione scelto a caso di 250 parole.

Lo studioso pubblica anche una tabella da utilizzare per evitare i calcoli della formula; la tabella fornisce direttamente i valori di difficoltà (grado di istruzione necessario) per varie combinazioni di lunghezza di parole e frasi, per punteggi criterio equivalenti al 35%, 45% e 55% di riempimenti corretti cloze.

Culhane e Ranking (1969) conducono uno studio per verificare la validità dei valori presi da Bormuth come punteggi di riferimento. L'ipotesi è la seguente: se uno studio che utilizza la stessa procedura ma materiali, test e soggetti diversi produce una serie di punteggi non sostanzialmente diversi da quelli di Bormuth, allora questi valori posso essere considerati come riferimento per l'interpretazione dei risultati. I due studiosi somministrano le due prove di comprensione (5 prove a scelta multipla e 5 cloze test) a un campione di 5 classi di 5° grado, per un totale di 105 studenti. Come testi criterio vengono scelti 5 articoli tratti dalla *World Book Encyclopedia*; la leggibilità dei materiali è misurata tramite il grafico di Fry, con i livelli che vanno dal quinto all'ottavo grado. Nella prova cloze non vengono considerati corretti i riempimenti con sinonimi. La Tabella 5 mostra le correlazioni tra i coefficienti di regressione ottenuti nelle due prove. La correlazione media tra le due prove di comprensione risulta 0,68.

Articoli	***n***	**Correlazione**
Bear	22	0,54
Mars	24	0,75
Stalin	22	0,63
Hitler	20	0,77
Jerusalem	17	0,71
Totale	105	0,68 (media)

Tabella 5. Coefficienti di correlazione tra test cloze e prove a scelta multipla.

La Tabella 6 mostra invece il confronto tra i punteggi ottenuti nello studio di Bormuth del 1967 e quelli di Culhane e Ranking del 1969. La differenza media tra i risultati è di 3,1 punti percentuali. Come si può notare, le differenze sono maggiori agli estremi, in particolare nei punteggi più elevati: lo scarto medio tra i risultati è di 6,2 punti percentuali per punteggi a scelta multipla che vanno da 50 a 70 e di 9,8 punti per punteggi che vanno da 75 a 100.

Punteggi dei test a scelta multipla	Punteggi cloze (Bormuth)	Punteggi cloze (Culhane e Ranking)	Differenza
50	19	10	+ 9
55	23	15	+ 8
60	27	22	+ 5
65	31	28	+ 3
70	35	35	0
75	38	41	- 3
80	42	48	- 6
85	46	54	- 8
90	50	61	- 11
95	53	67	- 14
100	57	74	- 17

Tabella 6. Confronto tra punteggi cloze e punteggi a scelta multipla.

Le percentuali cloze prese come riferimento da Bormuth per la valutazione della comprensione della lettura sono confrontate con quelle di Culhane e Ranking nella Tabella 7.

Punteggi criterio	Bormuth (1967)	Bormuth (1968)	Culhane e Ranking
75 %	38	44	41
90 %	50	57	61

Tabella 7. Corrispondenza tra punteggi cloze e punteggi criterio dei test a scelta multipla.

Questi risultati confermano la validità dei punteggi percentuali proposti da Bormuth. In particolare, per valori compresi tra 50 e 75 risultano più validi i punteggi ottenuti nello studio del 1967, per valori compresi tra 75 e 100 quelli ottenuti nello studio del 1968.

3.8. Automated readability Index (ARI)

Nel 1967 E. A. Smith e R. J. Senter creano una formula destinata ad applicazioni militari, *Automated readability Index* (ARI). L'Air Force (l'Aeronautica militare degli Stati Uniti) fa un ampio uso di materiali scritti come manuali, relazioni, studi, documenti di formazione del personale, ecc. La necessità di fornire documenti scritti in modo chiaro è evidente: la leggibilità influenza notevolmente il tempo necessario per estrarre le informazioni fondamentali e determina una maggiore probabilità che tali informazioni siano comprese e

utilizzate correttamente. Una comunicazione inadeguata porta inoltre con sé costi più elevati. Un indice di leggibilità tarato specificatamente per materiali tecnici dovrebbe fornire un mezzo veloce ed economico per la valutazione di questi documenti e, come sostengono gli autori, "dovrebbe contribuire in modo significativo all'efficienza di molte operazioni dell'Air Force" (p. 1).
La formula ARI utilizza una macchina da scrivere elettronica modificata con tre microinterruttori collegati a contatori (l'accessorio si chiama *Readability Index Tabulator*). Gli impulsi della macchina da scrivere attivano i contatori che registrano il numero di caratteri, il numero di parole e il numero di frasi contenute nel brano, calcolando in modo automatico la lunghezza media delle parole e delle frasi.
La formula è la seguente:

$$GL = 0{,}50\ (w/s) + 4{,}71\ (s/w) - 21{,}43$$

dove
GL = livello di istruzione (*grade level*)
w/s = parole per frase (*words per sentence*)
s/w = caratteri (o battute) per parola (*strokes per word*)

La formula può essere così semplificata:

$$ARI = (w/s) + 9\ (s/w)$$

L'indice GL produce un punteggio che indica il grado di istruzione, cioè l'età necessaria per comprendere il testo. La correlazione tra il livello di istruzione assegnato e la lunghezza delle frasi è 0,96; quella con la lunghezza delle parole 0,84; la correlazione tra lunghezza delle frasi e delle parole è 0,71. La correlazione multipla di queste variabili con il criterio è 0,98.
La formula semplificata ARI non determina invece livelli di istruzione: il numero risultante è associato al valore di leggibilità, senza un riferimento diretto ad un posizionamento di grado.
Smith e Kincaid (1970) validano con successo la formula ARI su materiali tecnici sia in modo manuale che con la versione automatica.

3.9. La formula SMOG

Nel 1969 G. Harry Mc Laughlin pubblica la sua formula SMOG, che egli ritiene essere veloce, semplice e più valida rispetto ai precedenti metodi di valutazione della leggibilità.
Il nome SMOG (dall'inglese *smoke* 'fumo' e *fog* 'nebbia') è un omaggio al *Fog Index* di Gunning, che per primo ha utilizzato il conteggio dei polisillabi per ottenere la misura della difficoltà semantica, ma è anche un riferimento alla città natale dell'autore: "il termine si riferisce anche al mio luogo di nascita, essendo lo smog apparso prima a Londra, sebbene, come per tante altre cose, da allora sia stato migliorato in diverse città americane" (Mc Laughlin 1969, p.641).
Il metodo per calcolare il sistema *SMOG grading*, cioè il grado di lettura richiesto per comprendere il testo valutato, è il seguente:

- Selezionare 10 frasi consecutive all'inizio del testo, 10 al centro e 10 alla fine. Considerare come una frase qualsiasi stringa di parole che termina con un punto fermo, un punto interrogativo o un punto esclamativo.
- Contare ogni parola di tre o più sillabe delle 30 frasi selezionate; se una parola polisillabica si ripete, contare ogni ripetizione.
- Stimare la radice quadrata del numero di vocaboli polisillabici contati, approssimando al quadrato perfetto più vicino. Ad esempio, se il conteggio è 95, il quadrato perfetto più vicino è 100, la cui radice quadrata è 10. Se il conteggio si trova a metà tra due quadrati perfetti, scegliere il numero più basso.
- Aggiungere 3 alla radice quadrata.

"Una formula di leggibilità è semplicemente un'equazione matematica che deriva dall'analisi di regressione. Questa procedura individua l'equazione che meglio esprime la relazione tra due variabili, che in questo caso sono la misura della difficoltà sperimentata da persone che hanno letto un dato testo e la misura delle caratteristiche linguistiche di quel testo. Questa formula può quindi essere usata per predire la difficoltà di lettura delle caratteristiche linguistiche di altri testi" (id., p. 640). Le misure linguistiche che hanno il maggior potere predittivo sono la lunghezza della parola e la lunghezza della frase, considerate rispettivamente come indicatori della difficoltà sintattica e di quella semantica. A differenza degli altri studiosi, Mc Laughlin crede però che queste due variabili vadano moltiplicate piuttosto che addizionate.

"Ciò che i precedenti ricercatori si sono fatti sfuggire è che la difficoltà semantica e sintattica interagiscono. Una lieve differenza nella lunghezza della parola o della frase tra due brani non indica lo stesso grado di differenza di difficoltà per brani difficili o testi semplici. Perciò una formula di leggibilità non dovrebbe avere la forma usuale:

$$Leggibilità = a + b\,(lunghezza\ della\ parola) + c\,(lunghezza\ della\ frase),$$

ma dovrebbe assumere la forma:

$$Leggibilità = a + b\,(lunghezza\ della\ parola\ x\ lunghezza\ della\ frase)$$

Dove a, b, c, sono costanti" (id., p. 640).

È possibile eliminare la moltiplicazione completa delle due variabili semplicemente selezionando un numero arbitrario di frasi e conteggiando il numero di sillabe in quelle frasi.

"Ho trovato una legge che collega il numero di sillabe in un brano alla percentuale delle parole polisillabiche, definite come parole di tre o più sillabe. Per praticità, il numero totale di sillabe in 100 parole può essere calcolato con questa regola generale: moltiplicare il numero di parole polisillabiche per 3 e aggiungere 112" (id., p. 641). È inoltre possibile rendere la costante *b* uguale a 1, tramite il semplice espediente di selezionare un numero adatto di frasi da valutare, cioè 30. La difficoltà di un brano può quindi essere misurata semplicemente contando le parole polisillabiche in 30 frasi.

I brani criterio utilizzati per la validazione della formula sono tratti da *Standard Test Lessons in Reading* di McCall e Crabbs (edizione del 1961). Ne derivano quattro equazioni di regressione che collegano il conteggio dei polisillabi (*p*) dei 390 brani analizzati ai punteggi (*g*) che hanno ottenuto gli studenti rispondendo alle domande su tali testi (Tabella 8).

Equazione di regressione		Correlazione	Errore Standard
(a)	$g = 6.2380 + 0.0785\, p$	0.713	1.4461
(b)	$g = 4.1952 + 0.8475\, \sqrt{p}$	0.709	1.4751
(c)	$g = 2.8795 + 0.9986\, \sqrt{p} + 5$	0.729	1.4446
(d)	$g = 1.0130\,(3 + \sqrt{p}) = 3.1291 + 1.0430\, \sqrt{p}$	0.985	1.5159

Tabella 8. Le quattro equazioni di regressione.

L'equazione (a) ha un coefficiente di correlazione con il criterio di 0,71 ma purtroppo non valuta la leggibilità al di sotto del sesto grado. Mc Laughlin prova quindi a considerare il quadrato del computo dei polisillabi, ottenendo le equazioni (b) e (c). In (b) rimane il problema del moltiplicatore fastidioso; in (c) questo è prossimo a 1 e dunque può essere eliminato ma sono coinvolte due addizioni. L'equazione (d) è un compromesso tra le precedenti.
La formula può essere così esplicitata:

$$SMOG\ grading = 3 + radice\ quadrata\ del\ conteggio\ dei\ polisillabi$$

L'errore standard delle previsioni fornite dalla formula SMOG è di circa 1,5 gradi.

3.10. La formula FORCAST

John S. Caylor, Thomas G. Sticht, Lynn C. Fox e J. Patrick Ford (1973), per conto dell'esercito degli Stati Uniti, conducono uno studio approfondito sulle abilità di lettura necessarie per i MOB (*Military Occupational Specialties* '*Specializzazione Occupazionale Militare'*), cioè sulle particolari esigenze di lettura richieste a seconda del campo di competenza militare[43].
Gli autori considerano le formule di leggibilità esistenti inadeguate per la valutazione di materiali tecnici dell'esercito, pertanto decidono di sviluppare una formula calibrata per tali applicazioni.
Per prima cosa selezionano le tipologie di posti di lavoro (MOS) da includere nella ricerca, individuandone 7 ad alta intensità[44]; in base a questa scelta, raccolgono i materiali tecnici, scegliendo come criterio 12 testi che le reclute devono aver compreso per potersi qualificare. La leggibilità dei brani è misurata secondo un indice di Flesch da loro modificato, con punteggi che vanno dal 6° al 13° grado (dalla prima media all'università).
La difficoltà dei testi è valutata tramite il cloze test; come punteggio criterio viene scelta una percentuale di 35% di completamenti corretti. I soggetti testati sono 395 reclute dell'esercito.

[43] Caylor et al. 1973
[44] Categorie MOS: Light Weapons Infantryman, Ground Control Radar Repairman, Wheel Vehicle Mechanic, Personnel Specialist, Armorer/Unit Supply Specialist, Medical Specialist, Military Policeman.

Gli studiosi analizzano 15 variabili linguistiche e trovano che il numero di monosillabi ha il più alto coefficiente di correlazione con il criterio (0,86). Ne deriva una formula basata su un solo fattore:

$$RGL = 20{,}43 - (0{,}11)(N)$$

dove
RGL = *Reading Grade Level*
N = conteggio dei monosillabi in un campione di 150 parole

Poiché l'aggiunta di una variabile legata alla lunghezza della frase non migliora il valore predittivo della formula, viene deciso di ometterla.
Viene pubblicata anche una versione semplificata della formula, chiamata FORCAST dal nome degli autori (FORd, CAylor, STicht).

$$FORCAST\ RGL = 20 - \left(\frac{N}{10}\right)$$

La formula è altamente correlata con l'indice di Flesch (0,92) e quello di Dale e Chall (0,94). I ricercatori effettuano anche una validazione incrociata, testando la formula su altre 365 reclute e su un altro campione di 12 testi tratti da materiali di lavoro degli stessi 7 MOS. La correlazione con l'indice di Flesch sale a 0,98 e quella con la formula di Dale e Chall a 0,95. Questi dati vengono giudicati idonei ai fini della validazione.
La formula è sviluppata per e in base a un corpus definito di materiale tecnico dell'esercito e a una data popolazione (giovani reclute); a differenza di altri indici di leggibilità di carattere più generale, non è destinata a valutare materiali scolastici, giornali o riviste e non è quindi dimostrata la sua applicabilità a tali tipologie testuali. Per lo stesso motivo, la formula non è in grado di misurare la difficoltà di testi al di sotto del sesto grado. Tuttavia, l'uso della sola variabile lessicale la rende adatta a valutare documenti molto brevi, come i testi delle pagine web.
Nel 1977 il Dipartimento della Air Force ha autorizzato l'uso di questo indice per la produzione di regolamenti tecnici comprensibili.

3.11. Navy Readability Indexes (NRI): la formula Flesch - Kincaid

Peter Kincaid, Lieutenant Robert P. Fishburne, Richard L. Rogers e Brad S. Chissom (1975) seguono l'esempio di Smith e Senter e calcolano nuove versioni di formule di leggibilità esistenti per testarle su materiali della Marina. Le formule scelte per le modifiche sono l'*Automated readability Index* (ARI), la formula *Reading Ease* di Flesch e il *Fog Index*[45].
Nella prima parte dello studio, i ricercatori determinano i livelli di lettura di 531 soggetti iscritti a quattro scuole di formazione tecnica della Marina. Tramite il test di lettura Gates-McGinitie viene valutata la comprensione di 18 brani tratti dai *Rate Training Manuals*, manuali di addestramento militare. I risultati sono utilizzati per calcolare il livello di posizionamento di grado dei brani e, in base a questi, vengono derivate le tre formule. Le

[45] Kincaid et al. 1975.

formule sono tarate specificatamente per la Marina e vengono chiamate *Navy Readability Index* (NRI).

ARI originale:

$$GL = 0{,}50\ (w/s) + 4{,}71\ (s/w) - 21{,}43$$

$$ARI[\text{versione semplificata}] = (w/s) + 9\ (s/w)$$

ARI nuova versione:

$$GL\ =\ 0{,}37\ (w/s)\ +\ 5{,}84\ (s/w)\ -\ 26{,}01$$

$$GL\ [\text{versione semplificata}] =\ 0{,}4\ (w/s)\ +\ 6\ (s/w) —\ 27{,}4$$

dove
GL = livello di istruzione (*grade level*)
w/s = parole per frase (*words per sentence*)
s/w = caratteri (o battute) per parola (*strokes per word*)

Fog Index originale:

$$Reading\ Grade\ Level = 0{,}4\ (lunghezza\ media\ delle\ frasi + percentuale\ di\ parole\ difficili)$$

Fog Index nuova versione:

$$GL\ =\ [\frac{\frac{easy\ words\ +\ 3\ (hard\ words)}{sentence}}{2} - 3]$$

dove
easy words = numero di parole di una o due sillabe su 100 parole
hard words = numero di parole con più di due sillabe su 100 parole
sentences = numero di frasi per 100 parole

Reading Ease originale:

$$\text{Reading Ease Score} = 206{,}835 - 0{,}846\text{W} - 1{,}015\text{S}$$

dove
w = indica il numero medio di sillabe per parola (*Word*)
s = indica il numero medio di parole per frase *(Sentence)*

Flesch ricalcolata o formula Flesch - Kincaid:

$$GL\ =\ 0{,}39\ (w/s)\ +\ 11{,}8\ (s/w)\ -\ 15{,}59$$

$$GL\ [\text{versione semplificata}]\ =\ 0{,}4\ (w/s)\ +\ 12\ (s/w) —\ 16$$

dove
GL = grade level
w/s = parole per frase (*words per sentence*)

s/w = sillabe per parola (*syllables per word*)

Nel corso dello studio vengono ricalcolati anche altri due indici: la formula di Farr, Jenkins e Paterson (FJP) e la formula FORCAST.

FJP originale:

$$\text{Reading Ease} = 1{,}599\text{os} - 1{,}015\text{s} - 31{,}517$$

FJP nuova versione:

$$GL = -0{,}307os + 0{,}387s + 22.05$$

dove:
os = numero di monosillabi (*one-syllable*) ogni 100 parole
s = numero medio di parole per frase *(sentence)*

FORCAST originale:

$$RGL = 20{,}43 - (0{,}11)(N)$$

FORCAST nuova versione:

$$GL = 25{,}31 - 0{,}24\ (N)$$

dove
RGL = *Reading Grade Level*
N = conteggio dei monosillabi in un campione di 150 parole

Gli studi hanno verificato che attualmente il materiale tecnico della Marina è scritto a un livello medio di difficoltà (12° grado) che è ben al di sopra delle capacità di lettura del personale che deve leggerlo (il cui livello medio varia tra il 9° e il 10° grado).
La formula Flesch-Kincaid è stata adottata come standard di leggibilità per le polizze assicurative e per i documenti utilizzati nell'assistenza sanitaria e in altre industrie; è stata inoltre incorporata in molti programmi di elaborazione di testi, come Microsoft Word[46]. Anche il Dipartimento della Difesa ha stabilito punteggi Flesch-Kincaid massimi per i propri contratti.

3.12. La formula di Coleman e Liau

Sempre nel 1975, Meri Coleman (figlia di E. B. Coleman) e T. L. Liau presentano una formula molto simile a quella di Bormuth, che impiega le stesse variabili ed è costruita con la stessa tecnica cloze. La loro formula è però tarata su studenti universitari ai primi anni di corso.
L'equazione si basa sulla misurazione di 36 brani di circa 150 parole ognuno, calibrati in base ai punteggi cloze da Miller e Coleman (1967).

[46] Si fa riferimento alla versione in inglese del programma; oltre alla formula Flesch-Kincaid, la leggibilità è valutata anche tramite la formula Reading Ease di Flesch. Nella versione italiana, oltre all'indice GULPEASE, tarato specificatamente sulla lingua italiana, viene utilizzato invece l'indice Fog.

$$Estimated\ cloze\ \% = 141{,}8401 - 0{,}214590L + 1{,}079812S$$

dove
Estimated cloze % = percentuale di completamenti cloze corretti;
L = numero di lettere su 100 parole
S = numero di frasi su 100 parole

Per quanto riguarda la validità, gli autori segnalano un coefficiente di correlazione di 0,92 con i brani criterio. Suggeriscono inoltre l'utilizzo di uno scanner ottico come aiuto per il conteggio.
Gli studiosi forniscono anche una tabella (pag. 284) e una formula per tradurre le percentuali cloze in livelli scolastici (con valori basati sullo *Standard Test Lessons in Reading* di McCall e Crabbs del 1961):

$$Grade\ level = -27{,}4004\ estimated\ cloze\ \% + 23{,}06395$$

La correlazione tra la percentuale cloze e il livello scolastico è - 0,88.

Nel 1976, Liau, Bassin, Martin e Coleman modificano le precedenti formule di Coleman (1965, 1971), cercando di migliorare i 36 brani criterio e le stesse 32 variabili usate dallo studioso[47]. Ne derivano diverse centinaia di formule computerizzate, tra cui gli autori ne selezionano quattro:

$$Estimated\ cloze\ \% = 159{,}76 - 0{,}24\ Let$$

$$Estimated\ cloze\ \% = 141{,}84 - 0{,}21\ Let + 1{,}08\ Sent$$

$$Estimated\ cloze\ \% = 39{,}84 - 0{,}10\ Let + 1{,}36\ Sent + 0{,}66\ 1\_Syl$$

$$Estimated\ cloze\ \% = 43{,}49 - 0{,}10\ Let + 1{,}22\ Sent + 0{,}67\ 1\_Syl - 0{,}44\ Cord\ Conj$$

dove
Estimated cloze % = percentuale di completamenti cloze corretti;
Let = numero di lettere su 100 parole
Sent = numero di frasi su 100 parole
1_Syl = numero di monosillabi su 100 parole
Cord Conj = numero di congiunzioni coordinanti su 100 parole

Come si nota, la seconda formula corrisponde a quella già impiegata da Coleman e Liau (1975).

3.13. La formula di Fry per i testi brevi

Quasi trent'anni dopo la pubblicazione del suo grafico, Fry (1990) propone una nuova formula, destinata a valutare la leggibilità di testi molto brevi.

[47] Liau et al.1976.

Lo studio di Fry si fonda sulla ricerca condotta nel 1977 da E. Dale e J. O'Rourke e pubblicata nel volume *The Living Word Vocabulary*[48]: gli autori forniscono il livello di grado di 43.000 voci diverse, corrispondenti a circa il 99% di tutte le parole usate al di sotto del 13° grado; il valore è dato per ogni accezione della parola.
La formula è progettata per misurare brani che contengono dalle 40 alle 99 parole, a condizione che questi contengano almeno 3 frasi, ma può essere utilizzata anche per testi lunghi dalle 100 alle 300 parole:

$$Readability = \frac{Word\ Difficulty\ +\ Sentence\ Difficulty}{2}$$

Il procedimento per determinare la leggibilità è così descritto:

- Selezionare almeno 3 parole chiave necessarie per la comprensione del testo;
- Per ognuna di queste cercare il livello di istruzione corrispondente nel *Living Word Vocabulary* (cercare per ogni accezione del termine);
- Calcolare la media delle 3 parole chiave più difficili: in questo modo si ottiene la difficoltà delle parole (*word difficulty*);
- Contare il numero di parole per in ogni frase e assegnare ad ognuna il livello di grado corrispondente utilizzando la tabella specifica (vedi Tabella 9); i valori della tabella sono ottenuti in base ai punteggi medi di difficoltà delle frasi del grafico di Fry (1977).
- Calcolare la media del livello di istruzione di tutte le frasi: in questo modo si ottiene la difficoltà delle frasi (*sentence difficulty*);
- Calcolare la media della difficoltà della frase e delle parole: si ottiene così il valore di leggibilità per qual brano.

La formula misura la leggibilità per livelli di difficoltà che vanno dal 4° al 12° grado; Fry raccomanda di riportare qualsiasi punteggio al di sotto del 4° grado come "4° grado o inferiore" (il motivo è che il *Living Word Vocabulary* non misura la difficoltà sotto tale soglia) e qualsiasi punteggio al di sopra del 12° grado come "12° grado o superiore" (data la variabilità osservata a livelli superiori).

Parole per frase	**Grade Level**
≤ 6.6	1
8.6	2
10.8	3
12.5	4
14.2	5
15.8	6
18.2	7

[48] E. Dale, J. O Rourke, *The Living Word Vocabulary, the Words We Know: A National Vocabulary Inventory*, Elgin, Illinois, Dome, 1976 (fuori commercio); il volume è stato ristampato nel 1981: *The Living Word Vocabulary: A National Vocabulary Inventory*, World Book-Childcraft International, 1981.

Parole per frase	Grade Level
20.4	8
22.2	9
23.2	10
23.6	11
24.3	12
25.0	13
25.6	14
26.3	15
27.0	16
> 27	17

Tabella 9. Lunghezza della frase e livelli di istruzione corrispondenti.

3.14. La nuova formula di Dale e Chall

Anche Dale e Chall (1995) sviluppano una nuova versione della loro formula. Gli autori utilizzano le stesse variabili impiegate nella formula originale (1948), ovvero la lunghezza delle frasi come misura della difficoltà sintattica e il numero di parole non comuni come difficoltà semantica ma effettuano diverse modifiche, riuscendo ad ottenere una correlazione multipla di 0,92 (invece di 0,70).
La formula è la seguente:

$$reading\ grade\ score\ = 0{,}1579x_1 + 0{,}0596x_2\ + 3{,}6365$$

dove
x_1 indica il numero di parole fuori dalla lista di Dale
x_2 indica la lunghezza media della frase
3,6365 è una costante

Viene impiegato un nuovo set di testi criterio, utilizzando la procedura cloze per stimare la difficoltà di comprensione della lettura. Il corpus comprende:

- 32 brani testati da Bormuth (1971) su studenti dal 4° al 12° grado.
- 36 brani testati da Miller e Coleman (1967) su 479 studenti universitari.
- 80 brani testati da MacGinitie e Tretiak (1971) su studenti universitari e laureati.
- 12 brani tecnici testati da Caylor et al. (1973) su 395 tirocinanti dell'aeronautica militare.

Anche la lista delle 3.000 parole più frequenti viene aggiornata e sono migliorate le regole per il conteggio delle parole familiari e non familiari.
La nuova versione include inoltre una duplice scelta per la presentazione della difficoltà dei testi, sia come punteggio cloze sia come livello di istruzione, che viene esteso dal 1° grado (nella formula precedente partiva dal 4°) fino alla laurea.

La convalida incrociata viene effettuata tramite:

- il test di lettura Gates-McGinitie
- il programma *Diagnostic Assessments of Reading and Trial Teaching Strategies* (DARTTS).
- il sistema di valutazione *The National Assessment of Reading Progress*.
- la formula Spache.
- il grafico di Fry.
- valutazioni da parte di un gruppo di docenti sul livello di lettura di 50 brani di letteratura.

3.15. Le formule commerciali: Lexile Framework, Degrees of Reading Power (DRP) e Advantage Open Standard (ATOS)

A partire dagli anni Ottanta, anche le grandi aziende commerciali sviluppano, con l'ausilio del computer, nuove e più sofisticate formule di leggibilità. Il sistema di misurazione *Lexile Framework*, il sistema *Degrees of Reading Power* (DRP) e la formula di leggibilità *Advantage Open Standard* (ATOS) sono tutti strumenti che misurano sia il grado di leggibilità dei testi sia il livello di lettura o di istruzione degli studenti; impiegano variabili tradizionali di lunghezza delle frasi e difficoltà del vocabolario ma, essendo informatizzati, sono in grado di valutare grandi campioni di testi o l'intero contenuto di libri. Le società forniscono inoltre, a pagamento o meno, liste di libri classificati per grado di leggibilità così da permettere agli insegnanti l'abbinamento con il livello di lettura degli studenti. Ad esempio, ATOS ha classificato 25.000 libri, il DRP ha una lista di 15.000 libri e il Lexile ha una lista di più di 26.000 libri commerciali classificati.[49]

3.15.1. Lexile Framework

Nel 1987, i fondatori della MetaMetrics, società che si occupa di educazione, pubblicano un nuovo sistema per valutare la comprensione della lettura, *Lexile Framework*[50]. Le variabili impiegate sono la lunghezza media della frase e la frequenza della parola, considerata come il numero di volte che un dato termine compare nell'*American Heritage Intermediate Corpus* (AHI). Il corpus AHI, creato da Carroll, Davies e Richman (1971), contiene circa 5 milioni di parole provenienti da una vasta gamma di materiali scolastici destinati a studenti dal 3° al 9° grado.
Come criterio vengono scelti 66 brani tratti dal test di comprensione della lettura *Peabody Individual Achievement Test* (Dunn e Markwardt 1970). La correlazione con le variabili è di 0,92.
Il sistema Lexile valuta la leggibilità su una scala che va da 0 a 2000; trattandosi di uno standard chiuso, è necessario utilizzare un test di lettura specifico Lexile (o comunque un test approvato da MetaMetrics) per la corrispondenza studente/libro.

[49] Secondo I dati riportati da Fry 2002.
[50] A. J. Stenner, D. R. Smith, I. Horabin, M. Smith, *Fit of the Lexile Theory to Sequenced Units from Eleven Basal Series*, MetaMetrics, Inc, 1987; si veda anche Stenner et al. 1988a, 1988b, 1997, 1998.

3.15.2. Degrees of Reading Power (DRP)

Nel 1981 il College Entrance Examination Board adotta il sistema *Degrees of Reading Power* (DRP)[51] per la comprensione della lettura, sviluppato dalla *Touchstone Applied Science Associates* (ATAS), società che offre prodotti educativi e didattici negli Stati Uniti e in Canada.
Il DRP utilizza la formula *Mean Cloze* per il calcolo automatico di Bormuth (1966) e prevede punteggi su una scala da 0 (facile) a 100 (difficile); può essere utilizzata per valutare sia la leggibilità di un dato testo sia le abilità di lettura degli studenti.
La formula originale di Bormuth:

$$R = 0{,}886593 - 0{,}083640\,(LET/W) + 0{,}161911\,(DLL/W)^3 - 0{,}021401\,(W/SEN) + 0{,}000577\,(W/SEN)^2 - 0{,}000005\,(W/SEN)^3$$

dove
R = punteggio Mean Cloze
LET/W = lettere per parola
DLL/W = numero di parole appartenenti alla lista di Dale (3.000 parole più frequenti) diviso il numero di parole totali (*Dale Long List Words*)
W/SEN = parole per frase

La formula DRP:

$$DRP = (1 - R)\,x\,100$$

3.15.3. Advantage-TASA Open Standard (ATOS)

Nel 2000 i ricercatori dello School Renaissance Institute, assieme alla società *Touchstone Applied Science Associates* (ATAS), sviluppano la formula *Advantage-TASA Open Standard* (ATOS)[52]. L'obiettivo è quello di creare una formula "aperta", disponibile e gratuita per gli insegnanti, facile da usare e in grado di essere utilizzata con tutti i test di lettura standard a livello nazionale. Il sistema informatizzato valuta la leggibilità di interi libri e non solo di campioni di testo.
Le variabili impiegate sono 3: il numero di parole per frase (correlazione o r2 = 0,897), il livello di istruzione medio delle parole (r2 = 0,891) e il numero di caratteri per parola (r2 = 0,839). La formula produce punteggi a livello di grado ma viene fornita anche una tabella di conversione per le scale usate nei sistemi DRP e Lexile.

[51] College Entrance Examination Board, *Degrees of reading power* (DRP). Princeton, NJ: College Entrance Examination Board, 1980 e successive; B. I. Koslin, S. Zeno, S. Koslin, *The DRP: An effective measure in reading*, New York: College Entrance Examination Board, 1987; S. M. Zeno, S. H. Ivens, R. T. Millard, R. Duvvuri, The *educator's word frequency guide*, Brewster, NY: Touchstone Applied Science Associates, 1995
[52] School Renaissance Institute, *The ATOS readability formula for books and how it compares to other formulas*, Madison, WI: School Renaissance Institute, Inc., 2000; T. Paul, *Guided Independent Reading*, Madison, WI: School Renaissance Institute, 2003

Gli autori riportano anche tabelle che consentono agli studenti di valutare la propria zona di sviluppo prossimale (*Zone of Proximal Development* o ZPD)[53], un concetto teorico proposto dallo psicologo russo Lev Vygotsky nel 1978. Secondo Vygotsky, la zona di sviluppo prossimale è la distanza tra il livello attuale di sviluppo, dove la soluzione dei problemi è indipendente, e il potenziale livello di sviluppo, in cui la soluzione dei problemi può essere raggiunta grazie all'aiuto di altre persone, come la guida di un adulto o la collaborazione con pari che abbiano livelli di competenza maggiore. Ai bambini dovrebbero essere proposti problemi leggermente superiori alle loro capacità, cioè a quel livello di difficoltà ottimale che produce il maggior guadagno in termini di apprendimento.

[53] School Renaissance Institute, *ZPD guidelines: Helping students achieve optimum reading growth*. Madison, WI: School Renaissance Institute, Inc., 1999

4. La leggibilità in lingue diverse dall'inglese

La maggior parte degli studi sulla leggibilità è condotta negli Stati Uniti e riguarda la lingua inglese. Per quanto riguarda le altre lingue[54], Klare (1974, 1984) sottolinea che gran parte delle prime ricerche è condotta negli Stati Uniti a beneficio di studenti di lingua inglese che studiano lingue straniere. Nascono così lo studio di Tharp (1939) sul francese, sette formule per lo spagnolo (Spaulding 1951, Patterson 1972, Thonis 1976, Garcia 1977, Gilliam et al. 1980, Vari-Cartier 1981 e Crawford 1984), strumenti per l'ebraico (Nahshon 1957), il tedesco (Walters 1966, Schwartz 1975), il russo (Rock 1970), il cinese (Yang 1970) e il vietnamita (Nguyen e Henkin 1982).
Le prime misure di leggibilità sviluppate in Europa sono semplicemente adattamenti della formula *Reading Ease* di Flesch: Kandel e Moles (1958) tarano l'indice sulla lingua francese, Huerta (1959) produce una versione spagnola, Douma (1960) e Brouwer (1963) adattano la formula alla lingua olandese, De Landsheere presenta una versione per il francese (1963) e una per il tedesco (1970).
L'attività di ricerca per lo sviluppo di strumenti più originali inizia in Europa alla fine degli anni Sessanta. Si trovano così ricerche sul finlandese (Wiio 1968), francese (De Landsheere 1966, Henry 1973, 1979, Richaudeau 1979), danese (Togeby 1971), svedese (formula Lix di Björnsson 1968 e 1983, Platzach 1974), olandese (van Hauwermeiren 1972, Zondervan, van Steen e Gunneweg 1976, Staphorsius e Krom 1985), tedesco (in Germania: Groeben 1972, Nestler 1977, Dickes e Steiwer 1977; in Austria: Bamberger 1973), spagnolo (in Venezuela: Gutiérrez Polini et al. 1972; in Spagna: Rodriguez 1981, Rodríguez Diéguez 1983).
Più recentemente, vengono effettuate ricerche che applicano i nuovi metodi computazionali a diverse lingue, come il cinese (Lau 2006, Chen et al. 2013), il tedesco (Vor Der Brück e Hartrumpf 2007), il francese (François e Fairon 2009), l'arabo (Al-Khalifa e Al-Ajlan 2010), il giapponese (Tanaka-Ishii et al. 2010), il tailandese (Daowadung e Chen 2011) e lo svedese (Sjöholm 2012)[55].

4.1. Ricerche sulla leggibilità dei testi in lingue straniere negli Stati Uniti

4.1.1. Francese

Uno dei primi studi sulla difficoltà dei testi in una lingua diversa dall'inglese è quello di Tharp (1939). Lo studioso si occupa di valutare la difficoltà di testi in lingua francese destinati a studenti di lingua inglese e, nonostante non sviluppi una vera e propria formula, propone un *Indice di Difficoltà* che combina due fattori, la densità e la frequenza delle parole. La misura della difficoltà è data dividendo l'indice di frequenza (calcolato utilizzando il *Basic French Vocabulary*) per la densità, che a sua volta si ottiene dividendo il numero di *running words* (*tokens*, parole totali del testo) per il numero di *burden words* (numero di *non-cognate words*)[56]. Maggiore è il valore della densità, più facile è il testo.

[54] Per una sintesi delle ricerche sulle formule di leggibilità in lingue diverse dall'inglese cfr. Klare 1974, 1984 e Rabin 1988.
[55] I nuovi metodi di valutazione automatica della leggibilità sono trattati nel Capitolo 6.
[56] Le *cognate words* sono i derivati della stessa famiglia lessicale, i termini che hanno la stessa radice.

4.1.2. Spagnolo

Dal momento che lo spagnolo viene spesso insegnato come lingua straniera nelle scuole statunitensi e a causa della vicinanza di paesi di lingua spagnola e della conseguente immigrazione, negli Stati Uniti vengono sviluppati diversi strumenti per la valutazione dei testi in tale lingua.
Nel 1951 Spaulding pubblica due formule per stimare la difficoltà di lettura di materiali didattici in spagnolo come lingua seconda. I materiali usati nello studio provengono da testi impiegati negli esami di dottorato; le variabili considerate sono la lunghezza media della frase, come misura della complessità sintattica per entrambe le formule, la frequenza media delle parole e la densità dei brani come indice di complessità del vocabolario. La densità si basa sul numero di parole nel testo che non sono presenti nell'elenco di Buchanan (1927) delle 1500 parole più usate nello spagnolo.

$$(1)\ Difficoltà\ =\ 4{,}115\ (FI)\ +\ 0{,}154\ (ASL)\ -\ 2{,}383$$

$$(2)\ Difficoltà\ =\ 0{,}1609\ (ASL)\ +\ 33{,}18\ (D)\ +\ 2{,}20$$

dove
FI = indice di frequenza
ASL = lunghezza media della frase misurata in parole
D = densità

Il range di difficoltà varia da 10 a 200, dove 10 indica un brano facile e 200 un brano difficile. La prima formula è leggermente più accurata rispetto alla seconda (i coefficienti di correlazione sono rispettivamente 0,90 e 0,87), ma la seconda risulta più facile da calcolare (la misurazione della densità è più semplice rispetto al calcolo della frequenza). In un articolo successivo, Spaulding (1956) presenta una versione leggermente modificata del secondo indice (cambiano i decimali)[57].
La formula è utilizzata da *Pan American Union* per stabilire la difficoltà di materiali didattici preparati per adulti latino-americani con capacità di lettura limitate.
La procedura di Spaulding è adattata da Patterson (1972) per l'uso da parte di operatori religiosi che si occupano di lettori con ridotte capacità di lettura. A sua volta Thonis (1976) utilizza la ricerca di Patterson per effettuare una conversione dei valori di difficoltà di Spaulding in livelli effettivi di istruzione (*grade level*), ma dato che la procedura seguita non risulta chiara, a questo studio non viene data molta credibilità.
Altri quattro studi (Garcia 1977, Gilliam et al. 1980, Vari-Cartier 1981 e Crawford 1984) si basano invece sul grafico di leggibilità di Fry (1968, 1977), che impiega come variabili la lunghezza delle parole misurata in sillabe e la lunghezza delle frasi misurata in parole.
Garcia (1977) propone di modificare i valori sull'asse delle ascisse e delle ordinate del grafico in modo da riflettere la lunghezza media di frasi e parole in spagnolo. L'autore

[57] Per produrre un grado di difficoltà compreso tra 20 e 200 i valori della formula sono moltiplicati per 10: *Difficulty = 1,609 (ASL) + 331,8 (D) + 22,0*. Nell'articolo del 1956 sono pubblicate per la prima volta anche le istruzioni per l'applicazione della formula 2, un grafico e l'elenco delle parole per il calcolo della densità (*Density Word List*).

suggerisce l'uso del cloze test anche se la sua ricerca ha rivelato che la procedura è adatta solo nei casi di allievi che studiano lo spagnolo avanzato.
Gilliam, Peña e Mountain (1980) utilizzano una versione adattata del grafico di Fry per determinare la leggibilità di 13 libri di testo e 9 libri per ragazzi con gradi da 1 a 3; gli autori indicano che, per stabilire una corrispondenza in termini di livelli di istruzione tra le due lingue, è necessario sottrarre 67 dal conteggio medio delle sillabe.
Vari-Cartier (1981) individua due tipologie di problemi nell'applicazione del grafico allo spagnolo: in primo luogo, il grafico non è progettato per stimare la leggibilità di materiali di prosa che abbiano un numero medio di sillabe superiore a 182; questo può rappresentare un problema in quanto la struttura delle parole in spagnolo è caratterizzata da un alto numero di sillabe e, come sottolineano anche Gilliam et al. 1980, non è raro contare in un testo di 100 parole un numero medio di sillabe che va dalle 175 alle 200. Le regole per la sillabazione in inglese e spagnolo sono simili ma se si conta il numero di sillabe nello stesso testo in ognuna delle due lingue risulterà un conteggio diverso, probabilmente più alto per lo spagnolo. In secondo luogo, il grafico di Fry determina punteggi che corrispondono a livelli di istruzione convenzionali, che non riflettono i diversi gradi di apprendimento di una lingua straniera.
Vari-Carter progetta dunque un nuovo grafico, detto *Fry Readability Adaptation for Spanish Evaluation* (FRASE), aumentando il valore massimo del conteggio medio di sillabe sull'asse orizzontale e regolando i parametri di leggibilità in modo da rendere conto dei diversi livelli di studio delle lingue straniere (principiante, intermedio, intermedio avanzato, avanzato); in questo modo il nuovo grafico potrebbe essere applicato anche allo studio di altre lingue, oltre allo spagnolo.

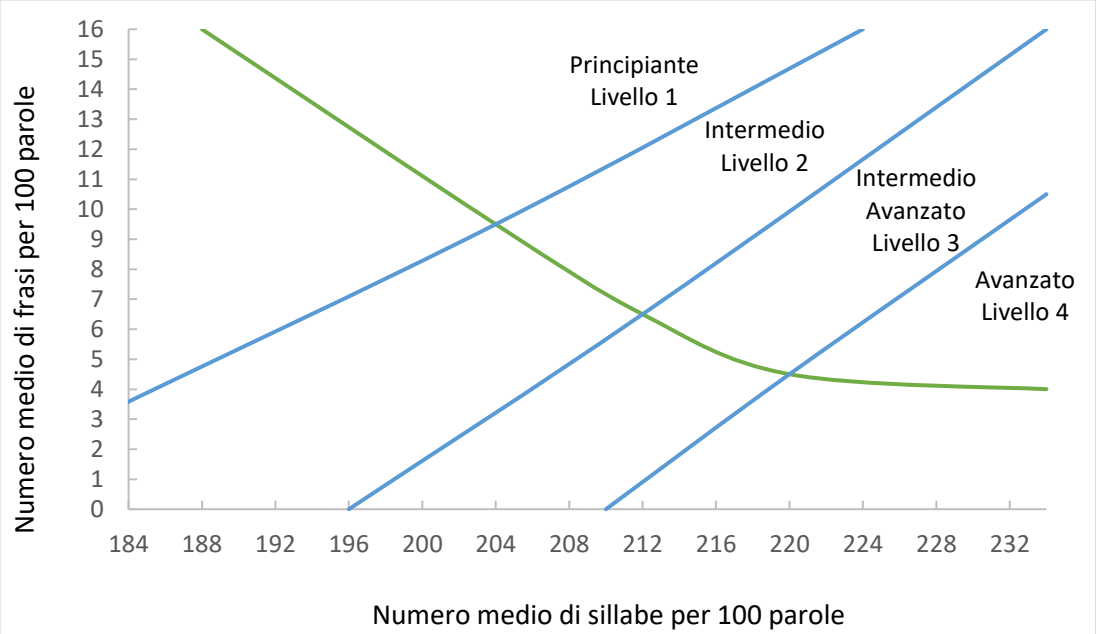


Figura 3. Il grafico FRASE

La ricerca di A. N. Crawford (1984) è sostenuta dal Department of Education degli Stati Uniti, in base al *Bilingual Education Act*. Lo studio è condotto su un corpus che comprende dieci serie diverse di testi in spagnolo elementare (che vanno dal 1° al 6° grado) impiegate

negli Stati Uniti, in America Latina e in Spagna. Lo studioso propone un nuovo grafico di leggibilità[58] basato sulla seguente equazione di regressione:

$$Grade\ level\ =\ [numero\ di\ frasi\ per\ 100\ parole\ \times\ (-0{,}205)] + (numero\ di\ sillabe\ per\ 100\ parole\ \times\ 0{,}049)\ -3{,}407$$

4.1.3. Ebraico

Nahshon (1957) valuta 11 variabili linguistiche e le combina in 8 formule di leggibilità per testi di prosa in ebraico. Riportiamo la formula più lunga e quella più breve:

$$GS\ =\ 0{,}1638\,x_1\ +\ 0{,}0334\,x_2\ -0{,}1382x_3\ -\ 0{,}0682x_4\ +\ 6{,}749$$

$$GS\ =\ 0{,}236x_1 +\ 0{,}1338x_2 -\ 3{,}305$$

Dove:
GS = livello di istruzione (*grade level*) in cui uno studente israeliano può comprendere un brano senza aiuto esterno;
x_1 = percentuale di parole diverse difficili;
x_2 = lunghezza media delle frasi in parole;
x_3 = percentuale di parole concrete;
x_4 = percentuale di verbi

La prima formula produce una correlazione di 0,935, mentre la seconda di 0,868.

4.1.4. Tedesco

T W. Walters (1966) sviluppa 33 formule di leggibilità destinate a valutare testi di teologia in tedesco. Le formule che presentano valori più alti di correlazione multipla sono quella a due variabili (0,92), a tre variabili (0,96) e a quattro variabili (0,97).

$$Y\ =\ 801{,}12 - 40{,}77\ (F87) -\ 172.32\ (F113)$$

$$Y\ =\ 904{,}30 - 32{,}09\ (F87)\ -\ 330{,}17\ (F104)\ -\ 119.66\ (F113)$$

$$Y\ =\ 384{,}89 - 5{,}75\ (F79A)\ -\ 240{,}16\ (F100) +\ 12.94\ (F111A)\ -\ 39.99\ (F126)$$

dove:
Y = indice di difficoltà
F87 = numero medio di segmenti verbali per frase
F113 = densità di modifica (*density of modification*) in unità nominali
F104 = percentuale di unità astratte /A/
F79A = numero di sostantivi
F100 = percentuale di nomi astratti
F111A = numero di unità nominali
F126 = numero medio di *clause units* per frase

[58] Crawford 1984 e 1989.

I punteggi vanno da 100 (molto difficile) a 550 (molto facile); si tratta di una formula particolare, che impiega un criterio particolare e che dunque non è adatta ad un impiego generale.
Schwartz (1975) adatta il grafico di Fry alla lingua tedesca per valutare i materiali didattici in tedesco elementare. Come set campione utilizza una serie di letture di base in tedesco occidentale, risalenti alla Seconda Guerra Mondiale; l'autrice nota che le parole tedesche sono più lunghe rispetto a quelle inglesi e determina che la differenza sia di circa 25-37 sillabe. Il numero di frasi invece è molto vicino ai corrispondenti livelli di grado.

4.1.5. Cinese

Shou-jung Yang (1970) intraprende un ampio progetto sulla leggibilità dei testi in cinese. Come criterio utilizza 85 brani in cinese moderno, somministrati gli studenti dei primi due anni delle scuole superiori di Taiwan. Tra le 39 variabili linguistiche studiate, individua le tre con il più alto valore predittivo: il fattore *parola* (inteso come quantità di parole che fanno parte di una lista di 5600 parole 'semplici'), il fattore *carattere* e quello *frase*. La correlazione multipla è di 0,80.
Ne derivano una formula "lunga" a 7 variabili e una versione "breve" con le 3 variabili citate:

$$(1)\ Y = 13.90963 + 1.54461\,(FULLSEN) + 3.01497\,(WORDLIST) - 2.52206\,(STROKES) - .29089\,(COUNT\text{-}5) + .26193\,(COUNT\text{-}12) + .99363\,(COUNT\text{-}22) - 1.64671\,(COUNT\text{-}23)$$

$$(2)\ Y = 14.95961 + 39.07746\,(WORDLIST) - 2.48491\,(STROKES) + 1.11506\,(FULLSEN)$$

dove:
Y = livello di difficoltà;
FULLSEN = proporzione di frasi complete;
WORDLIST = proporzione di parole che si trovano nella lista di 5.600 parole semplici;
STROKES = numero medio di battute per carattere;
COUNT-5 = percentuale di caratteri in gruppi di 5 battute;
COUNT-12 = percentuale di caratteri in gruppi di 12 battute;
COUNT-22 = percentuale di caratteri in gruppi di 22 battute;
COUNT-23 = percentuale di caratteri in gruppi di 23 battute

4.1.6. Russo

Rock (1970) sviluppa un grafico di leggibilità per aiutare gli insegnanti di russo. L'apprendimento di questa lingua è infatti molto difficile per gli studenti di lingua inglese e i docenti hanno bisogno di materiali adatti alle capacità dei lettori.
Il grafico non include come variabile la lunghezza della frase poiché le frasi russe tendono a presentare poche difficoltà sintattiche; il fattore linguistico considerato rilevante per la difficoltà è invece il vocabolario. Il grafico si basa su una lista di vocaboli di uso comune che

compaiono almeno in metà dei libri di testo russi in uso nelle scuole superiori statunitensi e sulla percentuale di parole che sono risultate sconosciute per un campione rappresentativo di studenti che ha letto e valutato un determinato materiale di lettura.

4.1.7. Vietnamita

Nguyen e Henkin (1982) sviluppano una formula di leggibilità per la lingua vietnamita. La formula, sostengono gli autori, è di particolare interesse per scrittori, redattori, educatori e fornitori di servizi sociali che creano materiali destinati ai rifugiati vietnamiti in altri paesi.
Gli studiosi selezionano un campione di 20 brani di circa 300 parole ciascuno da romanzi, riviste e libri di testo vietnamiti dal 4° grado al livello universitario. Il livello di leggibilità (RL) è dato dalla lunghezza media della frase (SL) e dalla lunghezza media della parola (WL):

$$RL = 2WL + 0{,}2SL - 6$$

Nel computo delle parole si tiene conto anche dei vari segni, come i segni tonali o i trattini.

4.2. Le formule di leggibilità in Europa e nel resto del mondo

4.2.1. Spagnolo

In Spagna, F. Huerta inizia nel 1950 una serie di studi volti ad adeguare la formula di Flesch allo spagnolo e nel 1959 presenta una versione provvisoria del suo indice:

$$\text{Lecturabilidad} = 206{,}84 - 0{,}60\ \text{P} - 1{,}02\ \text{F}$$

dove
P = numero medio di sillabe per 100 parole;
F = numero medio di frasi per 100 parole;

N. Lopez Rodriguez (1981) propone una formula di leggibilità per la lingua spagnola. Il criterio scelto per la comprensione della lettura è il cloze test. La studiosa misura 26 indicatori linguistici e costruisce diverse formule utilizzando varie combinazioni di variabili (da 4 a 12).
Sul lavoro di Lopez Rodriguez si basa la ricerca di Rodríguez Diéguez (1983), che aggiunge alla lista altre otto variabili (per un totale di 34); come criterio viene misurata la comprensione di 123 testi tramite la procedura cloze.

Ricerche sulla valutazione dei testi in lingua spagnola vengono svolte anche in Venezuela; molti di questi progetti venezuelani sono condotti presso l'Università di Chicago con la supervisione di John Bormuth.
L. E. Gutiérrez Polini (1972) crea la prima formula originale per lo spagnolo sviluppata al di fuori degli Stati Uniti. L'indice, che utilizza il cloze come criterio, valuta i libri di testo fino alla prima media e non sembra adatto a misurare materiali per adulti.

$$C = 95{,}2 - \frac{9{,}7\ L}{P} - \frac{0{,}35\ P}{F}$$

dove
C = comprensibilità del testo
L = numero di lettere
P = numero di parole
F = numero di frasi

Come spiega Rabin (1988) questa ricerca è sostenuta dal Dipartimento di Ricerca Educativa del Ministero della Pubblica Istruzione del Venezuela, in risposta alla grande necessità di proporre materiali didattici adatti al livello di lettura degli studenti; purtroppo però molti dei compatrioti di Gutiérrez non hanno compreso fino in fondo (o non hanno accettato) il concetto di misurazione della leggibilità e la procedura non è mai stata ampiamente utilizzata.
Anche la formula sviluppata da N. Rodríguez Trujillo (1980) nasce nell'ambito di uno studio supportato dal Ministero della Pubblica Istruzione. La formula è convalidata con l'uso della procedura cloze fino al 6° grado[59].

$$Comprensibildad = 95,2 - (9,7\ x\ L/W) - (0,35\ x\ W/S)$$

dove
L = numero di lettere
W = numero di parole
S = numero di frasi

A. Morles si è occupato in più sedi di adattare le prove cloze alla lingua spagnola; nello studio del 1981 l'autore si pone il problema dell'interpretazione dei punteggi ottenuti in questi test, non esistendo in spagnolo alcun criterio per la determinazione di tali risultati. Morles suggerisce di prendere come riferimento il modello usato per la lingua inglese e di impostare come livello minimo di comprensione della lettura una percentuale di risposte corrette del 58%. La percentuale di soggetti che hanno raggiunto o superato tale valore determina il livello di comprensibilità del materiale.

4.2.2. Tedesco

La prima formula originale in tedesco sviluppata in Europa è quella di Fucks (1955). Il livello di difficoltà dei testi è dato dal prodotto della lunghezza delle parole e della lunghezza della frase; questo strumento produce risultati simili a quelli del grafico di Fry ma viene considerato inadeguato per il tedesco, probabilmente per il fatto che in questa lingua le parole lunghe sono più frequenti rispetto all'inglese.
De Landsheere (1970) sviluppa una versione della formula *Reading Ease* di Flesch anche per la lingua tedesca, utilizzando gli stessi principi impiegati in quella tarata per il francese (De Landsheere 1963). Per determinare la difficoltà dei testi, Groeben (1972) utilizza il livello di astrazione delle parole, Nestler (1977) impiega invece il livello concettuale delle parole, identificando tre categorie: parole generalmente conosciute, parole difficili e parole *professionali* rare.

[59] I dettagli sul processo di sviluppo della formula non sono descritti dall'autore.

Dickes e Steiwer (1977) costruiscono tre indici di leggibilità: la formula con otto variabili ha una correlazione multipla di 0,91 con i punteggi ottenuti tramite il test cloze su 60 testi in tedesco; quella con sei variabili ha una correlazione di 0,89 e quella a tre variabili di 0,87. Quest'ultima formula è molto simile a quella di Flesch.

In Austria, Bamberger lavora a un progetto per valutare testi in lingua tedesca sia con metodi oggettivi che soggettivi[60]. Lo studioso utilizza un "profilo di leggibilità" composto da cinque variabili non linguistiche (contenuto, organizzazione, stampa, stile e motivazione) in combinazione con una serie di formule di regressione che valutano la difficoltà linguistica. Il metodo è applicato a diverse centinaia di libri in una convalida incrociata: i risultati mostrano che nel 70% dei casi il livello di istruzione prodotto dal profilo è simile a quello derivante dall'uso delle formule.
Seguendo il metodo additivo di Björnsson (cfr. 4.2.6)4.2.6, Bamberger e Vanecek (1982, 1984) elaborano una tecnica per la lingua tedesca. Utilizzano come criterio 120 libri di narrativa per bambini e 200 libri di saggistica per bambini ed elaborano una tabella di conversione dei punteggi in livelli di istruzione. Alla formula originale aggiungono altri fattori linguistici, come la percentuale dei monosillabi, delle parole polisillabiche o il numero di parole difficili, ottenendo correlazioni con il criterio ancora più alte.

4.2.3. Francese

Nel 1958 Kandel e Moles propongono un adattamento della formula *Reading Ease* di Flesch ai testi in francese. De Landsheere presenta una versione della formula di Flesch tarata sulla lingua francese (1963) e una formula originale (1966), usando la lista di frequenza delle parole di Verlee (1937-39), sostituita in seguito con la lista di Gougenheim (1967).
Nel 1973 Henry, un allievo di De Landsheere, pubblica una fomula specifica per la lingua francese, in 3 versioni: la prima, ritenuta dall'autore quella "ideale" o "più valida" è molto complessa e troppo lunga da applicare, in quanto contiene otto variabili; la seconda contiene cinque variabili ed essendo computerizzata è adatta per essere utilizzata nei centri di ricerca. Questa formula viene modificata nel 1979, con l'aggiunta di un'ulteriore variabile, cioè la percentuale di parole concrete. La terza versione è una formula breve destinata all'uso manuale da parte degli insegnanti ed è costruita su tre variabili: numero medio di parole per frase, lessico non di base (numero di parole assenti dalla lista di Gougenheim del 1967), proporzione di segni attivi (pronomi, punti esclamativi, virgolette o altri segnalatori di discorso diretto). Tutte e tre le formule utilizzano il cloze test come criterio per la convalida, con coefficienti che vanno da 0,84 a 0,93 per i gradi 5-6, da 0, 78 a 0,90 per i gradi 8-9 e da 0,70 a 0,83 per i gradi 11-12.
Richaudeau studia il rapporto tra linguaggio e memoria, basandosi sul presupposto che la lunghezza delle parole influisce sul loro livello di memorizzazione. Nel 1979 propone una formula sperimentale, detta *formule d'efficacité linguistique*, calcolata in tre tempi. Con *efficacité linguistique* l'autore intende il rapporto tra il numero di parole esatte che il lettore ricorda dopo aver letto una data frase e il numero totale di parole che compongono quella frase.

[60] Bamberger 1973, Bamberger e Vanecek 1982, 1984; Bamberger e Rabin 1984.

4.2.4. Olandese

Douma (1960) pubblica una versione della formula *Reading Ease* di Flesch tarata sulla lingua olandese, tenendo conto del fatto che le parole e le frasi risultano più lunghe del 10% rispetto alla lingua inglese.

$$Ease = 206.84 - 0.77sw - 0.93ws$$

Dove:
sw = numero di sillabe su 100 parole
ws = parole per frase

Anche Brouwer (1963) impiega una versione adattata della formula di Flesch, scegliendo come criterio un corpus di 25 libri per bambini:

$$Ease = 195 - 2/3\ sw - 2ws$$

Van Hauwermeiren (1972) sviluppa 6 formule utilizzando diverse combinazioni di variabili; i coefficienti di correlazione con il criterio cloze vanno da 0,60 a 0,67.
Zondervan, van Steen e Gunneweg (1976) valutano la leggibilità di testi di saggistica destinati ai gradi da 3 a 6 e sviluppano una formula per ognuno di questi livelli. Staphorsius e Krom (1985) creano un indice per l'uso manuale a 5 variabili e una formula automatizzata a 2 variabili (lunghezza delle parole in lettere e lunghezza della frase in parole); entrambi i metodi valutano materiali di saggistica per i gradi da 3 a 6.

4.2.5. Hindi

Bhagoliwal (1961) applica le formule di Flesch (1948), di Farr, Jenkins e Paterson (1951) e di Gunning (1952) a 31 racconti brevi in Hindi. Utilizza questi indici in quanto prevedono il conteggio delle sillabe; non può invece ricorrere a formule che prevedono liste di parole perché in hindi non è disponibile alcun elenco. La formula di Farr, Jenkins e Paterson risulta essere la migliore dal momento che non comporta il conteggio delle parole polisillabiche, che presenta problemi per la lingua hindi.

4.2.6. Svedese

La maggior parte delle formule di leggibilità ha la forma di un'equazione di regressione. Björnsson (1968a, 1968b, 1983) è il primo ad abbandonare questa tipologia a favore di una formula additiva, in cui i fattori linguistici sono semplicemente sommati e il risultato confrontato con una serie di criteri predeterminati. Björnsson sviluppa una formula semplice, detta Lix, abbreviazione di *Läsbar-hetsindex* ('indice di leggibilità' in svedese)[61]:

$$Lix = lunghezza\ della\ frase + lunghezza\ della\ parola$$

La lunghezza della frase è misurata in parole; la lunghezza della parola è considerata come la percentuale di parole con più di sei lettere. I punteggi vanno da 20 (molto facile) a 60 (molto difficile). La Tabella 10 mostra due diversi esempi di interpretazione dei punteggi:

[61] Il sito ufficiale della formula Lix è www.lix.se

Valore LIX	Descrizione 1	Valore LIX	Descrizione 2
20	very ease	20-25	Children's books
30	easy	31-35	Fiction
40	average	40-45	Newspapers
50	hard	50-55	Science reports
60	Very hard	60	Government texts, law texts, ecc.

Tabella 10. Due diverse interpretazioni dei valori di LIX.

Nel corso degli anni Sessanta e Settanta, Björnsson testa con il suo metodo migliaia di libri e conduce indagini sulla leggibilità in diverse lingue (svedese, danese, inglese, francese, tedesco, finlandese, ecc.).

Lingua	SL	WL	Lix
Svedese	16	23	39
Danese	19	22	40
Inglese	19	21	40
Tedesco	17	29	45
Francese	19	27	46
Finlandese	12	48	60

Tabella 11. Confronto tra lunghezza della frase (SL), lunghezza delle parole (WL) e Lix (SL + WL).

Lo studio del 1974 è condotto sulla lingua inglese. Come criterio sono scelti 100 testi di varia natura, i cui livelli di difficoltà sono giudicati da due gruppi di 14 persone ciascuno; la correlazione tra le valutazioni medie dei due gruppi risulta piuttosto alta (0,99). Sulla base dei risultati, l'autore sostiene che, contrariamente a quanto si crede, la valutazione della difficoltà dei testi da parte di giudici può essere affidabile se si rispettano tre condizioni: se le valutazioni sono effettuate da un numero sufficientemente alto di persone, se i brani sono relativamente lunghi e se la gamma di difficoltà del corpus di testi è ampia. In media, gruppi di 6 persone danno un coefficiente di correlazione di 0,94, gruppi di 12 di 0,97 e gruppi di 24 o più persone di 0,99. Con gruppi più grandi, otterremo probabilmente le stesse correlazioni medie.
Inizialmente Björnsson sviluppa un'equazione di regressione, basandosi sul calcolo della correlazione multipla e utilizzando i 100 testi come criterio. Tuttavia, dividendo i testi in due parti (da 1 a 50 o da 51 a 100), nota che si ottengono equazioni differenti e diversi coefficienti di validità: questo significa che le equazioni di regressione sono strettamente dipendenti dalla composizione del criterio e che non sono adatte come formule di leggibilità (cfr. Björnsson, 1983).

Testi	Equazioni di regressione
1 - 100	$0{,}15\,WL + 14\,SL - 0{,}86$
1 - 50	$0{,}15\,WL + 13\,SL - 0{,}67$
51 - 100	$0{,}16\,WL + 16\,SL - 1{,}47$

Tabella 12. Equazioni di regressione per Lix.

Lo studioso adotta quindi il metodo additivo. La formula Lix è ottenuta tramite la semplice somma della lunghezza delle parole e delle frasi. La correlazione tra le variabili e il criterio risulta di 0,92.
Nella ricerca del 1983, Björnsson valuta la leggibilità di 250 giornali in 11 lingue, tra cui l'italiano.

Lingua	SL	WL	Lix
Swedish	17	30	47
Norwegian	20	28	48
Danish	22	29	51
English	25	27	52
French	23	32	55
German	22	37	59
Italian	30	35	65
Spanish	35	32	67
Portuguese	36	34	70
Finnish	14	58	72
Russian	18	47	65

Tabella 13. Leggibilità dei giornali in 11 lingue.

La lingua italiana è valutata insieme allo spagnolo e al portoghese, in quanto appartenenti alla stessa famiglia delle lingue romanze. Il francese è invece considerato singolarmente. L'autore afferma di aspettarsi risultati simili dalle tre lingue e valori di leggibilità abbastanza vicini a quelli del francese. In realtà i punteggi risultano piuttosto diversi, anche se tutti molto alti (in media da 65 a 70) e comunque distanti dal francese; probabilmente le differenze sono dovute alla lunghezza della frase.

Giornali	SL	WL	Lix
Italiano			
Corriere della sera	31	36	67
La Stampa	30	35	65
Stampa sera	28	34	62
Spagnolo			
El Pais	37	33	70
Pueblo	35	32	67
La Vanguardia	32	32	64
Portoghese			
Diàrio de Noticias	36	35	71
Tempo	36	34	70
Diàrio Popular	34	34	68

Tabella 14. Confronto tra giornali italiani, spagnoli e portoghesi.

Della leggibilità dei testi svedesi si occupa anche Platzack (1974), che conduce vari studi sull'influenza di fattori fisici, sintattici, semantici e contestuali sul livello di difficoltà dei materiali scritti. Analizza per esempio l'uso della punteggiatura, l'ordine della frase, la lunghezza delle frasi e delle parole, la presenza dei pronomi relativi, ecc.

4.2.7. Danese

Nel corso del 1960 gli editori di vari giornali danesi si interessano a migliorare l'uso dei quotidiani nelle scuole e decidono quindi di rivolgersi all'Istituto Danese di Ricerca Educativa per far valutare il livello di difficoltà linguistica di un certo numero di periodici. Jesper Florander e Mogens Jansen (1966) si stanno occupando di questa ricerca, quando giungono dalla Svezia notizie degli studi di Björnsson. Dato che le due lingue sono molto simili, viene deciso di adattare la formula Lix ai testi in danese e viene istituito un Comitato Lix ufficiale. Il Comitato pubblica tre resoconti ufficiali sull'andamento delle ricerche (Jakobsen 1971, 1976, 1983) e alla fine degli anni '80 risulta ancora attivo (cfr. Rabin 1988).
La formula di leggibilità per il danese è simile a quella originale:

$$Lix = Ml + Lo$$

dove
Ml = lunghezza media del significato (lunghezza della frase)
Lo = percentuale di parole lunghe (parole con più di sei lettere)

Dal 1970 in Danimarca viene fatto un uso continuo della formula Lix, valutando materiali didattici, libri per bambini e molti libri di narrativa per adulti.

4.2.8. Coreano

Park (1974) sviluppa una formula per la lingua coreana che contiene cinque variabili: numero di parole semplici, numero di parole diverse, numero di diverse parole difficili, numero di frasi semplici e quantità di pronomi. La formula è ideata per valutare materiali destinati ai gradi 2-9. Come criterio utilizza 32 libri di lingua e di scienze sociali che il Ministero dell'Istruzione richiede nelle scuole coreane. L'indice risulta più predittivo per livelli di istruzione più bassi.

4.2.9. Inglese

Anche se le indagini sulla leggibilità in lingua inglese sono condotte principalmente negli Stati Uniti, una rassegna sugli studi in altri paesi sarebbe incompleta senza menzionare anche le ricerche condotte sull'inglese in Australia e nel Regno Unito.
Anderson (1965, 1967, 1972) sperimenta l'utilizzo della procedura cloze per valutare i livelli di leggibilità dei libri utilizzati nelle scuole australiane e inizia a interessarsi alla formula Lix di Björnsson. La formula è infatti facile da usare e da interpretare e può essere applicata sia a materiali in lingua inglese che in altre lingue. Anderson (1983) presenta una tabella che converte i punteggi Lix in livelli di istruzione e rende l'applicazione dell'indice ancora più veloce: per evitare fraintendimenti con la formula Lix chiama la propria versione Rix.
Le due variabili considerate in Lix sono la lunghezza della frase (misurata in parole) e la lunghezza delle parole; questo secondo fattore non è però misurato tramite i metodi più tradizionali (conteggio delle sillabe, conteggio dei caratteri, numero di parole monosillabiche o polisillabiche, ecc.) ma è dato dalla percentuale di parole che hanno 6 o più lettere. Anderson nota che il conteggio delle parole lunghe è più accurato e due volte più veloce rispetto al computo delle sillabe; evita inoltre tutta una serie di problemi legati alla sillabazione o all'individuazione delle sillabe in abbreviazioni o numeri. La formula Rix è la seguente:

$$Rix = \frac{numero\ di\ parole\ lunghe}{numero\ di\ frasi}$$

dove
numero di parole lunghe = numero di parole con più di 6 lettere
La correlazione tra le due formule è ovviamente quasi perfetta (0,99).

I due lavori sulla leggibilità più importanti nel Regno Unito sono quelli di Gilliland (1972) e Harrison (1980). Gilliland compie un'analisi più teorica delle variabili linguistiche, fonologiche e fisiche che determinano la difficoltà dei testi. La ricerca di Harrison è invece più sperimentale e prevede una misurazione pratica della leggibilità nelle classi; la sua analisi include i risultati di due studi governativi relativi all'adeguatezza dei libri di testo britannici: l'indagine della commissione Bullock sull'insegnamento dell'inglese (Department of Education and Science, 1975) e il progetto Effective *Use of Reading* (Lunzer e Gardner, 1979).

5. Studi di leggibilità in Italia

Mentre negli Stati uniti le ricerche sulla leggibilità dei testi e sulla comprensione della lettura nascono già negli anni Venti, con lo sviluppo di diverse formule di leggibilità, in Italia tali problemi iniziano a imporsi all'attenzione generale solo alla fine degli anni Sessanta. L'Italia risulta in ritardo anche rispetto agli altri paesi europei: se in questi paesi le prime tarature della formula *Reading Ease* di Flesch si hanno già a partire dalla fine degli anni Cinquanta o primi anni Sessanta, il primo adattamento all'italiano risale invece al 1972 da parte di Roberto Vacca. "In Italia, la formula di Flesch arriva all'inizio degli anni settanta, quando negli Stati uniti d'America era diventata ormai lo strumento di controllo dello standard di leggibilità, imposto dalle leggi della maggior parte degli stati, per i testi di interesse generale e pubblico, a cominciare dalle polizze assicurative e finire agli articoli dei quotidiani e ai testi delle varie amministrazioni destinati ad un pubblico ampio e differenziato" (Piemontese 1996, p. 35). In quegli anni si ha anche la prima indagine sulla comprensione della lettura in Italia, condotta dal Consiglio Nazionale delle Ricerche (CNR), in collaborazione con l'Associazione internazionale per la valutazione del profitto scolastico (IEA = International Association for the Evaluation of Educational Achievement)[62]. Allo stesso modo, mentre in Europa le prime formule originali sono sviluppate a partire dalla fine degli anni Sessanta, la prima formula tarata sulla lingua italiana si ha soltanto alla fine degli anni Ottanta.

"Rispetto ad altre nazioni europee ed extraeuropee, in Italia, infatti, la preoccupazione e l'attenzione alla reale fruibilità dei testi scritti, sia quelli destinati a situazioni didattiche sia quelli destinati alla comunicazione extrascolastica, si sono poste con qualche ritardo che ora si cerca di superare. Questo ritardo, che non è l'unico e forse neppure il più grave, affonda le sue radici nella diversa situazione sociale, culturale e politica dell'Italia che giunge all'unità politica e linguistica solo dopo la metà del XIX secolo. […] Venute meno le ragioni che spiegano il ritardo dell'Italia rispetto ad altre nazioni e acquisite alcune linee di tendenza nel campo degli studi linguistici (studio della lingua in rapporto alla società, nascita e affermazione della statistica linguistica, adattamento dell'italiano di vecchi indici di leggibilità di origine anglo-americana e studio di nuovi tarati sull'italiano, messa a punto della lista di frequenza del lessico italiano e del *Vocabolario di Base*), il problema della fruibilità dei testi scritti si è finalmente potuto porre in tutta la sua dignità scientifica, oltre che per la sua valenza politica, civile e democratica" (Piemontese 1991, pp. 152-154).

[62] I risultati sono presentati negli *Annali della Pubblica istruzione*, quaderno n. 5, 1977.
Negli Stati Uniti il *National Institute of Education* ha istituito il *National Assessment of Educational Progress* (NAEP) con lo scopo di monitorare lo stato e l'evoluzione dei profitti scolastici. Il NAEP ha effettuato 4 grandi rilevazioni tra gli anni '70 e '80, esaminando 250.000 studenti. In Italia, invece, non esiste una tradizione di rilevazione di abilità su grandi campioni. La prima indagine è quella IEA *Six Subject* (1969-71 ma pubblicata nel 1977) che comprende la comprensione della lettura nel pacchetto di prove somministrate. Una nuova rilevazione viene fatta nel 1983 utilizzando le stesse prove di quella precedente.

5.1. La formula di Vacca

Così come si deve a Flesch l'aver reso noto il concetto di leggibilità negli Stati uniti, si deve a Roberto Vacca, ingegnere, matematico e scrittore, la divulgazione degli studi di Flesch e l'introduzione delle formule di leggibilità in Italia.
"Qualcuno si scandalizzerà che la leggibilità di uno scritto possa essere misurata da un numero. Ma non dobbiamo scandalizzarci dei tentativi di misurare qualcosa di oggettivo. Dobbiamo, piuttosto, ricordarci di non trarre dalle misure che facciamo conclusioni generali ingiustificate. Dobbiamo spiegare che cosa misuriamo e come lo misuriamo. […] La misura del coefficiente di leggibilità non ci dice e non ci può dire se un testo è giusto, civile, razionale, scritto bene, né se è pieno di idiozie e falsità. Però da quando ho letto il libro di Flesch (nel 1973) io ho cominciato a scrivere in modo diverso: più chiaro e – spero – migliore. Il coefficiente di leggibilità di Flesch misura almeno uno sforzo che un autore fa per farsi capire. Io, anzi, sono convinto che indichi pure in che misura è riuscito a farsi capire, almeno entro larghi limiti. […] Il coefficiente di leggibilità somiglia un po' al quoziente di intelligenza. È una misura imprecisa di qualche cosa che non si sa bene cosa sia, però qualche significato lo ha" (Vacca 1978).

Vacca (1972, 1978) propone una versione modificata dell'indice *Reading Ease* di Flesch (1949)[63]. La taratura adottata dallo studioso si basa sulla misurazione manuale della leggibilità tramite l'indice di Flesch di un certo numero di campioni di testi italiani.
La formula *Reading Ease* originale:

$$\text{Reading Ease Score} = 206{,}835 - 0{,}846\text{W} - 1{,}015\text{S}$$

Vacca ottiene una nuova formula, eliminando la seconda e terza cifra decimale dei parametri e tenendo conto del fatto che in media le parole italiane sono più lunghe di quelle inglesi (hanno circa il 40% di sillabe in più):

$$\text{IL} = 206 - 0{,}6\text{W} - \text{S}$$

dove
IL = Indice di leggibilità
W = numero medio di sillabe per 100 parole;
S = numero medio di parole per frase;
206 è una costante fissa, scelta per fare in modo che i valori oscillino da 0 a 100.

Nel 1981 Vacca pubblica la formula in un suo libro[64] e Valerio Franchina la inserisce nel suo word processor (il primo realizzato in Italia nel 1972). Il programma premette il conteggio delle parole e delle sillabe ed è predisposto per effettuare la sillabazione a fine riga, secondo le regole della grammatica italiana.

[63] Per indicazioni specifiche sull'Indice di Flesch e l'adattamento di Vacca cfr. Vacca 1972, 1978, Franchina e Vacca 1986, De Mauro e Piemontese 1986, Fiorucci 1986, Lucisano e Piemontese 1988, Lucisano 1992, Piemontese 1996 e De Mauro e Chiari 2005.
[64] Roberto Vacca, *Come imparare più cose e vivere meglio*, Arnoldo Mondadori, 1981.

Nel 1986 propone un'altra variante della formula, usando come tecnica di taratura il confronto tra campioni di testi appartenenti ad un'unica opera (il suo libro *Rinascimento Prossimo Venturo*) nelle due versioni, italiana e inglese[65]. Essendo scritti simultaneamente da un autore perfettamente bilingue, i testi dovrebbero avere gli stessi indici di leggibilità.
Vacca e Franchina analizzano due campioni (italiano e inglese) di 22 testi, corrispondenti alle prime pagine di ciascuno dei 22 capitoli del libro. Come si vede nella tabella seguente, in italiano si ha un numero di sillabe maggiore del 35,6% rispetto all'inglese ma un numero di parole inferiore del 5,3%.

Caratteristiche	**Inglese**	**Italiano**
Numero delle frasi	406	401
Numero delle parole	6.499	6.154
Numero delle sillabe	10.186	13.821

Tabella 15. Confronto tra i 22 testi in inglese e in italiano.

Gli autori calcolano la leggibilità con la formula di Vacca e ottengono punteggi di 57,351 per i testi inglesi e 57,373 per quelli italiani. La correlazione è di 0,9092. Franchina applica ai dati un procedimento di regressione lineare a 3 parametri e determina nuovi coefficienti per la formula italiana, che risulta:

$$\mathrm{IL} = 216{,}88 - 0{,}61785\mathrm{W} - 1{,}282\mathrm{S}$$

La formula viene poi arrotondata a:

$$\mathrm{IL} = 217 - 0{,}6\mathrm{W} - 1{,}3\mathrm{S}$$

Il coefficiente di correlazione sale a 0,9398. Secondo gli esperimenti condotti dal Gruppo Universitario Linguistico Pedagogico (GULP) la formulazione del 1972 risulta però più attendibile.

5.2. Applicazioni della formula di Flesch – Vacca

Le occasioni di riflessione sul tema della comprensibilità e della leggibilità dei testi si moltiplicano a partire dagli anni Settanta e soprattutto nel corso degli anni Ottanta. La prima versione della formula di Vacca (1972) viene usata in vari studi e con diversi obiettivi; la sua applicazione sistematica a testi in lingua italiana ne evidenzia gli aspetti positivi e negativi.
Tra i vari dibattiti e i lavori svolti in quegli anni si segnalano: il dibattito aperto tra il 1976 e il 1978 da alcuni quotidiani e periodici italiani sulla semplificazione della comunicazione; le ricerche che hanno portato alla definizione del *Vocabolario di Base* della lingua italiana e alla nascita della collana dei Libri di Base (1980); l'analisi della leggibilità dei Libri di Base curata da Tiziana Fiorucci (1982); il XIX Congresso Internazionale della Società di Linguistica Italiana del novembre 1985 dedicato al problema della percezione, comprensione e

[65] Cfr. Franchina e Vacca 1986.

interpretazione, visto dalla parte del ricevente (De Mauro, Gensini, Piemontese 1985); l'incontro di studio *Leggibilità e comprensione* organizzato dalla cattedra di Filosofia del Linguaggio dell'Università La Sapienza e tenutosi a Roma il 26-27 giugno 1986 (De Mauro, Piemontese, Vedovelli 1986); l'analisi della leggibilità di manuali scolastici da parte di Anna Thornton (1984); le analisi di leggibilità svolte dalla cooperativa Spazio Linguistico (Palombi e Raponi 1984, Palombi 1986); l'analisi di leggibilità di testi giuridici e politici ad opera dell'Istituto di Documentazione Giuridica del CNR di Firenze (Martino e Bianucci 1986); le ricerche condotte nell'ambito di seminari intercattedra di Filosofia del linguaggio e Pedagogia dell'Università La Sapienza sulla leggibilità di testi scolastici e sul livello di comprensione degli studenti (Lucisano e Piemontese 1986); la formazione (sotto la supervisione di De Mauro e Piemontese) di un gruppo di studenti (Gruppo H) fra il 1984 e il 1987 alla redazione di testi di alta leggibilità, diretti in particolare a persone con deficit intellettivo (Piemontese e Vedovelli 1988); l'analisi da parte del Gruppo H, con il coordinamento di Piemontese e Tiraboschi, di una serie di libretti informativi del sindacato.

L'8 settembre 1979 il quotidiano *La Repubblica* pubblica un dossier dal titolo *Ma chi ci capisce?* dedicato all'oscurità del linguaggio politico e giornalistico italiano[66]. Il dossier, che ripropone il tema del dibattito aperto negli anni precedenti da vari quotidiani e periodici italiani sull'opportunità o meno della semplificazione della comunicazione, comprende un articolo di Emilia Passaponti che presenta l'indice di Flesch e i risultati dell'applicazione della formula (nella versione tarata da Vacca) a numerosi articoli di giornale e romanzi usciti nel 1979. I punteggi risultano piuttosto bassi, con valori molto differenziati e che raramente superano 40: "Il problema della circolazione della informazione è molto complesso: nessuno ha in mano la ricetta o la formula magica. E, tuttavia, l'attenzione ai livelli di leggibilità, specie per i quotidiani, è una condizione necessaria, anche se non sufficiente" (Passaponti 1980).
Sempre alla fine degli anni settanta, Cornoldi e il gruppo di ricerca MT dell'Istituto di Psicologia di Padova mettono a punto un pacchetto di prove oggettive per la verifica della comprensione della lettura per tutte le classi della scuola dell'obbligo (dalla prima elementare alla terza media): le prove di lettura MT[67]. Nel costruire le prove gli autori analizzano le diverse caratteristiche linguistiche dei testi, impiegando la formula di Flesch e il LIF.

5.2.1. La leggibilità dei Libri di Base

I dati sulla lettura e l'istruzione negli anni Ottanta mostrano che il numero di lettori in Italia è molto esiguo e che spesso i contenuti rivolti al grande pubblico sono così poco comprensibili che risultano difficili persino per gli strati più scolarizzati. "Il fatto è che la situazione italiana è caratterizzata da un forte ritardo per quanto riguarda l'attenzione alla trasparenza del linguaggio, che ha invece negli altri paesi una storia molto più lunga. […] All'abitudine straniera ad una scrittura semplice coerente o chiara si contrappone una

[66] *Ma chi ci capisce?* in «La Repubblica dossier», n. 33, 8 settembre 1979, con articoli di G, Bocca, T. De Mauro, S. Gensini, E. Passaponti. Dal momento che non è stato possibile rintracciare il documento, le informazioni sono tratte da Passaponti 1980 e Thornton 1992. Vedere anche Piemontese 1996 (p. 65).
[67] Cfr. Cornoldi, Colpo e Gruppo MT (1981, 1986), Cornoldi e Colpo (1981, 1998).

tradizione tutta italiana di oscurità, che, trasmessa dalla scuola, ha contribuito a tenere molti lontani dalla carta stampata. Se ormai la maggioranza degli italiani condivide la conoscenza della lingua parlata, quella scritta si ripropone, ancora una volta, come mezzo di discriminazione. Il mondo della stampa che potrebbe svolgere una funzione fondamentale nella trasmissione e diffusione della cultura rimane ancora riservato a pochi" (Fiorucci 1982, p. 40).

Dall'aprile del 1980 esce presso gli Editori Riuniti una collana diretta da Tullio de Mauro, chiamata Libri di base, che ha come obiettivo quello di produrre volumi in grado di rivolgersi e di soddisfare le esigenze culturali di un pubblico molto vasto. Ogni volume è affidato a uno specialista; prima della pubblicazione il testo viene sottoposto ad una riscrittura, in modo da raggiungere una più alta leggibilità e comprensibilità. Come livello di istruzione medio di riferimento viene scelta la III media, cioè l'ultima classe della scuola dell'obbligo[68]. "L'uso dei criteri stilistici ricavati dall'applicazione sistematica della formula di Flesch (periodare breve e organizzazione opportuna dei contenuti) e la preferenza accordata a parole di uso comune, usate anche per spiegare (e non per sostituire) anche le parole di uso meno comune o tecnico-specialistico necessarie, hanno costituito la base delle tecniche di redazione usate nei dei Libri di base" (Piemontese 2005, p. 390). Nel 1982 la collana dei Libri di base conta già 47 volumi, ciascuno con una tiratura media di 16.000 copie e una vendita media di 12.800 copie.

Parallelamente, prende avvio una ricerca rivolta alla costruzione di una lista di parole che un lettore in possesso della terza media possa sicuramente conoscere. Vengono esaminate le liste di frequenza dell'italiano scritto e, da queste, sono eliminati i vocaboli che non risultano compresi dalla maggior parte dei lettori con tale livello di istruzione (si tratta soprattutto di parole di tradizione letteraria); all'elenco vengono aggiunte parole che raramente si trovano scritte ma che sono ben note a tutti (ad esempio *dentifricio*). Il risultato è il *Vocabolario di Base della lingua italiana* (VdB), una lista di 6690 vocaboli diversi, pubblicata in appendice al volume di De mauro, *Guida all'uso delle parole* (Libri di base n. 3)[69].

Fiorucci (1982) verifica il livello di leggibilità di 40 diversi libri di base, impiegando la formula di Flesch nella versione adattata all'italiano proposta da Vacca (1972). I valori ottenuti con la formula di Flesch vanno da 0 (molto difficile) e 100 (molto facile). Poiché dall'analisi risulta che i testi italiani danno spesso punteggi negativi, o comunque raramente superiori a 80, Fiorucci propone una tabella di valori più adeguata all'italiano:

[68] Oggi l'istruzione obbligatoria è estesa fino ai 16 anni. L'adempimento dell'obbligo scolastico è disciplinato dalla Circolare Ministeriale n. 101 del 30 dicembre 2010 (che afferma che l'obbligo di istruzione riguarda la fascia di età compresa tra i 6 e i 16 anni), dal Decreto Ministeriale n. 139 del 22 agosto 2007 (che afferma che l'istruzione obbligatoria deve essere impartita per almeno 10 anni) e dalla Legge n. 296 del 27 dicembre 2006 (che afferma: "L'istruzione impartita per almeno dieci anni è obbligatoria ed è finalizzata a consentire il conseguimento di un titolo di studio di scuola secondaria superiore o di una qualifica professionale di durata almeno triennale entro il diciottesimo anno d'età").

[69] De Mauro 1980.

Valore	Interpretazione
≤ 0	molto difficile
0 - 30	difficile
30 - 40	abbastanza difficile
40 - 50	standard
50 - 60	abbastanza facile
60 - 70	facile
80 - 90	molto facile

Tabella 16. Adattamento all'italiano dei punteggi di Flesch (cfr. Fiorucci 1982, p.48).

Dei 40 volumi analizzati, 25 presentano un indice di leggibilità superiore a 40, 5 sono molto vicini a 40, 7 hanno valori compresi tra 36 e 39 e solo 3 sono inferiori a 35 (il più basso ha un punteggio di 30,65). Dal controllo risulta quindi che la maggior parte dei volumi è scritta in modo da risultare comprensibile a lettori con livelli di istruzione medio-bassi.
"I risultati concordarono nel confermare la validità della formula, che, nel suo impianto generale, si mostra tanto più adatta all'italiano scritto, caratterizzato, rispetto ad altre lingue europee, da un periodo molto complesso e da un'abbondanza di sinonimi. Inserita nel progetto editoriale, non solo fornisce delle indicazioni per conseguire i fini che i Libri di base si propongono, ma costituisce uno strumento efficace e di facile uso per un rapido controllo dei risultati raggiunti" (Fiorucci 1982).

5.2.2. La leggibilità dei manuali scolastici

Nel 1984 Anna Thornton, collaboratrice della cooperativa Spazio Linguistico, pubblica sulle pagine di *Riforma della Scuola* l'analisi di leggibilità di vari libri di testo in adozione nelle scuole italiane[70]. La formula impiegata per la valutazione è l'adattamento italiano dell'indice di Flesch.
Come esempi, sono riportati i dati relativi all'analisi di manuali di letteratura italiana, filosofia e biologia. Per quanto riguarda i libri di letteratura, l'analisi è condotta soltanto su campioni di testi redatti dai curatori e non, ovviamente, su brani letterari o di letteratura critica classica; va tenuto presente che si tratta di una percentuale minima (che va dal 6% al 18%) rispetto all'intero volume e che probabilmente è proprio questo ad incidere sul risultato positivo della leggibilità. Tutti i testi analizzati, infatti, tranne Basile, si collocano nella fascia dei valori *standard* (che va da 40 a 50). Risultati peggiori si hanno per i volumi di filosofia, che risultano in media nella fascia dell'*abbastanza difficile* (da 30 a 40). I manuali di biologia sembrano essere una via di mezzo: dei 17 manuali analizzati, 10 si collocano nell'area del *difficile* (valori inferiori a 40), 6 presentano punteggi *standard* e 1 arriva addirittura alla fascia dell'*abbastanza facile* (da 50 a 60)[71].

[70] Cfr. Thornton 1984a, 1984b, 1984d, 1984e. Anche il GISCEL Lombardia ha condotto dal 1986 al 1992 una ricerca sull'analisi della leggibilità dei manuali di scienze (cfr. Guerriero 1988, Zambelli 1994).
[71] Per i punteggi si fa riferimento alla tabella proposta da Fiorucci (1982)

Testi	Flesch
Manuali di letteratura	
Pazzaglia	49,85
Gianni	48,54
Ceserani	48,40
Salinari	48,40
Giudice	46,61
Petronio	46,26
Marchese	41,01
Basile	32,62
Manuali di filosofia	(media vol.)
Dal Pra	44,47
Lamanna	41,22
Perone-Ferretti-Ciancio	36,19
Moravia	35,19
Adorno-Gregory-Verra	33,41
Vegetti-Alessio-Fabietti-Papi	32,62
Giannantoni	31,16
Merker (a cura di)	30,77
Voltaggio	27,20
Manuali di biologia	(solo vol. 1)
Longo-Longo	51,7
Battaglini-Totaro	42,93
Oram	42,38
Falaschi-Galizzi	40,16
Caramiello-Lomagno	38,2
Montalenti-Giacomini	36,89
Curtis	36,76
D'Alessandro	36,55
Capanna-Mainardi-Sparvoli	34,27
Aloj	32,68
Terrenato	27,8

Tabella 17. Valori di leggibilità dei manuali scolastici.

5.2.3. Le analisi di leggibilità della cooperativa Spazio Linguistico

La cooperativa Spazio Linguistico raccoglie ricercatori e collaboratori della cattedra di Filosofia del Linguaggio dell'Università La Sapienza di Roma (prof. Tullio De Mauro) e si occupa di applicare metodologie e strumenti per la misurazione della leggibilità in diversi settori di studio. Dal 1981 al 1986 i principali settori d'intervento della cooperativa sono tre: l'editoria scolastica (semplificazione di manuali scolastici, controllo della leggibilità nella redazione di testi per la scuola), la comunicazione aziendale (interventi commissionati da aziende private) e la comunicazione di massa (ricerche e interventi svolti per enti pubblici, grandi associazioni o istituti di ricerca).

L'esperienza di Spazio Linguistico mette in luce un aspetto interessante, cioè il fatto che nel processo di valutazione della leggibilità e semplificazione del materiale è importante tenere conto non solo delle abilità linguistiche dei lettori ma anche dei diversi scopi della comunicazione. E che, in base a questi, cambia proprio il modo in cui si deve considerare il "problema leggibilità": "Una prima considerazione, direttamente legata alla valutazione della leggibilità, riguarda il diverso valore che il concetto stesso di leggibilità assume nei diversi ambiti comunicativi e, soprattutto, il suo carattere *non monofunzionale*. […] Sembra dunque produttivo progettare interventi per la valutazione della leggibilità e la semplificazione non solo in relazione alle abilità linguistiche dei destinatari, ma anche in relazione agli *scopi* della comunicazione" (Palombi 1986, p. 128-129).

Uno dei progetti della cooperativa consiste nell'analisi della leggibilità dei giornali aziendali dell'ANCC (Associazione Nazionale delle Cooperative di Consumo) per gli anni 1981 e 1982. Lo strumento impiegato per la misurazione della leggibilità è l'indice di Flesch, nell'adattamento di Vacca.

Dall'applicazione della formula emergono diverse perplessità, tra cui il fatto che l'indice (o meglio, il fattore che considera la lunghezza delle parole in sillabe) non sembra riflettere la complessità lessicale dei testi, data dalla percentuale delle parole che appartengono al *Vocabolario di Base*.[72]. La tabella seguente riporta i risultati dell'analisi della leggibilità. La prima colonna riporta la percentuale delle parole non appartenenti al VdB; la seconda il valore medio di leggibilità misurata con l'indice di Flesch, la terza il numero medio di sillabe su 100 parole (S) e l'ultima il numero medio di parole per frase (P).

Extra VdB	F	S	P
16,81	47,4	224,8	23,6
10,53	44,5	236	19,1
16,48	42,9	230.9	24,3
15,49	31,5	251	24
18,78	31,2	238,4	31,2
15,76	29,4	426,4	28,4
17,21	27,7	258	27,4

Tabella 18. Analisi della leggibilità dei giornali dell'ANCC.

[72] Il vocabolario di base di una lingua è formato principalmente dalle parole con maggiore frequenza che, di solito, sono anche le più brevi.

Il risultato atteso è una relazione di proporzionalità inversa tra parole che non appartengono al *Vocabolario di Base* e indice di Flesch: all'aumentare del primo valore dovrebbe calare il secondo. Quello che emerge è invece una difformità tra i due punteggi. "La discordanza fra presenza di vocabolario di base e valori dell'indice di Flesch deriva semplicemente dal fatto che la prima non sembra sufficientemente rappresentata dal calcolo della lunghezza media in sillabe delle parole. In altri termini, *ad un minore numero medio di sillabe non corrisponde una maggiore presenza di vocabolario di base*. Questo non vuol dire, a nostro giudizio, che sia smentita la valutazione secondo la quale le parole più brevi di una lingua sono, mediamente, anche le più usate e conosciute. Indica piuttosto che *l'indice di calcolo utilizzato nella formula di Flesch non è forse il più adatto ad esprimere tale correlazione*" (Palombi 1986, p. 131).

5.2.4. Due parole, il gruppo H e la redazione di testi ad alta leggibilità

Nel 1984, ancora nell'ambito della cattedra di Filosofia del Linguaggio dell'Università La Sapienza di Roma, viene attivato un seminario che si occupa di formare gli studenti alla redazione di testi ad alta leggibilità rivolti in particolare a portatori di deficit intellettivo. "Il problema del rapporto tra leggibilità, comprensibilità e comprensione dei testi si complica ulteriormente quando la diversità del lettore affonda le sue radici in forme di *ritardo mentale* (RM) più o meno lieve. Salvo che in ambienti neuropsichiatrici e clinici, anche questo problema è rimasto abbastanza ai margini della ricerca e della didattica, in Italia" (Piemontese 1991, pp. 157-158). Il corso, a cui partecipano decine di studenti, si conclude con la costituzione di un gruppo di 13 persone, chiamato Gruppo H, che sotto la supervisione di De Mauro e Piemontese continua ad occuparsi di problemi di comprensibilità e leggibilità dei testi.

L'attività del gruppo H prevede da un lato l'individuazione di criteri per la redazione di testi di facile lettura e dall'altro l'applicazione e la verifica di questi; vengono quindi prodotti testi ad alta leggibilità (in base alle regole stabilite) e viene controllato il loro livello di comprensione in soggetti portatori di ritardo mentale lieve o medio-lieve. "Alla individuazione e taratura dei criteri per la redazione di testi scritti di alta leggibilità e comprensibilità si è giunti attraverso un faticoso e paziente lavoro di analisi della leggibilità di testi informativi e formativi: quotidiani e periodici, da una parte, e manuali scolastici, dall'altra" (id., p. 158)[73].

[73] Cfr. Vedovelli 1986, Piemontese-Tiraboschi 1986, Piemontese-Vedovelli 1988, Piemontese 1991 e 1996.

Nel 1985 il gruppo H conduce una ricerca, promossa dal Centro di Formazione Professionale del Comune di Scandicci, sui problemi di comprensione del linguaggio da parte di soggetti con handicap mentale. I risultati vengono presentati al XIX Congresso della SLI (Società Linguistica Italiana) dello stesso anno.

L'indagine è condotta su un piccolo gruppo di giovani tra i 19 e i 30 anni; i testi proposti per le prove di comprensione sono tratti da materiale di formazione tecnica in falegnameria e articoli di giornale di vario genere. Viene effettuata una verifica iniziale della leggibilità dei testi mediante l'indice di Flesch adattato all'italiano e una valutazione della presenza del vocabolario di base. A questa prima fase, segue un processo di riscrittura e semplificazione linguistica del materiale; si ha poi una nuova verifica dell'effettiva facilità di lettura e comprensibilità dei brani prodotti, con una nuova misurazione dell'indice di Flesch.

La percentuale di risposte errate è molto alta (dal 27% al 50%), a cui si aggiunge un'altra grande percentuale di domande lasciate senza risposta da parte di alcuni soggetti. "Questo comportamento

I criteri per la redazione di testi di facile lettura risultano i seguenti[74]:

1. I testi non devono superare le 150-250 parole complessive.
2. Le parole devono essere tratte il più possibile dal *Vocabolario di Base*. Quelle che non appartengono alla lista devono sempre essere spiegate con parole del VdB.
3. Le frasi devono essere brevi e semplici (preferire le coordinate alle subordinate).
4. È preferibile ripetere il soggetto e l'oggetto invece di ricorrere ai pronomi, che possono generare equivoci. Il pronome *che* va usato possibilmente solo in funzione di soggetto e non di complemento oggetto.
5. Ripetere quando necessario le parole-chiave.
6. I verbi devono essere prevalentemente di modo finito.
7. Tra i modi verbali, preferire l'indicativo al congiuntivo.
8. Evitare il più possibile la forma passiva.
9. Tra i tempi dell'indicativo usare il presente, il passato prossimo e il futuro semplice.
10. Evitare le doppie congiunzioni e le doppie negazioni.
11. Porre attenzione nell'uso dei connettivi, per non causare ambiguità.
12. Evitare le nominalizzazioni e le personificazioni.

Il progetto di lavoro si concretizza nella pubblicazione nel 1989 di un periodico informativo ad alta leggibilità, chiamato *Due parole. Mensile di facile lettura*[75], rivolto a portatori di ritardo mentale lieve ma che viene utilizzato anche nei corsi di italiano per stranieri, nei corsi per lavoratori adulti italiani e stranieri, da comunità di anziani e nella scuola dell'obbligo, "tutte categorie di persone che hanno sete di informazione attuale sulla vita italiana e i fatti del mondo e non possono certa soddisfarla alla fonte del nostro giornalismo gassato e barocco" (De Mauro 1994, p. 18).
Il periodico è composto da 8 pagine; gli argomenti trattati nella rivista sono scelti in base a due criteri: sono tematiche vicine alla quotidianità e all'esperienza dei lettori e sono notizie che durano nel tempo. La prima pagina è dedicata alla notizia principale del mese, la seconda allo spettacolo (musica, cinema, ecc.), la terza alla vita in casa (istruzioni, ricette e

---

non sembra assimilabile direttamente a quello di mancata comprensione, ma ad uno stadio dove la componente emotivo-motivazionale agisce come un filtro che non spinge il soggetto a ricercare con autonomia la comprensione dl testo e la realizzazione della prova" (Vedovelli 1986). Si registra un miglioramento delle prestazioni di comprensione per quanto riguarda i testi tecnici riscritti e semplificati; per i testi comuni non vi è regolarità e, anzi, in alcuni casi, succede il contrario. "Noi interpretiamo questo dato come conferma del fatto che nessun tecnicismo (sia esso una prova strutturata o una tecnica di semplificazione) può da solo attivare i processi di comprensione e di interpretazione se non si allaccia ai sistemi motivazionali che si strutturano in un contesto più ampio. […] In altri termini, è difficile dire se sia stata verificata positivamente l'ipotesi che i testi semplificati linguisticamente rendano la comprensione più facile, quale che sia il suo contesto e lo stato motivazionale del soggetto" (id.).
Nel 1985-86 viene stabilita una collaborazione con gli operatori di due Centri per la Formazione Professionale di Roma e una cooperativa, il Centro sociale al Parco, che segue giovani portatori di handicap psichico. Lo studio prevede una verifica della comprensione di testi ad alta leggibilità tramite prove della lettura e questionari a scelta multipla (con domande poste nella sola forma affermativa). I risultati mostrano che una scrittura attenta ai problemi dei destinatari può portare ad un miglioramento nella comprensione e che è effettivamente possibile redigere testi accessibili a chi ha un lieve o medio ritardo mentale.

[74] Cfr. Piemontese e Tiraboschi 1986.

[75] Indicazioni più specifiche sul mensile *Due parole* si trovano in Piemontese 1996 (cap. 6 e 7).

consigli pratici), le successive alla politica italiana, alla politica estera e allo sport. L'ottava e ultima pagina è dedicata alla cultura[76].
Dopo otto anni di pubblicazione, nel 1997 *Due parole* viene sospeso a causa degli elevati costi di stampa e spedizione. Nel 1998 alcuni dei suoi redattori fondano *Parlar chiaro. Associazione per la semplificazione della comunicazione di interesse pubblico* e qualche anno più tardi la rivista *dueparole* torna ad essere pubblicata in formato telematico sul sito www.dueparole.it.

5.2.5. La leggibilità di testi politici e giuridici

L'Istituto di Documentazione Giuridica (I.D.G.) del CNR di Firenze si occupa di analizzare la leggibilità di testi politici[77]. Il materiale della ricerca è costituito dalle trascrizioni di interventi dei tre Segretari politici dei maggiori partiti italiani (Berlinguer, Craxi, De Mita) negli anni 1980-83. Gli interventi sono effettuati in varie sedi: parlamento, interviste alla stampa, comizi, congressi di partito, Comitati Centrali e Consigli Nazionali, conferenze stampa televisive, appelli tv agli elettori. Il corpus è costituito da circa 130.000 occorrenze.
La leggibilità è valutata con la formula di Flesch nell'adattamento di Vacca[78], facendo riferimento per i punteggi alla tabella proposta da Fiorucci (1982).

Materiale	**Berlinguer**	**Craxi**	**De Mita**	**Totale**
Intervista alla stampa	56,11	54,27	60,90	57,09
Conferenza stampa tv	50,77	53,13	51,37	51,76
Appelli tv	48,48	46,86	56,57	50,64
Congressi di partito	39,38	35,50	43,92	39,60
Parlamento	37,92	39,77	33,60	37,10
C. C. e C. N.	37,20	34,51	39,52	37,08
Comizi	34,70	39,91	32,29	35,63
TOTALE	43,50	43,42	45,45	44,12

Tabella 19. Analisi della leggibilità degli interventi politici (valutata con l'indice di Flesch).

I risultati mostrano che i coefficienti di leggibilità variano a seconda dell'ambito in cui il politico parla e dell'uditorio a cui si rivolge: valori medi tra 50 e 60 (abbastanza facile) per

[76] Esiste in Svezia un periodico informativo equivalente a *Due parole*, chiamato *8 Sidor. Lättlästa Nyheter* ('8 pagine. Notizie di facile lettura'). Il giornale, pubblicato per la prima volta nel 1985, ha gli stessi destinatari, gli stessi obiettivi, le stesse tematiche e le stesse caratteristiche linguistiche e testuali di quello italiano. La differenza sta nella diversa origine dell'iniziativa: mentre quello italiano è il prodotto di studi e ricerche svoltosi in ambito accademico, quello svedese è nato per iniziativa del governo stesso.
Altre esperienze simili sono avviate in Finlandia con il quindicinale *Selko-Uutiset* ('Notizie chiare'), anch'esso promosso dal governo, in Norvegia con il periodico *Klar Tale. Lettlest ukeavis* ('Discorso Chiaro. Settimanale di facile lettura'), in Belgio con il mensile di lingua francese *L'essentiel. L'actualité simple comme bonjour* e quello di lingua olandese *Wa Blieft*.
[77] Cfr. Martino e Bianucci 1986.
[78] Per l'applicazione della formula effettuata tramite procedure elettroniche cfr. Mercatali et al. 1979.

appelli tv, conferenze stampa, interviste alla stampa, valori medi tra 30 a 40 (abbastanza difficile) per congressi, Comitati Centrali e Consigli Nazionali, comizi e parlamento.
L'ipotesi degli studiosi è che il linguaggio dei politici sia in un certo modo funzione dell'uditorio. Nel gruppo di testi con una leggibilità più alta, il politico parla principalmente agli elettori, all'opinione pubblica; nell'altro gruppo si rivolge invece agli altri membri del partito o agli avversari. L'unica eccezione è rappresentata dai comizi, per i quali ci si aspetterebbe una maggiore leggibilità; in realtà questo valore è dovuto al fatto che si tratta principalmente di interventi diretti ad affiliati e simpatizzanti e non di comizi elettorali.
Gli autori manifestano alcune perplessità per questi risultati. Sembra infatti che queste differenze sostanziali tra i testi siano determinate principalmente da una diversa lunghezza dei testi, che è appunto uno dei due parametri della formula. Le conferenze stampa, le interviste o gli appelli in televisione hanno di solito una struttura dialogica, che vede l'alternarsi di più persone: questo determina naturalmente frasi più brevi (domanda e risposta). L'obiezione è che "non necessariamente un periodo breve è più leggibile di uno ben più lungo in quanto entrano in gioco molti fattori, primo fra tutti la scelta dei vocaboli" (Martino e Bianucci 1986, p. 58).
Per questo motivo, gli studiosi decidono di applicare un altro indice di leggibilità che tiene conto anche della scelta del lessico e adattano all'italiano la formula di Dale e Chall (1948).
La formula di Dale e Chall è costruita con due variabili, la lunghezza media delle frasi espressa in numero di parole e il numero di parole non comuni, ovvero non presenti nella lista di Dale delle 3.000 parole più frequenti:

$$reading\ grade\ score\ = 0{,}1579x_1 + 0{,}0496x_2\ + 3{,}6365$$

dove
x_1 indica il numero di parole fuori dalla lista di Dale
x_2 indica la lunghezza media della frase
3,6365 è una costante

Il parametro di complessità sintattica è lo stesso della formula di Flesch, cioè la lunghezza delle frasi. La difficoltà semantica, misurata da Flesch in termini di lunghezza delle parole in sillabe, è invece per Dale e Chall data dalla familiarità delle parole[79]. Per l'adattamento alla lingua italiana è necessario sostituire la lista delle parole di Dale con il *Vocabolario di Base*.
Tra i vocaboli comuni, oltre ai nomi propri, vengono inserite anche le sigle, i numerali, le esclamazioni, le parole tronche, le parole vuote, le parole composte (solo quando entrambe le componenti risultano inserite nel VdB), gli avverbi che terminano in *–mente* derivati da verbi, sostantivi o aggettivi presenti nel VdB.

[79] In realtà anche questo criterio, nonostante superi alcune obiezioni mosse per l'indice di Flesch, non è totalmente esente da perplessità. Si consideri ad esempio il fatto che lo stesso vocabolo possa avere più di un'accezione e che tra queste solo una sia considerata comune: non è però detto che chi scrive adoperi questa accezione e non un'altra meno familiare.

Materiale	Berlinguer	Craxi	De Mita	Totale
Intervista alla stampa	5,53	5,54	5,49	5,52
Conferenza stampa tv	5,64	5,70	5,57	5,64
Appelli tv	5,52	5,95	4,72	5,46
Congressi di partito	6,43	6,47	6,15	6,35
Parlamento	6,48	6,88	6,72	6,69
C. C. e C. N.	6,45	6,56	6,49	6,50
Comizi	6,70	6,34	6,56	6,53
TOTALE	6,22	6,33	6,22	6,09

Tabella 20. Analisi della leggibilità degli interventi politici (valutata con l'indice di Dale e Chall).

In questo caso, un punteggio basso indica una maggiore leggibilità.
Come già emerso per la formula di Flesch, si rilevano differenze in relazione alle sedi di intervento: il gruppo che presentava un indice piuttosto alto di Flesch (alta leggibilità) presenta valori bassi di Dale e Chall (tra 5,46 e 5,64); viceversa, il gruppo di testi che risultava più difficile e aveva un punteggio Flesch più basso, riporta valori più elevati con questo secondo indice (tra 6,35 e 6,69).
Il coefficiente di correlazione tra le due formule, che in questo caso deve assumere un valore negativo dato che le scale di punteggi sono rovesciate[80], risulta – 0,97. La correlazione tra i due parametri dell'indice (parole non familiari e lunghezza delle frasi) è di 0,70.
L'Istituto di Documentazione Giuridica del CNR si occupa anche della leggibilità di testi giuridici[81]. Il corpus di testi analizzati comprende 51.000 documenti facenti parte della banca dati di testi giuridici gestita dall'I.D.G. e che conta un totale di 130.000 articoli tratti da 300 periodici dal 1970.
La leggibilità è valutata in modo automatico con l'ausilio di due programmi appositi che applicano la formula di Flesch e forniscono altri indici statistici[82]. Il valore medio di leggibilità ottenuto su tutti i documenti è di 29.

5.2.6. La redazione di testi didattico-scientifici da parte del CUD

Il Consorzio per l'Università a Distanza (CUD) si occupa di redigere dispense per la formazione universitaria[83]. Si tratta di materiali didattici specialistici, tecnici, rivolti a studenti universitari; sono testi multimediali, progettati per la formazione a distanza e prevedono esercizi di autovalutazione del livello di apprendimento dei contenuti. Proprio per queste caratteristiche, l'obiettivo del CUD è quello di produrre brani altamente comprensibili. "Lo studio autonomo comporta la necessità di «fare tutto da solo»: in questa

[80] La formula di Flesch indica una maggiore leggibilità quanto più elevato è il valore che risulta dall'applicazione dell'indice; nella formula di Dale e Chall, invece, la leggibilità risulta maggiore quanto più basso è il valore dell'indice.
[81] Si veda Martino et al. 1986, Mercatali 1986.
[82] Per i criteri di riconoscimento automatico delle sillabe e delle frasi crf. Mercatali 1979 e 1986.
[83] Cfr. Cimatti 1986 e Vedovelli 1991.

situazione è fondamentale che il tempo dedicato allo studio sia tutto incentrato sulla assimilazione dei contenuti senza dispersioni dovute a difficoltà testuali. L'ipotesi è che una buona leggibilità sia uno dei capisaldi di un sistema di istruzione a distanza" (Cimatti 1986, pp. 123-124).
Nel caso di documenti multimediali come quelli CUD, il problema della leggibilità non è riferibile al solo contenuto dei testi ma riguarda tutta una serie di accorgimenti che influiscono sulla facilità di lettura, come quelli grafico-visivi (impaginazione, dimensione e corpo del testo), testuali (colonna dei commenti, glossari e indici analitici), multimediali (integrazione tra testo scritto e altri media, come un software didattico, videocassette, audiocassette).
In realtà il lavoro del CUD non è la redazione vera e propria di testi didattico-scientifici quanto la loro riscrittura. I testi formativi sono infatti elaborati da uno specialista del settore e i redattori CUD si occupano di riscriverli in modo da trasformarli in materiale per l'insegnamento a distanza. Questo solleva ovviamente diverse problematiche[84], tra cui il metodo di misurazione della leggibilità: "Gli strumenti esistenti per misurare la leggibilità di un testo, indice di Flesch in testa, ci sono di poco o nullo aiuto. Si tratta infatti di formule tutte incentrate sullo scritto mentre nei nostri testi abbondano grafici, tabelle, formule, programmi di computer. La nostra pagina standard è ricca di integrazioni per le quali non disponiamo, a nostra conoscenza, di algoritmi che ne calcolino il valore comunicativo e l'aiuto che possono dare alla comprensione" (id., p. 125)

5.2.7. Il progetto *La lingua italiana: uno strumento per il made in Italy*

Il progetto *La lingua italiana: uno strumento per il made in Italy* nasce con l'obiettivo di realizzare corsi di lingua italiana per gli stranieri facendo ricorso a strumenti avanzati[85]. La ricerca prevede lo studio e l'analisi di tre linguaggi settoriali (storia dell'arte, musica ed economia) e sulla base di questi la produzione di materiali didattici: vengono costituite tre banche dati, una per ogni ambito e tre percorsi formativi di italiano L2. Per l'analisi linguistica vengono elaborati dal CNUCE di Pisa (con la collaborazione di Eugenio Picchi) una serie di programmi informatici che consentono di descrivere il comportamento semantico, morfosintattico e testuale di tutti i singoli elementi del corpus testuale.
Vi è anche una procedura automatica per il calcolo della leggibilità tramite l'indice di Flesch: la verifica della leggibilità consente di selezionare, all'interno delle banche dati, i testi più adeguati alle capacità e alle competenze degli studenti stranieri.

[84] Per le procedure standard di riscrittura e i diversi problemi affrontati dai redattori cfr. Cimatti 1986.
[85] Cfr. Vedovelli et al. 1989 e Barni e Peccianti 1991. Il progetto si inserisce nel complesso dei cosiddetti *Giacimenti Culturali* promossi dal Ministero dei Beni Culturali e Ambientali (ex Art. 15 della Legge 41/1986). In breve, si tratta di una serie di progetti, finanziati appunto dal Ministero e affidati ad aziende informatiche, volti a valorizzare economicamente, tramite operazioni di catalogazione elettronica, raccolta e descrizione, il patrimonio archeologico, archivistico, storico e artistico italiano. Il progetto *La lingua italiana: uno strumento per il made in Italy* è gestito dalla Società Video/Italia in collaborazione con la Scuola di Lingua e Cultura italiana per Stranieri di Siena. Il direttore scientifico è Tullio De Mauro; fanno parte del Comitato Scientifico Massimo Vedovelli (per la parte linguistica), Simonetta Lux (per l'arte), Lorenzo Arruga (per la musica) e Pier Luigi Nuti (per l'economia).

Non potendo disporre di un programma per la sillabazione automatica, gli autori effettuano un adattamento alla formula e sostituiscono il conteggio delle sillabe con il conteggio delle lettere.
La formula così modificata diventa:

$$I = 206 - 0{,}267L - S$$

dove
L = numero medio di lettere su 100 parole;
S = numero medio di parole per frase;

Il nuovo coefficiente deriva dal rapporto esistente tra la lunghezza di una parola misurata in lettere e il numero di sillabe che la compongono; il rapporto è calcolato su un campione di circa 20.000 parole.

5.2.8. Critiche alla formula di Flesch-Vacca

La prima validazione della versione italiana dell'indice di Flesch è condotta da Maria Corda Costa, che misura la correlazione tra i valori di leggibilità ottenuti attraverso la formula e i punteggi di comprensione della lettura da parte di ragazzi di terza media nell'anno scolastico 1982-1983[86].
Il campione scelto per la ricerca è composto da 231 studenti di terza media in 12 classi di 6 scuole romane e di 1 scuola di Terracina. Le prove di comprensione della lettura (TCL) a risposta chiusa sono otto, cinque delle quali già utilizzate nell'indagine internazionale IEA del 1969-1971.
I brani sono analizzati tramite l'adattamento italiano della formula di Flesch (FLESCH) e tramite un ulteriore indice di complessità sintattica (IC) che tiene conto di diversi parametri, come il numero medio di proposizioni per periodo, numero di proposizioni subordinate su periodo, congiuntivi e condizionali su indicativi, modi infinitivi su subordinate.
Sui risultati delle prove di comprensione e i due indici vengono calcolate le correlazioni di rango:
FLESCH – TLC: 0,45
FLESCH – IC: 0,62
TLC – IC: 0,30
Le correlazioni tra le tre misure risultano poco significative, probabilmente anche a causa del numero ristretto di testi considerati. Tuttavia, è possibile fare alcune considerazioni: l'indice di Flesch può essere considerato un buon indicatore del livello di complessità sintattica; l'indice di Flesch si correla meglio con i risultati delle prove di quanto non faccia l'indice di complessità; l'indice di complessità da solo non è un buon predittore della difficoltà dei brani.

Nonostante la formula di Flesch-Vacca riesca a fornire indicazioni utili sulla difficoltà dei testi scritti e sia un facile e rapido strumento per il controllo della leggibilità, tra i ricercatori emergono diversi dubbi sulla sua validità. Si evidenziano innanzitutto problemi nella sua

[86] Cfr. Maria Corda Costa 1984.

applicazione, come il conteggio automatico delle sillabe o il computo di cifre, sigle, abbreviazioni e simboli.
"L'automazione di questo indice si scontra con la difficoltà di progettare un esauriente algoritmo di sillabazione delle parole. Tutti gli algoritmi di sillabazione finora adottati dai sistemi automatici di scansione di un testo lasciano alcuni margini di errore, che possono essere più o meno gravi a seconda dei casi; in particolare il punto critico più rilevante nella sillabazione delle parole rimangono i dittonghi, per i quali la scansione in sillabe è legata alla collocazione dell'accento nella parola" (Amizzoni e Mastidoro 1993). Flesch suggerisce di contare il numero di sillabe di date e numeri in riferimento alla pronuncia. L'alto numero di sillabe contenute nelle date (ad esempio "1995") o cifre (come "3500 euro") determina però bassi valori di leggibilità anche se il testo è molto semplice. Per quanto riguarda le sigle, i simboli chimici e le abbreviazioni, Flesch non fornisce indicazioni sul conteggio.
Si pone inoltre il problema dell'individuazione automatica della frase: non sempre infatti la punteggiatura ha valore univoco nel segnalare l'articolazione dei periodi. Di norma il programma informatico viene istruito in modo da contare una frase ogni volta che incontra un punto fermo; ma anche sigle e abbreviazioni terminano con un punto ed è difficile effettuare un riconoscimento tra i due segni in modo automatico. Vari ricercatori rilevano inoltre l'ambiguità della nozione stessa di frase e propongono di sostituirla con quella di periodo (Palombi e Raponi 1984, Palombi 1968).
Altre obiezioni mosse alla formula riguardano più in generale la misurazione della leggibilità. Viene criticato il fatto che l'indice di Flesch (ma vale per ogni altra formula) non tiene conto di altri fattori che influenzano il processo di comprensione, come il livello culturale e la preparazione del lettore, il vocabolario impiegato, la correttezza ortografica, grammaticale e sintattica del testo, la struttura logica, l'impaginazione, la dimensione e il tipo di caratteri impiegati, la presenza di tabelle, immagini, grafici o di accorgimenti volti a facilitare la decodifica, come titoli, sottotitoli, sottolineature, grassetti, ecc.
"Il concetto di comprensione è sicuramente più ampio di quello di leggibilità. Testi leggibili nell'accezione tecnica dell'espressione (cioè con periodi di 20/30 parole, termini appartenenti al vocabolario di base o, se esterni ad esso, definiti precedentemente, struttura sintattica semplificata) non sono sinonimi di testi comprensibili. La comprensibilità di un testo […] è data dal prodotto di molteplici fattori: a) la leggibilità (in termini di Flesch); b) l'integrazione con altri media (se esistono); c) l'organizzazione testuale complessiva (cioè il modo in cui viene regolato il flusso di informazioni verso lo studente); d) la capacità di interagire con il lettore, di coinvolgerlo. […] Evidentemente *leggibile* non è sinonimo di comprensibile, ma è sicuramente molto difficile che un testo poco leggibile sia facilmente comprensibile" (Cimatti 1986, p. 127).
Ci sono poi altri fenomeni particolarmente diffusi nell'italiano scritto che influiscono sulla difficoltà dei testi e che andrebbero dunque considerati, ad esempio l'eventuale presenza di nominalizzazione (Palombi e Raponi 1984, Palombi 1968).

5.2.9. Leggibilità e prove di comprensione della lettura: verso un nuovo indice

"Dal crescente disagio avvertito nei confronti dell'utilizzazione della formula di Flesch da parte di tutti coloro che l'avevano sperimentata nacque l'esigenza di un confronto sul tema della leggibilità e della comprensibilità dei testi" (Thornton 1992). Nel 1986 viene

organizzato a Roma il convegno *Leggibilità e comprensione*, a cui partecipano tutti i ricercatori, i redattori e i gruppi di studio interessati al tema della leggibilità e della riscrittura[87].

In occasione dell'incontro, Roberto Vacca propone il nuovo adattamento della formula all'italiano e vengono illustrati i risultati dei già citati studi sull'uso sistematico dell'indice di Flesch. Uno dei principali aspetti emergenti dal dibattito è la necessità di individuare criteri unitari, standard per l'applicazione della formula di Flesch alla lingua italiana. "Questo convegno potrebbe essere la sede per gettare le basi di una specie di Commissione di controllo che fissi i parametri per il calcolo dell'indice e ne controlli l'applicazione in modo da ottenere una specie di Formula di Flesch a «Denominazione d'origine controllata» da usare in modo uniforme uguale per tutti. Ciò ci consentirebbe di disporre di dati omogenei completamente confrontabili" (Mercatali 1986).

Durante l'incontro di studio vengono presentati anche i primi risultati di un ampio progetto di ricerca iniziato nel 1985 e portato avanti da un gruppo di ricercatori, studenti e insegnanti del dell'Istituto di Filosofia dell'Università La Sapienza di Roma, in collaborazione con l'IBM. Il gruppo di lavoro, denominato GULP (Gruppo Universitario Linguistico Pedagogico), è coordinato da Pietro Lucisano e si avvale della consulenza di un Comitato scientifico composto dai professori Maria Corda Costa, Tullio de Mauro, Mario Gattullo ed Aldo Visalberghi[88]. L'obiettivo della ricerca è "verificare se i livelli medi di comprensione dei testi scritti misurati con le prove di comprensione a scelta multipla (TCL) e con le prove di *cloze*, forniscono una graduatoria di rango di facilità di lettura che correli con i dati risultanti dalle misurazioni sul testo. Riteniamo, infatti, insufficienti i metodi seguiti finora per l'adattamento degli indici di leggibilità alla lingua italiana, poiché quest'ultimi fanno riferimento solo ai rapporti tra indici nelle diverse lingue, senza alcuna considerazione della variabilità che strutture linguistiche diverse offrono ai processi di lettura e comprensione" (Lucisano e Piemontese 1986, p. 28).

Lo studio prevede quindi da una parte le misurazioni sul lettore tramite prove di comprensione della lettura (TCL e cloze test) e dall'altra misurazioni sul testo tramite l'uso di formule come quella di Flesch o la ricerca di altri possibili predittori della leggibilità e la rilevanza del vocabolario di base.

I brani scelti come campione sono tratti da 4 ambiti disciplinari: testi storico-politici, testi narrativi e letterari, testi di divulgazione scientifica, testi di istruzioni. Per il try out sono utilizzati soltanto 9 brani di storia[89]. Le prove vengono somministrate a soggetti di quattro popolazioni: quinta elementare (A), terza media (B), quinta superiore (C) e adulti (D). La popolazione degli adulti è stratificata per età, sesso e livello di istruzione. Agli studenti delle elementari vengono proposti solo i primi 5 testi mentre agli altri gruppi è somministrato l'intero pacchetto.

I testi sono estratti dai libri di testo degli studenti e, per quanto riguarda gli adulti, da un quotidiano e dallo statuto dei lavoratori; sono scelti secondo determinate indicazioni:

---

[87] Cfr. De Mauro et al. 1986

[88] Il gruppo GULP è composto da G. Asquini, G. Benvenuto, A. Columba, F. D'Antonis Onofri, M. de Grafenstein, M. Drigo, R. Morani, M. E. Piemontese, S. Pierdonati, F. Pietrobelli, A. Salerni, D. Scalet, M. T. Siniscalco, G. M. Tavanti, M. T. Tiraboschi.

[89] Il try out è effettuato dagli studenti di Pedagogia II e Filosofia del linguaggio nell'ambito di un'esercitazione intercattedra tenuta nell'anno accademico 1985-86.

- hanno senso compiuto, cioè non contengono un numero elevato di informazioni presupposte né rimandi espliciti a parti di testo precedenti o seguenti;
- hanno uguale lunghezza per popolazione con oscillazione tra le 250 e le 400 parole;
- hanno indici di Flesch differenziati in modo che all'interno dei testi di ciascuna popolazione esista una differenza di circa 4-10 punti nella scala di Flesch e che analoga differenza esista tra le popolazioni.

La tabella seguente riporta, per ciascun testo, l'argomento, il valore dell'indice di Flesch (indicato con F) dei singoli campioni, il valore medio della leggibilità e la deviazione standard.

Testo	F1	F2	F3	media	dev. st.	Argomento
0	51	72,6		67	10,8	Scuola Antica Roma
1	55,8	39,4		47,5	8,2	Rivoluzione industriale
2	46	52,4		49,2	3,2	Patrizi e plebei
3	33,8	46	38,2	39,3	5,0	Restaurazione
4	49,2	52,1	38,3	46,5	5,9	Conquista West
5	25	24,8	14,3	21,3	5	Resistenza
6	27,3	35	38,6	33,6	4,7	Indipendenza Africa
7	39	34,7	37,3	37	1,8	Craxi - De Mita
8	30,1	23,8	9,5	17,8	6,1	Statuto lavoratori
F1 – valore della leggibilità del 1° campione secondo Flesch						
F2 – valore della leggibilità del 2° campione secondo Flesch						
F3 – valore della leggibilità del 3° campione secondo Flesch						

Tabella 21. Valori di leggibilità, deviazione standard e argomento dei testi campione.

Per misurare la comprensione della lettura vengono utilizzate le prove a scelta multipla e il test cloze. I problemi delle domande a scelta multipla riguardano sia la formulazione delle domande (evitare che presentino difficoltà maggiori di quelle del testo) sia l'individuazione dei punti cruciali su cui costruire le domande. Il metodo seguito per l'individuazione dei punti nodali dei testi è quello di una lettura individuale e collettiva da parte del gruppo del Laboratorio (15 persone) e della successiva verifica con gli studenti dell'esercitazione intercattedra (25 persone).

Le prove di cloze sono costruite sugli stessi testi; viene usato il metodo classico, quello della cancellazione ogni 5 parole, per un totale di 25 cancellazioni per brano. La prima lacuna è collocata dopo circa 10 righe di testo. "La principale difficoltà posta dall'uso del cloze, che dagli anni '50 è la tecnica di riferimento più in uso negli Stati Uniti per la taratura degli indici di leggibilità, è l'assenza di sperimentazioni in lingua italiana" (Lucisano e Piemontese 1986, p. 31).

Nel corso dell'indagine vengono utilizzati anche due questionari: il primo raccoglie informazioni circa la provenienza socioculturale di ogni studente, le sue letture e il suo

atteggiamento verso la lettura; il secondo chiede agli insegnanti di valutare la difficoltà delle prove sottoposte agli studenti, secondo un criterio personale.
Per quanto riguarda l'applicazione dell'Indice di Flesch, emergono gli stessi problemi individuati da altri ricercatori, in particolare nel computo delle sillabe, delle date, dei numeri, delle sigle, delle abbreviazioni e dei simboli. "Per le date, anche se non lo riteniamo giustificato dal punto di vista teorico, ci siamo attenuti alle indicazioni di Flesch, di sillabare cioè le date come se fossero scritte in lettere. Per quanto riguarda, invece, le sigle le abbiamo considerate composte dal numero di sillabe che vengono pronunciate nella lettura della sigla stessa; per le abbreviazioni, invece, abbiamo preferito considerare l'intero vocabolo (es.: art. letto come articolo). Un problema di rilevazione non trascurato è costituito dall'equazione personale del rilevatore nel senso che ogni rilevatore forniva un numero di sillabe diverso a volte anche in modo sensibile" (id.).
Ciascun testo è poi confrontato con il *Vocabolario di Base* (nell'edizione del 1985), in modo da verificare se esiste una relazione tra l'appartenenza o meno a questo e l'indice di Flesch. Anche l'uso del VdB pone dei problemi: "Innanzitutto, nella sua forma attuale, il Vocabolario di Base, essendo una lista di parole, non dà informazioni sull'accezione nella quale un termine è considerato. In alcuni casi è assente anche l'indicazione della categoria grammaticale nella quale è considerato il termine del Vocabolario di Base. Un problema analogo si pone per i verbi, i participi e gli aggettivi sostantivati, per le locuzioni nominali e avverbiali che assumono un significato complessivamente diverso da quello delle singole parole che lo compongono. Dubbi di classificazione hanno posto anche i verbi riflessivi sia per il possibile cambiamento di uso e/o significato sia per la possibilità di essere scomposti in due parole. Problemi pongono anche l'uso del *si* passivante e del *si* impersonale, i verbi legati a pro complementi, i verbi composti, la sinonimia tra *venire* e *essere* nelle forme passive, i verbi causativi e servili, forme verbali derivate da verbi presenti nel Vocabolario di Base ma usate con accezioni diverse o come aggettivi. Frequenti dubbi per la classificazione si sono avuti a proposito degli alterati" (id., p. 32).
Per quanto riguarda il confronto tra le prove di comprensione a scelta multipla e le prove di cloze, è stata verificata una sostanziale omogeneità dei risultati: le due prove quindi risultano avere una buona correlazione tra loro. L'ordinamento dei testi in base alla difficoltà risultante dalle due prove e quella ottenuta in base all'indice di Flesch invece corrispondono solo in parte; sui testi più complessi, la formula di Flesch sembra perdere valore predittivo.
Gli autori considerano infine le correlazioni di rango tra le prove di comprensione, l'indice di Flesch e la percentuale nel testo di alcune classi di parole. Correlazioni significative con la comprensione sono ricavate da un indice che considera le frequenze del vocabolario non di base, i nomi propri, le sigle e le date (NSD). Emerge inoltre con le parole vuote la tendenza a correlare negativamente con la comprensione; fanno eccezione gli articoli che, più dello stesso Flesch, costituiscono un predittore significativo della leggibilità.
Tali risultati valgono almeno in relazione ai testi analizzati e devono comunque essere verificati su un campione più ampio di testi e lettori. "Riteniamo in conclusione che l'uso di indici per la rilevazione della leggibilità non possa prescindere da un lavoro di ricerca sulla effettiva comprensione dei testi, ricerca che deve essere condotta sia su grandi campioni con tecniche di rilevazioni «quantitative», sia, in parallelo, su piccoli gruppi, attraverso

interviste ed analisi «qualitative» di quelli che sono gli ostacoli alla comprensione" (id., p. 36).

5.3. La formula GULPEASE

"I problemi relativi all'uso della formula di Flesch e del vocabolario di base come strumenti per la predizione della leggibilità dei testi e per la produzione di testi più leggibili sono stati affrontati in una serie di esercitazioni di ricerca intercattedra svolte dalle cattedre di «Filosofia del Linguaggio» [prof. De Mauro] e di «Pedagogia» [prof.ssa Corda Costa] dell'Università di Roma «La Sapienza» a partire dal 1985" (Lucisano e Piemontese 1988, p. 114).

Negli anni successivi il gruppo GULP continua la propria ricerca, con l'obiettivo di tarare la formula di Flesch e mettere a punto una nuova formula di leggibilità basata sulla lingua italiana. Il progetto si svolge con l'intervento di IBM Italia e prevede anche lo sviluppo di un programma informatico in grado di calcolare la formula in modo automatico. Risultato di questi studi è lo sviluppo dell'indice GULPEASE, la prima formula di leggibilità per l'italiano[90].

Il modello di ricerca è quello classico (cfr. Miller 1972 o Klare 1984) e può essere sintetizzato in 6 fasi:

1. selezione di un set di testi criterio
2. misurazione del livello di difficoltà di questi testi per determinate popolazioni
3. misurazione delle variabili linguistiche dei testi
4. verifica delle correlazioni tra variabili linguistiche e livelli di difficoltà dei testi
5. costruzione di una formula attraverso una regressione lineare
6. validazione della formula su nuovi set di testi

5.3.1. Scelta del campione

Le popolazioni considerate nello studio sono gli anni terminali di ciascun ciclo scolastico, per un totale di 850 soggetti, di cui 274 studenti di quinta elementare (A), 301 studenti di terza media (B) e 275 studenti dell'ultimo anno delle superiori (C). A causa delle risorse limitate la ricerca è circoscritta al campionamento di studenti della sola città di Roma. Le scuole interessate sono 28 di cui 10 elementari, 9 medie e 9 superiori. "Abbiamo scelto le scuole sulla base di criteri di diversificazione socioeconomica e verificato la rispondenza del campione all'ipotesi attraverso l'analisi della varianza applicata ad un indice di cultura familiare ricavato da un questionario al quale sono stati sottoposti tutti gli studenti del campione. Questo indice tiene conto dei livelli di studio dei genitori e dei familiari conviventi e dell'occupazione dei genitori. Abbiamo dunque verificato con una procedura statistica, che si chiama analisi della varianza, se le scuole avevano effettivamente popolazioni differenti per provenienza sociale e la nostra scelta è risultata corretta" (Lucisano 1992, p.56).

[90] Per ulteriori indicazioni sull'Indice Gulpease cfr. Lucisano e Piemontese 1988, Lucisano 1992, Piemontese 1996, De Mauro e Chiari 2005.

Nella seconda fase (II anno, 1988), condotta in collaborazione con l'Ufficio Studi del Provveditorato di Roma, è considerato invece un campione di 51 classi di terza media, per un totale di 768 studenti.

Popolazione	**classe**	**n. classi**	**n. studenti TCL**	**n. studenti cloze**	**n. classi cloze**	**n. studenti cloze**
A	V elemen.	10	232	42		
B	III media	9	250	51	51	768
C	V super.	9	208	67		
totale		28	690	160	51	768
			I anno: 1986-87		II anno: 1988	

Tabella 22. Campione considerato nello studio.

5.3.2. Scelta dei testi criterio

Come criterio vengono scelti brani relativi a tre ambiti disciplinari: testi storico-politici, testi narrativi e letterari, testi di divulgazione scientifica. Il set è composto da 9 brani per ogni tipologia. I brani sono estratti da libri di testo del campione considerato e presentano le stesse caratteristiche di quelli impiegati nella ricerca precedente, ovvero:

- hanno senso compiuto, cioè non contengono un numero elevato di informazioni presupposte né rimandi espliciti a parti di testo precedenti o seguenti;
- hanno uguale lunghezza per popolazione con oscillazione tra le 250 e le 400 parole
- hanno indici di Flesch differenziati in modo che all'interno dei testi di ciascuna popolazione esista una differenza di circa 4-10 punti nella scala di Flesch e che analoga differenza esista tra le popolazioni.

Agli studenti della scuola elementare vengono proposti solo i primi 5 testi mentre agli altri gruppi è assegnato l'intero pacchetto. La somministrazione delle prove avviene in due momenti successivi: le prove di storia nel 1986 e quelle di scienze e lettere nel 1987, sempre nel periodo marzo-aprile.

5.3.3. Prove di comprensione

La comprensione della lettura è valutata tramite prove a scelta multipla (TCL), prove di cloze e prove di competenza lessicale[91]. Il metodo seguito per la costruzione delle domande a scelta multipla è lo stesso dell'indagine precedente, ovvero l'individuazione dei punti essenziali dei testi da parte di un gruppo di 15 lettori esperti e la successiva verifica da parte di 25 studenti universitari.

Le prove di cloze sono costruite usando il metodo classico della cancellazione sistematica ogni 5 parole, per un totale di circa 30 cancellazioni per brano. La prima lacuna è collocata dopo il primo capoverso. I test cloze sono assegnati solo ad un sottocampione e

[91] Esempi di testi delle prove a scelta multipla e cloze utilizzati nella ricerca si trovano in Lucisano 1992.

riguardano il set di testi storici. Data l'assenza di ricerche sul cloze test in lingua italiana "ci è sembrato dunque necessario procedere in via preliminare ad una taratura di questo strumento sui testi di tipo storico. Un problema nell'utilizzazione del cloze è dato dal fatto che il punteggio di comprensione considerato come percentuale di riempimenti esatti appare visibilmente più basso del punteggio di facilità ottenuto considerando la percentuale di risposte corrette con prove di comprensione della lettura a scelta multipla sugli stessi testi" (Lucisano 1992, p. 64).

I due tipi di prove risultano avere una buona correlazione tra loro (0,93). Gli studiosi calcolano dunque un'equazione che metta in corrispondenza i punteggi cloze e quelli ottenuti dalle prove a scelta multipla:

$$Facilità\ TCL = 22.14 + Cloze\ x\ 1.07$$

La corrispondenza tra i punteggi delle due prove permette di utilizzare il solo cloze come misura della comprensione della lettura nella fase di convalida delle formule. Per la validazione incrociata sono utilizzati soltanto i brani di tipo storico e di divulgazione scientifica, integrando ciascun set di nove testi con altri 15, scelti anche questi ultimi in modo da presentare diversi livelli di difficoltà.

Un altro problema del cloze riguarda la scelta del metodo di correzione delle prove. "Una volta praticato il buco in un testo il lettore può riempirlo proprio con la parola che era stata tolta, oppure con una parola che rappresenta un sinonimo della parola cancellata o ancora con una parola che magari non ha niente a che vedere con la parola cancellata, ma che collocata nel buco mantiene inalterato il senso della frase e infine può riempire il buco in modo non corretto. Le ricerche svolte in altri contesti linguistici affermavano unanimemente che tra la correzione rigida, quella cioè che accetta come buona solo la reintegrazione della parola cancellata, e la correzione svolta accettando i sinonimi non esiste differenza" (Lucisano 1992, p. 65-66). Per verificare questa ipotesi, il gruppo di lavoro calcola il punteggio cloze ottenuto nei due diversi modi su tutti i 48 testi usati per la validazione e la loro correlazione: nei 24 testi di divulgazione scientifica la correlazione è di 0.9969 (323 soggetti), nei 24 testi storici è di 0.9777 (328 soggetti). "Dunque risulta confermata anche per l'italiano l'indicazione di una sostanziale equivalenza dei due metodi di correzione anche se le due strategie di correzione forniscono indicazioni di livello di facilità lievemente differenti. Nei 48 testi esaminati la media dei riempimenti esatti è 11.41, quella dei riempimenti con sinonimi 1.67 e quella dei riempimenti con parole accettabili 1.70. Se si considera sempre il totale dei 48 testi dunque lo scarto tra la percentuale dei riempimenti calcolata considerando solo le esatte (38.04) e quella calcolata considerando insieme alle esatte i sinonimi e le accettabili (46.16) è di 8 punti" (id., p. 66).

Le prove di competenza lessicale sono 4 e servono come ancoraggio tra le diverse popolazioni: tre prove sono messe a punto da un gruppo di lavoro del GULP[92] e la quarta è una prova standardizzata utilizzata nell'indagine IEA *Six Subject* del 1970. Nel corso dell'indagine vengono utilizzati anche due questionari per la raccolta di informazioni: il primo circa la provenienza socioculturale degli studenti e le loro abitudini di lettura; il secondo rivolto agli insegnanti, con la richiesta di valutare la difficoltà delle prove sottoposte agli studenti, secondo un criterio personale. Gli studiosi non rilevano

[92] D'Antonis Onofri e Salerni 1987.

correlazioni tra l'interesse che hanno gli studenti per un dato testo e i punteggi riportati nelle prove di comprensione.

5.3.4. Misurazione delle variabili linguistiche

Sono diverse le variabili linguistiche che possono essere scelte per la valutazione della facilità di lettura. La misurazione di alcune di queste presenta alcune difficoltà, legate al computo della lunghezza delle parole: si deve infatti stabilire a priori se misurare la lunghezza delle parole in sillabe (come Flesch) o in lettere (come fanno altri autori, ad esempio Coleman e Liau). In questa ricerca vengono seguiti entrambi gli approcci ed emerge che la misura delle parole in lettere presenta correlazioni più alte con la comprensione dei testi. La scelta di considerare il numero di lettere consente inoltre di superare tutte quelle difficoltà legate al calcolo automatico della sillabazione delle parole.
Altre problematiche sono legate alla definizione della lunghezza quando si considerano sigle, date, numeri e abbreviazioni. Nello studio precedente il gruppo di ricerca aveva deciso di seguire le indicazioni di Flesch e sillabare le date e i numeri come se fossero scritte in lettere. Le abbreviazioni erano invece considerate nella forma estesa. Nel corso dell'analisi emerge però l'esigenza di definire nuove regole univoche, che possano meglio adattarsi alla lingua italiana[93]. Le nuove norme prevedono che ciascun numero, simbolo o abbreviazione sia considerato come una parola. Per misurare la lunghezza devono essere considerati il numero di sillabe/lettere che sono presenti nel numero o abbreviazione. Ciascun simbolo chimico è considerato come una sillaba e per misurarne la lunghezza in lettere è sufficiente misurare il numero di lettere da cui è composto il simbolo.

Tipologia	N. parole	N. sillabe	N. lettere
1986	1 parola	4 sillabe	4 lettere
18	1 parola	2 sillabe	2 lettere
321	1 parola	3 sillabe	3 lettere
CGIL/C.G.I.L.	1 parola	4 sillabe	4 lettere
USL/U.S.L.	1 parola	3 sillabe	3 lettere
UNESCO/U.N.E.S.C.O.	1 parola	6 sillabe	6 lettere
H20	1 parola	3 sillabe	3 lettere
NaCl	1 parola	2 sillabe	4 lettere

Tabella 23. Norme per misurare la lunghezza di abbreviazioni, sigle, simboli e numeri.

Per quanto riguarda le frasi, viene deciso di considerare convenzionalmente il solo punto fermo come delimitatore di frase. Altre variabili linguistiche considerate come possibili indicatori di difficoltà sono:

- parole appartenenti al *Vocabolario di Base* fondamentale;
- parole appartenenti al *Vocabolario di Base* ad alto uso;

[93] Della definizione di una serie di regole convenzionali per l'applicazione della formula di Flesch si è occupato il gruppo H. Le norme sono poi riprese dal GULP per la taratura di una nuova formula per l'italiano.

- parole appartenenti al *Vocabolario di Base* ad alta disponibilità;
- parole non appartenenti al *Vocabolario di Base*;
- nomi propri;
- parole vuote;
- sigle, abbreviazioni e simboli;
- date.

Le parole vuote contengono diverse categorie di parole:

- articoli determinativi e indeterminativi;
- congiunzioni coordinanti e subordinanti;
- pronomi personali;
- aggettivi e avverbi quantificatori;
- pronomi e aggettivi indefiniti;
- pronomi e aggettivi interrogativi;
- pronomi relativi;
- pronomi e aggettivi dimostrativi;
- pronomi e aggettivi possessivi;
- preposizioni semplici e articolate;
- procomplementi;
- avverbi di tempo e luogo di significato indeterminato.

Come già detto in precedenza, l'utilizzazione del *Vocabolario di Base* (nella forma del 1985) presenta alcune difficoltà in quanto, essendo una lista di parole, non fornisce informazioni sull'accezione nella quale un termine è considerato e, in alcuni casi, non presenta neppure l'indicazione della categoria grammaticale delle parole di base. Problemi analoghi si pongono per i verbi, i participi e gli aggettivi sostantivati, per le locuzioni nominali e avverbiali quando assumono un significato diverso da quello delle singole parole che lo compongono. "Molti di questi problemi sono stati legati anche al tentativo di realizzare un programma di computer in grado di computare la maggior parte di queste variabili. Allo stato attuale abbiamo messo a punto un programma in grado di risolvere solo i problemi più elementari e di calcolare formule semplici come Flesch o Coleman e Liau" (Lucisano e Piemontese 1988, p. 117).
Oltre all'indice di Flesch vengono calcolati anche l'indice ARI *(Automated Readability Index)* e la formula di Coleman e Liau. Le tre formule di leggibilità hanno in comune alcune specificità: sono tutte di facile applicazione, anche in vista di un'automazione e considerano tutte le stesse due variabili, cioè la lunghezza delle parole e la lunghezza delle frasi. La differenza è che la formula di Flesch misura la parola in numero di sillabe mentre gli altri due indici valutano la lunghezza in base al numero di lettere. Sia la formula ARI che quella di Flesch tendono a sovrastimare la difficoltà di lettura, mentre la formula di Coleman e Liau tende a sottostimarla.

La tabella seguente mostra per ciascun testo della ricerca i valori di leggibilità e i punteggi risultanti dalle prove di comprensione:

Testo	Flesch	Ari	Coleman	Punt. A	Punt. B	Punt. C	Argomento
H0	54.71	8.8	12.0	77	88	90	scuola antica Roma
H1	53.78	14.2	12.7	53	68	89	patrizi e plebei
H2	42.35	15.1	14.1	50	63	88	rivoluzione industriale
H3	38.22	17.1	15.3	55	73	88	conquista west
H4	39.00	16.9	13.7	45	63	84	restaurazione
H5	22.00	24.5	15.9		43	71	Craxi e De Mita
H6	27.85	21.27	15.6		46	71	indipendenza africana
H7	38.75	15.5	14.7		55	78	resistenza
H8	21.41	20.1	18.3		56	75	statuto lavoratori
S0	73,90	6.4	8.7	85	79	92	adattamento del cammello
S1	46.63	14.6	13.7	59	76	92	le foreste equatoriali
S2	51.31	15.6	12.8	43	65	86	il cuore
S3	30.08	17.7	11.9	48	69	92	l'energia nucleare
S4	40.96	16.5	14.6	50	70	95	organismi e ambiente fis.
S5	47.09	16.1	12.5		39	69	il magnetismo
S6	41.02	14.8	15.2		33	63	lo sviluppo cellulare
S7	37.78	16.5	15.2		60	92	l'intelligenza artificiale
S8	40.08	15.1	15.3		58	89	i danni del fumo
L0	68.54	11.1	9.4	82	87	99	Raccontino morale
L1	82.45	4.6	6.9	53	61	83	R. Kipling *Libro della giungla*
L2	86.23	3.6	5.6	38	75	91	G.C. Vamba *Giamburrasca*
L3	44.92	15.3	13.0	39	59	84	L. Sciascia *Una storia vera*
L4	83.66	14.9	13.5	41	60	79	I. Silone *Uscita di sicurezza*
L5	2.14	35.7	14.3		46	67	S. Petronio *L'attività letteraria*
L6	36.51	18.6	14.1		37	62	M. Tobino *L'armistizio*
L7	51.36	13.7	11.5		49	81	I. Calvino *Le città invisibili*
L8	36.50	16.4	15.4		33	55	E. Morante *La storia*

Tabella 24. Testi utilizzati nella ricerca, valori delle formule, punteggi medi delle prove e argomenti dei testi.

Le correlazioni tra i punteggi riportati nelle prove da ciascuna popolazione risultano costanti. Questo significa che è possibile limitare la rilevazione ad una sola popolazione, in questo caso la B (terza media). "Il rapporto costante tra le prove ci consente di fare ricorso ad una equazione di regressione in modo da poter trarre dai dati di una popolazione informazioni sulle probabili prestazioni di un'altra" (Lucisano e Piemontese 1988, p. 119). Grazie a questa semplificazione, è possibile lavorare su un campione più ampio di scuole e di soggetti: 51 classi di terza media per un totale di 768 studenti.

5.3.5. Costruzione della formula

Sono poi calcolate le correlazioni tra le variabili linguistiche dei testi e i punteggi delle prove di comprensione. Come si può osservare nella Tabella 25 le correlazioni più alte sono ottenute dalle parole piene del *Vocabolario di Base*. Una buona correlazione si ha anche con la lunghezza delle frasi (numero di frasi su 100 parole). Diversamente da quanto rilevato in precedenza sui soli testi di storia, non risultano correlazioni significative con le parole vuote. Da notare il fatto che, mentre le parole fondamentali hanno correlazioni di segno positivo con i valori di comprensione, sia le parole ad alto uso che quelle ad alta disponibilità hanno invece correlazioni negative.

Variabili	**Correlazioni**
Virgolette e trattini	0,1734
Densità di parole	0,2235
Sillabe su 100 parole	- 0,3668
Lettere su 100 parole	- 0,4370
Frasi su 100 parole	0,4831
Parole di base	0,5531
Fondamentali %	0,5314
Alto uso %	- 0,3963
Alta disponibilità %	- 0,2105
Parole piene %	0,6253
Parole vuote %	- 0,0084
Generiche %	- 0,1763
Quantificatori %	0,1481
Tempo/luogo %	0,2052
Parole ndb %	- 0,5087
Nomi propri %	- 0,2335
Sigle, abbreviazioni, simboli %	- 0,1490
Totale non di base %	- 0,5560
Congiunzioni %	0,1654
Coordinanti %	0,3264

Variabili	Correlazioni
Subordinanti %	- 0,3241
Articoli	0,0351
Indeterminativi %	- 0,0404
Determinativi %	- 0,0396
Partitivi %	0,3680
Avverbi %	0,2460

Tabella 25. Correlazioni dei punteggi della popolazione B con le variabili linguistiche-

Per la costruzione della formula vengono dunque impiegate le variabili che risultano avere un più alto valore predittivo per questa data popolazione:

- indicatori di difficoltà lessicale
  - parole di base piene (BP)
  - lettere su 100 parole (LP)
- indicatori della complessità sintattica
  - percentuale di congiunzioni subordinanti (CS)
  - numero di frasi su 100 parole (FR)

In base a queste sono calcolate le equazioni di regressione multipla e vengono ricavate 4 formule:

$$SOLOBASE = 17{,}08578 + (0{,}96764BP)$$

$$GULPBASE = 3{,}75592 + (0{,}89620BP)(3{,}64779FR)$$

$$GULPSINT = 11{,}61527 + (0{,}93091BP) + (3{,}17459FR) - (6{,}13819CS)$$

$$GULPLTFR = 89{,}32615 + (3{,}00583FR) + (-0{,}08397LP)$$

Dell'ultima formula è ricavata anche una versione semplificata, chiamata GULPEASE:

$$GULPEASE = 89 - LP/10 + 3\,FR$$

Sono quindi calcolati i vari indici e le loro correlazioni con i valori di comprensione (Tabella 26).

	PUNT. B	SOLOBASE	GULPSINT	GULPBASE	GULPTFR	GULPEASE	FLESCH 72	FLESH 86
PUNT. B	1,0000							
SOLOBASE	0,6253**	1,0000						
GULPSINT	0,8047**	0,7771**	1,0000					
GULPBASE	0,7515**	8321**	0,9339**	1,0000				
GULPTFR	0,5184*	0,1377	0,6267**	0,6261**	1,0000			
GULPEASE	0,5179*	0,1398	0,6020**	0,6184**	0,9991**	1,0000		
FLESCH 72	0,4709*	0,2256	0,5198*	0,5964**	0,8339**	0,8355**	1,0000	
FLESCH 86	0,4576*	0,2333	0,5362*	0,6187**	0,8391**	0,8377**	0,9959**	1,0000
N. di casi: 27 Significatività: * = 0,01 ** 0,001								

Tabella 26. Correlazione fra formule e valori di comprensione.

Tutte le formule presentano buone capacità predittive. Delle due formule di Flesch negli adattamenti di Vacca, la prima versione (1972) risulta essere la migliore. La correlazione più significativa è data dall'indice GULPSINT (0,81) che impiega come variabili la percentuale di parole di base piene (BP), la lunghezza delle frasi (FR) e la percentuale di congiunzioni subordinanti (CS). La formula GULPEASE risulta però la più adatta, perché facile da applicare e facilmente accessibile sia per il calcolo manuale che automatico. La lunghezza delle parole (in lettere) costituisce la variabile lessicale, mentre la lunghezza delle frasi è la variabile sintattica.

5.3.6. Validazione della formula

La validazione della formula GULPEASE viene effettuata su un nuovo campione di testi tratti prevalentemente da manuali scolastici di storia e di scienze, con lunghezza variabile tra le 170 e le 200 parole. Le nuove prove (48 test cloze) sono somministrate ad un campione di 768 soggetti, provenienti da 51 classi di terza media di 32 scuole diverse. Le prove sono costituite da testi con 30 buchi, secondo il metodo di cancellazione tradizionale.
Per ogni studente sono calcolati tre punteggi:

- un punteggio cloze (clz) secondo il quale viene attribuito 1 punto a ciascun riempimento corretto;
- un punteggio ponderato secondo il quale vengono attribuiti 3 punti alle risposte esatte, 2 punti ai sinonimi e 1 punto ai riempimenti accettabili (clza);
- un punteggio calcolato attribuendo 1 punto a ciascun riempimento che risulta accettabile (clzb).

La Tabella 27 riporta i punteggi medi per ciascun tipo di prova (in percentuale di riempimenti sul totale di 30 buchi) e i valori di leggibilità.

Arg.	cod.	clz	clza	clzb	Flesch	GULPEASE	GULPSINT
H	0	42,67	53,22	62,33	62,61	71,24	67,86
H	1	53,33	56,67	59,33	59,67	61,69	62,36
H	2	43,67	53,78	64,67	47,77	57,84	51,41
H	3	41,00	43,89	46,33	42,34	55,69	55,45
H	4	33,67	40,00	45,00	46,95	56,69	40,17
H	5	30,00	46,00	59,67	36,57	53,62	53,81
H	6	23,33	25,56	29,00	31,24	52,00	53,43
H	7	20,00	23,44	27,00	42,48	55,63	60,28
H	8	30,00	31,56	33,00	21,49	49,37	65,27
H	9	28,00	35,67	44,33	57,83	61,39	55,98
H	10	34,00	37,89	41,67	34,16	52,86	42,80
H	11	24,33	33,67	44,33	30,76	52,04	41,04
H	12	33,67	37,56	43,33	14,40	50,55	58,27
H	14	32,33	34,22	36,33	45,95	57,83	62,60
H	16	32,67	36,89	41,33	47,62	60,71	73,76
H	18	26,00	35,22	45,33	46,07	57,53	59,03
H	20	39,00	36,11	68,00	58,95	64,89	62,03
H	21	36,00	38,22	40,33	23,80	51,15	54,20
H	23	31,00	33,89	37,33	0,72	46,72	51,64
H	24	20,67	25,89	32,33	8,80	49,79	43,25
H	25	38,67	43,11	47,00	55,92	62,82	57,97
H	26	47,00	52,56	57,67	51,60	59,68	69,68
H	27	30,67	34,00	36,67	62,03	62,43	39,75
H	28	20,67	28,33	37,33	3,24	47,13	46,79
S	0	52,00	55,11	58,67	77,06	75,02	74,25
S	1	43,33	51,67	64,33	44,66	56,33	58,81
S	2	54,67	63,78	74,33	53,77	59,69	58,74
S	3	33,00	35,11	36,67	48,38	57,65	50,71
S	4	35,33	42,89	49,33	43,25	56,11	48,04
S	5	34,33	37,78	42,33	56,89	60,24	36,40
S	6	26,67	27,44	28,67	40,40	55,06	38,17
S	7	32,67	35,00	38,33	37,28	56,56	73,57

Arg.	cod.	clz	clza	clzb	Flesch	GULPEASE	GULPSINT
S	8	29,67	36,00	41,67	44,21	58,51	57,20
S	9	25,00	29,22	33,33	50,76	60,44	56,53
S	10	39,00	42,11	46,00	56,00	61,16	63,81
S	11	32,00	36,78	41,33	36,23	53,63	41,64
S	14	38,67	42,67	46,00	34,42	52,96	45,21
S	19	37,67	44,22	49,00	24,96	51,07	28,63
S	20	55,33	60,33	66,33	57,10	60,54	29,53
S	22	36,33	40,11	43,33	39,99	54,97	33,46
S	23	32,00	38,22	44,33	47,65	57,14,	45,33
S	24	43,33	46,67	50,00	60,68	64,11	60,12
S	25	40,33	44,11	48,67	53,48	59,95	58,36
S	26	49,00	59,78	67,33	66,79	64,43	57,94
S	28	49,67	54,33	58,33	66,80	68,04	57,35
S	29	22,67	24,22	27,00	21,84	49,46	49,70
S	30	27,67	35,44	41,67	15,32	47,63	60,70
S	31	29,67	34,11	39,00	4,80	48,30	41,01

Tabella 27. Punteggi medi e indici di leggibilità per i testi di storia (H) e scienze (S).
clz = % riempimenti corretti; clza = % riempimenti ponderata; clzb = % riempimenti accettabili.

Sono poi considerate le correlazioni tra i punteggi ottenuti nelle prove e le variabili linguistiche del testo. A differenza di quanto osservato in precedenza, le percentuali di termini appartenenti al *Vocabolario di Base* (VdB), e in particolare di parole di base piene (BP), non presentano correlazioni significative con i punteggi cloze. Questo mette in crisi la formula GULPSINT che include tra i suoi parametri la percentuale di parole piene. Correlazioni importanti (anche se non altissime) si hanno invece con la lunghezza delle parole in lettere (LP) e con la lunghezza media delle frasi (FR), le due variabili che costituiscono la formula GULPEASE.

	Storia			Scienze		
	PCLZ	PCLZA	PCLZB	PCLZ	PCLZA	PCLZB
VdB	0,48	0,43	0,39	0,16	0,23	0,26
BP	0,36	0,38	0,46	0,11	0,20	0,22
FR	0,49	0,47	0,52	0,44	0,37	0,30
LP	- 0,38	- 0,39	- 0,41	- 0,72	- 0,68	- 0,62
CS	- 0,6	0,06	0,19	0,02	0,04	0,01

Tabella 28. Correlazione tra punteggi cloze e variabili linguistiche nelle prove di storia e scienze.
VdB = vocabolario di base; BP = parole di base piene; FR = n. di frasi su 100 parole; LP= n. di lettere su 100 parole; CS = congiunzioni subordinanti.

La Tabella 29 mostra i coefficienti di correlazione delle formule con i punteggi delle prove cloze. Analizzando le differenze tra i testi storici e i testi di divulgazione scientifica si nota che mentre nei testi scientifici il vocabolario di base non ha alcun valore predittivo, esso acquista una certa significatività nei testi storici. “In questo senso crediamo sia possibile affermare che mentre per i testi che utilizzano prevalentemente la lingua standard è utile poter fare riferimento ad un vocabolario di base, questo è molto meno utile nei testi che fanno ampiamente uso di linguaggi speciali, in cui spesso i vocaboli di base sono utilizzati con significati particolari. Il risultato conferma quello di altre ricerche sulla leggibilità: in qualche misura sembra che il potere predittivo delle formule tenda a diminuire al crescere della complessità delle formule. Un'ipotesi da considerare inoltre è la possibilità di realizzare formule di leggibilità specifiche per diversi tipi di testo” (Lucisano 1992, p. 81).

Formule	**24 testi storia**	**24 testi scienze**	**Totale 48 testi**
Flesch	0,53	0,67	0,61
GULPEASE	0,55	0,66	0,61
GULPSINT	0,41	0,16	0,20
SOLOBASE	0,48	0,16	0,32
GULPBASE	0,47	0,57	0,52

Tabella 29. Correlazione delle formule con i punteggi delle prove cloze sui testi storici e scientifici.

La formula di Flesch e la formula GULPEASE presentano valori di correlazione molto simili. Se però osserviamo i grafici di regressione delle due formule (Figura 4) si nota che l’indice GULPEASE ha una maggiore stabilità ed una maggiore affidabilità della formula di Flesch, la quale presenta forti oscillazioni e tende a sovrastimare la difficolta dei testi.

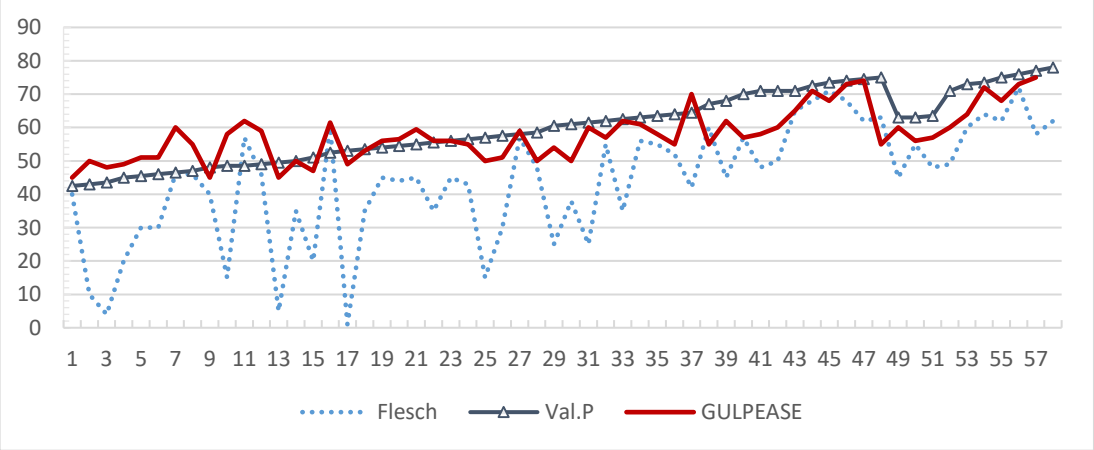


Figura 4. Confronto tra la formula di Flesch e la formula GULPEASE

5.3.7. Interpretazione dei risultati

Insieme alla formula viene fornita una scala di leggibilità per l'interpretazione dei risultati, con punteggi che vanno da 100 (leggibilità massima) a 0 (leggibilità nulla). Se si fa riferimento alla facilità/difficoltà del testo i punteggi indicano:

- > 80 molto facile
- 79 - 60 facile
- 59 - 50 difficile
- 49 - 35 molto difficile
- < 35 quasi incomprensibile

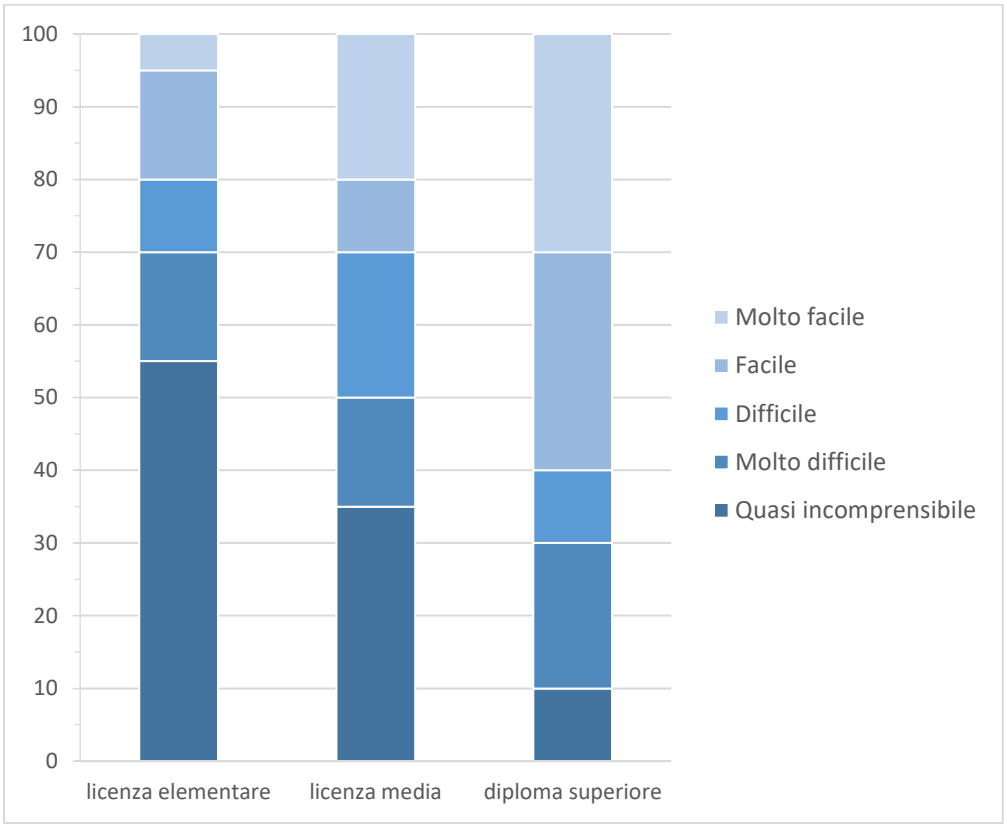


Figura 5. Scala dei valori GULPEASE.

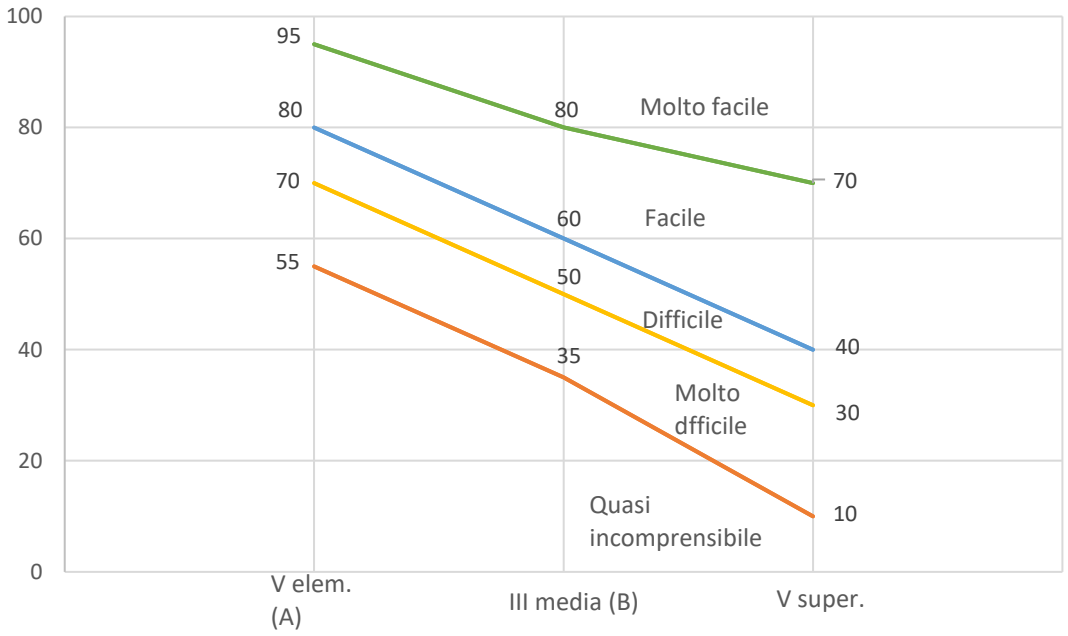


Figura 6. Leggibilità dei testi in base ai valori della formula GULPEASE.

Se si considerano i vari livelli di scolarizzazione dei lettori:

- un punteggio superiore a 80 indica un testo che può essere compreso facilmente da lettori che hanno un'istruzione elementare (A);
- un punteggio superiore a 60 indica un testo che può essere compreso facilmente da lettori che hanno un'istruzione media (B). Il lettore è in grado di leggere il testo in modo autonomo (livello di lettura indipendente);
- un punteggio tra 59 e 40 indica che il lettore ha bisogno di aiuto per comprendere il testo (livello di lettura scolastica);
- un punteggio inferiore a 40 indica un testo che può essere compreso facilmente solo da lettori che hanno un'istruzione superiore (C). Il lettore proverà nel leggere il testo la frustrazione di non essere in grado di comprenderlo (livello di frustrazione).

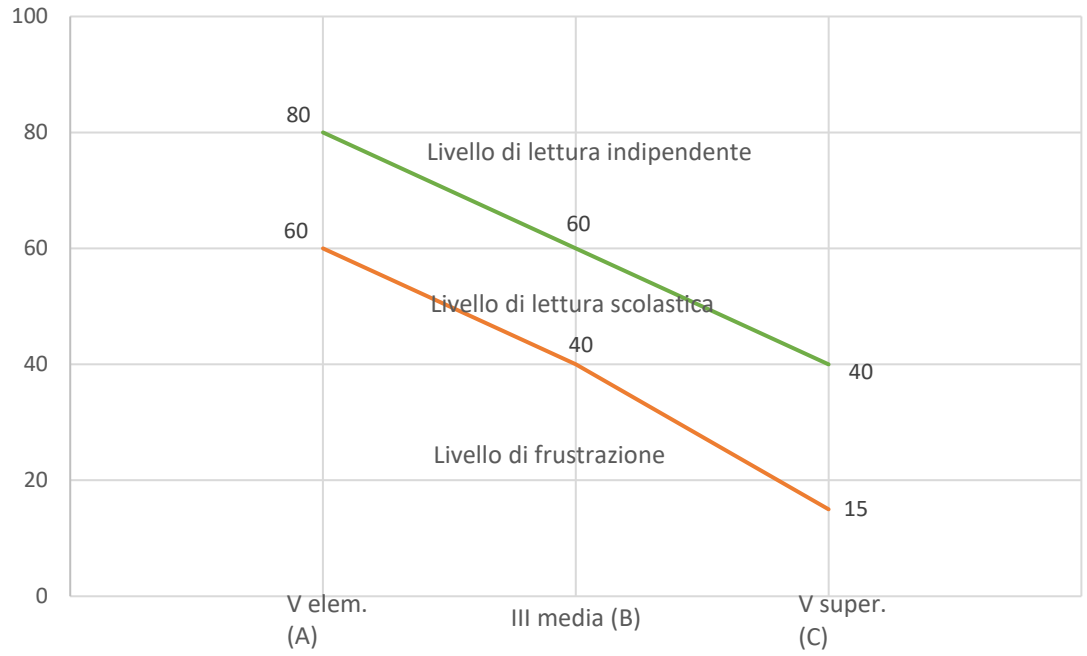


Figura 7. Tipi di lettura in base ai valori della formula GULPEASE.

Si deve tenere presente che la scala è tarata su bambini e ragazzi in età scolare e mancano invece verifiche sistematiche ed estese su gruppi di adulti. I valori di riferimento sono comunque accettabili anche quando applicati a lettori adulti, specialmente a quelli con un'istruzione medio-bassa o bassa.

5.3.8. Applicazioni della formula GULPEASE: Èulogos Censor e Corrige!Leggibilità

Della realizzazione di strumenti che calcolano in automatico la formula GULPEASE e la composizione del lessico di un testo si occupano due tesi di laurea svolte con il prof. Tullio De Mauro ed Emanuela Piemontese presso la cattedra di Filosofia del linguaggio dell'Università di Roma La Sapienza: quella di Maurizio Amizzoni (*Calcolo automatico della leggibilità: l'indice Gulpease*, 1991) e quella di Nicola Mastidoro (*Rilevamento automatico del tasso di vocabolario di base*, 1991)[94]. I risultati delle due tesi hanno portato alla realizzazione di una nuova versione automatica della formula, l'AUTOGULP, e di un programma che effettua in automatico l'analisi lessicale di un testo partendo dal *Vocabolario di Base*. I due software sono poi riuniti in un sistema integrato, Èulogos, che consente sia l'analisi automatica della leggibilità di un testo che quella statistico-lessicale[95].

[94] M. Amizzoni, *Calcolo automatico della leggibilità, l'indice Gulpease*, tesi di laurea in Filosofia del linguaggio, Facoltà di lettere e Filosofia, Università di Roma La sapienza, 1991; N. Mastidoro, *Rilevamento automatico del tasso di vocabolario di base*, tesi di laurea in Filosofia del linguaggio, Facoltà di lettere e Filosofia, Università di Roma La sapienza, 1991

[95] Cfr. Mastidoro 1992. Èulogos è un marchio registrato che fa parte di Èulogos SRL di proprietà di Nicola Mastidoro. L'algoritmo di scansione del testo e di calcolo dell'indice GULPEASE è stato realizzato da Maurizio Amizzoni. Il sito di riferimento è attualmente www.eulogos.it
Il programma Èulogos è stato impiegato per lo spoglio elettronico del corpus delle prime cinque annate (1989-1994) del periodico *Due parole* (si veda Piemontese 1996°, 1996b).

All'interno del programma è registrato il *Vocabolario di Base* di De Mauro e grazie alla lemmatizzazione automatica è possibile confrontare il lessico del testo analizzato per verificarne l'appartenenza o meno al VdB. Durante l'analisi del testo il sistema calcola anche l'indice GULPEASE, utilizzando un particolare algoritmo che misura le due variabili considerate (lunghezza delle frasi e lunghezza delle parole). Sono inoltre possibile altre funzioni, come la generazione di liste di frequenza con calcolo dell'indice di dispersione, l'indice d'uso dei lemmi, la gestione di aspetti morfologici come alterazioni, abbreviazioni, ecc.
A partire dal 1999 viene attivato un servizio automatico in rete, Èulogos *Censor*, che fornisce a chi ne fa richiesta una serie di servizi, come l'analisi della leggibilità di un testo secondo l'indice GULPEASE e il confronto del lessico del testo analizzato con il *Vocabolario di Base*[96]. Il servizio funziona per posta elettronica: mandando un testo al server, si riceve indietro (via email) il risultato dell'analisi, che comprende varie statistiche generali, il valore della leggibilità, le liste di frequenza delle parole del testo. La leggibilità è calcolata sia sull'intero testo che per ogni singola frase. *Censor* fornisce gratuitamente risultati parziali e in abbonamento offre l'insieme completo delle statistiche. Il servizio conta diverse migliaia di utenti, prevalentemente dalla Pubblica Amministrazione; viene inoltre utilizzato in diverse collaborazioni scientifiche ed editoriali[97].
Attualmente *Censor* è stato sostituito da un altro servizio automatico, *Corrige!Leggibilità*, che impiega la tecnologia Imprimatur, un sistema avanzato di trattamento automatico del linguaggio[98]. A differenza del precedente, *Corrige!Leggibilità* è a pagamento ma è possibile effettuare una prova gratuita e controllare testi fino a 5.000 parole (complessive). Il sistema genera il resoconto di leggibilità, un documento in formato HTML che offre i dati di leggibilità e di lessico, con informazioni sia a livello di frase che di singola parola.

Non è possibile fornire un resoconto di tutti gli studi in cui è applicata la formula GULPEASE[99]. Se l'indice di Flesch viene impiegato principalmente in ricerche condotte in ambito accademico o comunque svolte da specialisti, la formula GULPEASE ha invece una diffusione su larga scala. Chiunque può utilizzare la formula come controllo della leggibilità dei propri testi, per accertarsi che il livello sia adeguato al livello di lettura dei destinatari scelti come riferimento, oppure per effettuare studi di valutazione di materiali esistenti, come articoli di giornali, riviste, documenti amministrativi, medici, ecc. Oltre al servizio *Corrige!Leggibilità*, è possibile trovare altri siti web che effettuano il controllo della leggibilità tramite l'indice. In alternativa, è possibile usare Microsoft Word; la formula è

[96] Per indicazioni su Censor cfr. Lucisano 1992, Mastidoro e Amizzoni 2005.
[97] Il sistema Censor è stato impiegato per valutare e migliorare la leggibilità del sussidiario per le scuole elementari *Il cosmonauta – Sussidiario ad alta leggibilità*, in collaborazione con la casa editrice Elmedi. È stato inoltre usato per le valutazioni del *Premio Chiaro!* (Dipartimento della Funzione Pubblica, 2003), destinato alle pubbliche amministrazioni che hanno dimostrato di scrivere testi particolarmente chiari e accessibili.
[98] Il servizio è accessibile all'indirizzo www.corrige.it.
[99] Si segnalano, a titolo di esempio, l'applicazione della formula ai sussidiari della scuola elementare (Piemontese e Cavaliere 1997), ai testi per la scuola secondaria di primo e secondo grado, sia di italiano che di altre materie, come la fisica, la storia dell'arte, la geografia, ecc. (Giscel Piemonte 1997).

infatti incorporata nel programma di elaborazione di testi, il quale, oltre ad effettuare il controllo ortografico, consente di visualizzare le statistiche di leggibilità.

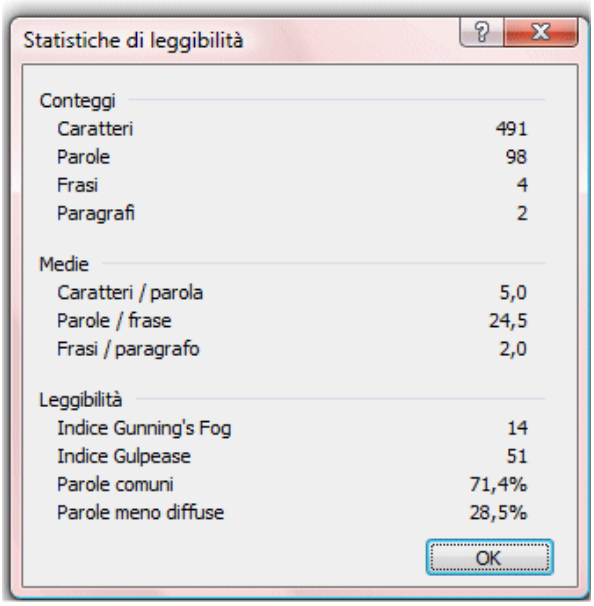


Figura 8. Statistiche di leggibilità su Microsoft Word.

Tra le varie applicazione dell'indice GULPEASE, segnaliamo lo studio condotto compiuto dall'Istituto per le Nuove Tecnologie Genesio nel 2000 sulla leggibilità di alcuni siti web. Il campione analizzato comprende siti istituzionali (Ministero degli Esteri, Governo italiano), il sito dell'azienda informatica Microsoft e i siti di tre testate editoriali italiane (*La Gazzetta dello Sport*, *La Repubblica*, *Il Corriere della Sera*). Per ogni sito sono stati analizzati almeno dieci testi; la leggibilità è valutata sia in base all'indice GULPEASE che in base a quello di Flesch-Vacca.

Come si può osservare (Tabella 30), i siti della pubblica amministrazione e il sito della Microsoft presentano punteggi piuttosto bassi. Le testate editoriali mostrano livelli di leggibilità più elevati ma si mantengono comunque in una fascia medio-bassa.

Sito web	**GULPEASE**	**Flesch-Vacca**
Ministero degli Esteri	41,7	23
Governo italiano	45,7	36,7
Microsoft	43,3	38,4
Gazzetta dello Sport	53,3	61,8
Repubblica	49,2	52
Corriere della Sera	51,4	56,6

Tabella 30. Valori di leggibilità dei siti web analizzati da Genesio.

5.4. Altri studi italiani

5.4.1. Indice di Leggibilità per Varietà Testuali (ILVAT)

All'università degli Studi di Torino è stato sviluppato ILVAT, un indicatore di leggibilità variabile che consente di analizzare diverse varietà testuali[100].
La leggibilità di un testo è valutata in base a un set di parametri elaborati specificatamente per quel tipo di testo; in pratica esiste un indice di leggibilità specifico per ogni varietà testuale.
È possibile utilizzare i parametri esistenti e disponibili nel database oppure estrarne di nuovi da un corpus di addestramento.

5.4.2. La leggibilità dei testi matematici

Athanasios Gagatsis (1995, 1999) si pone il problema di come misurare le leggibilità dei testi matematici. Si domanda, in particolare, quali siano gli aspetti misurabili del linguaggio scritto che possano essere associati alla leggibilità dei testi matematici. Il suo lavoro è influenzato dalla ricerca effettuata da Kane et al. (1974) per la lingua inglese.
La familiarità delle parole matematiche è una variabile che influisce molto sulla comprensione della lettura. Kane et al. (1974), ad esempio, hanno costruito per l'inglese alcune tabelle di familiarità di parole e simboli matematici. Per determinare il grado di familiarità, hanno presentato 100 termini matematici a 350 allievi e 18 simboli ad altri 250 studenti, appartenenti in totale a 36 scuole diverse. L'indice di familiarità è costruito in base alla percentuale di studenti che conoscevano quel dato termine o simbolo. Gli autori hanno anche effettuato una verifica riguardo alla validità del cloze test in questo ambito: hanno sottoposto agli alunni 22 testi matematici, valutandone la comprensione tramite cloze test e prove a scelta multipla e hanno dimostrato una sostanziale equivalenza tra i due metodi.
Kane et al. (1974) hanno infine valutato la comprensione di 70 testi da parte di 2400 studenti tra i 13 e i 18 anni e misurato ben 110 variabili quantitative diverse. In base alle correlazioni tra queste e i risultati dei test hanno costruito due formule di leggibilità per i testi matematici.

$$Leggibilità\ 1 = -0{,}23x - 0{,}53y + 61{,}88$$

dove:
x = numero di parole matematiche assenti dalla lista di parole familiari (parole matematiche che hanno una familiarità dell'80%);
y = numero di parole diverse con più di 3 sillabe.

$$Leggibilità\ 2 = -0{,}15A + 0{,}10B - 0{,}42C - 0{,}17D + 35{,}52$$

dove:
A = parole che sono assenti dalla lista delle 3000 parole di Dale e dalla lista di parole familiari degli autori;
B = numero di cambiamenti dalla lingua naturale alla lingua simbolica e viceversa;

[100] L'indicatore ILVAT è disponibile all'indirizzo:
http://www.corpora.unito.it/cgi-bin/lingue/ilvat/ilvat_index.pl?var=ILVAT

C = numero dei diversi termini matematici assenti dalla lista di familiarità + numero dei diversi simboli matematici assenti dalla lista di simboli con una familiarità del 90%;
D = numero dei punti interrogativi.

Il coefficiente di correlazione della prima formula è pari a 0,60; quella della seconda è di poco inferiore (0, 55).

A differenza di Kane et al. (1974), Gagatsis non presenta una vera e propria formula di leggibilità, quanto una proposta di metodologia da seguire per la misurazione della leggibilità dei testi matematici in italiano e spagnolo.
La questione sollevata da Gagatsis è ripresa da D'Amore e Fandiño Pinilla (2016), nell'ambito di uno studio congiunto tra l'Università di Bologna e l'Universidad Distrital Francisco José de Caldas di Bogotá, in Colombia. L'obiettivo della loro ricerca è "creare una formula del tipo "test di chiusura", indipendente dal livello scolastico, che misuri la difficoltà nella quale si trova uno studente per comprendere un brano tratto da un testo di matematica adatto al livello scolastico nel quale egli si trova" (p. 7).
Le prove di comprensione, effettuate soprattutto in Italia, sono sottoposte a 656 allievi di 39 classi (463 studenti di scuola primaria, 66 di scuola secondaria di primo grado e 127 di secondo grado).
La valutazione delle risposte delle prove cloze tiene conto delle tipologie di parole cancellate. «L'esperienza derivata dallo studio della bibliografia e dalle prove euristiche più volte ripetute provano che la tipologia delle parole cancellate incide sulla comprensione del testo; che sono più identificabili parole della lingua naturale, che non parole di carattere logico, che non termini tecnici della matematica; e che dunque, nel valutare la difficoltà di uno studente nel ricreare il brano tratto da un testo a lui adatto, l'incidenza dell'errore debba essere "pesata"» (id.). Non vengono cancellate né le formule né i segni di punteggiatura.
Per la definizione della formula gli autori partono da una formula esistente[101] e ne ritoccano i coefficienti, tarandola su testi di tipo matematico e in base agli indici stabiliti.
Consideriamo un brano T che comprende *n* parole. Il numero delle parole cancellate è Int(*n*/5), cioè la parte intera del numero razionale *n*/5. Le parole cancellate appartengono a diverse categorie:

a. Parole del linguaggio naturale (né di carattere logico, né tecnico);
b. Tecnicismi matematici;
c. Termini di carattere logico: connettivi (non, e, o, implica, ecc.), quantificatori (nessuno, alcuni, tutti, ecc.), termini deduttivi (poiché, siccome, dimostra, ecc.).

Quindi:

$$a + b + c = Int(n/5)$$

dove *n*, Int(*n*/5), *a*, *b*, *c* sono tutti numeri naturali.
L'indice di difficoltà del testo (m_T) può essere così calcolato:

[101] Non viene specificato di quale formula si tratti; possiamo supporre che sia una versione modificata dell'indice di Flesch.

$$m_T = a \times 0{,}1 + b \times 0{,}3 + c \times 0{,}4$$

Siano:

- *a'.* le parole di tipo *a* che il soggetto S riconosce in forma corretta (*a'* ≤ *a*);
- *b'.* le parole di tipo *b* che il soggetto S riconosce in forma corretta (*b'* ≤ *b*);
- *c'.* le parole di tipo *c* che il soggetto S riconosce in forma corretta (*c'* ≤ *c*).

è possibile calcolare *l'indice di comprensione di T da parte di S* (r_{TS}) con la seguente formula:

$$r_{TS} = (a - a') \times 0{,}1 + (b - b') \times 0{,}3 + (c - c') \times 0{,}4$$

Se $r_{TS} = 0$, si considera la comprensione del brano T da parte di S perfetta.
Se $0 \leq r_{TS} < m_T/2$ si considera la comprensione del brano accettabile o positiva;
se $m_T/2 \leq r_{TS} < m_T$ si considera la comprensione del brano insufficiente o negativa.

5.5. Il cloze test in italiano

La procedura cloze nasce negli Stati uniti come nuovo metodo per misurare la leggibilità dei testi e viene poi largamente usata con altre finalità, in particolare per valutare la comprensione di un brano scritto in lingua madre o in lingua straniera.
In Italia il cloze viene inizialmente impiegato nell'insegnamento delle lingue straniere, ad esempio come test di ingresso per l'assegnazione degli studenti nelle varie classi, ma ben presto è adottato anche come tecnica per l'insegnamento in lingua madre e tutt'oggi è ampiamente usato per valutare le competenze degli allievi e per misurare la loro comprensione dei testi[102].
In italiano si pone da subito il problema terminologico: c'è chi accoglie il prestito inglese *cloze*, chi lo traduce con *esercizi di completamento,* chi con *esercizi di chiusa*. D'accordo con Marello (1984, 1989) scegliamo di mantenere il termine *cloze* sia per indicare la procedura sia il testo bucato, precisando di volta in volta se si tratta del modello "classico", cioè quello a intervalli regolari, o di quello "mirato", in cui le cancellazioni sono effettuate in base a criteri testuali o grammaticali. Ricorrere a termini come *esercizi di completamento dei testi* è poco pratico; impiegare il termine *chiusa* non è invece adeguato perché, come ci dice Marello (1984), non è del tutto conforme al significato di *cloze* così come inteso da Taylor, in quanto gli esercizi di chiusa, utilizzati soprattutto dagli insegnanti di lingua inglese, sono semplicemente il completamento di frasi. "Chi traduce con *esercizi di chiusa* si ricollega alle prove dei gestaltisti menzionate da Taylor, ma trascura il fatto che Taylor non ha usato *closure*, traducibile con *chiusa*, *chiusura*, ma si è servito di una nuova parola. Il tecnicismo *cloze* merita o una traduzione diversa da quella riservata a *closure* o l'ingresso in italiano come prestito. Questa seconda soluzione mi sembra la migliore, perché anche i termini *esercizi di completamento di testi* o *di integrazione di testi* si possono confondere con pratiche diverse quali lo sviluppo di «testi disidratati» o il fornire finali di racconti lasciati in sospeso. Inoltre *completamento* e *integrazione* sono da sempre adoperati per esercizi su frasi ed è perciò necessario esplicitare ogni volta che si tratta di completamento di lacune *in*

[102] Del cloze test in italiano si sono occupati Marello 1984, 1989a, 1989b, 1991, De Grafenstein e Pierdonati 1985, Lucisano 1989, 1992 e più recentemente Chiari 2002.

testi". Orientarsi sul termine *cloze* sembra dunque la scelta migliore: "cloze è breve, non è sfacciatamente straniero, indica con sicurezza una procedura applicata a testi e non a frasi, perciò accoglierlo è più facile che creare una nuova parola" (Marello 1989, p. 7).
Marello (1984, 1989, 1991) propone alcuni suggerimenti per la costruzione delle prove in italiano.
Il cloze classico prevede di cancellare le parole ad intervalli regolari, di solito una ogni 5; è il modello usato nella valutazione della leggibilità nell'inglese in quanto permette di valutare la competenza testuale del lettore nella sua totalità e sembra essere una misura equilibrata anche per la lingua italiana. Per essere un valido test di comprensione, il brano dovrebbe contenere almeno 50 lacune: se si cancella una parola ogni 5, il testo dovrebbe essere quindi di almeno 250 parole, a cui si aggiungono due righe iniziali, di solito fornite interamente. Lucisano (1989) ritiene invece sufficienti 25-30 buchi e propone come lunghezza dei testi un minimo di 180 parole e un massimo di 600. Il fatto di lasciare qualche riga iniziale o frase introduttiva senza buchi serve come elemento di facilitazione perché "in qualche modo consente l'attivazione di schemi normali di comprensione di quel testo e solo in un secondo momento introduce l'aspetto di prova" (p. 162).
Sono da preferire brani tratti da giornali, libri e riviste a testi creati appositamente dall'insegnante, che presenterebbero una concentrazione di aggettivi, avverbi, verbi, ecc. superiore rispetto a materiale non prettamente scolastico. Anche brani letterari e poesie vanno esclusi dalla selezione. È importante riscrivere il testo scelto, adottando la stessa lunghezza per tutti gli spazi bianchi.
Oltre a questo modello esiste il cloze "mirato", ovvero tutta una serie di varianti in cui sono cancellate determinate parole o parti del discorso, indipendentemente dalla posizione che occupano. La cancellazione delle sole lettere, ad esempio, è una variante rivolta principalmente ai bambini delle elementari, per valutare la loro competenza ortografica. Si possono inoltre cancellare le sillabe oppure, salendo di difficoltà, i morfemi grammaticali o gli affissi.
Un'altra possibilità è la cancellazione di classi di parole: si possono eliminare le parole "vuote", come preposizioni e articoli o le parole "piene", come verbi, aggettivi, nomi, avverbi. Preposizioni e articoli sono più facili da integrare, in quanto appartengono ad insiemi chiusi e numericamente limitati. La difficoltà sale in modo progressivo con le lacune di avverbi e aggettivi fino ad arrivare all'occultamento di nomi e verbi, che risultano i più complessi. Marello consiglia di non cancellare tutti i verbi, lasciando al loro posto quelli fondamentali per la comprensione del testo e, se serve, di fornire agli studenti una lista di alcuni o tutti i verbi, magari nella forma di infinito. La cancellazione dei verbi può servire agli insegnanti come controllo della padronanza da parte degli studenti dell'accordo soggetto-predicato e della sintassi del periodo. Così come per i verbi, è sconsigliabile cancellare tutti i nomi ed omettere soltanto quelli che possono essere recuperati dal contesto o tramite inferenze.
Cloze che prevedono la cancellazione di proforme[103] e connettivi sono utili per valutare la capacità di cogliere nessi logici e temporali e, in generale, la competenza pragmatica e

[103] Con *proforme* l'autrice intende quelle espressioni linguistiche che stanno al posto di parole già dette/scritte (proforme anaforiche) o che saranno dette/scritte entro breve (proforme cataforiche), oppure espressioni che riassumono il contenuto preposizionale di quanto è stato detto o sarà detto (cfr. Marello 1984).

testuale del lettore. L'omissione dei deittici mette invece alla prova la capacità di inserire un testo nel suo contesto extralinguistico. È inoltre possibile presentare una prova in cui siano stati tolti i segni di interpunzione o la suddivisione in paragrafi. Esiste infine la variante del cloze di parole-chiave, la cui omissione segue un criterio semantico: si cancellano parole, indipendentemente dalla loro appartenenza a una qualche categoria linguistica, in base al tipo di verifica delle nozioni stabilito dal docente.

Per quanto riguarda la correzione delle prove, il punteggio è calcolato in base alla percentuale di buchi riempiti in modo corretto. Gli errori di ortografia, i numeri scritti in cifre anche se nel testo si trovano in lettere e viceversa, le abbreviazioni sciolte sono di solito valutati come completamenti corretti. Secondo Marello (1984) non devono essere considerati errori i completamenti con sinonimi o iperonimi accettabili sia dal punto di vista sintattico, che semantico frasale e testuale. Gli studi del gruppo GULP confermano la sostanziale equivalenza tra i due metodi di correzione, ovvero tra una correzione più rigida, che accetta il completamento con la sola parola cancellata e una correzione che accetta anche il completamento con i sinonimi. "Ai fini della misurazione della leggibilità del testo accettare come reintegrazioni valide anche quelle contenenti sinonimi non cambia sostanzialmente i risultati" (Marello 1989, p. 9).

Quando si costruisce una nuova prova cloze, potrebbe essere utile verificarne la validità. Lucisano (1989) suggerisce di somministrare la nuova prova insieme ad una già sperimentata, ad esempio una prova di comprensione della lettura a scelta multipla e di calcolare la correlazione tra i punteggi ottenuti. Una buona correlazione consente inoltre di stabilire una corrispondenza tra i punteggi ottenuti nei due tipi di prove.

Negli Stati Uniti, Earl Rankin e John Bormuth conducono molti esperimenti per dimostrare l'esistenza di una correlazione tra i punteggi cloze e i risultati ottenuti in altre prove di comprensione[104]. Bormuth (1967, 1968) stabilisce una corrispondenza tra i due punteggi: percentuali di risposte corrette di 50%, 75% e 90% nei test a scelta multipla equivalgono a percentuali di 35%, 45% e 55% di completamenti cloze corretti. Questi valori vengono presi come punteggi di riferimento per indicare i vari livelli di lettura degli studenti: livello di frustrazione (50% di risposte corrette nei test a scelta multipla e 35% di riempimenti corretti in un cloze), livello di lettura scolastica (75% e 45%) e livello di lettura indipendente (90% e 55%). La validità di queste percentuali è confermata dallo studio di Culhane e Ranking (1969).

Il primo studio italiano (l'unico finora rintracciato) in cui si calcola tale correlazione è quello del gruppo GULP (cfr. Lucisano e Piemontese 1988). Il coefficiente ottenuto è di 0,93; la corrispondenza tra i due punteggi è data dall'equazione:

$$Facilità\ TCL = 22.14 + Cloze\ x\ 1.07$$

Le ricerche sul cloze test in lingua italiana sembrano avere quasi esclusivamente finalità didattiche; il test è ampiamente usato per misurare la comprensione dei testi in lingua madre e in lingua straniera e per valutare le competenze degli studenti ma risultano quasi del tutto assenti studi che riguardano l'applicazione della procedura alla leggibilità dei testi. Il fatto che gli unici dati a disposizione provengano da una sola ricerca solleva alcune domande.

[104] Si vedano Bormuth 1967, 1968, Rankin 1959, 1965, Culhane e Rankin 1969, Rankin e Dale 1969.

È possibile confermare la validità del cloze come criterio per la costruzione di una formula di leggibilità in italiano? Esistono altri studi che confermano le correlazioni tra i punteggi ottenuti da cloze test e prove a scelta multipla? Le percentuali assunte come riferimento (50%, 75% e 90% di risposte corrette nei test a scelta multipla) per indicare il livello di lettura degli studenti (lettura indipendente, lettura scolastica, lettura frustrante) seguono il modello anglosassone o sono state verificate anche in italiano? L'equivalenza tra queste e le percentuali di completamenti cloze corretti (35%, 45% e 55%) è la stessa anche in italiano? O varia a seconda del coefficiente di correlazione?

6. Nuovi approcci al tema della leggibilità

Con l'avvento di metodi computazionali sempre più sofisticati e una crescente disponibilità di nuove fonti di dati e applicazioni per il web e i social media, le ricerche sulla valutazione della leggibilità dei testi si sono evolute in modo significativo a partire dalla prima metà degli anni 2000.

Le tradizionali formule di leggibilità, come l'indice di Flesch, sono state impiegate per decenni sui testi scritti; tuttavia, si è registrato un passaggio da tali misure tradizionali a favore di nuovi approcci alla valutazione della leggibilità, che utilizzano modelli di previsione avanzati basati sull'apprendimento automatico. "These new approaches are dynamic and oriented towards both traditional and non-traditional texts: They can learn to evolve automatically as vocabulary evolves, adapt to individual users or groups, and exploit the growing volume of deep knowledge and semantic resources now becoming available online. In addition, non-traditional domain areas like the Web and social media offers novel challenges and opportunities for new forms of content, serving broad categories of tasks and user populations" (Collins-Thompson 2014, p. 4).

Nonostante i tradizionali indici di leggibilità siano relativamente facili da calcolare e siano stati applicati con successo in molti campi, sono molte le obiezioni che sono state rivolte a tali tecniche: non tutti gli studiosi concordano infatti nel ritenere che le formule costituiscano uno strumento efficace e molti criticano il valore o la validità della misurazione stessa della leggibilità[105]. Le critiche principali riguardano il fatto che le formule non tengono conto di diversi fattori che influenzano il processo di comprensione come il vocabolario impiegato, la correttezza ortografica, grammaticale e sintattica del testo, la struttura logica, l'impaginazione, la dimensione e il tipo di caratteri impiegati, la presenza di tabelle, immagini, grafici o di accorgimenti volti a facilitare la decodifica, come titoli, sottotitoli, sottolineature, grassetti, ecc. Non tengono inoltre conto delle caratteristiche che riguardano il lettore, come il suo livello culturale, la sua preparazione, la sua motivazione, il suo interesse, ecc. È possibile anche che alcuni fattori, nonostante possano rappresentare dei buoni indicatori di difficoltà, vengano lasciati fuori dalle formule perché troppo complessi da misurare.

Un altro limite evidente è che esse sono progettate, eccetto alcuni casi sporadici, esclusivamente per l'analisi di testi scritti e male si adattano al contesto del web e delle informazioni online. Alcuni studi recenti hanno dimostrato l'inaffidabilità di questi strumenti nel valutare pagine web e altri tipi di documenti non tradizionali[106]. Uno dei problemi è innanzitutto la dimensione e la varietà del campione di testi presenti sul web: è possibili trovare testi abbastanza lunghi ma anche brani di poche parole, affiancati da immagini e video o corredati di tabelle ed elenchi. Le classiche formule di leggibilità sono state sviluppate per valutare generalmente brani o porzioni di testo di almeno 100 parole e risultano invece inattendibili nel caso di testi più brevi. Le pagine web possono inoltre presentare una struttura sintattica diversa da quella dei documenti tradizionali; l'individuazione stessa dei confini della frase diventa problematica, in quanto la presenza di

[105] Si veda al riguardo Klare 1976, Coupland 1978, Bruce et al. 1981, Ambruster et al. 1985, Manzo 1986.

[106] Cfr. Si e Callan 2001, Collins-Thompson e Callan 2004 e 2005, Feng et al. 2009.

numerosi collegamenti ipertestuali potrebbe confondere gli algoritmi che conteggiano le frasi. Infine, nel caso dell'estrazione automatica di informazioni da documenti web si rende necessario ripulire i dati ed eliminare il rumore.

Recentemente i ricercatori hanno dimostrato un rinnovato interesse per la ricerca sulla leggibilità; l'intento di superare le suddette limitazioni, insieme ai progressi compiuti nel campo del *Machine Learning* e lo sviluppo di efficienti tecniche di *Natural Language Processing* (NLP)[107], hanno infatti contribuito alla nascita di nuovi approcci alla valutazione della leggibilità. Da una parte, l'opportunità di sfruttare nuove risorse computazionali e grandi quantità di dati ha consentito ai ricercatori di esplorare una più ampia varietà di caratteristiche linguistiche e sperimentare variabili più complesse, dall'altra, l'uso di modelli di previsione più sofisticati basati sull'apprendimento automatico ha permesso di costruire strumenti e algoritmi avanzati per la misurazione della leggibilità. Uno dei vantaggi di questi nuovi modelli è che possono essere riadattati facilmente in base a nuovi dati e a diverse applicazioni: "that makes data-driven methods well suited for use in classification of readability, as languages change rapidly and so are the types of text to analyze" (Larsson 2006, p. 13).

In questo capitolo, dopo una breve panoramica sul *machine learning* e alcuni algoritmi di apprendimento automatico, verranno presentati gli approcci più recenti alla misurazione della leggibilità, sia per la lingua inglese, che come sempre è la lingua da cui parte l'impulso alla ricerca, sia per le altre lingue, tra cui l'italiano. Vedremo, in particolare, che la tendenza è ormai lo sviluppo di strumenti rivolti a misurare in modo automatico testi e risorse presenti sul web.

6.1. Il machine learning

Il *machine learning*, o *apprendimento automatico*, si occupa di progettare algoritmi che consentono la costruzione di sistemi in grado di apprendere dati[108]. Gli algoritmi di apprendimento permettono al computer di imparare a svolgere un compito, a partire da un

[107] Il *Natural Language Processing* (NLP), in italiano *Trattamento Automatico del Linguaggio* (TAL), è un settore della linguistica computazionale che si occupa dello "studio dei sistemi informatici per la comprensione e generazione del linguaggio naturale" (Grishman 1986 p. 4), ovvero di sviluppare programmi e sistemi informatici che, attraverso l'elaborazione automatica del linguaggio, siano in grado di estrarre informazioni da documenti testuali. "Fin dalle prime apparizioni dei calcolatori, negli anni '40, prese piede e si diffuse l'idea di utilizzarli per compiere elaborazioni su lingue come l'inglese, il francese, il russo, cioè quelle che in una parola si sogliono chiamare le *lingue naturali*. […] Naturalmente, queste elaborazioni presero forme diverse. Da un lato, il calcolatore venne utilizzato come uno strumento per isolare le parole di un testo, ordinarle alfabeticamente, contarle, ed acquisire, così, i dati per l'elaborazione linguistico statistica. D'altro lato, il calcolatore venne anche utilizzato come uno strumento per costruire modelli del processo umano di strutturazione, interpretazione e comprensione delle frasi di una lingua. […] Si può dire, al giorno d'oggi, che per Natural Language Processing s'intende, ormai, l'insieme degli studi e delle realizzazioni collegate col secondo modo di utilizzazione di un calcolatore, cioè come simulatore del comportamento linguistico umano" (Ferrari 1991, p. 5-6).

Sul *Natural Language Processing* (NLP) si veda Manning e Shutze (1999); un'introduzione italiana all'argomento è fornita da Ferrari 1991. Per una panoramica più recente sul *text mining* si vedano Bolasco 2013 e Melucci 2013.

[108] Una trattazione approfondita dell'argomento è fornita in Mitchell 1997, Russell e Norvig 2003, Alpaydin 2004.

campione di dati o esempi rappresentativi di come si svolge quel compito. Un esempio è lo sviluppo di programmi in grado di classificare automaticamente un messaggio di posta elettronica come spam, basandosi su un insieme di messaggi classificati manualmente come tali. Il programma si avvale dell'esperienza estratta dai dati osservati per migliorare le sue performance, cioè il suo comportamento di fronte a nuovi input: "Si dice che un programma apprende dall'esperienza E rispetto ad una classe di compiti T e alla misurazione della performance P, se la sua performance sui compiti in T, così come misurata da P, migliora con l'esperienza E" (Mitchell 1997).

L'insieme dei dati linguistici di partenza è detto *training corpus* ed è costituito principalmente da corpora linguistici, annotati o meno. La metodologia di apprendimento prevede generalmente l'uso della distribuzione statistica dei dati nel corpus per la costruzione di un modello generale, che sarà poi verificato su un nuovo set di dati, il *test corpus*.

Gli algoritmi di apprendimento automatico possono essere raggruppati in tre categorie, in base alla struttura dei dati del *training corpus* e alle metodologie impiegate:

- Apprendimento supervisionato
  Il corpus di apprendimento è già etichettato: vengono forniti al computer dei documenti precedentemente classificati in modo manuale da un esperto del dominio. Sono adatti a risolvere problemi di classificazione: l'obiettivo è quello di generare automaticamente un classificatore per una data categoria osservando le caratteristiche di questi documenti e determinando quindi le caratteristiche che un nuovo documento dovrebbe avere per essere classificato sotto quella data categoria.
- Apprendimento non supervisionato
  Vengono forniti al computer input non classificati precedentemente; l'obiettivo è quello di trovare automaticamente una struttura degli input forniti. Sono adatti a compiti di organizzazione di dati, come il *clustering*, cioè il raggruppamento di elementi in base a caratteristiche simili.
- Apprendimento per rinforzo
  Anche in questo caso il corpus di apprendimento non è etichettato; la differenza rispetto all'apprendimento non supervisionato è la diversa metodologia impiegata. Il computer potrebbe essere considerato come un alunno che interagisce con un'insegnante, l'ambiente esterno: di fronte a un nuovo input la macchina viene "premiata" o "punita" in base alle scelte compiute e in questo modo impara e migliora la propria performance. Russell e Norvig (2003) utilizzano infatti il termine *ricompensa* come sinonimo di *rinforzo*.

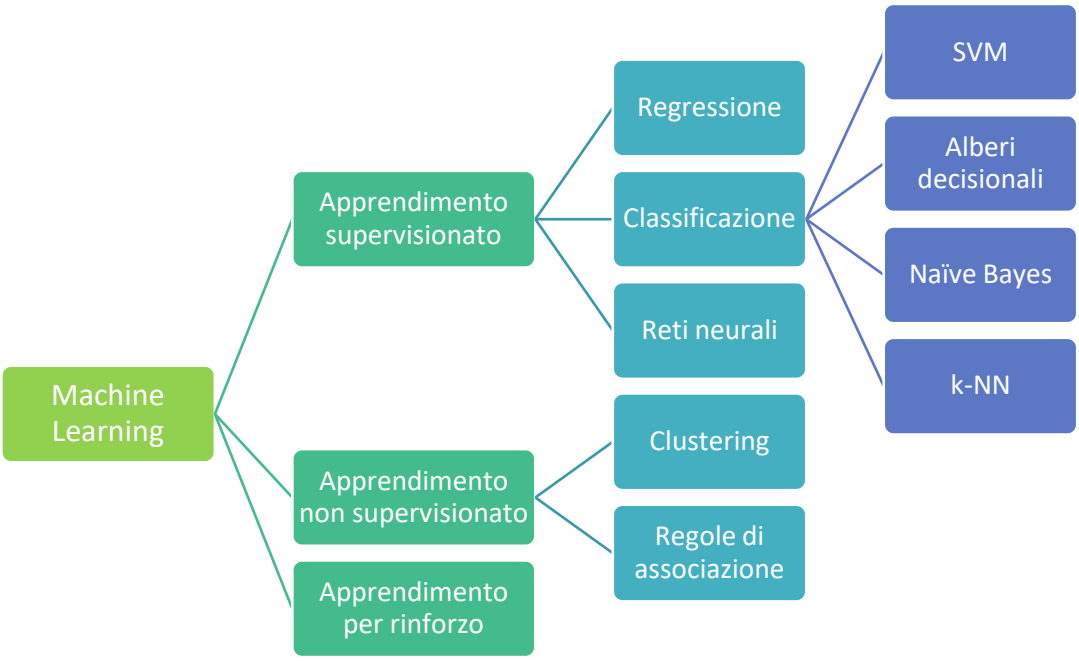


Figura 9. Algoritmi di apprendimento automatico.

Il *machine learning* inizia a diffondersi a partire dagli anni Novanta[109] e in poco tempo diventa il paradigma dominante in molti campi di applicazione: oltre al già citato filtraggio anti-spam della posta elettronica, il metodo è impiegato nei motori di ricerca, nelle traduzioni automatiche, nel riconoscimento vocale, nei sistemi di visione artificiale (*computer vision*) come il riconoscimento facciale, il riconoscimento di immagini o di caratteri, nella guida automatica di veicoli e più in generale nei sistemi robotici, nella bio-sorveglianza, ecc.[110]

[109] In realtà il termine *machine learning* è stato coniato da Arthur Samuel nel 1959 (cfr. Samuel 1959). Lo scienziato americano aveva sviluppato un programma in grado di giocare a dama: effettuando centinaia di partite contro se stesso, il sistema aveva così imparato a giocare ad alti livelli. Il metodo applicato da Samuel era quello dell'apprendimento per rinforzo.

[110] Ormai tutte le più grandi aziende tecnologiche, come Amazon, Apple, Facebook, Google, Microsoft, ecc. investono nel settore dell'intelligenza artificiale e dell'apprendimento automatico. Oltre ad utilizzare questi strumenti per i loro prodotti, molti di questi "giganti" li vendono come servizi per le aziende e i privati. Ad esempio Google propone: *Google Cloud Machine Learning Engine*, un servizio di *machine learning* che offre modelli predefiniti e un servizio per generare i propri modelli personalizzati; *API Cloud Vision* e *API Cloud Speech* che consentono l'analisi di immagini e il riconoscimento vocale; *API Google Natural Language* che si occupa di analisi testuale e può essere impiegato per il riconoscimento di entità, analisi delle opinioni, ecc.
Anche Amazon fornisce tutta una serie di prodotti e servizi per l'apprendimento automatico. Tra questi si segnalano: *Amazon Machine Learning* e *Amazon SageMaker*, che forniscono strumenti per creare modelli di apprendimento automatico senza dover apprendere tecnologie e algoritmi complessi; *Amazon Rekognition*, che si occupa di analizzare immagini e video (ad es. consente il riconoscimento facciale) basandosi sull'apprendimento profondo; *Amazon Comprehend*, servizio di elaborazione del linguaggio naturale (NLP) che usa l'apprendimento automatico per trovare informazioni e relazioni nel testo.

6.2. Apprendimento non supervisionato

6.2.1. Clustering

Il *clustering* (raggruppamento) è un metodo non supervisionato che si occupa di organizzare i dati in gruppi (*clusters*), in base a un criterio di somiglianza, o in termini di distanza o vicinanza. A differenza della classificazione, i gruppi non sono conosciuti a priori. Un esempio di questi algoritmi si ha nei motori di ricerca, ad esempio se si vogliono raggruppare i risultati di una ricerca per identificare i vari temi trattati o se si vogliono raggruppare gli utenti in base alle loro query per identificare i diversi profili del pubblico.
Se l'assegnazione è univoca, cioè se un elemento appartiene a un solo *cluster*, si parla di *clustering esclusivo* o *hard clustering*. Se un elemento può essere inserito in più di un gruppo, si parla invece di *soft clustering* o *fuzzy clustering*.
Esistono diverse tecniche di *clustering*; in particolare, si distingue tra metodo *parzionante* (detto anche *non gerarchico* o *k-clustering*) e metodo *gerarchico*:

- *clustering parzionante*: può essere realizzato tramite il noto algoritmo *k-means* (o delle *k-medie*)[111]. L'algoritmo crea k gruppi e assegna, spesso in modo casuale, k dati che divengono i punti centrali dei gruppi (centroidi o punti medi). Procede dunque ad assegnare gli altri dati ai centri dei cluster più vicini; utilizza quindi i dati per ricalcolare una nuova media (cioè nuovi centroidi) per ogni cluster, e così via, finché l'algoritmo non converge[112]. All'interno dello stesso cluster dunque i dati saranno simili tra loro rispetto a un criterio (metrica).

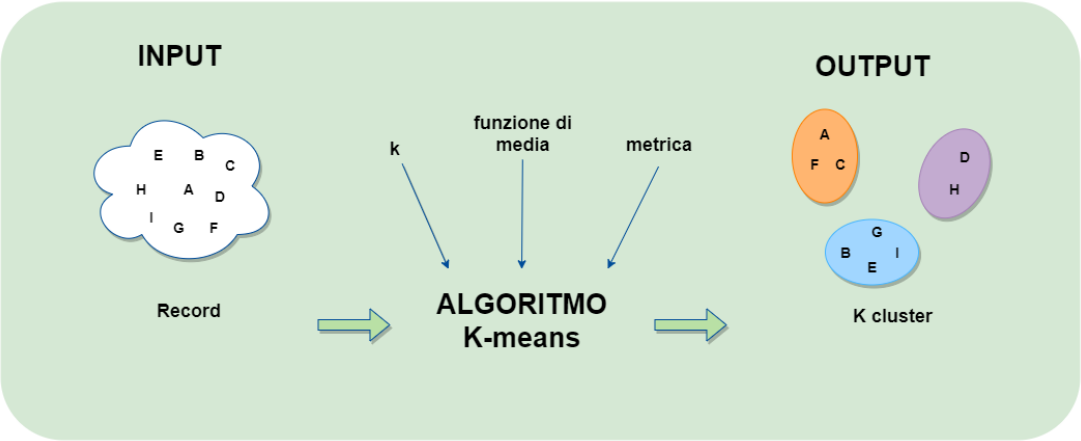


Figura 10. Schema dell'algoritmo K-means (rielaborazione da Zazzaro 2009).

- *Clustering gerarchico*: l'algoritmo costruisce una gerarchia di cluster, o suddividendo il dataset via via in sotto gruppi (approccio divisivo) o inserendo i dati in cluster e combinando a coppie i vari gruppi (approccio agglomerativo). Una rappresentazione grafica del processo può essere fornita da un diagramma ad albero o *dendogramma*.

[111] Cfr. MacQueen 1967.
[112] In un algoritmo iterativo, la *convergenza* indica la possibilità di giungere a un risultato in un numero finito di passi, o attraverso l'individuazione del risultato vero o attraverso una sua approssimazione attendibile. È prassi comune scegliere un criterio di conclusione ed eseguire l'algoritmo finché il risultato raggiunto non soddisfa il criterio prestabilito.

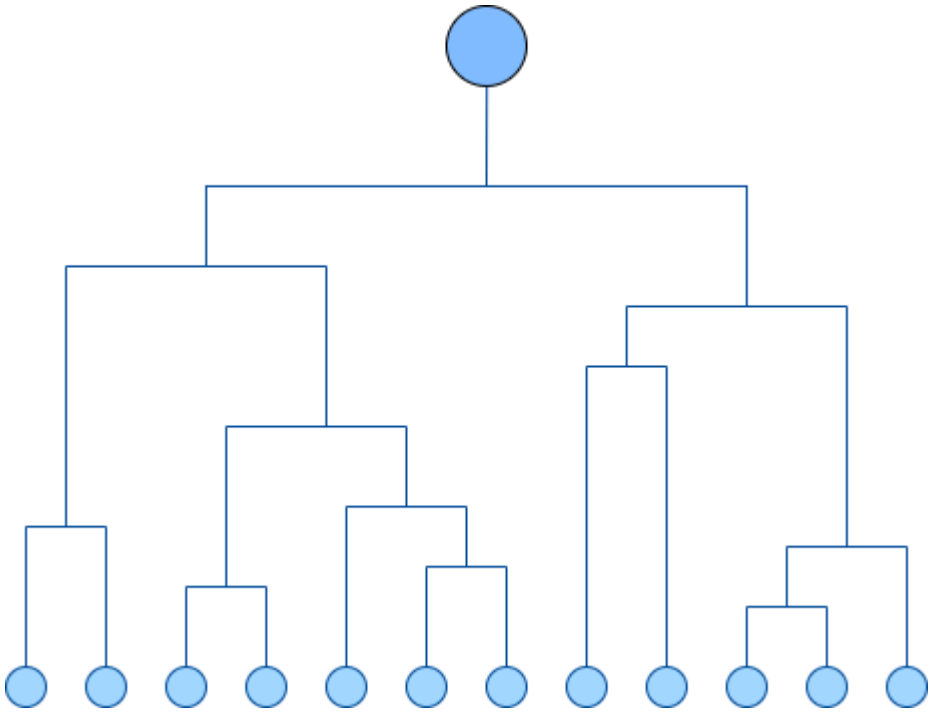


Figura 11. Esempio di dendogramma.

6.2.2. Regole di associazione

Le regole di associazione (o regole associative) si collocano tra i metodi di apprendimento non supervisionato e sono volte a identificare regolarità e relazioni tra i dati[113]. Sono spesso impiegate nelle analisi di transazioni commerciali (*market basket analysis*)[114]. Ad esempio, attraverso la raccolta dei dati delle transazioni online o l'analisi degli scontrini dei supermercati, è possibile individuare quali sono i prodotti che hanno la maggiore probabilità di essere comprati insieme.

6.3. Apprendimento Supervisionato

I modelli di apprendimento supervisionato si basano su corpora di dati già etichettati. A partire dall'analisi del training corpus precedentemente classificato, si estrae il modello in grado di assegnare un classificatore ad un nuovo set di dati non classificato. Quando l'etichetta dei dati è di tipo testuale (ad esempio spam/non spam) si parla di *classificazione*; quando è di tipo numerico si parla di tecniche basate sulla *regressione*.

[113] Il concetto di regola associativa è stato introdotto per la prima volta nell'articolo di Agrawal et al. (1993).

[114] Si tratta di una metodologia orientata all'identificazione delle relazioni esistenti tra i prodotti acquistati dai consumatori. Serve da supporto per decidere il posizionamento dei prodotti negli scaffali, per focalizzare le offerte promozionali, ecc.

6.3.1. Regressione

L'analisi di regressione è un modello che si occupa di descrivere (se esiste) la relazione tra una variabile dipendente (o variabile di risposta) e una o più variabili indipendenti[115]. Può essere usata per individuare la natura della relazione tra due variabili o in funzione predittiva, per prevedere i possibili valori che può assumere la variabile dipendente in funzione di quella indipendente.
Ad esempio, supponiamo di avere un set di dati relativi a un gruppo di persone di cui conosciamo altezza e peso. Verifichiamo che esiste una relazione lineare tra le due variabili: possiamo quindi classificare nuovi soggetti in un *range* di peso conoscendo solo la loro altezza, o viceversa. L'output è di tipo quantitativo.

6.3.2. Classificazione automatica

La *classificazione automatica* (o *categorizzazione automatica*)[116], che appartiene al gruppo dei metodi supervisionati, prevede tutta una serie di tecniche rivolte alla costruzione automatica di classificatori di documenti testuali, cioè di algoritmi in grado di etichettare i documenti scritti in un linguaggio naturale in un insieme di categorie[117]. A partire da un corpus di dati già etichettato viene costruito il modello di classificazione, che sarà poi utilizzato per assegnare i nuovi dati alle categorie predefinite[118].
Si possono trovare innumerevoli applicazioni della classificazione, come i filtri anti-spam per la posta elettronica, i metodi di indicizzazione automatica di pagine web, strumenti destinati ai motori di ricerca, al web semantico e alla creazione di ontologie, tecniche di disambiguazione automatica, attribuzioni automatiche di paternità a documenti scritti, ecc.
È possibile operare varie distinzioni nell'ambito della classificazione (Figura 12):

- Single-Label e Multi-Label
  - Nella classificazione *single-label* (a etichetta singola), ogni documento appartiene esattamente ad una categoria, o meglio, una sola categoria può essere assegnata ad ogni documento. Un caso particolare di *single-label* è la classificazione *binaria*.
  - Nella classificazione *multi-etichetta* le categorie si sovrappongono e il documento può appartenere a più categorie.
- Classificazione binaria e Multi-classe

[115] Il *coefficiente di regressione* è stato introdotto in statistica nel XIX secolo da Francis Galton, fondatore dell'eugenetica; studiando le relazioni tra le stature di padri e figli constatò che l'altezza media dei figli tornava a regredire verso la media.
[116] In inglese *Text Classification* o *Text Categorization* (TC). La differenza tra il concetto di *classificazione* e quello di *categorizzazione* non risulta molto chiara: la maggior parte degli studiosi li considera come sinonimi.
[117] Della classificazione si è occupato in più sedi Sebastiani (1999, 2002, 2005, 2005b, 2006).
[118] Esistono attualmente diversi repository in cui è possibile reperire dataset già etichettati disponibili per la ricerca. Attualmente, la raccolta più utilizzata nell'ambito della classificazione automatica dei testi è *Reuters 21578* dell'agenzia di stampa americana Reuters. Si tratta di un corpus di 21.578 notizie divise per genere in 118 categorie, disponibile al link http://trec.nist.gov/data/reuters/reuters.html. Nel 2000 è stata rilasciata una versione più estesa, contenente circa 1 milione di articoli e chiamata RCV1 (per i documenti in inglese) e RCV2 (per la collezione multilingue).

- Nella classificazione *binaria* l'output include solo due classi: ogni documento può essere assegnato alla categoria *c* o al suo complemento $\bar{c}$;
- Nella classificazione *multi-classe* è possibile avere più di un output (cioè più classi) ma è possibile associare il documento ad una sola classe.

❖ Classificazione Hard e Ranking

- Nella classificazione *hard* il classificatore assegna ad un documento una classe distinta. La procedura è completamente automatizzata.
- Nella classificazione *soft* o *ranking*, per un dato documento, il classificatore ordina le classi di appartenenza del documento secondo una stima di probabilità. Questo metodo è utile soprattutto nei sistemi non completamente automatizzati (classificazione semiautomatica), ad esempio quando la qualità dei dati di addestramento è bassa o quando il training set non può essere considerato rappresentativo.

Generalmente i risultati della classificazione sono valutati tramite alcune misure, l'*accuratezza* (*accuracy*), la *precisione* (*precision*) e il recupero (*recall*), detto anche *copertura* o *richiamo*.
Nell'*information retrieval*, la *precisione* indica il numero di documenti effettivamente rilevanti recuperati sul totale dei risultati, cioè sul totale di tutti i documenti recuperati; il *recupero* misura il numero di documenti rilevanti recuperati diviso il numero totale di documenti rilevanti esistenti (che dovrebbero cioè essere recuperati). Nella classificazione, la *precisione* (P) indica il numero di documenti etichettati correttamente come appartenenti a una data classe (*veri positivi*) diviso il numero totale di documenti etichettati (che comprende sia i *veri positivi* che i *falsi positivi*, cioè i documenti etichettati erroneamente come appartenenti a una data classe): $P = vp/(vp + fp)$.
I *falsi positivi* rappresentano quello che viene chiamato *rumore*.
Il *recupero* rappresenta il numero di *veri positivi* diviso il numero totale dei documenti che appartengono a quella data classe (*veri positivi* e *falsi negativi*, cioè tutti i documenti che non sono stati etichettati come appartenenti a quella classe ma che dovrebbero esserlo): $R = vp/(vp + fn)$.
La combinazione dei punteggi di *precisione* (P) e di *recupero* (R) è una misura chiamata *punteggio F* ed è così calcolata: $F = 2\,PR/(P + R)$. Il *punteggio F* rappresenta la *media armonica* tra P e R, cioè il rapporto tra la media geometrica e la media aritmetica.
L'*accuratezza*, che rappresenta una valutazione complessiva della qualità del modello, è data dal numero totale di predizioni corrette (cioè dalla somma dei *veri positivi* e *veri negativi* sul totale dei casi): $A = vp + vn/(vp + fp + vn + fn)$.

Gli algoritmi di classificazione sono tantissimi e molti possono essere combinati tra loro (*multiclassificatori*); tra i principali si segnalano i classificatori Bayesiani, gli alberi di decisione, il *K-nearest neighbors* e il Support Vector Machine.

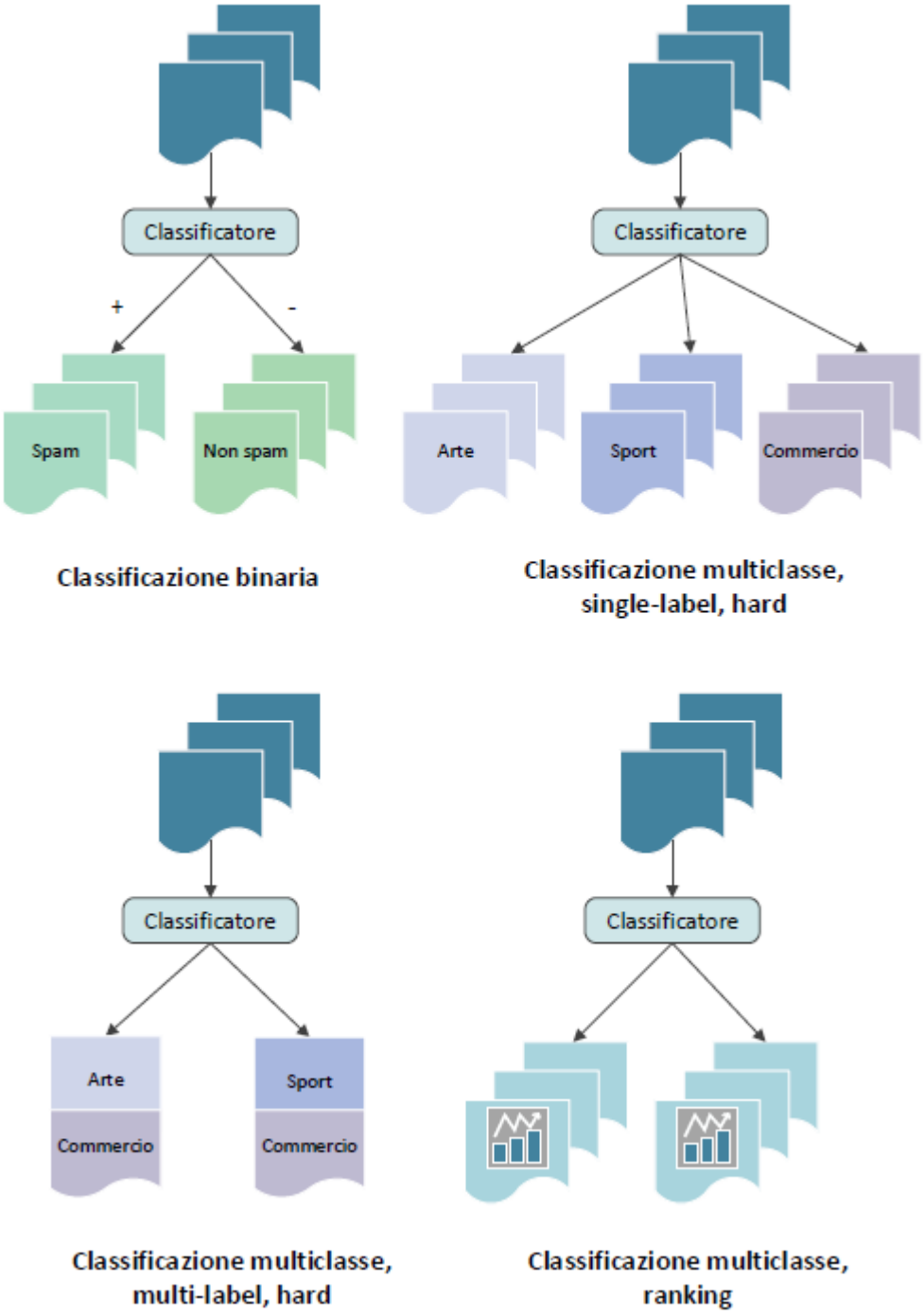


Figura 12. Tipi di classificazione (rielaborazione da Qi e Davison 2009).

6.3.2.1. Naïve Bayes

Si tratta di classificatori bayesiani (cioè basati sull'applicazione del teorema di Bayes)[119] semplificati, costruiti su modelli probabilistici. Si assume che la presenza o assenza di una determinata caratteristica (*feature*) di una classe in un documento testuale sia

[119] Il teorema, proposto da Thomas Bayes, è alla base di tutti i sistemi moderni di inferenza probabilistica e definisce la probabilità condizionata (o a posteriori). Viene impiegato per calcolare la probabilità di una causa che ha scatenato l'evento verificato. L'equazione è la seguente: $P(A|B) = (P(B|A) \times P(A)) \ / \ (P(B)\ )$.

indipendente dalla presenza o meno di qualsiasi altra caratteristica, dato il valore della classe. Tale assunzione, chiamata *indipendenza condizionale delle classi*, ha lo scopo di semplificare i calcoli ed è proprio per questo che l'algoritmo prende il nome di *naïve* ('ingenuo').
Uno dei campi di applicazione del metodo bayesiano è il filtro anti-spam della posta elettronica. Supponiamo di aver stabilito un insieme di feature con cui descrivere i messaggi da filtrare (*k* parole di un vocabolario prefissato) e di avere già a disposizione un set di addestramento (cioè un corpus di email già analizzate). Per ogni messaggio del corpus si costruisce un vettore di *k* cifre binarie per indicare la presenza o l'assenza della feature (cioè della parola) nel messaggio. Ad esempio, se si considera il messaggio "vincita sicura" e come dimensione del vettore k=5 (*vincita*, *acquista*, *conto*, *sicura*, *ciao*), si avrà:

$$x = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \begin{matrix} vincita \\ acquista \\ conto \\ sicura \\ ciao \end{matrix}$$

Dove 1 indica la presenza della parola e 0 l'assenza. Si tratta quindi, una volta osservato un vettore di caratteristiche, di attribuire una probabilità al messaggio, cioè decidere se l'email di cui sono state analizzate le feature è più probabilmente un messaggio di spam oppure no.
Un'alternativa è rappresentata dallo schema *multinomiale.* In questo caso si determina se un messaggio è o meno spam con probabilità a priori e poi si generano le feature.

6.3.2.2. Alberi di decisione

Si tratta di un modello predittivo costruito come un albero in cui ogni nodo interno è etichettato tramite termini, le foglie rappresentano le categorie e le ramificazioni l'insieme delle proprietà che portano a quelle categorie.
Gli alberi di decisione consentono di sviluppare sistemi di classificazione in grado di prevedere o classificare osservazioni future in base a un insieme di regole decisionali. Ad esempio, è possibile creare una struttura ad albero che classifica il rischio di credito, dividendo i dati in classi di interesse (prestiti a basso rischio e prestiti ad alto rischio) in base a vari fattori (età, tipo di impiego, ecc.): utilizzando tali dati sarà possibile generare regole che potranno essere utilizzate per classificare ulteriori casi.
Un altro esempio potrebbe essere la classificazione di documenti in base alla presenza o meno di determinate parole. La regola decisionale potrebbe essere: *Dato un documento x, se questo contiene le parole y e z, allora il documento può essere assegnato al gruppo g*. Questo tipo di classificazione richiede però un esperto di dominio che determini le regole per l'assegnazione in categorie.

6.3.2.3. Support Vector Machine (SVM)

Le *macchine a vettori di supporto* (SVM), chiamate anche *macchine kernel*, sono un metodo di apprendimento, usato sia nella classificazione che nella regressione, in grado di

rappresentare funzioni non lineari complesse[120]. L'algoritmo di apprendimento utilizza una funzione, chiamata *kernel*, per mappare uno spazio di punti di dati che non sono linearmente separabili in altro modo.
Supponiamo di dover ripartire in due categorie in insieme di *n* dati in ingresso, rappresentati da *n* punti nello spazio: un punto per ogni dato.

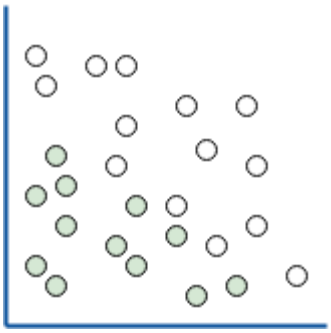


Figura 13. Dati di input.

La classificazione basata sulle SVM consiste nel far passare una curva per separare le due categorie, in modo che da una parte stiano tutti e solo i punti corrispondenti ai dati di una classe e dall'altra i punti corrispondenti ai dati dell'altra classe.

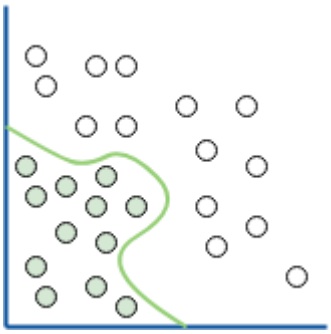


Figura 14. La curva separa i dati un due classi.

Dal momento che il calcolo di una curva che separi esattamente un insieme di punti è difficile, è possibile trasformare i dati in modo che il separatore possa essere tracciato come un iperpiano[121]. La funzione matematica utilizzata per la trasformazione è nota come *funzione Kernel*.

[120] Le *macchine a vettori di supporto* furono introdotte da Vapnik negli anni '90 (cfr. Vapnik 1995) e divulgate nel campo della classificazione da Joachims (1998).
[121] L'iperpiano è un *concetto geometrico che rappresenta l'estensione a spazi a più dimensioni dei concetti di retta e di piano* (Treccani online). In uno spazio bidimensionale (come nel piano cartesiano), l'iperpiano è una retta che soddisfa un'equazione lineare (del tipo $y=a+bx$) e divide lo spazio in due parti (semipiani). In uno spazio monodimensionale (come una retta), corrisponde ad un punto che separa lo spazio, cioè la retta, in due semirette. In uno spazio tridimensionale è un insieme di punti che soddisfa un'equazione lineare (del tipo $y=a+a_1x_1+a_2x_2$) e separa lo spazio in due semispazi.

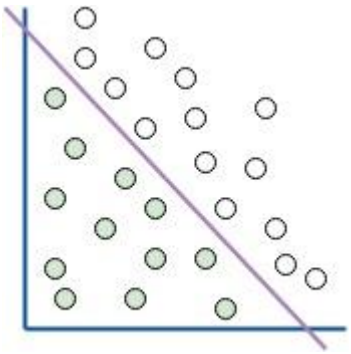


Figura 15. Iperpiano.

È possibile quindi utilizzare le caratteristiche dei dati trasformati per prevedere il gruppo di appartenenza di nuovi dati in ingresso.

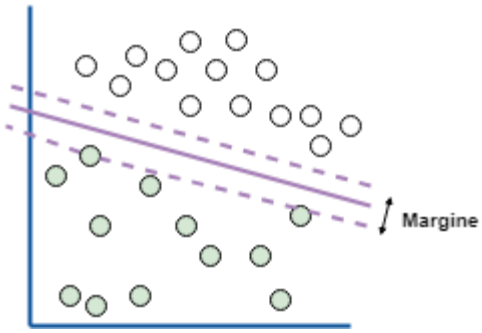


Figura 16. Linee marginali e vettori di supporto.

Oltre alla linea che separa le categorie (separatore), un modello SVM contiene anche linee marginali che definiscono lo spazio tra le due categorie. I punti che si trovano ai margini sono noti come *vettori di supporto* (così chiamati perché "sostengono" il separatore).
L'approccio SVM viene impiegato in molti campi, dall'estrazione di informazioni al riconoscimento di immagini e testo, al riconoscimento facciale e vocale, alla bioinformatica, al rilevamento delle intrusioni, all'identificazione di pedoni, ecc.

6.3.2.4. Classificatori basati su modelli statistici del linguaggio

Il *modello statistico del linguaggio* è un sistema che consiste nell'assegnazione di una misura di probabilità a sequenze di parole (o di fonemi, sillabe, lettere, ecc.). "Poche parole di poche lettere sono utilizzate molto di frequente per esprimere quasi tutti i concetti di un testo. In termini di probabilità, le poche parole o lettere molto usate sono facilmente prevedibili o molto probabili" (Melucci 2013, p. 63). Questo sistema viene impiegato per il riconoscimento vocale[122], la traduzione automatica, il riconoscimento di caratteri (OCR), il correttore ortografico, ecc.

[122] Google ha pubblicato una ricerca (cfr. Chelba et al. 2012) che mostra il funzionamento del suo sistema di riconoscimento vocale automatico. Google ha utilizzato un modello linguistico basato sia su trigrammi che su 5 grammi, costruito su un corpus di dati di allenamento formato da 230 miliardi di parole. Il modello è stato testato attraverso indagini casuali effettuate tramite un'applicazione di ricerca vocale in Android. La ricerca mostra come la precisione di sistemi di questo tipo possa aumentare all'aumentare della quantità dei dati di allenamento e delle dimensioni del modello.

In genere si presuppone che la probabilità di una parola in un testo dipenda principalmente dalle *n* parole precedenti: questo modello è chiamato modello n-gramma (*n-gram*). Quando un *n-gramma* è di lunghezza 1, cioè *n*=1, si parla di *unigramma* (modello *unigram*); in questo caso, si presuppone che la probabilità di una parola dipenda soltanto dalla probabilità di questa nel documento. Quando *n*=2 si parla di *digramma*, quando *n*= 3 di *trigramma*, per n ≥ 4 si parla di *n-gramma*.
Una delle applicazioni del modello *n-gram* è il PoS *tagging* (*Part of Speech* tagging), cioè l'assegnazione di una categoria morfo-sintattica a una parola in un testo. Per effettuare questo tipo di annotazione in modo automatico è infatti possibile usare un tagger di tipo *n-gram*. Un tagger di di tipo *n-gram* utilizza un corpus di apprendimento già etichettato per determinare la parte del discorso più probabile per quel contesto. Se si usa un tagger di tipo *unigram*, ad ogni occorrenza verrà associata un'etichetta (*tag*) con maggiore probabilità per quel token (ad esempio verbo, aggettivo, ecc.). Se si usa un tagger di tipo *n-gram* si considerano invece, oltre al tag del token considerato, anche i tag delle (n-1) occorrenze precedenti.

6.3.2.5. K-nearest neighbors (k-NN)

Si tratta di un algoritmo utilizzato per la classificazione di documenti che si basa sulle caratteristiche dei documenti vicini a quello considerato. L'idea dei modelli ai *vicini più prossimi* (*nearest neighbors*) è che le proprietà di un particolare punto di input *x* sono probabilmente simili a quelle dei punti nelle immediate vicinanze (Russell e Norvig 2003).
Per ogni nuovo documento si ricercano nel set di addestramento, contenente dati già classificati, i k documenti che risultano più simili. L'assegnazione ad un dato gruppo avviene in base al gruppo di appartenenza della maggioranza dei k vicini. Osservando ad esempio la Figura 17, se k=3 si considerano i 3 oggetti più vicini (due triangoli e un quadrato): il documento andrà nella classe che risulta in maggioranza, in questo caso la classe dei triangoli. Se k=8 si considerano gli 8 oggetti più vicini: il documento verrà classificato nella classe nella classe dei triangoli (5 triangoli contro 3 quadrati).

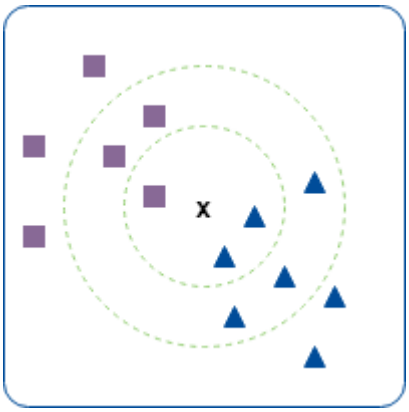


Figura 17. Scelta del numero k di *nearest-neighbors.*

6.3.2.6. Reti neurali artificiali

Le *reti neurali artificiali* (*Artificial Neural Network* o ANN), spesso chiamate semplicemente *reti neurali* (NN), sono un modello matematico-informatico per l'interconnessione dei dati

che si ispira al funzionamento del cervello umano. Possono essere immaginate come grandi reti di oggetti, detti *nodi*, che si connettono tra loro: vi sono dei nodi in ingresso, dei nodi di output e una serie di nodi intermedi, organizzati in più livelli. Ogni nodo elabora i dati ricevuti e trasmette il risultato ai nodi successivi: in questo modo la rete si evolve e adatta la propria struttura in base agli ingressi ricevuti.
Le reti neurali sono particolarmente adatte per individuare analogie e somiglianze in grandi insiemi di dati. I campi di applicazione includono il riconoscimento delle immagini, dei gesti, dei caratteri e quello vocale, i sistemi di controllo come il controllo dei veicoli, diagnosi mediche, applicazioni finanziare, ecc.
Le più recenti ricerche sull'apprendimento non supervisionato e sull'apprendimento profondo (*deep learning*)[123] sono orientate verso la creazione di modelli di reti neurali che si basano su dati in ingresso non etichettati (che è l'approccio tipico del metodo non supervisionato).
Uno degli studi più interessanti è condotto da Google, che ha tentato di costruire un modello di riconoscimento di volti fornendo alla macchina una collezione di immagini non etichettate[124]. I laboratori Google X, cioè quella divisione in cui si sviluppano i progetti più avanzati, hanno realizzato una rete neurale artificiale formata da 1.000 macchine, per un totale di circa 16.000 processori; il modello ha simulato oltre un miliardo di connessioni neurali. A questa enorme rete sono stati forniti come set di dati in ingresso 10 milioni di video di YouTube, sotto forma di fotogrammi.
Senza dare alcuna indicazione sugli oggetti rappresentati nei video, gli studiosi hanno lasciato il sistema agire in maniera autonoma per circa una settimana. Alla fine del processo di apprendimento è risultato che la rete non solo era in grado di riconoscere i volti con un'accuratezza dell'81%, ma che era in grado di imparare (e dunque riconoscere) anche il concetto di "gatto" e di "corpo umano", senza essere stata istruita in precedenza.

6.3.3. Classificazione automatica di pagine web

A partire dagli anni Duemila, il web ha subito una fortissima accelerazione e si è evoluto da un *repository* di informazioni a una piattaforma potente che supporta un'ampia gamma di servizi e applicazioni. Il fatto che siano messi a disposizione un'enorme quantità di dati fa sì che le attività di recupero e organizzazione delle informazioni siano sempre più fondamentali.
"In questa società dell'informazione, le organizzazioni (come le aziende, i centri di ricerca, le banche, i centri di analisi statistica, etc. etc.) hanno a disposizione enormi quantità di dati, non solo relativi a sé stesse, ma anche riguardanti l'ambiente nel quale si trovano ad agire. Infatti, si stima che, approssimativamente, ogni 1100 giorni, nel mondo, le informazioni memorizzate in formato elettronico raddoppino di volume; siamo sommersi dai dati provenienti dalle fonti più disparate: dati numerici provenienti, ad esempio, dai satelliti, oppure da sensori di qualsiasi natura, come quelli che rilevano movimenti tellurici o dati meteorologici; e dati testuali, non strutturati, provenienti, ad esempio, da siti web, agenzie stampa, e-mail, forum, mailing list, newsgroup, etc. […] Dunque, se da un lato le

[123] L'apprendimento profondo (*deep learning*) è una sotto area del *machine learning* che si basa su multipli livelli di rappresentazione dei dati. Fa uso delle reti neurali profonde, cioè dotate di molti strati (di profondità).
[124] Cfr. Q.V. Le et al. (2011).

aziende, oppure gli enti di ricerca (etc.), hanno a disposizione una enorme quantità di dati dettagliati e di testi, dall'altro risulta sempre più difficile districarsi tra le informazioni rilevanti e quelle superflue. È così emersa l'esigenza di creare dei metodi di scoperta automatica di conoscenza nelle grandi basi di dati; metodi capaci, ad esempio, di discernere le informazioni utili dal rumore. In breve, i dati vengono sottoposti ad un processo di analisi al fine di trasformarli in conoscenza utile alle aziende per supportare decisioni e intraprendere soluzioni più efficaci, ovvero veloci, economicamente sostenibili e tecnologicamente possibili" (Zazzaro 2009, p. 38).

La disciplina che si occupa di sviluppare e applicare algoritmi per estrarre automaticamente informazioni dalle risorse presenti sul web è chiamata *Web Mining*. "L'obiettivo del Web Mining trova giustificazione nell'opinione diffusa che l'informazione presente nel Web è sufficientemente strutturata da consentire una efficace applicazione di tecniche statistiche e di apprendimento automatico"[125]. Tra queste si inserisce la classificazione automatica delle pagine web (o *Web page categorization*), che, come la classificazione automatica dei testi, è un metodo di apprendimento supervisionato[126].

La classificazione delle pagine web consiste nell'assegnazione di una pagina web a una o più categorie predefinite in base a una serie di dati di addestramento precedentemente etichettati. L'assegnazione in categorie può avvenire in base a vari criteri: classificazione del contenuto (ad esempio in base all'argomento o al soggetto, come *arte*, *sport*, *commercio*, ecc.), classificazione in base alla funzione o al ruolo svolto dalla pagina (ad esempio se si tratta di una pagina di un sito personale, di una pagina del corso di dottorato, ecc.), classificazione dei giudizi (conosciuta come *sentiment analysis,* si basa sulle opinioni degli utenti), classificazione del genere (in base alla forma, allo stile, ai destinatari)[127], classificazione *per contesto* (si analizza la struttura dei documenti web per estrarre informazioni contestuali sui documenti; tali contesti sono quindi usati per classificare i documenti)[128], ecc.

La classificazione delle pagine web può essere suddivisa in piatta (*flat*) e gerarchica. Nella classificazione *flat* le categorie si trovano sullo stesso piano, in quella gerarchica sono organizzate in una struttura ad albero e ogni classe può contenere un certo numero di sottocategorie. Solitamente, le gerarchie sono definite manualmente.

[125] Etzioni 1996, citato in Convertino et al. 1998.

[126] Per lo stato dell'arte della ricerca sulla classificazione delle pagine web cfr. Qi e Davison 2009 e Qi 2012.

[127] Cfr. zu Eissen e Stein 2004.

[128] Cfr. Attardi et al. (1998, 1999).

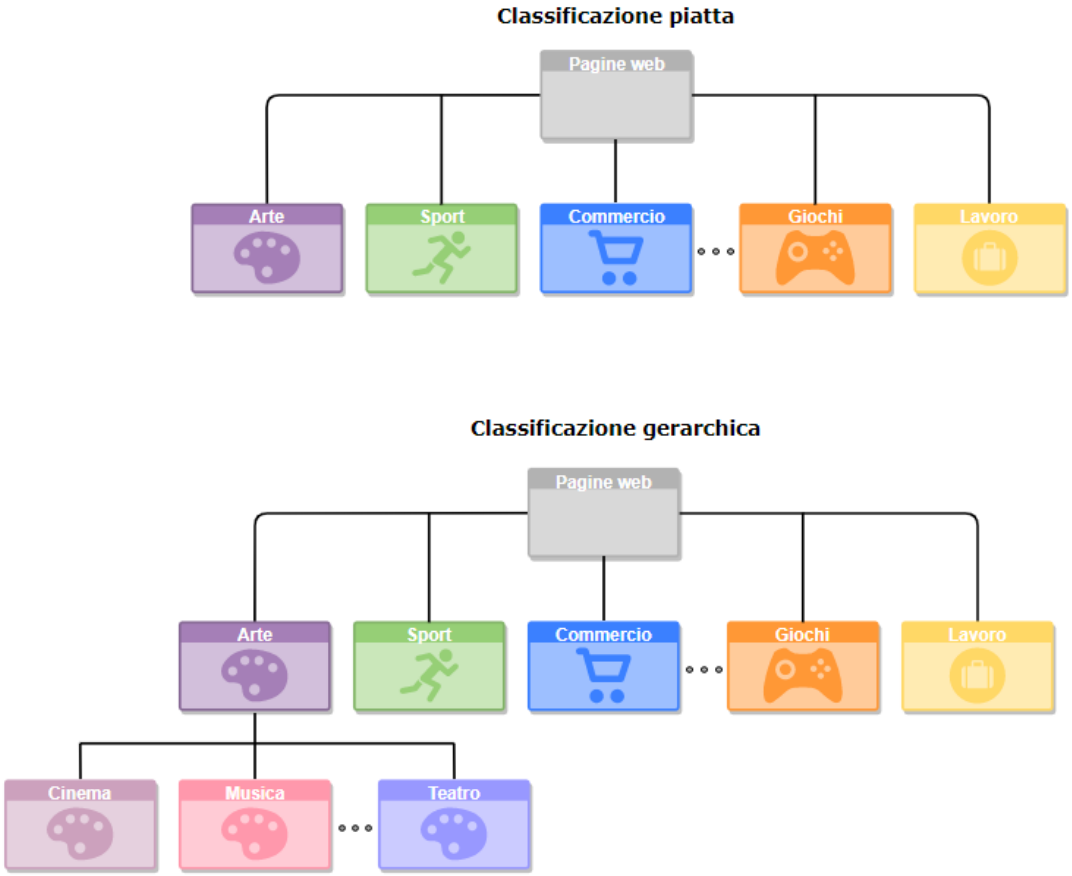


Figura 18. Classificazione piatta e gerarchica delle pagine web.

Per quanto riguarda le tecniche di classificazione, si utilizzano gli stessi algoritmi impiegati nella classificazione automatica dei testi (k-nn, SVM, modello *n-gram*, ecc.), riadattati per rispondere alle esigenze del web. Nella categorizzazione delle pagine web devono infatti essere considerate le caratteristiche e le funzionalità aggiuntive proprie del web (e dunque anche i problemi che ne derivano), come la marcatura, i collegamenti ipertestuali, la presenza di immagini, audio e video, ecc.

6.3.3.1. Learning to rank

La classificazione automatica di pagine web ha molte applicazioni, tra cui il recupero dei documenti. Gli algoritmi di *machine learning* vengono infatti utilizzati per raffinare e migliorare la qualità dei risultati delle ricerche. Quando un motore di ricerca risponde ad una interrogazione (*query*) da parte di un utente, recupera le pagine web corrispondenti alla ricerca e restituisce i risultati in ordine di rilevanza: questo processo di classificazione è detto *ranking* e si basa su vari criteri, come il contenuto delle pagine, la loro struttura[129], la frequenza con cui gli utenti seguono i link suggeriti, ecc. L'applicazione di algoritmi di

[129] Un esempio di algoritmo di ranking che utilizza la struttura dei collegamenti ipertestuali è *PageRank*. Si tratta di uno degli algoritmi impiegati da Google per il posizionamento: le pagine web sono classificate per importanza in base alla quantità e qualità dei loro collegamenti ipertestuali, a prescindere dal contenuto testuale presente.

apprendimento automatico alla procedura di *ranking* è chiamata *learning to rank* (o *machine-learned ranking*)[130].
I dati di addestramento sono costituiti dalle interrogazioni e dai documenti associati a queste, insieme al grado di rilevanza per ciascuna corrispondenza. La costruzione del *training corpus* può essere manuale, cioè effettuata da valutatori umani (*raters*) che controllano i risultati e ne determinano la rilevanza, o automatica, ad esempio tramite l'analisi dei click degli utenti (*click through*) o l'analisi delle sequenze di ricerca (*query chains*).

6.3.3.1. Un esempio italiano: Webclass

Si tratta di un sistema, sviluppato presso il Dipartimento di Informatica dell'Università di Bari, che offre servizi di assistenza al reperimento di informazione sul web[131]; funziona come un intermediario tra l'utente che naviga nel web tramite il sistema e il motore di ricerca. WebClass si occupa di classificare automaticamente le pagine web in base al loro contenuto testuale: i classificatori sono costruiti in base a un *profilo di interessi* degli utenti relativo alle pagine web.
La procedura avviene in due fasi: in una prima fase di formazione, gli utenti navigano nel web e addestrano il sistema fornendo una serie di pagine web classificate in formato HTML, dove le classi corrispondono agli argomenti di interesse degli utenti. "Si assume che la rilevanza di una pagina Web dipenda essenzialmente dal suo contenuto testuale, quindi criteri di giudizio "esterni", come il carattere di novità del documento o l'affidabilità dell'informazione, ed eventuali contenuti di tipo non testuale della pagina, non vengono considerati. Inoltre, si assume che l'utente sia in grado di specificare un insieme di classi corrispondenti ai vari argomenti di interesse, e di fornire un insieme di esempi significativi per ciascuna delle classi (insieme di addestramento)" (Convertino et al. 1998, p. 5). L'allenamento viene eseguito tramite di un'interfaccia grafica che supporta contemporaneamente funzioni di navigazione e funzioni tipiche di un sistema di apprendimento, come l'impostazione dei parametri, l'estrazione e la selezione delle caratteristiche, la definizione del training set e del test set, la generazione del classificatore e la valutazione.
In un secondo momento il sistema assiste l'utente nella navigazione, effettuando la classificazione in modo automatico. La classificazione automatica richiede la soluzione di due problemi: la definizione di un *linguaggio di rappresentazione* da utilizzare per descrivere le pagine HTML e la costruzione di un classificatore in grado di categorizzare nuove pagine web in base alle classi definite dagli utenti.
Per la classificazione WebClass si serve di varie tipologie di algoritmi (alberi di decisione, *k*-nn, centroidi, Naïve Bayes e SVM).

[130] Sull'argomento si veda Liu 2009 e Li 2011.
[131] Per informazioni sul sistema WebClass si veda Convertino et al. 1998, Esposito et al. 1999, Esposito et al. 2000, Ceci et al. 2000, Ceci et al. 2003, Ceci e Malerba 2003, De Luise et al. 2007.
La pagina web di riferimento è http://www.di.uniba.it/~malerba/software/webclass/.

6.4. La valutazione automatica della leggibilità

A partire dagli anni 2000 comincia a diffondersi un nuovo tipo di approccio al tema della leggibilità, che prevede l'applicazione di tecniche di apprendimento automatico per la predizione della difficoltà dei testi. Questi nuovi metodi di valutazione sono rivolti alla costruzione di un modello che permetta di classificare in modo automatico un insieme di documenti testuali in base al loro livello di leggibilità.
Il processo comprende diverse fasi:

- definizione di un corpus di apprendimento;
- selezione delle caratteristiche linguistiche da analizzare;
- estrazione automatica delle caratteristiche dai dati;
- selezione dell'algoritmo di apprendimento;
- creazione del modello;
- validazione del modello.

Per prima cosa, viene costruito un corpus di allenamento, rappresentativo di quell'aspetto che si intende valutare (un particolare genere testuale, un insieme di testi destinati ad un pubblico specifico, una certa varietà linguistica, ecc.). Ad ogni testo del corpus viene assegnato un livello di leggibilità *gold standard*, cioè di riferimento: è su questi livelli di leggibilità che si baserà il modello. Il livello può essere assegnato in vari modi, ad esempio tramite valutatori umani esperti; in altri casi, soprattutto per la lingua inglese, è possibile trovare dei set di dati già etichettati. La stabilità e l'affidabilità del modello dipendono dalla quantità di dati utilizzati (Larsson 2006). Possono essere impiegate diverse scale di misurazione: nella maggior parte degli studi considerati, i livelli *gold standard* indicano i livelli di comprensione della lettura di una data popolazione e si basano sul sistema scolastico americano che prevede una divisione in 12 gradi (livelli di istruzione).
La seconda fase prevede la selezione di quelle caratteristiche linguistiche che dovranno essere analizzate da ciascun testo. Si tratta di scegliere un insieme di caratteristiche che potrebbero essere dei buoni predittori della leggibilità. Le caratteristiche possono essere di tipo lessicale, sintattico, semantico, ecc.
Una volta effettuata la selezione, si procede con l'estrazione automatica delle caratteristiche: si trasforma ogni testo in un vettore di caratteristiche numeriche che servirà da input per l'algoritmo di apprendimento. L'algoritmo crea quindi il modello: impara cioè, in base agli esempi forniti, ad associare ogni vettore di caratteristiche che rappresenta un testo al livello di leggibilità definito per quel testo.
L'ultima fase prevede la validazione del modello su un nuovo set di dati. Il modello ottimizzato viene applicato a un nuovo corpus per stimare la sua capacità di predizione, cioè per valutare se il sistema è in grado di prevedere correttamente il livello di leggibilità dei nuovi testi. La qualità del modello dipende da una serie di fattori coinvolti nel processo: la scelta del set di dati di allenamento, la scelta di un algoritmo di apprendimento efficiente e la selezione delle caratteristiche linguistiche da estrarre dai dati.

Per quanto riguarda i diversi metodi di valutazione automatica della leggibilità, è possibile fare una prima distinzione in base al tipo di approccio utilizzato: la valutazione della leggibilità può essere trattata come un compito di classificazione (assegnazione del

documento a una specifica classe o livello di leggibilità), un compito di ranking (assegnazione del documento a una posizione all'interno di una scala di leggibilità) o come un problema di regressione (i livelli o i punteggi si trovano in un intervallo continuo). La classificazione è l'approccio più utilizzato, ad esempio in studi come quelli di Si e Callan 2001, Liu et al. 2004, Collins-Thomson e Callan 2004, Schwarm e Ostendorf 2005, Heilman et al. 2007, Al-Kalifa e Amani 2010, Aluisio et al. 2010, Chen 2013. Il metodo di ranking è adottato da Inui e Yamamoto 2001, Pitler e Nenkova 2009, Tanaka-Ishii et al. 2010. Il modello di regressione è invece utilizzato da Kate et al. 2010 e François e Fairon 2012.
Un'ulteriore distinzione può essere operata a seconda delle caratteristiche linguistiche considerate.
Gli studi analizzano tutta una serie di funzionalità collegate alla leggibilità, le quali possono essere raggruppate in diverse categorie: caratteristiche lessicali, sintattiche, semantiche e relative alle parti del discorso. Le caratteristiche lessicali e semantiche si riferiscono agli aspetti associati al vocabolario dei testi, come la difficoltà o la familiarità delle parole; vengono utilizzate ad esempio:

- la frequenza relativa delle parole;
- la presenza o assenza in una data lista di parole;
- la ricchezza lessicale (rapporto tipi/repliche);
- la lunghezza delle parole;
- il numero di parole funzionali;
- il numero di pronomi;
- il modello statistico del linguaggio (fornisce la distribuzione delle probabilità delle parole nel testo).

La complessità sintattica, valutata tramite la lunghezza delle frasi, è una delle metriche più usate nelle formule di leggibilità tradizionali; gli studi più recenti considerano un insieme più ampio di parametri per valutare la complessità e sono in grado di analizzare anche le strutture delle frasi più complesse, tramite strumenti specifici, chiamati *parser* (analizzatori). Le principali caratteristiche sintattiche considerate sono:

- lunghezza delle frasi;
- numero di frasi verbali;
- numero di frasi nominali;
- numero di subordinate;
- numero di frasi preposizionali.

Anche le relazioni che esistono tra i vari elementi della frase influiscono sulla leggibilità: una buona organizzazione e la coerenza dei contenuti contribuiscono infatti a rendere un testo più leggibile. Le tradizionali formule di leggibilità non sono in grado di cogliere questi aspetti, che sono invece considerati dai lavori più recenti. Le variabili misurate sono:

- la coesione;
- la coerenza;
- le relazioni tra le parti del discorso.

Per valutare questi aspetti si considerano ad esempio:

- i connettivi;
- la continuità degli argomenti;

- la densità delle idee;
- il numero di pronomi;
- il numero di articoli determinativi;
- la sovrapposizione delle parole.

Nella maggior parte degli studi è impiegata una combinazione delle diverse caratteristiche: Si e Callan (2001) e Collins-Thompson e Callan (2004) utilizzano modelli statistici del linguaggio di tipo *unigram* combinati con altre caratteristiche, di tipo sintattico o semantico. Liu et al. (2004) e Schwarm e Ostendorf (2005) impiegano l'algoritmo SVM per combinare le caratteristiche sintattiche con quelle semantiche. Kate et al. (2010) usano algoritmi di regressione per combinare caratteristiche sintattiche, lessicali e modelli linguistici specifici per generi testuali. François e Fairon (2012) considerano ben 46 parametri linguistici diversi (lessicali, sintattici, semantici, oltre a parametri relativi al francese come L2).
I metodi si differenziano tra loro anche in base al campo di applicazione e ai destinatari. Schwarm e Ostendorf (2005), Heilman et al. (2007) e Peterson e Ostendorf (2009) si occupano di classificare il livello di lettura di testi scritti destinati a studenti di L2. Altri studi si concentrano sulla valutazione del livello di lettura di pagine web, come Si e Callan (2001) e Collins-Thompson e Callan (2004). Wang (2006) misura la leggibilità delle informazioni presenti nei siti web di assistenza sanitaria. Liu et al. (2004) determinano il livello di lettura dei risultati delle query dei motori di ricerca. Miltsakaki e Troutt (2007) progettano un'applicazione per valutare la leggibilità dei testi sul web e classificarli in base al loro contenuto tematico.
In questa sezione presentiamo alcuni dei principali approcci alla valutazione automatica della leggibilità.

6.4.1. Si e Callan 2001

Si e Callan (2011) propongono un metodo per stimare la leggibilità, intesa come difficoltà di lettura, di pagine web educative. La leggibilità viene considerata come un problema di classificazione: i classificatori dei vari livelli di lettura sono creati come combinazioni lineari di un modello statistico del linguaggio di tipo *unigram* e di un modello che renda conto delle caratteristiche linguistiche del testo.
L'ipotesi di partenza è che la misurazione della leggibilità sarebbe più accurata se le formule prendessero in considerazione anche le informazioni sul contenuto dei documenti. Le formule tradizionali considerano infatti soltanto le caratteristiche "di superficie", come la lunghezza delle frasi o delle parole. Questi parametri non sono però adatti a valutare pagine web destinate alla didattica, ad esempio per il fatto che contengono testi molto brevi per cui il livello di difficoltà è spesso sottostimato. Gli autori propongono quindi un approccio che tenga conto sia delle caratteristiche linguistiche sia del contenuto.
Per quanto riguarda la valutazione del contenuto, l'ipotesi è che un modello statistico del linguaggio come quello *unigram* sia in grado di acquisire informazioni sul contenuto relative alla difficoltà di lettura. Come corpus di allenamento viene scelto un campione di 91 pagine web di educazione scientifica, scritte sia da studenti con vari gradi di istruzione ed età, sia da adulti. I testi coprono vari livelli di lettura: materna-2° grado, 3°-5° grado, 6°-8° grado; i livelli di lettura sono indicati dalla fonte o sono dedotti in base all'età degli autori dei testi.

In base alle caratteristiche del corpus, gli autori scelgono la lunghezza dei testi come parametro linguistico da considerare; la lunghezza delle parole (in sillabe) ed altre metriche, come il numero dei monosillabi o delle parole polisillabiche, non risultano in questo caso parametri rilevanti.
La terza ipotesi di Si e Callan è che la distribuzione normale possa essere usata per modellare la distribuzione della lunghezza della frase per ogni livello di leggibilità.
Per combinare i due modelli (modello *unigram* e modello per la lunghezza della frase) viene scelta una combinazione lineare, realizzata tramite l'algoritmo EM[132]; questo tipo di algoritmo è utilizzato spesso nei modelli lineari quando i dati di addestramento contengono vari tipi di informazione.
Gli esperimenti hanno mostrato che questo metodo di misurazione della leggibilità è più accurato rispetto alle formule tradizionali, come ad esempio l'indice Flesch-Kincaid.

6.4.2. Inui e Yamamoto 2001

Lo studio di Inui e Yamamoto (2001) si inserisce nel contesto di una ricerca più ampia sulla semplificazione di testi giapponesi destinati all'assistenza alla lettura; i testi sono rivolti in particolare a studenti non udenti delle scuole secondarie di primo grado che presentano difficoltà nella lettura e nella scrittura. Le persone non udenti tendono infatti ad avere difficoltà nel comprendere frasi passive, causali, relative, frasi scisse, ecc. Scopo della ricerca è sviluppare un sistema di semplificazione testuale che sia in grado di trasformare in modo automatico un dato documento in uno più semplice e comprensibile, tramite parafrasi di tipo lessicale e sintattico.
Per sviluppare questo sistema è innanzitutto necessario costruire un modello che si occupi di classificare un determinato insieme di parafrasi in base al loro livello di leggibilità.
Per quanto riguarda il giapponese, esistono diversi studi che si sono occupati della misurazione della leggibilità, soprattutto in ambito tecnico o in contesti ingegneristici; "the readability criteria proposed in those works are, however, based mainly on simple statistics such as sentence length, depth of embedding, and the Kanji/Kana ratio, analogous to Flesch's readability measurement (Flesch, 1948); they are far less sophisticated than the criteria we present in this paper. Furthermore, none of those works took into account the language prociency of a particular population segment such as deaf people, aphasic people, or second-language learners" (Inui e Yamamoto 2001, p. 2).
Per la costruzione del modello, gli autori si servono di un questionario, sottoposto a 240 insegnanti di giapponese e inglese delle scuole per non udenti. I questionari, composti da 510 domande, hanno lo scopo di raccogliere i dati per la valutazione della leggibilità: ai docenti è chiesto di confrontare una data frase con altre possibili parafrasi per quella frase e di valutare la leggibilità di ognuna di queste. Per realizzare il questionario, gli autori selezionano 50 aspetti morfosintattici che possono influenzare la comprensione di una frase per le persone non udenti. Per ognuno di questi parametri raccolgono alcune frasi di esempio e poi costruiscono un set di parafrasi per ciascuna, togliendo ovviamente dalle frasi alternative l'elemento che procura difficoltà (vedi Figura 19). Per minimizzare

[132] L'algoritmo EM (*expectation-maximization*) viene impiegato per la stima di massima verosimiglianza dei parametri di un modello probabilistico. Per informazioni si veda Dempster 1977.

l'incidenza del lessico sulla difficoltà, gli studiosi limitano il vocabolario delle frasi a un set di 2000 parole di base (NIJL 1991)[133].

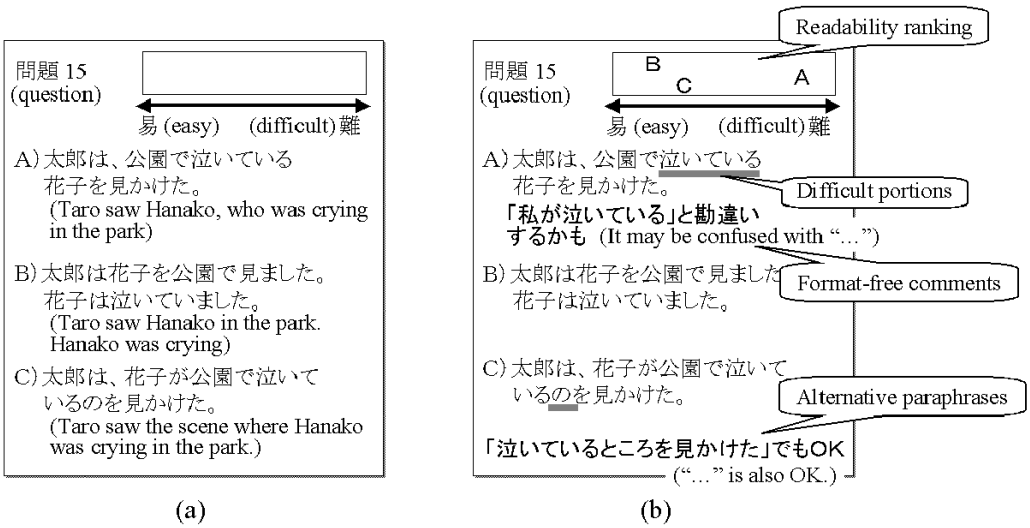


Figura 19. Esempio di questionario. L'elemento di difficoltà è in questo caso la frase relativa (A), che viene eliminata sia in (B) che in (C). L'immagine è tratta da Inui e Yamamoto 2001.

I risultati del questionario servono come dati di apprendimento per la costruzione di un modello che classifica le parafrasi in base al loro livello di leggibilità. Per semplificare, il compito di classificazione può essere scomposto in una serie di confronti tra due elementi selezionati dal set: il classificatore deve valutare quale dei due elementi è più leggibile o se le frasi hanno la stessa leggibilità. Per l'apprendimento del classificatore è possibile usare diverse tecniche di modellazione automatica; gli autori ne hanno sperimentati due: un metodo che si basa su regole di classificazione e un metodo basato sulla classificazione SVM.

Usando i due metodi, gli studiosi hanno condotto una convalida incrociata dei dati raccolti e ottenuto dei risultati promettenti: entrambi i modelli hanno infatti ottenuto una precisione superiore all'88%; in particolare, il modello basato su SVM ha ottenuto una precisione del 95%.

6.4.3. Liu et al. 2004

Il lavoro di Liu et al. (2004) si concentra sul riconoscimento automatico dei livelli di lettura degli utenti in base alle interrogazioni nei motori di ricerca.

I tradizionali indici di leggibilità sono stati sviluppati per valutare generalmente brani o porzioni di testo di almeno 100 parole o 10 frasi[134] e divengono inaffidabili nel caso di testi più brevi, come invece è tipico delle domande poste nei vari motivi di ricerca. In questo studio, la leggibilità è trattata come un problema di classificazione: tramite algoritmi di apprendimento automatico, le ricerche degli utenti nei motori di ricerca sono classificate in

[133] The National Institute for Japanese Language (NIJL), *Nihongo Kyôiku-no tame-no Kihon-Goi Chôsa* (*The basic lexicon for the education of Japanese*), Shuei Shuppan, Giappone, 1991.

[134] L'unica eccezione è la formula creata da Fry 1990 (cfr. 3.13). Sviluppata appositamente per valutare testi scritti brevi (tra le 40 e le 100 parole), può essere applicata anche alla misurazione della leggibilità di pagine web.

base alle proprie caratteristiche sintattiche (lunghezza della frase, lunghezza delle parole, ecc.), in modo da determinare, per ciascuna domanda, il livello di lettura.
Il corpus di apprendimento è composto da 3 sotto corpora: il primo è un set di domande raccolte da una scuola elementare nel giugno del 2003 e il cui livello di lettura è il 6° grado; si tratta di domande sollevate dagli studenti su vari argomenti discussi in classe nelle lezioni di scienze. Il secondo set è composto da un campione raccolto casualmente tra le interrogazioni inviate al motore di ricerca Excite il 20 dicembre 1999; il terzo set è composto dalle domande archiviate tra il 1996 e il 2002 dal servizio di consulenza di Mad Scientist[135]. Le statistiche del corpus sono sintetizzate nella Tabella 31:

Grade Level	N. di query	N. medio di parole per query	N. medio di caratteri per query	N. medio di sillabe per query
6	407	6,93	4,22	1,36
7-9	2.508	9,13	4,64	1,51
10-12	3.374	9,20	4,78	1,57
Undergraduate	2.414	9,23	4,87	1,60
Graduate	1.669	9,24	4,85	1,59
Excite	1.999	3,35	5,86	1,83

Tabella 31. Statistiche del corpus per livelli di istruzione.

Per costruire il modello di apprendimento, gli studiosi usano il software LIBSVM (Chang e Lin 2001), che implementa l'algoritmo SMO per il *Support Vectore Machine* (SVM)[136]: il classificatore impara a classificare le query nei vari livelli di istruzione in base alle diverse caratteristiche linguistiche che derivano dallo studio del training corpus. Le caratteristiche studiate sono sia di tipo sintattico (lunghezza della frase, lunghezza delle parole in sillabe, lunghezza delle parole in caratteri, ecc.) che semantico (frequenza di sequenze di *unigrammi*, *digrammi* e *trigrammi*); vengono considerati anche i livelli di leggibilità, misurati tramite l'indice di Flesch-Kincaid, l'indice SMOG e l'indice Fog.
I risultati mostrano che l'approccio basato sul *Support Vectore Machine* offre una maggiore accuratezza rispetto alle classifiche formule di leggibilità (Tabella 32) e può raggiungere una precisione anche superiore all'80% nel riconoscere il livello di lettura.

Categorie	Flesch-Kincaid	SMOG	FOG	SVM
6 + 7-9	40,0000	13,7931	14,1379	93,4483
6 + 10-12	24,9337	10,6101	10,6101	95,7560

[135] MadSci Network è un sito che offre un servizio di risposte a domande poste dagli utenti su tematiche riguardanti principalmente materie scientifiche: http://www.madsci.org/
[136] LIBSVM (Chang e Lin 2001) è un pacchetto software, sviluppato presso la National Taiwan University, utilizzato per costruire modelli di classificazione e regressione basati su *Support Vector Machines*. Il software consente di selezionare i parametri, effettuare una classificazione multiclasse e costruire quindi il modello. Offre anche un modulo avanzato per la convalida incrociata. LIBSVM implementa l'algoritmo *Sequential Minimal Optimization* (SMO).

Categorie	Flesch-Kincaid	SMOG	FOG	SVM
6 + undergrad.	19,5730	14,2349	14,2349	86,1210
6+ graduate	20,8739	19,4175	19,4175	92,7184
10-12 + graduate	11,5308	0	0	66,6004

Tabella 32. Confronto dei valori di accuratezza (%) tra gli indici di leggibilità e l'approccio SVM.

Per il lavoro futuro, gli studiosi propongono un possibile implemento del loro metodo in modo che possa essere incorporato direttamente nei sistemi dei motori di ricerca. Gli autori hanno infatti classificato i vari livelli di lettura partendo dalle interrogazioni degli utenti, ma il sistema potrebbe essere applicato al modello di recupero delle informazioni per fare in modo che i risultati delle query possano essere personalizzati in basi al livello di istruzione dell'utente.

6.4.4. Collins-Thompson e Callan 2004

Collins-Thompson e Callan (2004, 2005) presentano un approccio alla stima automatica del livello di lettura di pagine web che si basa sul modello statistico del linguaggio.
La maggior parte delle formule classifiche si è concentrata su due fattori linguistici, considerandoli come maggiormente predittivi della difficoltà testuale: la variabile sintattica e la variabile semantica. La difficoltà sintattica è spesso misurata in funzione della lunghezza della frase; la variabile semantica, intesa come misura della difficoltà del vocabolario, è spesso valutata tramite il confronto con liste di parole. L'ipotesi degli autori è che ognuna di queste liste possa essere concepita come un modello di linguaggio semplificato.
Questo tipo di approccio presenta diversi vantaggi rispetto alle tecniche tradizionali, che generalmente sono costruite su due o tre parametri: i modelli statistici del linguaggio sono infatti in grado di individuare modelli più dettagliati dell'uso delle singole parole; la precisione aumenta quando si valutano testi molto brevi o pagine web. I modelli statistici consentono inoltre di ottenere una distribuzione della probabilità tra tutti i modelli dei livelli di lettura e non solo una previsione sui singoli livelli.
Il corpus di addestramento è composto da 500 pagine web in inglese, etichettate in base al livello di difficoltà di lettura; i livelli di lettura, assegnati dagli autori stessi dei siti, corrispondono ai 12 livelli di istruzione americani. I documenti sono tratti da diverse aree, come narrativa, saggistica, storia, scienze, ecc.
Prima di definire il modello per la classificazione, gli autori esaminano le frequenze delle parole nel corpus, per vedere se esistono delle linee di tendenza. Com'è prevedibile, le parole più difficili sono introdotte nei testi dei livelli di istruzione più avanzati. Le Figure 20-22 mostrano l'andamento della frequenza di alcune parole tratte dal corpus: come si può osservare, parole concrete come *red* tendono a divenire meno frequenti nei livelli di istruzione più alti, mentre parole come *determine* a divenire più probabili. Altri termini, come *perimeter*, aumentano la loro frequenza in uno specifico range di grado, probabilmente in corrispondenza del periodo in cui sono studiati a scuola.
"Our main hypothesis in this work is that there are enough distinctive changes in word usage patterns between grade levels to give accurate predictions with simple language

models, even when the subject domain of the documents is unrestricted" (Collins-Thompson e Callan 2004).

Figura 20. Probabilità della parola *red*.

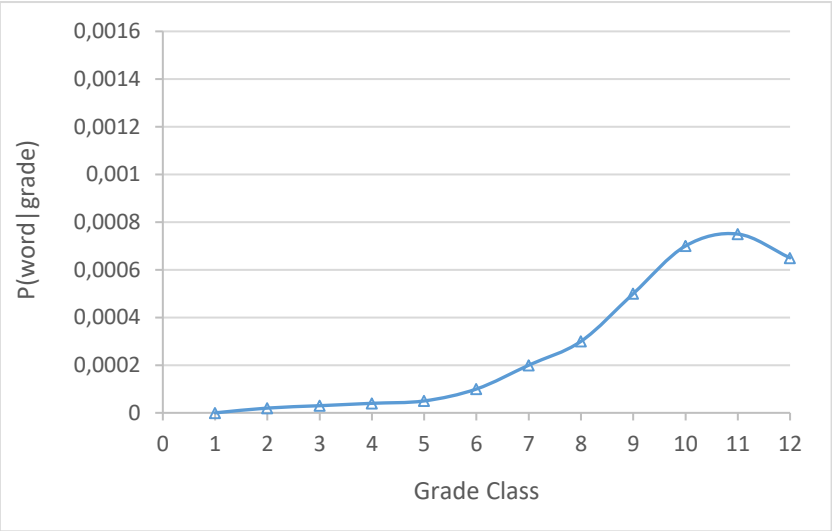


Figura 21. Probabilità della parola *determine*.

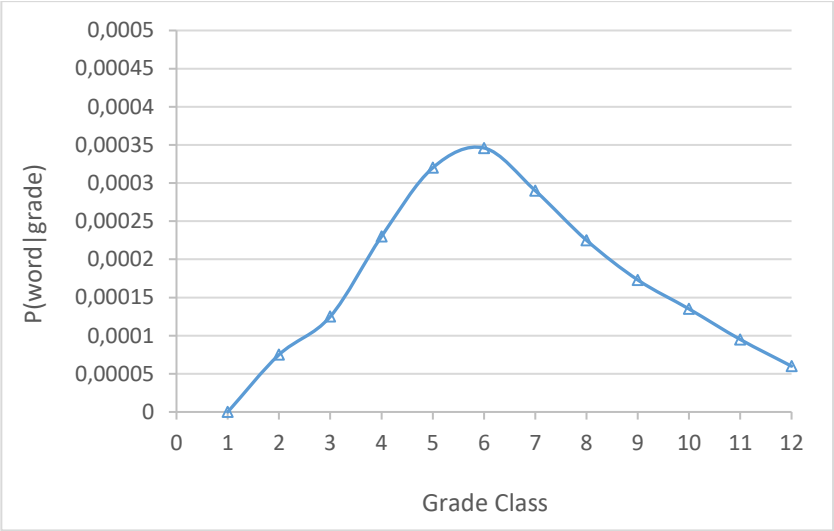


Figura 22. Probabilità della parola *perimeter*.

Il modello proposto da Collins-Thompson e Callan può essere considerato come una generalizzazione dell'approccio basato sul vocabolario, in cui modelli di linguaggio multipli (uno per ciascun livello di istruzione) raccolgono informazioni più dettagliate sull'uso del linguaggio. A differenza dello studio preliminare di Si e Callan (2004), in questo caso la componente sintattica è considerata poco predittiva e non viene analizzata.
Il modello, chiamato dagli autori *Smoothed Unigram*, si basa su un'estensione del metodo multinomiale Naïve Bayes, che consente di combinare vari modelli statistici del linguaggio per stimare il livello di lettura (cioè la classe) più plausibile per un dato testo. I modelli sono di tipo *unigram* e presuppongono che la probabilità di un *token* sia indipendente dai *token* circostanti, data una specifica classe di modello di linguaggio. Ogni modello è definito da una lista di parole (*types*) e dalla loro probabilità individuale. Nonostante si tratti di un modello debole, il vantaggio è che può essere addestrato anche con una piccola quantità di dati (e con dati relativamente poco classificati) e risulta avere una buona precisione.
Per verificarne la flessibilità, gli autori testano il modello anche su un corpus di 189 pagine web in francese, etichettate in 5 livelli di difficoltà. Lo studio preliminare dimostra che, con cambiamenti minimi, il classificatore può essere rieducato anche alla valutazione di documenti in francese. Per entrambe le lingue, il classificatore mantiene una buona correlazione con il livello di istruzione etichettato (tra 0,63 e 0,79) per tutti i set di testi.
Alcune variabili semantiche tradizionali, come il rapporto *type-token*, ottengono un'alta correlazione per quanto riguarda i brani etichettati come commerciali, ma le stesse statistiche non risultano coerenti per quanto riguarda le pagine web. Il modello *Smoothed Unigram* risulta comunque avere una maggiore accuratezza rispetto alle tecniche tradizionali su testi molto brevi (meno di 10 parole) e pagine web.

6.4.5. Schwarm e Ostendorf 2005

Schwarm e Ostendorf (2005) propongono un metodo per valutare in modo automatico il livello di lettura di testi scritti destinati a studenti che studiano l'inglese come lingua seconda. Il loro approccio prevede l'uso di algoritmi SVM per combinare le caratteristiche di modelli linguistici tradizionali e modelli statistici di tipo *n-gram*. Lo studio si inserisce in un

più ampio progetto di ricerca che si occupa di sviluppare strumenti di supporto per insegnanti di lingue, ad esempio sistemi di semplificazione automatica di testi; questi strumenti possono essere utili non soltanto per studenti stranieri ma anche per tutti coloro che presentano scarse abilità di lettura o difficoltà di apprendimento.
Nonostante il modello *Smoothed Unigram* di Collins-Thompson e Callan (2004) abbia ottenuto dei buoni risultati e risulti più accurato delle metriche tradizionali di leggibilità, Schwarm e Ostendorf sostengono che i modelli statistici del linguaggio possano ottenere prestazioni migliori acquisendo sia informazioni di tipo semantico che sintattico.
Il corpus di addestramento è formato da circa 2400 articoli, tratti da *Weekly Reader*, una rivista educativa in cui è possibile trovare testi rivolti ai vari livelli di istruzione, e copre i livelli di lettura dal secondo al quinto grado (Tabella 33). Come supplemento al corpus, sono impiegati anche altri set di testi, in modo da avere un numero più ampio di classi di lettura: il corpus dell'Enciclopedia Britannica[137], che contiene sia i testi della versione completa dell'enciclopedia sia i testi corrispondenti della Britannica Elementary, la versione rivolta ai bambini e un archivio di notizie della CNN, sia in forma estesa che in versione ridotta[138] (Tabella 34).

Livello	**Numero di articoli**	**Numero di parole**
2°	351	71,5 k
3°	589	444 k
4°	766	927 k
5°	691	1 M

Tabella 33. Statistiche del corpus *Weekly Reader.*

Corpus	**Numero di articoli**	**Numero di parole**
Britannica	115	277 k
Britannica Elementary	115	74 k
CNN	111	51 k
CNN ridotta	111	37 k

Tabella 34. Statistiche del corpus dell'Enciclopedia Britannica + CNN.

Gli autori costruiscono un classificatore per ogni livello di lettura. Il classificatore non si occupa di categorizzare i documenti nelle varie classi (cioè nei vari livelli di lettura) ma si occupa invece, per ciascuna classe, di decidere se un documento appartiene o meno a quella classe.
Il modello statistico impiegato è il *trigramma*, che risulta più accurato rispetto all'*unigramma* e al digramma. L'algoritmo SVM utilizza diverse caratteristiche linguistiche, come la lunghezza della frase, la lunghezza delle parole, la leggibilità misurata con l'indice di

[137] Barzilay e Elhadad (2003).
[138] Tratte dal sito della Western/Pacific Literacy Network: http://literacynet.org/cnnsf/

Flesch-Kincaid e il sistema Lexile, ecc. La combinazione del modello statistico con il *Support Vector Machine* fornisce i risultati migliori.
Il classificatore è testato con successo su un ulteriore corpus di articoli tratti dall'edizione "Kidspost" del *Washington Post* del 2005, che coprono i livelli dal terzo all'ottavo; gli autori si ripropongono di sperimentare il loro metodo anche su lingue diverse dall'inglese e di riuscire ad incorporarlo in un sistema di recupero delle informazioni sul web, come strumento di supporto per gli insegnanti di lingua straniera.

6.4.6. Larsson 2006

Larsson (2006) crea un modello per classificare testi svedesi in livelli di leggibilità. Il sistema può servire come strumento di supporto per gli insegnanti o può essere integrato in un sistema di recupero delle informazioni; è creato per lo svedese ma può essere applicato anche ad altre lingue.
Dal momento che per la lingua svedese non è disponibile un corpus annotato in diversi livelli di lettura, il set di addestramento utilizzato nella ricerca è assemblato dall'autore e comprende tre diversi corpora (e quindi tre livelli di leggibilità):

- Giornali del mattino
  Il set è costituito da articoli tratti due giornali svedesi, Uppsala Nya Tidning (UNT) e Svenska Dagbladet (SvD) e utilizzati nel progetto SCARRIE[139]. I testi sono scritti da professionisti e sono rivolti ad un pubblico di lettori adulti: il livello può essere considerato come *difficile*.
- Testi delle scuole superiori
  Il corpus comprende 418 testi scritti da studenti di 16-18 anni nel corso degli esami. Esistono ovviamente delle differenze, ma in generale i testi possono essere classificati nello stesso livello di difficoltà, il livello *medio*.
- Giornali di facile lettura
  Il set è costituito da 787 testi tratti dal sito di *Sesam*, un giornale di facile lettura. Gli articoli sono scritti da professionisti ma sono indirizzati specificamente a persone che presentano difficoltà nella lettura: il livello può essere quindi considerato *semplice*.

Il corpus è ridimensionato in modo che ogni sottocorpora abbia lo stesso numero di testi (418).
Il sistema progettato da Larsson utilizza LIBSVM, il software integrato per la classificazione tramite SVM; il software comprende varie funzioni, tra cui la selezione dei parametri, la formazione del modello, la classificazione multiclasse e un modulo per la convalida incrociata.
Per la costruzione del modello si considerano diverse caratteristiche linguistiche: frequenza delle parole (*unigram*), lunghezza della frase, profondità sintattica (proporzione di frasi complesse), numero di frasi preposizionali, numero di congiunzioni subordinanti, numero di parole difficili (parole con più di 6 lettere), numero medio di vocali per frase (= numero di sillabe per parola), quoziente nominale (NQ)[140], quoziente nomi/pronomi, numero di

[139] Cfr. Dahlqvist 1999.
[140] Inteso come numero di nomi, preposizioni e participi diviso il numero di pronomi, verbi e avverbi. Misura la quantità di informazioni che ci sono in un testo. Il valore normale è 1,0.

attributi per frase nominale, lunghezza dell'espressione (*phrase*)[141], numero di articoli determinativi (misura l'astrattezza della frase). La Tabella 35 mostra i risultati della classificazione per ciascuna caratteristica.

Caratteristiche	Totale	Facile	Medio	Difficile
profondità sintattica	61,04	81,92	32,50	68,67
lunghezza della frase	58,02	**93,97**	2,40	79,49
frasi preposizionali	63,45	86,75	28,92	74,70
cong. subordinanti	52,61	62,65	59,03	36,14
parole difficili	62,25	81,92	28,91	75,90
vocali	62,25	43,37	72,29	71,08
NQ	**69,73**	55,41	69,67	**84,12**
nomi/pronomi	69,44	60,24	**74,70**	73,39
attributi	61,04	43,37	72,28	67,47
lunghezza espressione	59,04	46,99	54,21	75,90
articoli determinativi	49,40	73,49	53,01	21,69

Tabella 35. Punteggi di copertura (*recall*) per ciascuna caratteristica.

Come si osserva, la caratteristica migliore risulta il quoziente nominale (NQ). In generale, esistono molte differenze tra i tre livelli di leggibilità, ad esempio la lunghezza della frase ha un punteggio di recupero di circa 94% nel classificare il livello *facile* ma solo del 2,4% nel classificare il livello *medio*. Questi punteggi sono utilizzati per la combinazione delle caratteristiche più rilevanti; i risultati sono mostrati nella Tabella 36.

Combinazione	Caratteristiche combinate	Recall
Tutte le caratteristiche	Tutte	**88,26**
10 migliori caratteristiche	Tutte tranne articoli determinativi	87,86
9 migliori caratteristiche	Tutte tranne articoli determinativi e cong. subordinanti	86,96
7 migliori caratteristiche	NQ, Nomi/pronomi, frasi preposizionali, vocali, parole difficili, profondità sintattica, attributi.	86,87
8 migliori caratteristiche	Tutte tranne articoli determinativi, cong. subordinanti e lunghezza della frase	86,57
6 migliori caratteristiche	NQ, Nomi/pronomi, frasi preposizionali, vocali, parole difficili, profondità sintattica	85,37
2 migliori / liv. leggibilità	NQ, lunghezza della frase, frasi preposizionali, vocali, attributi	85,37
Migliori / liv. leggibilità	NQ, lunghezza della frase, Nomi/pronomi	83,68
3 migliori caratteristiche	NQ, Nomi/pronomi, frasi preposizionali	82,69
2 migliori caratteristiche	NQ, Nomi/pronomi	69,85

Tabella 36. I risultati della convalida incrociata di 40 volte delle varie caratteristiche.

[141] Inteso come numero medio dei costituenti per *espressione*.

La combinazione di tutte le caratteristiche risulta avere il valore di copertura più alto (88,26); con la sola eccezione della combinazione di 8 caratteristiche, aumentando il numero di caratteristiche, aumenta anche il punteggio di recupero. Viene dunque effettuata una nuova convalida incrociata di 40 volte, stavolta con le possibili combinazioni di 10 caratteristiche. I risultati sono mostrati nella Tabella 37.

Caratteristiche combinate	**Recall**
Tutte tranne lungh. espressione	**88,76**
Tutte tranne NQ	88,06
Tutte tranne frasi prep.	88,06
Tutte tranne articoli det.	87,86
Tutte tranne parole difficili	87,86
Tutte tranne nomi/pronomi	87,86
Tutte tranne profondità int.	87,56
Tutte tranne vocali	87,36
Tutte tranne lunghezza frase	87,26
Tutte tranne attributi	86,57
Tutte tranne cong. Sub.	86,47

Tabella 37. I risultati della convalida incrociata di 10 caratteristiche.

In base ai dati raccolti, vengono selezionati 4 diversi modelli:

- *notPhrase-model*: modello basato sulla combinazione di tutte le caratteristiche tranne la lunghezza dell'espressione (valore di *recall* 88,76);
- *all-model*: modello basato sulla combinazione di tutte le caratteristiche (88,26);
- *notPP-model*: modello basato sulla combinazione di tutte le caratteristiche tranne le frasi preposizionali (88,06);
- *notNQ*-model: modello basato sulla combinazione di tutte le caratteristiche tranne NQ (88,06).

I 4 modelli sono quindi usati per la classificazione del corpus (Tabella 38).

Modello	**Precisione**	**Copertura**	**Punteggio F**
notPP-model	**90.21**	**89.56**	**89.88**
notPhrase-model	88.93	88.35	88.64
All-model	88.90	88.35	88.62
notNQ-model	88.55	87.95	88.25

Tabella 38. Risultati della classificazione tramite i 4 modelli.

Larsson riporta 3 tipologie di punteggio e ne fornisce una spiegazione (p. 22): *il valore di precisione*, definito come "the fraction of documents that actually turns out to be correct in the group of documents that the model has declared as a class", mostra l'affidabilità del

modello nella classificazione; *il valore di copertura*, definito come "the fraction of documents correctly predicted by the model compared to what actually should be detected", misura la quantità di documenti rilevati; la combinazione del punteggio di precisione e copertura è una misura chiamata *punteggio F* ed è così calcolata:

$$2 \times Precision \times Recall/(Recall + Precision)$$

Il modello che risulta avere le migliori prestazioni è quello *not-PP*, cioè il modello che considera tutte le caratteristiche tranne le frasi preposizionali.

6.4.7. Wang 2006

Wang (2006) si è occupato di valutare la difficoltà delle informazioni presenti nei siti web che si occupano di salute[142]. L'algoritmo SVM è usato per classificare i documenti in due livelli di difficoltà principali: testi di facile lettura, destinati a un pubblico con un basso livello di competenze sanitarie (4°-6° grado) e testi per un pubblico generico, destinati a un pubblico con un livello medio di alfabetizzazione sanitaria (6°-8° grado). A differenza dello studio di Liu et al. (2004), che stima la leggibilità a livello della frase, in questo caso l'approccio SVM è impiegato a livello del documento. L'accuratezza della classificazione è confrontata tramite diversi set di caratteristiche:

- *caratteristiche linguistiche di superficie*
  Comprendono le metriche usate solitamente nelle formule tradizionali: numero medio di parole per frase, numero medio di caratteri per parola e numero medio di sillabe per parola.
- *difficoltà delle parole*
  Dato che l'autore non è stato in grado di trovare un indicatore affidabile per quanto riguarda la difficoltà delle parole appartenenti al dominio medico, vengono usati parametri più generali. Si considerano *facili* le parole che appartengono alla lista di Dale e Chall (1995), *difficili* le parole polisillabiche, con 3 o più sillabe.
- *modello statistico del linguaggio di tipo unigram.*

Per verificare se le performance dei set di caratteristiche sono coerenti con vari metodi di apprendimento automatico, oltre all'approccio SVM sono utilizzati anche gli alberi decisionali e il metodo Naïve Bayes[143]. Gli algoritmi sono forniti tramite il software open source Weka[144].

---

[142] Per i diversi studi che hanno esaminato il livello di leggibilità di siti web che trattano di salute cfr. capitolo 8.2.

[143] Zheng et al. 2002 si è occupato di classificare articoli di notizie mediche tramite l'applicazione di due metodi di apprendimento automatico: gli alberi decisionali e il metodo Naïve Bayes. Il classificatore è in grado di classificare gli articoli, distinguendo tra materiale medico e non, con un'accuratezza del 92%.

[144] Weka (Waikato Environment for Knowledge Analysis) è un software open source sviluppato dall'Università di Waikato in Nuova Zelanda nel 1993. Si tratta di una raccolta di algoritmi di apprendimento automatico per attività di data mining. Gli algoritmi possono essere applicati direttamente a set di dati: Weka contiene strumenti per la pre-elaborazione dei dati, la classificazione, la regressione, il clustering, le regole di associazione e la visualizzazione. È anche adatto per lo sviluppo di nuovi schemi di apprendimento automatico.
Il sito di riferimento è: https://www.cs.waikato.ac.nz/ml/weka/

Il set di addestramento è formato da documenti tratti da alcuni siti web di assistenza sanitaria, etichettati in base ai due livelli di lettura: i testi di facile lettura sono raccolti dal sito di MedlinePlus[145], quelli generali dal sito Familydoctor[146].
L'indice di Flesch (F) e l'indice di Flesch-Kincaid (FK), misurati tramite Microsoft Word, sono utilizzati per filtrare i documenti[147]: testi che presentano punteggi FK superiori all'8° grado e punteggi F inferiori a 60 sono esclusi dai materiali di facile lettura in quanto considerati troppo difficili. testi con valori FK oltre il 10° grado e valori F inferiori a 50 sono esclusi dai materiali per un pubblico generico. In totale, risulta un corpus formato da 79 articoli di facile lettura e 95 articoli di livello di lettura medio.

Categorie	**Flesch Reading Ease**	**Flesch-Kincaid**
Facile lettura (79)	65,86 (51,5 – 86,1)	6,66 (3,7 – 8,3)
Livello medio (95)	62,25 (46,4 – 91,4)	7,87 (4,8 – 9,9)

Tabella 39. Punteggi di leggibilità dei due corpora.

I set di caratteristiche sono confrontati con una validazione incrociata di 10 volte. L'accuratezza della classificazione dei tre set e delle loro combinazioni è mostrata nella Tabella 40:

Set di caratteristiche	**Alberi decisionali**	**Naïve Bayes**	**SVM**
(1) Caratter. superficiali	66,81	66,34	62,72
(2) Difficoltà delle parole	67,18	66,68	64,67
(1) + (2)	73,41	75,55	76,82
(3) Modello Unigram	78,68	75,26	80,71
(1) + (2) + (3)	**79,72**	**76,18**	**84,06**

Tabella 40. Valori di accuratezza (%) della classificazione dei tre set.

Come si osserva, il metodo SVM che utilizza solo caratteristiche superficiali raggiunge una precisione del 62,72% e non sembra essere un buon indicatore di difficoltà per questo corpus. Se si considera la difficoltà della parola, l'accuratezza aumenta al 64,7%; una combinazione delle due migliora le prestazioni fino al 76,82% di accuratezza. Le caratteristiche *unigram* raggiungono una precisione dell'80,71%. La combinazione di tutte e tre i set di caratteristiche sembra essere l'opzione più efficace, con un'accuratezza dell'84,06%.
Per quanto riguarda i tre approcci di *machine learning*, i risultati mostrano che le prestazioni del metodo SVM sono inferiori quando si utilizzano i primi due set

[145] https://medlineplus.gov/
[146] https://familydoctor.org/
[147] Secondo l'indice di Flesch un testo è considerato *facile* se ottiene un punteggio superiore a 70, *standard* se ottiene un punteggio tra 60 e 70, *difficile* se il punteggio è inferiore a 60. La formula modificata Flesch-Kincaid fornice un punteggio in termini di livello di istruzione, con un range che va da 0 a 12 (livello universitario).

separatamente, ma superiori se sono usati in combinazione; sono invece migliori per quanto riguarda il modello *unigram* o una combinazione dei tre set.
"The combination of three feature sets are the most effective in classifying consumer health information into easy to read or general reading difficulty level in our corpus. Since three feature sets may capture different aspects of text difficulty, it is not surprising that their combination achieve the best performance." (Wang 2006).

6.4.8. Heilman et al. 2007

Heilman, Collins-Thompson, Callan e Eskenazi hanno sviluppato il sistema di tutoraggio REAP[148] che fornisce agli studenti di inglese L2 materiali di lettura appropriati al loro livello di lettura; i testi sono selezionati automaticamente dal web. Per migliorare il loro sistema gli studiosi si sono concentrati su un metodo di valutazione automatica della leggibilità che tiene conto sia delle caratteristiche lessicali che di quelle sintattiche.
Il loro studio (Heilman et al. 2007) si occupa di valutare il livello di lettura sia di testi in lingua madre (L1) sia in lingua seconda (L2). Per quanto riguarda i materiali in L1, la classificazione in 12 livelli (secondo il sistema di istruzione statunitense) avviene tramite l'uso di modelli statistici del linguaggio di tipo *unigram*, in combinazione con l'analisi delle caratteristiche grammaticali. Il modello statistico utilizzato si basa una variazione del classificatore multinomiale Naïve Bayes.
Per i testi in L2, la difficoltà lessicale è stimata tramite il modello *unigram*; la struttura sintattica delle frasi è analizzata tramite la combinazione di modelli statistici e i più tradizionali alberi sintattici. Le caratteristiche grammaticali prese in considerazione sono 12, tra cui l'uso del passivo, le frasi relative e alcuni tempi verbali e sono classificate tramite l'algoritmo k-NN (con k=12).
L'approccio utilizza due corpora di dati già etichettati: per il corpus L1 si utilizzano 362 testi raccolti dal web, classificati in 12 livelli da insegnanti delle scuole elementari (cfr. Collins-Thompson e Callan 2005); per il corpus L2 si utilizzano documenti tratti da 4 libri di testo che coprono i livelli da 2 a 5 (da principiante ad avanzato).
I risultati mostrano che, per entrambi i corpora, l'approccio che impiega modelli statistici del linguaggio produce previsioni più accurate rispetto a quello che si basa sulle sole caratteristiche grammaticali (Tabella 41). La combinazione dei due metodi porta ad una precisione maggiore.
Si può inoltre osservare che le caratteristiche grammaticali sembrano influire maggiormente sulla difficoltà nei testi in lingua seconda rispetto a quelli in lingua madre.

Metodi	**L1 (12)**	**L2 (4)**
Modelli statistici del linguaggio	0.71	0,80
Caratteristiche grammaticali	0.46	0.55
Combinazione dei due	0,72	0.83

Tabella 41. Coefficienti di correlazione tra i metodi di valutazione e i due corpora.

[148] Heilman et al. 2006

In uno studio successivo (Heilman et al. 2008) gli autori riportano dei diversi risultati, mostrando che anche le sole caratteristiche grammaticali possono essere efficaci predittori della difficoltà; va precisato che in questa ricerca è impiegato un set più ampio di funzioni grammaticali.

6.4.9. Miltsakaki e Troutt 2007

Miltsakaki e Troutt (2007) svulippano un'applicazione di ricerca che individua e classifica i testi sul web. Il loro strumento, chiamato Read-X, è destinato a studenti e adulti con un basso livello di lettura e consente di filtrare i risultati delle interrogazioni nei motori di ricerca a seconda del livello di leggibilità definito dall'utente.
Le attività svolte da Read-X sono le seguenti:

- *Ricerca sul web*
  In base alle richieste degli utenti, Read-X esegue una ricerca sul web tramite Yahoo! Web Services e recupera tutti i documenti correlati a una delle parole chiave fornite dall'utente.
- *Estrazione del testo*
  Read-X estrae il testo dalle pagine web recuperate, eliminando il codice html.
- *Analisi* della leggibilità
  Read-X analizza la leggibilità di ciascun testo tramite tre indici di leggibilità (Lix, Rix e la formula di Coleman e Liau[149]); i livelli di classificazione sono *molto facile*, *facile*, *standard*, *difficile* o *molto difficile*. Le statistiche considerate sono: numero di frasi, numero di parole, numero di parole lunghe (con sette o più caratteri) e numero di lettere.
- *Classificazione tematica*
  Read-X classifica i risultati in base al loro contenuto tematico, in modo da aiutare l'utente nella disambiguazione. Per la classificazione automatica viene impiegato il classificatore Mallet (Machine Learning for Language Toolkit), un metodo di apprendimento automatico sviluppato presso il Dipartimento di Scienze informatiche dell'Università del Massachusetts. Il corpus di addestramento comprende 3 milioni di testi etichettati in 8 aree tematiche (arte, carriera e affari, letteratura, filosofia e religione, scienza, studi sociali, sport e salute, tecnologia).
- *Visualizzazione dei risultati*
  Read-X restituisce i risultati della leggibilità della classificazione tematica, è inoltre possibile visualizzare, modificare e salvare i vari testi recuperati.

[149] La formula Lix (Björnsson 1968) considera due variabili, la lunghezza della frase (in parole) e la lunghezza delle parole (% parole con più di 6 lettere). I punteggi vanno da 20 (molto facile) a 60 (molto difficile).
La formula Rix (Anderson (1983) è una variante della formula Lix e considera le stesse due variabili. La differenza tra le due formule è che Lix somma i due parametri (lunghezza frase + lunghezza delle parole), Rix ne calcola il quoziente (numero di parole lunghe/numero di frasi). La formula di Coleman e Liau (1975), simile alle precedenti, considera come variabili il numero di lettere su 100 parole e il numero di frasi su 100 parole. Dal momento che la formula fornisce come risultato le percentuali di completamenti cloze corretti, è necessaria una conversione in livelli scolastici (cfr. 3.12).

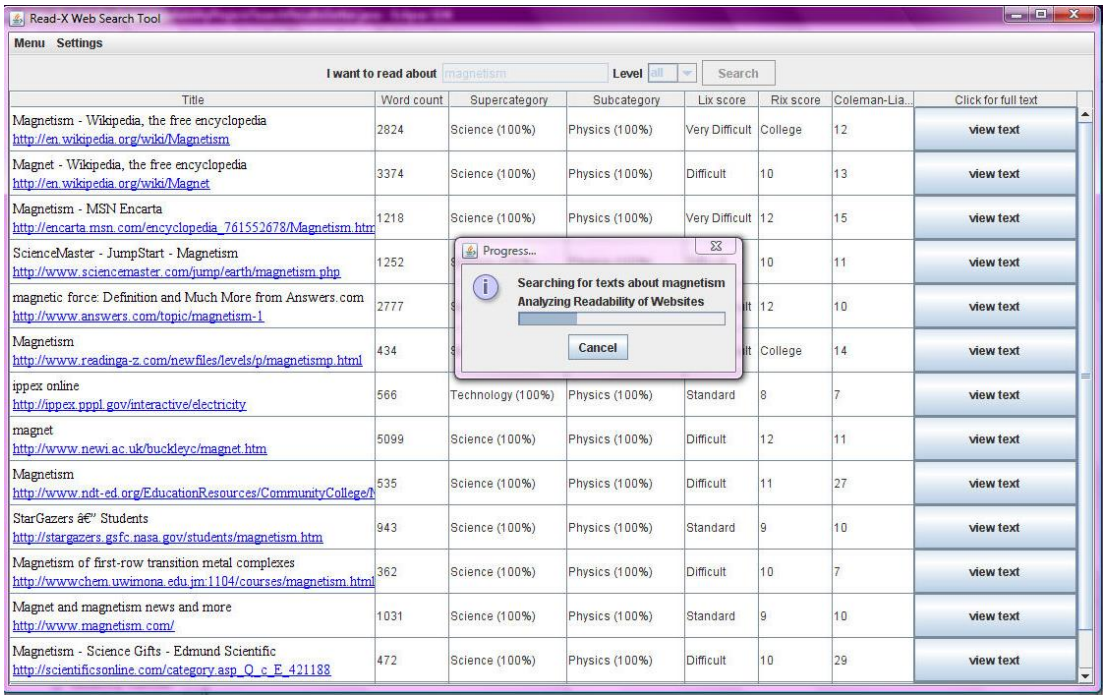


Figura 23. La visualizzazione dei risultati con Read-X. In questo caso la parola chiave cercata è *magnetism.* L'immagine è presa da Miltsakaki e Troutt (2007).

In uno studio successivo (Miltsakaki e Troutt 2008), gli autori presentano un secondo strumento, Toreador, che analizza la difficoltà del vocabolario in rapporto alla familiarità del lettore con il contenuto tematico. La difficoltà è calcolata in base al livello di istruzione o alle frequenze delle parole specifiche per quella classe tematica.

6.4.10. Pitler e Nenkova 2008

Pitler e Nenkova (2008) sviluppano un modello in grado di predire le valutazioni dei lettori sui livelli di leggibilità dei testi; il modello prende in considerazione la combinazione di diverse caratteristiche: lessicali, sintattiche e analisi del discorso.

Come corpus di addestramento vengono utilizzati 30 articoli del *Wall Street Journal*, rivolti ad un pubblico adulto istruito. Le variabili di leggibilità considerate sono il vocabolario, la sintassi, la coerenza e la coesione delle entità, la struttura del discorso. Per l'analisi del discorso viene impiegato il corpus annotato *Penn Discourse Treebank* (Prasad et al. 2008)[150].

La valutazione della leggibilità è considerata un compito di ranking e i vari livelli sono assegnati da un gruppo di lettori. "In the easier task of text quality ranking, entity coherence and syntax features also become significant and the combination of features allows for ranking prediction accuracy of 88%" (Pitler e Nenkova 2008, p. 186). Ogni articolo è valutato da almeno 3 lettori (studenti universitari) secondo un punteggio che va da 1 (il peggiore) a 5 (il migliore). Dalla valutazione risultano valori che vanno da 1,5 a 4,33, con un punteggio medio di 3,2.

[150] Il *Penn Discourse Treebank* è un corpus annotato, composto da articoli tratti dal *Wall Street Journal*, in cui sono marcate le relazioni tra le parti del discorso. Oltre ad etichettare i connettivi, il *Penn Discourse Treebank* annota anche il *senso* delle relazioni.

Sono quindi calcolate le correlazioni tra le varie caratteristiche linguistiche e le valutazioni dei lettori. La Tabella 42 mostra le correlazioni con le "metriche di base": numero medio di caratteri per parola, numero medio di parole per frase, numero massimo di parole per frase e la lunghezza dell'articolo. L'unico parametro significativo sembra essere la lunghezza dell'articolo.

Caratteristica	**Correlazione**
Caratteri per parola	r = - 0,0859
Parole per frase	r = 0,1637
Max. parole per frase	r = 0,0866
Lunghezza articolo (F_7)	**r = - 0,3713**

Tabella 42. Correlazioni con le caratteristiche di base.

Per lo studio delle caratteristiche del vocabolario dei testi, oltre al corpus di articoli del *Wall Street Journal* (WSJ), viene impiegata anche una raccolta di notizie da *AP News* (NEWS). Il vocabolario è analizzato tramite un modello linguistico del corpus del WSJ (F_5) e un modello statistico di tipo *unigram* del corpus NEWS (F_6). La stima della probabilità delle parole è analizzata in combinazione alla lunghezza degli articoli (F_7).

Caratteristica	**Correlazione**
F_5 (WSJ)	r = 0,3723
F_6 (NEWS)	r= 0,4497
F_5 (WSJ) + F_7	r= 0,3732
F_6 (NEWS) + F_7	**r= 0,6359**

Tabella 43. Correlazioni con le caratteristiche del vocabolario.

Entrambe le caratteristiche del vocabolario sono maggiormente correlate con i giudizi di leggibilità rispetto alle caratteristiche di base. Com'è prevedibile, i valori del modello NEWS sono più alti, dal momento che gli articoli del corpus sono di carattere più generale; in combinazione con la lunghezza degli articoli, questo elemento diviene piuttosto predittivo, con una correlazione di 0,63.
Per quanto riguarda le variabili sintattiche, sono considerate: l'altezza media degli alberi di analisi (F_1), il numero medio di frasi nominali (F_2), il numero medio di frasi verbali (F_3) e il numero medio di subordinate per frase (SBARs, F_4). Le correlazioni sono mostrate nella Tabella 44:

Caratteristica	**Correlazione**
F_1	r = 0,0634
F_2	r= 0,2189
F_3	**r= 0,4213**

Caratteristica	Correlazione
F_4	r= 0,3405

Tabella 44. Correlazioni con le caratteristiche sintattiche.

La coesione lessicale è misurata tramite 5 caratteristiche: il numero di pronomi per frase (F_{11}), il numero di articoli determinativi per frase (F_{12}), la somiglianza media del coseno (F_8), la sovrapposizione delle parole (F_9) e la sovrapposizione di parole soltanto su nomi e pronomi (F_{10}) tra coppie di frasi adiacenti.

Caratteristica	Correlazione
F_8	r = -0,1012
F_9	r = -0,0531
F_{10}	r = 0,0905
F_{11}	r = 0,2381
F_{12}	r = 0,2309

Tabella 45. Correlazioni con le caratteristiche di coesione.

Come si può vedere, nessuna di queste caratteristiche è correlata in modo significativo alla leggibilità. Lo stesso risultato si ottiene per la coerenza delle entità, misurata su 16 diverse caratteristiche (F_{17-32}) tramite il Brown Coherence Toolkit.
Infine, viene effettuata l'analisi del discorso. Le relazioni tra le parti del discorso sono studiate tramite un modello statistico del linguaggio (multinomiale), che usa in questo caso le relazioni invece delle parole; ogni relazione è annotata sia per il senso che per il modo in cui è realizzata, cioè se è implicita (F_{16}) o esplicita (F_{15}). Si considerano inoltre il numero di relazioni del discorso (F_{14}) e la funzione di log-verosomiglianza (F_{13}) combinata con il numero di relazioni.

Caratteristica	Correlazione
F_{13}	**r = 0,4835**
F_{14}	r = -0,2729
$F_{13} + F_{14}$	**r = 0,5409**
F_{14} + n. di parole	r = 0,3819
F_{15} (relazioni esplicite)	r = 0,1528
F_{16} (relazioni implicite)	r = 0,2403

Tabella 46. Correlazioni con le caratteristiche del discorso.

La funzione di log-verosomiglianza[151] delle relazioni del discorso presenta le correlazioni maggiori (0,48), soprattutto in combinazione con la lunghezza del testo (0,54).
In sintesi, le caratteristiche che risultano più predittive della leggibilità sono quelle collegate al vocabolario (F_6) e alle parti del discorso (F_{13}), seguite dalla lunghezza dei testi (F_7) e dal numero medio di frasi verbali (F_3). La combinazione dei primi tre elementi fornisce un valore di 0,5029.
Oltre a prevedere il livello di leggibilità di ogni singolo testo, il classificatore è allenato a considerare coppie di documenti per determinare quale dei due sia il migliore (compito di ranking). Per la classificazione è impiegata un'implementazione del metodo SVM e le prestazioni sono valutate tramite una convalida incrociata di 10 volte. "Our experiments indicate that discourse relations are the one class of features that exhibits robustness across these two tasks" (Pitler e Nenkova 2008, p. 186).

6.4.11. Peterson e Ostendorf 2009

Lo studio di Peterson e Ostendorf (2009) si basa sul lavoro preliminare di Schwarm e Ostendorf (2005), nel quale è presentato un metodo per la valutazione automatica del livello di lettura di testi scritti destinati a studenti stranieri o di L2. L'approccio, che prevede l'uso di algoritmi SVM per combinare le caratteristiche di modelli linguistici tradizionali e modelli statistici di tipo *n-gram*, è ripreso dagli autori anche in questo studio più recente.
Anche il corpus di addestramento è lo stesso della ricerca precedente: si tratta di circa 2400 articoli tratti da *Weekly Reader*, una rivista educativa in cui è possibile trovare testi che coprono i livelli di lettura dal secondo al quinto grado. La distribuzione degli articoli nel corpus è mostrata nella Tabella 47.

Livello	Numero di articoli	Numero di parole	Lunghezza media degli articoli (in parole)
2°	351	71,5 k	161,1
3°	589	444 k	151,4
4°	766	927 k	254,3
5°	691	1 M	314,4

Tabella 47. Statistiche del corpus *Weekly Reader.*

Come supplemento, sono impiegati anche gli altri due corpora, quello dell'Enciclopedia Britannica (versione completa e versione per bambini) e l'archivio di notizie della CNN (sia in forma estesa che in versione ridotta)[152]. A questi, viene aggiunto un terzo set, costituito da *newswire* (notizie in tempo reale) tratti dal corpus TIPSTER (Harman e Liberman 1993); nonostante il livello di lettura non sia indicato, gli autori presumono che sia superiore al 5°

[151] La *verosomiglianza* (L, *likelihood*) di un dato valore del parametro è la probabilità di ottenere i dati osservati se il parametro assume quello specifico valore. La stima di *massima verosomiglianza* indica il valore del parametro per il quale la probabilità di ottenere i dati osservati è massima. La *massima verosomiglianza* è calcolata tramite la *funzione di verosomiglianza*, che indica le probabilità di osservare il campione al variare dei valori del parametro. Il logaritmo naturale della *verosomiglianza* è chiamato *funzione di log-verosomiglianza*.
[152] Cfr. 6.4.5

grado. Questo set è impiegato come *negative training data*, per migliorare l'accuratezza delle prestazioni del rilevatore nel caso di testi con livello di lettura pari o superiore al 5° grado. La Tabella 48 mostra le statistiche dei corpora di supplemento.

Corpus	**Numero di articoli**	**Numero di parole**
Britannica	115	277 k
Britannica Elementary	115	74 k
CNN	111	51 k
CNN ridotta	111	37 k
TIPSTER	979	420k

Tabella 48. Statistiche dei corpora dell'Enciclopedia Britannica + CNN + TIPSTER.

Il problema della classificazione dei livelli di lettura dei testi viene affrontato da due punti di vista.

Da una parte, si tratta di un problema di rilevamento binario: viene costruito un classificatore per ogni livello di lettura; il classificatore decide se un dato documento appartiene o meno a quella classe. Dall'altra, il classificatore si occupa di categorizzare i documenti nelle varie classi, cioè nei vari livelli di lettura (secondo un modello di regressione). In entrambi i casi è utilizzato il metodo SVM; l'algoritmo utilizza diverse caratteristiche linguistiche, come la lunghezza della frase, la lunghezza delle parole, la leggibilità misurata con l'indice di Flesch-Kincaid, ecc. Viene inoltre combinato con tre modelli linguistici (*unigram*, *bigram* e *trigram*) formati sui 4 corpora di addestramento principali, per un totale di 12 modelli statistici.

I risultati del rilevamento basato sul *Support Vector Machine* sono mostrati nella tabella seguente:

Grado	**Precisione**	**Copertura**	**Punteggio F**
2	38	61	47
3	38	87	53
4	70	60	65
5	75	79	77

Tabella 49. Valori (in %) di precisione, copertura e punteggio F per il modello di classificazione basato su SVM.

Per ottimizzare la classificazione dei testi che presentano livelli di lettura più alti del 5° grado (cioè per ridurre il numero di falsi positivi per tali livelli), viene incluso anche il corpus TIPSTER. Per valutare le prestazioni del sistema su nuovi dati, si testa il rilevatore su un corpus di 30 articoli tratti dall'edizione *Kidspost* del *Washington Post* del 2005, con livelli che vanno dal 3° all'8° grado.

La Tabella 50 mostra i risultati della classificazione degli articoli *Kidspot*.

Grado	Weekly Reader	WR + TIPSTER
2	0	0
3	4	2
4	11	10
5	21	12
Non rilevati	0	12

Tabella 50. Numero di articoli *Kidspot* rilevati dal modello SVM addestrato sul solo *Weekly Reader* (WR) e sul set *Weekly Reader* + TIPSTER (esempi negativi).

Come si osserva, entrambi i classificatori hanno correttamente ignorato gli articoli di grado 2. Il classificatore addestrato con il solo corpus del *Weekly Reader* ha rilevato un numero maggiore di articoli per il grado 5, classificando in questo livello anche gli articoli con un grado maggiore; il classificatore addestrato anche con gli esempi negativi presenta invece una maggiore accuratezza, individuando un numero più basso di articoli per il grado 5 e lasciando invece 12 articoli come non classificati.
Le prestazioni dei due tipi di classificatori SVM sono poi confrontate con quelle del classificatore SVM basato sulla regressione e utilizzato per il secondo approccio (Figura 24). Vengono impiegate le stesse funzionalità e gli stessi dati di addestramento.

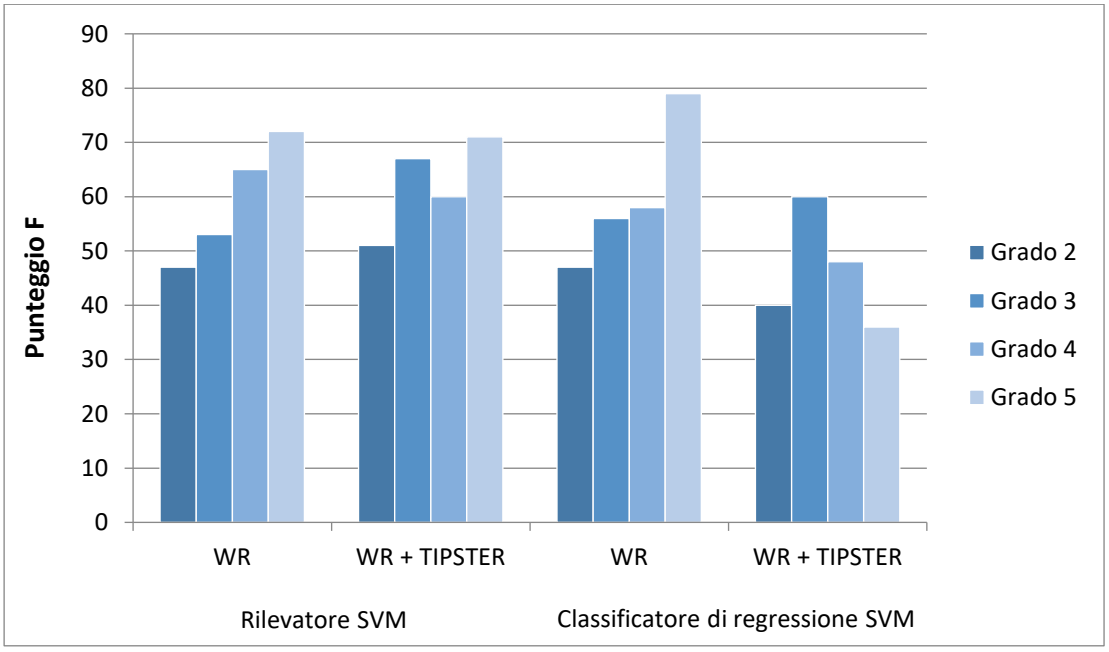


Figura 24. Confronto tra i punteggi F ottenuti utilizzando il rilevatore SVM e il classificatore di regressione SVM.

Il classificatore di regressione SVM addestrato con il solo corpus del *Weekly Reader* si comporta in modo simile al rilevatore SVM per quanto riguarda i gradi 2 e 3; le prestazioni peggiorano per quanto riguarda il grado 4 ma migliorano per il 5. Il classificatore di regressione addestrato con il WR e il corpus TIPSTER presenta invece valori più bassi per tutti i livelli tranne il terzo grado.

6.4.12. Kanungo e Orr 2009

Uno degli approcci standard utilizzati dai motori di ricerca online per rendere migliore l'esperienza degli utenti, in termini di recupero di pagine pertinenti all'interrogazione e riduzione del tempo di ricerca, consiste nel visualizzare un abstract del contenuto di ciascuna pagina nella SERP, cioè nella pagina che riporta la lista dei risultati (Figura 25)[153].

Figura 25. Pagina di riepilogo dei risultati della ricerca. Per ogni risultato viene visualizzato il titolo, l'URL e l'abstract.

Lo studio di Clarke et al. (2007) ha dimostrato che la leggibilità di questi abstract ha un impatto diretto sul comportamento degli utenti: una sintesi migliore ha maggiore possibilità di generare un clic da parte dell'utente[154]. La leggibilità di tali abstract è valutata periodicamente, ma poiché si tratta di un processo piuttosto dispendioso, viene effettuata almeno a cadenza trimestrale. La metodologia impiegata consiste prima nella raccolta di un corpus di query casuali e dei corrispondenti risultati recuperati e in seguito nella valutazione di questi da parte di giudici umani. È evidente che un processo di valutazione manuale non possa essere effettuato in tempo reale.

Kanungo e Orr (2009) propongono un approccio di apprendimento automatico per misurare la leggibilità degli abstract dei risultati delle ricerche sul web. Il modello può essere impiegato sia per il monitoraggio dei riassunti in tempo reale sia direttamente nel processo di generazione degli abstract.

Il metodo prevede la raccolta di un corpus di pagine recuperate dalla ricerca e i relativi giudizi da parte di valutatori umani, con punteggi che vanno da 1 (illeggibile) a 5 (facile da leggere). Vengono quindi estrapolate le caratteristiche da ciascun abstract e, in funzione di queste, viene creato un modello per la previsione dei giudizi. La modellazione è ottenuta

[153] SERP è l'acronimo per *Search Engine Results Page* ('pagina dei risultati del motore di ricerca').

[154] Altri studi sulla leggibilità degli abstract dei risultati delle ricerche sono condotti da Aula 2004, Radev e Fan 2000, Obendorf e Weinreich 2003, Kickmeier e Albert 2003, Rose et al. 2007.

tramite la regressione, utilizzando alberi decisionali a gradiente stocastico (Gradient Boosting Decision Tree, GBDT)[155].
Le caratteristiche considerate nello studio sono:

- La leggibilità, misurata tramite l'indice Fog, la formula di Flesch e quella di Flesch-Kincaid;
- Numero medio di caratteri per parola (CPWRD);
- Numero medio di sillabe per parola (SYLPWRD);
- Percentuale di parole complesse (PCMPLXWRDS);
- Numero di frammenti (NSNIP)[156];
- Se il testo inizia con punti di sospensione (BELLIP);
- Se il testo finisce con punti di sospensione (ELLIP);
- Quantità di lettere maiuscole (OAPFRAO);
- Quantità dei segni interpuntivi (PUNCFRAC): se ci sono troppi caratteri di punteggiatura, molto probabilmente si tratta di spam o di un documento non testuale;
- Quantità di *stop word* (STOPFRAC)[157];
- Quantità di *Query Word Hit* (HITFRAC): i lettori sono influenzati dalla presenza o dall'assenza di specifici termini nella query.

I dati di addestramento consistono in 5.382 valutazioni effettuate da 7 giudici nell'arco di un anno sugli abstract dei risultati delle ricerche effettuate con Yahoo! e Google. Le valutazioni sono suddivise casualmente in set di addestramento e set di prova.
La Tabella 51 mostra le correlazioni con le valutazioni ottenute dalle formule di leggibilità e dal modello costruito con gli alberi decisionali; viene effettuato un confronto anche con il modello di Collins-Thompson e Callan (2004).

Metodo	Correlazione
Fog	0,01572242
Kincaid	- 0,02689905
Flesch-Kincaid	0,02323278
Lineare	- 0,001198311
Collins-Thompson e Callan	0,0597
GBDT	0,6321157

Tabella 51. Correlazione tra i modelli e le valutazioni dei giudici.

Come possiamo notare, non sembrerebbe esserci una correlazione con i tradizionali indici di leggibilità; questo dipende probabilmente dal fatto che gli abstract sono molto brevi, contengono pochissimo testo e sono spesso costituiti da frammenti di frasi e non da frasi intere. Anche la correlazione con il modello di Collins-Thompson e Callan risulta trascurabile

[155] Cfr. Friedman 2001a e 2001b.
[156] I frammenti (*snippets*) possono essere frasi complete o parti di frasi.
[157] Sono le parole prive di significato, come articoli o congiunzioni.

(0,05). Il sistema costruito con gli alberi decisionali a gradiente stocastico risulta maggiormente correlato (0,63).
Per quanto riguarda le caratteristiche, le più rilevanti sono la quantità di lettere maiuscole (OAPFRAO), la quantità dei segni interpuntivi (PUNCFRAC) e la quantità di *stop word* (STOPFRAC): si tratta di elementi che non sono generalmente presenti nelle formule di leggibilità. L'influenza relativa delle caratteristiche è mostrata nella Figura 26.

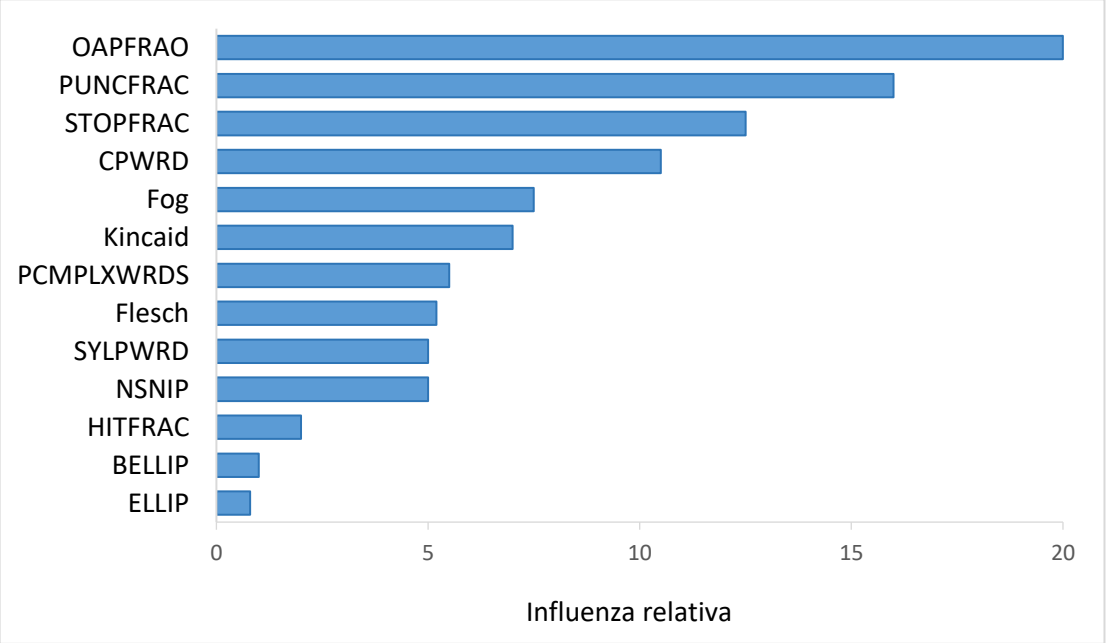


Figura 26. Influenza relativa delle caratteristiche.

6.4.13. Kate et al. 2010

Questo studio si inserisce nel *Machine Reading Program* (MRP) della DARPA[158], programma di ricerca dedicato allo sviluppo di sistemi in grado di acquisire conoscenze dai corpora in linguaggio naturale e renderle disponibili per l'elaborazione formale. L'approccio di Kate et al. (2010) prevede lo sviluppo di un sistema di valutazione automatica delle leggibilità di documenti basato su giudizi umani; a differenza dei lavori precedenti, il modello è costruito per la previsione della leggibilità e non dei livelli di istruzione associati ai testi. Inoltre, i testi utilizzati per l'addestramento non appartengono a un singolo dominio ma provengono da varie fonti e coprono diversi generi; questo consente la formazione di modelli linguistici specifici per ogni genere testuale.
Il corpus di addestramento è formato da 390 testi, tratti da giornali e newswire, weblog, post di newsgroup, trascrizioni manuali, output di traduzioni automatiche, articoli di Wikipedia, trascrizioni di sottotitoli. I testi sono distribuito in modo uniforme su 7 generi.
La leggibilità è valutata da due categorie di giudici: 8 giudici esperti di madrelingua inglese (linguisti e professionisti specializzati in analisi e annotazione linguistica) e 6 giudici non esperti di madrelingua inglese (insegnanti di inglese, redattori, scrittori e altre figure professionali che non possiedono competenze specifiche nell'analisi e nell'annotazione

[158] La *Defense Advanced Research Projects Agency* (DARPA, 'agenzia per i progetti di ricerca avanzata di difesa'), è un'agenzia governativa del Dipartimento della Difesa degli Stati Uniti che si occupa dello sviluppo di nuove tecnologie per uso militare.

linguistica). I punteggi assegnati vanno da 1 (livello basso) a 5 (livello alto): la leggibilità è definita come il "subjective judgment of how easily a reader can extract the information the writer or speaker intended to convey" (Kate et al. 2010, p. 548).
Il sistema è addestrato utilizzando una combinazione di varie caratteristiche: caratteristiche derivate dai modelli linguistici, dai parser sintattici e caratteristiche lessicali. Il suo compito è imparare, tramite un'analisi di regressione, a far corrispondere le valutazioni dei giudici con le caratteristiche linguistiche ritenute potenzialmente rilevanti per la leggibilità. Tra le variabili vengono inclusi anche i modelli linguistici specifici per ogni genere, così da vedere se questi hanno un qualche valore predittivo.
Trattandosi di dati di tipo numerico (e non testuale), gli autori scelgono di trattare il compito come un problema di regressione, arrotondando il punteggio previsto per ottenere il valore intero più vicino. Per ogni documento, il punteggio medio ottenuto dalle valutazioni dei giudici esperti, è preso come standard di riferimento (*gold strandard*). Gli algoritmi di regressione utilizzati sono forniti dal software open source Weka[159].
L'analisi sintattica è effettuata tramite il parser *Sundance* (Riloff e Phillips 2004), che misura la violazione delle regole grammaticali della lingua inglese tramite alcune caratteristiche (lunghezza delle frasi, numero di frasi nominali, numero di frasi verbali, ecc.), e il parser *English Slot Grammar* (McCord 1989), che esegue un'analisi linguistica più approfondita. Le caratteristiche lessicali considerate sono: percentuale di parole OOV (*out of vocabulary*), numero di parole funzionali, numero di pronomi, numero di parole conosciute (che si possono trovare in un dizionario inglese o in un dizionario geografico che contiene nomi di persone e luoghi).
Inizialmente gli studiosi conducono esperimenti per testare gli algoritmi di regressione utilizzando tutte le funzioni; successivamente, sono escluse varie serie di caratteristiche per determinare quale combinazione abbia il maggior valore predittivo.
La tabella seguente mostra i valori di correlazione ottenuti dalle diverse caratteristiche.

Caratteristiche	**Correlazione**
Lessicali	0,5760
Sintattiche	0,7010
Lessicali + sintattiche	0,7274
Modello del linguaggio	0,7864
Tutte	0,8173

Tabella 52. Confronto tra i diversi set di caratteristiche.

Come si osserva, il modello statistico del linguaggio presenta una correlazione più alta rispetto agli altri set. Le prestazioni migliorano quando si combinano tutte le caratteristiche. Sono poi confrontati i valori di correlazione ottenuti da modelli linguistici indipendenti dal genere e modelli specifici per ogni genere (Tabella 53).

[159] Cfr. Nota [144].

Modello linguistico	Correlazione
Modello indipendente dal genere	0,6978
Modello basato sul genere	0,7749
Combinazione dei due modelli	0,8173

Tabella 53. Confronto tra modello indipendente dal genere e modello basato sul genere.

I risultati mostrano che l'utilizzo di modelli specifici di genere per l'addestramento del sistema contribuisce a migliorare le previsioni delle leggibilità.

Il team SAIC conduce una valutazione ufficiale per conto di DARPA a cui partecipano, oltre a Kate e al. (A) altre due squadre (B e C). Il materiale da valutare consiste in 150 documenti tratti dai 390 testi di addestramento. Oltre alla correlazione, sono utilizzate due metriche aggiuntive: la prima calcola la differenza tra i punteggi ottenuti dalle valutazioni di giudici esperti e giudici inesperti e la differenza tra i punteggi ottenuti dalle valutazioni di giudici esperti e dalle valutazioni della macchina; la seconda (*target hits*) misura se il punteggio previsto per un documento rientra nell'intervallo (*width*) di valori per quel documento (che va dal punteggio più basso a quello più alto) e, nel caso in cui sia compreso, calcola un punteggio inversamente proporzionale a tale intervallo. Il punteggio finale dei *target hits* è calcolato facendo una media su tutti i documenti. La tabella 54 illustra i risultati della valutazione.

Sistema	Correlazione	Differenza assoluta media	Target hits
Sistema A	0,8127	0,4844	0,4619
Sistema B	0,6904	0,3916	0,4530
Sistema C	0,8501	0,5177	0,4641
Valori critici superiori	0,7423	0,0960	0,3713

Tabella 54. Confronto tra i punteggi ottenuti nella valutazione dei sistemi. I valori critici (CV) superiori sono i punteggi ottenuti dai giudici inesperti.

L'approccio di Kate et al. ottiene un buon punteggio di correlazione e in tutte le metriche supera i valori ottenuti dai giudici inesperti. Il sistema migliore risulta tuttavia essere il terzo (C).

6.4.14. Tanaka-Ishii et al. 2010

Tanaka-Ishii et al. (2010), dell'Università di Tokyo, presentano un approccio alternativo per la valutazione automatica della leggibilità che si basa sull'ordinamento (*sorting*)[160].
Il modello è implementato in due fasi: nella prima, l'algoritmo SVM genera un comparatore, che viene addestrato a giudicare la leggibilità relativa tra due dati testi. Nella seconda fase, viene applicato il comparatore per ordinare un corpus di testi. "The readability of a text is

[160] Questo approccio può essere inserito tra i metodi di ranking. Per il SVM viene usato il software LIBSVM.

assessed by searching for its position within the sorted texts. The norm is thus considered as the location of a text among an *ordered* set of texts. Our approach linguistically enhances assessment of the readability of a text as the *relative* ease compared to other texts, not as the absolute difficulty of the text" (Tanaka-Ishii et al. 2010, p. 205-206).

Questo metodo risolve il problema della mancanza di dati di addestramento etichettati: "large amounts of training data annotated with 12 school grades have not been at all easy to obtain on a reasonable scale. Another possibility might have been to manually construct such training data but humans are generally unable to precisely judge the level of a given text among 12 arbitrary levels. The corpora therefore have to be constructed from academic texts used in schools. The amount of such data, however, is limited, and its use is usually strictly limited by copyrights. Thus, it is crucial to devise a new method or approach that allows readability assessment by using only generally available corpora, such as newspapers" (id., p. 203-204).

La prima fase di costruzione del modello prevede la presenza di un corpus di addestramento, ma, dato che il comparatore giudica i testi soltanto rispetto a due livelli di leggibilità (facile o difficile), il materiale deve semplicemente essere annotato secondo uno di questi due livelli. È infatti più semplice determinare tra due testi quale sia il più difficile, che dover attribuire a un testo un valore di leggibilità tra 12 diversi livelli.

"Note that we do not claim that our model and method is better than existing methods. Although our method does compete well with previous methods, the classification approach used in any given scenario should remain the most natural, relevant method. The intention of this article is simply to propose an alternative way of handling readability assessment, especially when adequate training corpora annotated with multiple levels are not available" (id., p. 204).

Le caratteristiche considerate nel modello sono soltanto quelle legate al vocabolario; dal momento che alcune funzionalità non possono essere applicate al giapponese, gli autori si concentrano soltanto su quelle che sono disponibili (e dunque comparabili) in tutte le lingue. I fattori utilizzati sono la frequenza relativa delle parole (*fattore locale*), intesa come la frequenza di ogni parola diviso la frequenza del numero di parole nel testo, e il grado di leggibilità delle parole rispetto al vocabolario generale (*fattore globale*), che considera il *log frequency* delle parole, ottenuto da un corpus di grandi dimensioni[161].

L'approccio proposto dagli studiosi viene confrontato con i metodi di regressione e di classificazione.

Aspetti confrontati	**Regressione**	**Classificazione**	**Ordinamento**
Leggibilità	Punteggio	Classe	Ranking
Output	Continuo	Discreto	Discreto
Livelli dei dati di input richiesti	Multipli	Multipli	Due
Completezza	Alto	Alto	Discutibile
Velocità di valutazione	Veloce	Veloce	Lento

Tabella 55. Confronto tra i tre metodi: regressione, classificazione e ordinamento.

[161] Sono impiegati due web corpora, uno per l'inglese, costituito da 6 terabyte e uno per il giapponese, costituito da 2 terabyte. Il parametro *log frequency* è strettamente correlato al livello di familiarità delle parole.

Nei primi due metodi la leggibilità è valutata in termini di punteggi o classi (generalmente corrispondenti a livelli di istruzione); nell'ordinamento è rappresentata da una posizione all'interno del corpus di testi. Per quanto riguarda l'output, il metodo di regressione è continuo, in quanto i punteggi si trovano in un intervallo continuo; la classificazione e l'ordinamento sono entrambi discreti, in quanto costituiti da elementi isolati (classe o posizione). Come già osservato, la regressione e la classificazione richiedono dati etichettati in più livelli mentre il terzo metodo richiede un'annotazione di solo due valori (facile, difficile). Tuttavia, questo approccio presenta dei problemi per quanto concerne la completezza e la velocità di valutazione. Il sistema riporta infatti un valore relativo di leggibilità e manca di *assolutezza* nel determinare una norma. La lentezza nella valutazione è data dal fatto che la posizione del documento deve essere cercata dal comparatore e questo richiede un certo periodo di tempo prima che si ottenga una risposta.
Gli autori presentano inoltre la loro applicazione, *Terrace*, che si occupa di recuperare documenti con una leggibilità simile a quella di un dato testo di input. Il sistema, sviluppato in risposta a una richiesta mossa da docenti universitari di lingue straniere, funziona sia per l'inglese che per il giapponese, con possibili estensioni al cinese e al francese. La raccolta comprende 14.877 testi tratti dalla CNN.

6.4.15. Al-Kalifa e Amani 2010

Al-Kalifa e Amani (2010) presentano il loro prototipo, *Arability*, uno strumento per misurare in modo automatico la leggibilità di testi in arabo. Il problema della leggibilità nella lingua araba è ancora nelle prime fasi della ricerca e, come osservano gli autori, sono state sviluppate soltanto due formule, quella Dawood e quella Al-Heeti[162].
La Figura 27 illustra le diverse fasi di sviluppo del prototipo.

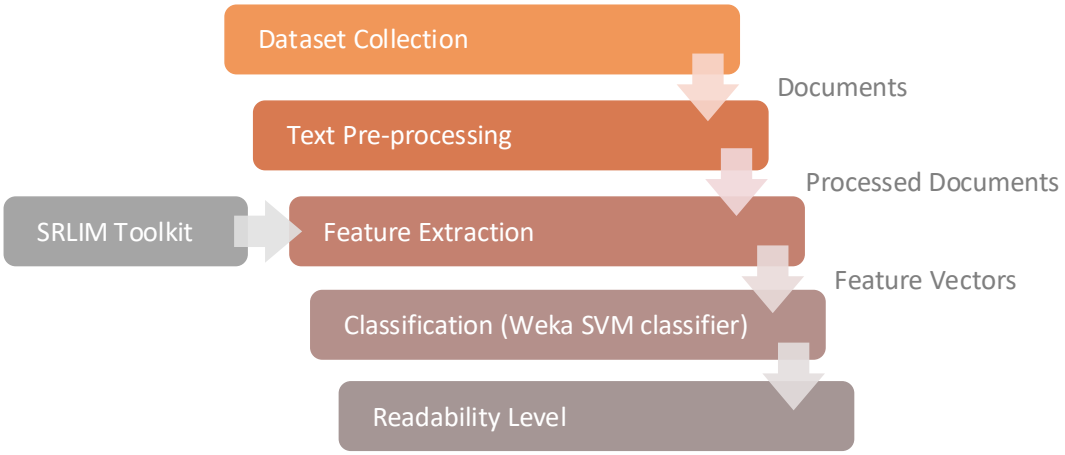


Figura 27. Fasi di sviluppo del prototipo (rielaborazione da Al-Kalifa e Amani 2010).

[162] La formula Dawood considera 5 variabili: la lunghezza media delle parole, la lunghezza media della frase, la frequenza delle parole, la percentuale di frasi nominali e la percentuale di sostantivi definiti; la formula Al-Heeti invece include una sola caratteristica, la lunghezza media delle parole.

Dal momento che non è disponibile per la lingua araba un corpus di testi già etichettati in livelli di lettura, gli autori raccolgono manualmente un set di addestramento; il corpus comprende 150 testi tratti da libri che si trovano nei programmi di studio delle scuole elementari, intermedie e secondarie dell'Arabia Saudita. Ad ogni ordine di scuola corrisponde un livello di leggibilità: facile, medio e difficile. La tabella seguente mostra le statistiche del corpus.

Livello di leggibilità	**N. testi**	**N. parole**	**Lunghezza media della frase**	**Lunghezza media delle parole**
Facile	50	4.729	3,95	4,38
Medio	50	24.810	6,37	4,56
Difficile	50	27.550	7,72	4,77

Tabella 56. Composizione del corpus.

Come si osserva, la lunghezza media delle parole e delle frasi aumenta via via che il livello di leggibilità diventa più difficile.
Le caratteristiche linguistiche considerate per lo sviluppo del modello sono:

- Lunghezza media della frase (numero medio di parole per frase);
- Lunghezza media delle parole (numero di lettere per parola);
- Numero di sillabe per parola (numero di vocali per parola),
- Frequenza delle parole (% di parole con frequenza <1);
- Punteggio di *perplessità* (*perplexity scores*) per il modello linguistico basato su bigrammi (un modello per ogni livello di leggibilità).

È creato un modello statistico basato su bigrammi per ogni livello di leggibilità. Il punteggio di *perplessità* indica la probabilità che ha un testo di appartenere a quella data classe (cioè a quel dato livello di leggibilità); una minore perplessità indica una probabilità maggiore. Questa funzionalità è misurata tramite il Toolkit SRILM[163], che è stato integrato nel prototipo in fase di estrazione delle caratteristiche, per generare modelli statistici specifici per ogni livello.
Per misurare le altre quattro caratteristiche, gli autori sviluppano un programma java in grado di calcolare, dato un testo di input, un vettore di caratteristiche e restituire i risultati in uno speciale formato, chiamato ARFF (Attribute-Relation File Format), compatibile con il software Weka.
Weka viene integrato nel prototipo e serve per la generazione di algoritmi di classificazione. Gli studiosi effettuano 5 esperimenti pilota per verificare quali caratteristiche (e le loro combinazioni) sono maggiormente predittive della leggibilità e quali algoritmi (SVM, alberi decisionali e Naïve Bayes) sono più adatti per questo compito di classificazione. Valutano inoltre l'affidabilità del prototipo, confrontandone l'accuratezza con le valutazioni da parte di giudici umani esperti. Infine, testano lo strumento su un nuovo set di dati. La Tabella 57 presenta i punteggi ottenuti dalle varie caratteristiche, utilizzando il classificatore SVM.

[163] SRILM (The SRI Language Modeling Toolkit) è un kit di strumenti sviluppato nel 1995 dalla SRI Speech Technology e Research Laboratory. Viene impiegato per la costruzione di modelli statistici del linguaggio.

Caratteristiche	Accurat.	Facile			Medio			Difficile		
		P	R	F	P	R	F	P	R	F
Lunghezza media della frase	66,67%	0,75	1	0,857	0,571	0,667	0,615	0,667	0,333	0,444
Lunghezza media delle parole	44,44%	1	0,167	0,286	0,429	0,5	0,462	0,4	0,667	0,5
N. sillabe per parola	38,89%	0	0	0	0,333	0,5	0,4	0,444	0,667	0,533
Frequenza dei termini	50%	0	0	0	0,571	0,667	0,615	0,455	0,833	0,588
Modello linguistico	61,11%	0,625	0,833	0,714	0,667	0,333	0,444	0,571	0,667	0,615

Tabella 57. Punteggi ottenuti da ciascuna caratteristica: valori di accuratezza (%), precisione (P), recupero (R) e Punteggio F.

La migliore funzionalità risulta la lunghezza media della frase, seguita dal modello statistico del linguaggio. La lunghezza delle parole e il numero medio di sillabe per parola hanno invece bassi livelli di accuratezza. Per quanto riguarda le combinazioni tra le varie caratteristiche, il miglior modello è dato dall'unione delle tre variabili principali (lunghezza media della frase + modello statistico del linguaggio + frequenza dei termini). Come si può vedere nella Tabella 58, la combinazione delle due migliori funzionalità raggiunge un grado di accuratezza del 73,33%, quella di tutte le caratteristiche ottiene un valore più basso (72,22%).

Caratt.	Accurat.	Facile			Medio			Difficile		
		P	R	F	P	R	F	P	R	F
Tutte	72,22%	1	1	1	0,6	0,5	0,545	0,571	0,667	0,615
ASL + LM + TF	77,78%	1	1	1	0,667	0,667	0,667	0,667	0,667	0,667
ASL + LM	73,33%	1	1	1	0,571	0,667	0,615	0,6	0,5	0,545

Tabella 58. Valori ottenuti dalla combinazione delle caratteristiche.
ASL = lunghezza media della frase; LM = modello del linguaggio; TF = frequenza dei termini.

Vengono quindi confrontati diversi algoritmi di apprendimento automatico per verificare quale di questi risulta essere il migliore per questo compito di classificazione (Tabella 59). Il modello Naïve Bayes risulta più accurato quando viene impiegata una singola caratteristica, il classificatore SVM è invece più accurato quando si considera una combinazione delle variabili. Le prestazioni degli alberi decisionali risultano basse per ogni caratteristica (e combinazione) considerata. In generale, il metodo SVM risulta il migliore, raggiungendo un valore di accuratezza del 77,78%. La lunghezza della frase è un buon predittore della leggibilità con qualsiasi tecnica impiegata.

Caratteristiche	SVM	Naïve Bayes	Alberi decisionali
Tutte	72,22	61,11	44,44
Lunghezza della frase	66,67	66,67	61,11
Lunghezza delle parole	44,44	50	33,33
N. sillabe per parola	38,89	50	33,33
Frequenza dei termini	50	61,11	44,44
Modello linguistico	61,11	50	44,44
ASL + LM + TF	**77,78**	55,55	44,44
ASL + LM	73,33	50	44,44

Tabella 59. Valori di accuratezza (%) della classificazione usando SVM, Naïve Bayes e Alberi decisionali.

L'affidabilità del prototipo è valutata tramite il confronto con le valutazioni da parte di giudici umani esperti. Il set utilizzato nell'esperimento è composto da 26 testi, valutati da 3 esperti della Princess Norah University (Arabia Saudita). I risultati sono illustrati nella tabella seguente.

Livello	Arability	Esperto 1	Esperto 2	Esperto 3	Media esperti
Facile	100%	88,89%	66,67%	88,89%	81,4%
Medio	0%	71,4%	28,5%	42,8%	47,5%
Difficile	70%	80%	90%	50%	73,3%

Tabella 60. Valori di accuratezza ottenuti dal prototipo e dai giudici esperti.

Come è evidente, il prototipo Arability risulta il metodo migliore per classificare i testi con un livello facile (accuratezza del 10%), tuttavia la precisione scende a 70% per il livello difficile ed è pari a zero per il livello medio. Probabilmente, questo risultato dipende dal set di dati utilizzati negli esperimenti[164]. Anche i punteggi delle valutazioni degli esperti risultano peggiori per questo livello.
Il prototipo è infine testato su un nuovo set composto da 6 testi tratti da un libro dello scrittore arabo Al-Manfaloti e 6 testi di storie per bambini raccolte da vari siti web. Il 100% dei testi di Al-Manfaloti è classificato da Arability come difficile e il 75% dei testi per bambini come facile. Gli studiosi ritengono questo esperimento valido come prova dell'affidabilità del modello.

6.4.16. Aluisio et al. 2010

Questo studio si inserisce nell'ambito del progetto di semplificazione testuale PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital, 'Semplificazione

[164] È infatti probabile che i tre livelli di istruzione dei testi del corpus non corrispondano esattamente ai tre livelli di leggibilità ipotizzati dagli autori. Anche una piccola variazione tra i livelli può causare un'errata classificazione.

testuale del portoghese per l'inclusione e l'accessibilità digitale')[165], che prevede lo sviluppo di metodi e strumenti di adattamento dei testi per migliorare la comprensibilità di materiali rilevanti pubblicati sul web, in particolare da siti governativi e agenzie di stampa. Nello specifico, il progetto si propone di fornire uno strumento di semplificazione automatica di contenuti web da implementare nei browser e uno strumento di *authoring* per guidare gli autori nella creazione di versioni semplificate di testi[166]. Il lavoro di Aluisio et al. (2010) presenta lo sviluppo di un metodo di valutazione automatica della leggibilità per la lingua portoghese da integrare nello strumento di *authoring* SIMPLIFICA*,* il quale offre la possibilità di semplificare i testi sia dal punto di vista lessicale che sintattico. L'utente (cioè l'autore dei contenuti) può scegliere quando e in che misura applicare le operazioni di semplificazione, in base al livello di istruzione del target di riferimento. Grazie a uno strumento di valutazione automatica della leggibilità, l'utente può verificare il livello di complessità sia del testo originale, sia delle versioni da lui modificate, fino ad arrivare al livello richiesto, cioè quello adeguato al lettore di destinazione. Lo strumento classifica i testi in 3 livelli di alfabetizzazione, definiti del *National Indicator of Functional Literacy* (INAF): elementare, base, avanzato.

La ricerca considera un set di 59 caratteristiche (Tabella 61) e comprende alcune *caratteristiche di base* (1-3, 9-11), cioè parametri tradizionali che sono misurati con semplici conteggi e non richiedono l'uso di strumenti o risorse esterne, e funzionalità più complesse. Le caratteristiche sono divise in 3 gruppi: il primo include un insieme di parametri (1-42) tratti da Coh-Metrix Port (cfr. 6.5.1); il secondo contiene caratteristiche sintattiche (43-49); il terzo gruppo (50-59) considera funzionalità derivate da modelli statistici del linguaggio *n*-gram, che considerano punteggi di probabilità e perplessità di unigrammi, bigrammi e trigrammi e il tasso di parole fuori dal vocabolario.

N.	Caratteristica	N.	Caratteristica
1	Number of words	31	Number of positive additive connectives
2	Number of sentences	32	Number of negative additive connectives
3	Number of paragraphs	33	Number of positive temporal connectives
4	Number of verbs	34	Number of negative temporal connectives
5	Number of nouns	35	Number of positive causal connectives
6	Number of adjectives	36	Number of negative causal connectives
7	Number of adverbs	37	Number of positive logic connectives
8	Number of pronouns	38	Number of negative logic connectives
9	Average number of words per sentence	39	Verb ambiguity ratio
10	Average number of sentences per paragraph	40	Noun ambiguity ratio
11	Average number of syllables per word	41	Adverb ambiguity ratio
12	*Flesch* index for Portuguese	42	Adjective ambiguity ratio

[165] Cfr. Aluisio et al. 2008.

[166] Gli strumenti di *authoring* sono software che servono per la produzione di contenuti, come presentazioni multimediali, e-book, tutorial, lezioni/verifiche e altri strumenti didattici usati nell'e-learning, siti web, ecc.

N.	Caratteristica	N.	Caratteristica
13	Incidence of content words	43	Incidence of clauses
14	Incidence of functional words	44	Incidence of adverbial phrases
15	Raw Frequency of content words	45	Incidence of apposition
16	Minimal frequency of content words	46	Incidence of passive voice
17	Average number of verb hypernyms	47	Incidence of relative clauses
18	Incidence of NPs	48	Incidence of coordination
19	Number of NP modifiers	49	Incidence of subordination
20	Number of words before the main verb	50	Out-of-vocabulary words
21	Number of high level constituents	51	LM probability of unigrams
22	Number of personal pronouns	52	LM perplexity of unigrams
23	Type-token ratio	53	LM perplexity of unigrams, without line break
24	Pronoun-NP ratio	54	LM probability of bigrams
25	Number of “e” (and)	55	LM perplexity of bigrams
26	Number of “ou” (or)	56	LM perplexity of bigrams, without line break
27	Number of “se” (if)	57	LM probability of trigrams
28	Number of negations	58	LM perplexity of trigrams
29	Number of logic operators	59	LM perplexity of trigrams, without line break
30	Number of connectives		

Tabella 61. Set di caratteristiche considerate.

La leggibilità (caratteristica n. 12) è valutata tramite un adattamento della formula di Flesch (Martins et al. 1996), l’unico strumento disponibile per il portoghese; questo indice è implementato anche in Coh-Metrix-Port. Le caratteristiche dei modelli linguistici sono ricavate da un corpus di 96.868 testi tratti dal giornale brasiliano *Folha de São Paulo*, nel periodo 1994-2005.

Per la creazione del modello, gli studiosi utilizzano tre diversi metodi di apprendimento automatico: classificazione, ranking e regressione. Come algoritmo di classificazione è scelto SVM, fornito dal Toolkit Weka (SMO); per il ranking viene usato un metaclassificatore, anch’esso fornito dal software Weka (usa l’algoritmo SMO per effettuare una classificazione binaria); per la regressione è impiegato un classificatore di regressione SVM (SMO-reg).

Per l’addestramento sono utilizzati 7 corpora semplificati, creati nell’ambito del progetto PorSimple. Il primo è composto da articoli di notizie generiche tratti dal giornale brasiliano *Zero Hora* (ZH *original*). I testi sono stati riscritti da un linguista, secondo due livelli di semplificazione: naturale (ZH *natural*), che corrisponde ad un livello di alfabetizzazione *di base* e forte (ZH *strong*), che corrisponde al livello *elementare*. Gli altri corpora contengono articoli di divulgazione scientifica provenienti da diverse fonti: la sezione *Caderno Ciência* del quotidiano brasiliano *Folha de São Paulo*, un quotidiano di ampia diffusione (*CC*

original) e le sue due versioni semplificate (CC *natural* e CC *strong*), testi di livello *avanzato* tratti dalla rivista *Ciência Hoje* (CH). La Tabella 62 mostra la composizione dei 7 corpora.

Corpus	Testi	Frasi	Parole	Parole per testo	Parole per frase.
ZH original	104	2184	46190	444,1 (133,7)	21,1
ZH natural	104	3234	47296	454,7 (134,2)	14,6
ZH strong	104	3668	47938	460,9 (137,5)	13,0
CC original	50	882	20263	405,2 (175,6)	22,9
CC natural	50	975	19603	392,0 (176,0)	20,1
CC strong	50	1454	20518	410,3 (169,6)	14,1
CH	130	3624	95866	737,4 (226,1)	26,4

Tabella 62. Statistiche dei sette corpora.

Sono quindi calcolate le correlazioni (Tabella 63) tra le diverse caratteristiche e il livello di istruzione previsto per i due corpora che contengono più versioni (originale, naturale e forte). Le correlazioni più alte sono ottenute dalle caratteristiche di base e da quelle sintattiche.

Caratteristiche	Corr.
Parole per frase	0,693
Incidenza delle apposizioni	0,688
Incidenza delle frasi	0,614
Indice di Flesch	0,580
Parole prima del verbo principale	0,516
Frasi per paragrafo	0,509
Incidenza delle frasi relative	0,417
Sillabe per parola	0,414
Numero di connettivi additivi positivi	0,397
Numero di connettivi causali negativi	0,388

Tabella 63. Le 10 caratteristiche con i più alti valori di correlazione.

Viene infine eseguita una convalida incrociata di 10 volte delle varie caratteristiche e combinazioni di queste, tramite i tre metodi di apprendimento automatico.
Le tabelle seguenti mostrano i risultati ottenuti con la classificazione standard (Tabella 64), con il ranking (Tabella 65) e la regressione (Tabella 66); sono riportati il punteggio F (F), la correlazione (Corr.) e l'errore medio assoluto (MAE).

Caratteristiche	Classe	F	Corr.	MAE
Tutte	Original	**0,913**	**0,84**	**0,276**
	Natural	**0,483**		
	Strong	**0,732**		
LM	Original	0,669	0,25	0,381
	Natural	0,025		
	Strong	0,221		
Base	Original	0,846	0,76	0,302
	Natural	0,149		
	Strong	0,707		
Sintattiche	Original	0,891	0,82	0,285
	Natural	0,32		
	Strong	0,74		
Coh-Metrix-Port	Original	0,873	0,79	0,290
	Natural	0,381		
	Strong	0,712		
Flesch	Original	0,751	0,52	0,348
	Natural	0,152		
	Strong	0,546		

Tabella 64. Risultati della classificazione standard.

Caratteristiche	Classe	F	Corr.	MAE
Tutte	Original	**0,904**	**0,83**	**0,163**
	Natural	0,484		
	Strong	**0,731**		
LM	Original	0,634	0,49	0,344
	Natural	**0,497**		
	Strong	0,05		
Base	Original	0,83	0,73	0,231
	Natural	0,334		
	Strong	0,637		
Sintattiche	Original	0,891	0,81	0,180
	Natural	0,382		
	Strong	0,714		
Coh-Metrix-Port	Original	0,878	0,80	0,183
	Natural	0,432		

Caratteristiche	Classe	F	Corr.	MAE
	Strong	0,709		
Flesch	Original	0,746	0,56	0,310
	Natural	0,489		
	Strong	0		

Tabella 65. Risultati ottenuti con il metodo del ranking.

Caratteristiche	Corr.	MAE
Tutte	**0,8502**	**0,3478**
LM	0,6245	0,5448
Base	0,7266	0,4538
Sintattiche	0,8063	0,3878
Coh-Metrix-Port	0,8051	0,3895
Flesch	0,5772	0,5492

Tabella 66. Risultati della regressione.

I valori di correlazione e il Punteggio F ottenuti tramite la classificazione e il ranking risultano molto simili, ma l'errore medio assoluto è inferiore nel secondo metodo. La regressione raggiunge i più alti valori di correlazione (0,85), tuttavia i tassi di errore risultano più alti rispetto agli altri modelli.

Per quanto riguarda i set di caratteristiche, possiamo osservare che la combinazione di tutte le caratteristiche raggiunge risultati migliori per tutti e tre i modelli. Le prestazioni dei vari sottoinsiemi di caratteristiche invece variano da metodo a metodo. In generale, le variabili sintattiche ottengono correlazioni più alte (0,82 - 0,81 - 0,80), seguite da quelle di Coh-Metrix-Port (0,79 - 0,80 - 0,80); le funzionalità del modello linguistico riportano valori più bassi (0,25 - 0,49 - 0,62).

In base ai risultati ottenuti, gli autori scelgono di usare la classificazione come metodo per valutare automaticamente la leggibilità all'interno dello strumento di semplificazione: "the linear classification is our simplest model, has achieved the highest F-measure and its correlation scores are comparable to those of the other models" (Aluisio et al. 2010, p.8).

6.4.17. Feng et al. 2010

Feng et al. (2010) conducono uno studio approfondito in cui confrontano una vasta gamma di possibili caratteristiche da impiegare nella valutazione automatica della leggibilità.

Gli autori valutano in che misura questi aspetti siano predittivi del livello di difficoltà dei materiali destinati a studenti della scuola primaria: "we treat readability assessment as a classification task and evaluate trained classifiers in terms of their prediction accuracy. To investigate the contributions of various sets of features, we build prediction models and examine how the choice of features influences the model performance" (Feng et al., p. 276).

Il corpus impiegato nella ricerca è costituito da 1433 testi tratti dalla rivista educativa *Weekly Reader*, etichettati in livelli di istruzione che vanno dal 2° al 5° grado (Tabella 67).

Grado	N. testi	Parole per testo	Parole per frase
2	174	128,27	9,54
3	289	171,96	11,39
4	428	278,03	13,67
5	542	335,56	15,28

Tabella 67. Statistiche del corpus.

Sono analizzati 4 set di caratteristiche: superficiali, sintattiche, relative alle parti del discorso, legate ai modelli statistici del linguaggio.
Le caratteristiche che riguardano il discorso si ispirano alla linguistica cognitiva; le parti del discorso sono individuate tramite un criterio semantico-concettuale, secondo il quale le parole che appartengono a una stessa categoria possiedono un contenuto semantico comune. Le parti del discorso sono divise in base al tipo di entità che denotano: i nomi designano persone, animali, cose, i verbi si riferiscono ad azioni o processi, gli aggettivi a qualità, ecc. Le entità sono importanti per la comprensione del testo, in quanto formano le componenti di base dei concetti e delle proposizioni su cui si costruisce l'elaborazione del discorso a un livello superiore (Feng et al. 2009). Queste caratteristiche sono divise in 4 sottoinsiemi:

- densità delle entità (% di entità per frase, per documento, ecc.);
- catene lessicali (sequenze di termini semanticamente correlati tra loro, come sinonimi, iperonimi; si misura la lunghezza della catena, cioè il numero di entità contenute nella catena o l'intervallo, cioè la distanza tra la prima e l'ultima entità, ecc.);
- inferenza della coreferenza (sono estratti entità e riferimenti pronominali che hanno lo stesso coreferente e vengono formate delle catene di coreferenza; di queste, si misura il numero, la lunghezza, ecc.);
- griglie di entità (sono tracciati modelli di distribuzione delle entità per ciascuna coppia di frasi adiacenti; si calcola quindi la probabilità di distribuzione di ogni modello all'interno del testo)[167].

Gli studiosi considerano inoltre le tradizionali parti del discorso, per verificare in che misura siano correlate alla difficoltà. Per ogni classe di parole (nomi, verbi, aggettivi, avverbi e preposizioni) sono valutate 5 caratteristiche: ad esempio, per la classe aggettivo sono misurate la percentuale di aggettivi (token) nel testo e la percentuale di aggettivi (type) nel testo, il numero medio di aggettivi per frase, il numero medio di aggettivi (type) per frase e il rapporto tra aggettivi (type) sul totale di tutti i type nel testo.

[167] Il modello della griglia delle entità (cfr. Barzilay e Lapata 2008) si basa sul presupposto che la distribuzione delle entità nei testi mostra alcune regolarità. La griglia è una matrice bidimensionale in cui una dimensione è rappresentata dalle entità salienti nel testo e l'altra corrisponde a ciascuna frase del testo. Ogni cella rappresenta il ruolo grammaticale (soggetto, oggetto o nessuno dei due) corrispondente a una specifica entità in una specifica frase.

Le caratteristiche sintattiche analizzate sono l'altezza media degli alberi di analisi, il numero di subordinate, il numero medio di frasi nominali, il numero medio di frasi verbali, il numero di frasi preposizionali, la lunghezza delle frasi in parole e in caratteri, ecc.
Per quanto riguarda i modelli statistici, sono studiati i punteggi di perplessità di unigrammi, bigrammi e trigrammi di parole.
Le caratteristiche superficiali sono quelle utilizzate nelle tradizionali formule di leggibilità; in particolare, sono prese in considerazione:

- il numero medio di sillabe per parola;
- la percentuale di parole polisillabiche nel testo;
- il numero di parole polisillabiche per frase;
- numero di caratteri per parola;
- numero di parole per frase;
- numero di parole difficili nel testo;
- il numero totale di parole nel testo;
- punteggio di leggibilità misurato tramite l'indice di Flesch-Kincaid.

Nel corso degli esperimenti, sono impiegati diversi modelli di apprendimento, tra cui la regressione lineare, la classificazione standard (Regressione Logistica e SVM), la classificazione/regressione ordinale (che presuppone che i livelli di istruzione siano ordinati)[168]. I risultati mostrano che il metodo della classificazione standard risulta avere una maggiore accuratezza rispetto agli altri modelli.
Rispetto ai set di caratteristiche (Tabella 68), tra i modelli addestrati con le funzionalità legate al discorso, i valori più alti sono ottenuti dalle caratteristiche relative alla densità delle entità (SVM: 59,63%, Regressione Logistica: 57,59%). La combinazione di tutte le variabili non migliora in modo significativo la precisione rispetto al solo parametro delle densità delle entità.
Le caratteristiche sintattiche riportano in generale punteggi più bassi; il miglior predittore risulta il numero di frasi verbali (SVM: 53,07%, Regressione Logistica: 48,67%). In questo caso la combinazione di tutte le caratteristiche aumenta il punteggio di precisione (SVM: 57,79%, Regressione Logistica: 54,11%). Tra le caratteristiche legate alle parti del discorso, quelle legate ai nomi presentano una maggiore precisione (SVM: 58,15%, Regressione Logistica: 57,01%). Anche in questo caso, l'utilizzo di tutte le funzionalità non comporta un aumento significativo della precisione rispetto al parametro più predittivo. Per quanto riguarda le caratteristiche superficiali, la lunghezza media della frase raggiunge il punteggio di precisione più alto (52,17).
Le caratteristiche legate ai modelli linguistici ottengono i punteggi più alti (SVM: 62,52%, Regressione Logistica: 62,14%), soprattutto considerando la combinazione di tutte le variabili (SVM: 68,38%, Regressione Logistica: 66,82%).

[168] Per la costruzione del modello sono impiegati l'algoritmo SMO (LIBSVM) e quello formito da Weka.

Set di caratteristiche	SVM	Regres. Log.
Discorso		
Densità delle entità	59,63	57,59
Catene lessicali	45,86	42,58
Inferenza coreferenza	40,93	42,19
Griglie di entità	45,92	42,14
Tutte	60,50	58,79
Sintattiche		
Altezza alberi di analisi	44,26	43,45
Numero di subordinate	44,42	43,50
Numero di frasi nominali	51,56	48,14
Numero di frasi verbali	53,07	48,67
Numero di frasi preposizionali	49,36	46,47
Tutte	57,79	54,11
POS		
Nomi	58,15	57,01
Verbi	54,40	55,10
Aggettivi	53,87	52,75
Avverbi	52,66	50,54
Preposizioni	56,77	54,13
Parole contenuto	56,84	56,18
Parole funzione	52,19	50,95
Tutte	59,82	57,86
Superficiali		
Parole per frase		52,17
Sillabe per parola		42,51
Combinazione di queste		53,04
Punteggio Flesch-Kincaid		50,83
Polisillabi per frase		45,70
Tutte e 8 le caratteristiche		52,34
Modelli linguistici		
LM	**62,52**	**62,14**
Sequenze solo testo	60,17	60,31
Sequenze solo POS	56,21	57,64
Testo/POS	60,38	59,00
Tutti	**68,38**	**66,82**

Tabella 68. Confronto tra i punteggi dei set di caratteristiche.

6.4.18. François e Fairon 2012

François e Fairon (2012) presentano una formula di leggibilità per il francese come lingua straniera (FFL), basata sull'uso di 46 caratteristiche linguistiche[169].

Il corpus di addestramento è composto da 1.852 documenti tratti da libri di testo FFL, pubblicati dopo il 2001 e destinati ad adulti e adolescenti. Come scala di punteggi, gli autori scelgono la suddivisione in livelli definita nel Quadro comune europeo di riferimento (QCER): A1 (Breakthrough); A2 (Waystage); B1 (Threshold); B2 (Vantage); C1 (Effective Operational Proficiency) e C2 (Mastery). Questa suddivisione è ormai divenuta il riferimento per l'insegnamento delle lingue straniere in Europa.

Inizialmente sono impiegati 6 diversi algoritmi di apprendimento automatico: due modelli di regressione logistica, uno multinomiale (MLR) e uno ordinale (OLR), alberi di classificazione, due modelli basati su alberi decisionali (*bagging* e *boosting*) e il metodo SVM. I modelli di regressione e l'algoritmo SVM risultano i migliori.

Il set di caratteristiche è classificato dagli autori in 4 famiglie: lessicale, sintattica, semantica e specifica del contesto FFL. Ognuno di questi gruppi è ulteriormente diviso in sottofamiglie, di cui segnaliamo le principali:

- Caratteristiche lessicali:
  - *Statistiche delle frequenze lessicali*. Come database per la frequenza è impiegato *Lexique3* (New et al. 2007), un lessico che include circa 50.000 lemmi e 125.000 forme flesse, la cui frequenza è ottenuta tramite l'analisi dei sottotitoli dei film.
  - *Percentuale di parole che non si trovano in una lista di riferimento*. Sono impiegati due elenchi di parole per il francese L2: la lista di Gougenheim (1964), ormai datata e una lista che fa parte del libro di testo *Alter Ego* del 2006.
  - *Lunghezza della parola*.
  - *Modelli n-gram*. Viene usato un approccio *unigram* basato sulle frequenze di *Lexique3* e un modello *bigram* più complesso formato sul corpus di Google n-gram e sul corpus di articoli di giornale tratti da Le Soir.
  - *Diversità lessicale*, misurata tramite il rapporto TTR (rapporto tipi/repliche).
  - *Orthographic neighborhood* ('quartiere'). Si tratta di una nuova variabile, basata sull'ipotesi che alcune caratteristiche dei "vicini ortografici" di una data parola influenzano la lettura di questa parola. Con "vicini ortografici" di una parola *x* gli autori intendono tutte le parole che hanno la stessa lunghezza di *x* e che variano di una sola lettera (ad esempio *fase* e *base*). La variabile comprende 13 parametri che tengono conto del numero e della frequenza dei vicini ortografici di tutte le parole in un testo.
- Caratteristiche sintattiche:
  - *Lunghezza della frase* (numero di parole per frase).
  - *Relazioni tra le parti del discorso*.
  - *Verbi*. Gli autori riprendono il set di funzionalità proposto da François (2009), che considera come parametri l'uso dei tempi e dei modi verbali in un testo.

[169] Questo studio riprende il lavoro di François (2009), che aveva già proposto una formula di leggibilità per il francese L2 usando il metodo di regressione.

- Caratteristiche semantiche:
  - *Livello di personalizzazione*: sono definite 13 variabili che tengono conto delle proporzioni dei pronomi personali nel testo.
  - *Densità concettuale*, misurata tramite il numero di proposizioni e di diversi argomenti in una frase.
  - *Coesione lessicale*.
- Funzionalità FFL:
  - *Multi-word expressions* (MWE): sono misurate il numero, la frequenza, la struttura sintattica e la lunghezza di MWE.
  - *Tipo di testo*: gli autori definiscono 5 variabili per identificare i dialoghi, come la presenza di virgole, il rapporto di punteggiatura, ecc.

Dallo studio emerge che la famiglia lessicale è quella più predittiva, seguita da quella sintattica. Le caratteristiche semantiche e quelle specifiche FFL non hanno invece ottenuto correlazioni significative. La percentuale di parole che non si trovano nella lista di *Alter Ego* è risultata il parametro migliore.
Le funzionalità che ottengono valori più alti sono combinate in una serie di modelli, come illustrato nella Tabella 69. Oltre ai modelli SVM e quelli basati sulla regressione (OLR e MLR), che selezionano le caratteristiche in modo automatico, sono aggiunti due modelli "esperti", le cui funzionalità sono selezionate manualmente: il primo (Exp1) comprende le migliori caratteristiche di ciascuna famiglia, il secondo (Exp2) le due migliori caratteristiche.

Modello	**Classificatore**	**Caratteristiche**
Exp1	OLR, MLR e SVM	Parole non presenti nella lista Alter Ego + Lunghezza frase + Coesione + Presenza virgole
Exp2	OLR, MLR e SVM	Parole non presenti nella lista Alter Ego + 90° percentile di forme flesse + Lunghezza frase + Presenza participio presente + Coesione + % di pronomi personali + Presenza virgole + Proporzione di MWE che hanno la struttura nome+aggettivo
Auto-OLR	OLR	Parole non presenti nella lista Alter Ego + Lunghezza frase + Presenza participio presente + modello unigram
Auto-MLR	MLR	Parole non presenti nella lista Alter Ego +Presenza del condizionale + Presenza dell'imperativo + Presenza participio presente + Presenza participio passato + Presenza congiuntivo imperfetto + Presenza congiuntivo presente + Presenza virgole + TTR (type-token ratio) + Lunghezza (in parole) del 90 ° percentile della frase + % frasi più lunghe di 30 parole + n. medio di vicini ortografici più frequenti
Auto-SVM	SVM	Tutte le 46 variabili

Tabella 69. Selezione delle caratteristiche dei diversi tipi di modelli.

Per il confronto sono considerati anche altri due modelli: il modello *Random*, basato su una classificazione casuale e il modello *Baseline*, le cui variabili sono quelle delle classiche formule di leggibilità (numero di lettere per parola e numero di parole per frase).

I valori di accuratezza e di correlazione multipla ottenuti da ciascun metodo sono illustrati nella tabella seguente.

Modello	Classificatore	Correlazione	Accuratezza
Random	/	/	16,6
Baseline	SVM	0,62	34,0
Exp1	RLM	0,70	39,4
Exp2	SVM	0,73	40,8
Auto-OLR	OLR	0,71	39,6
Auto-SVM	SVM	0,73	49,1

Tabella 70. Confronto tra i migliori modelli Exp1, Exp2 e Auto + Random + Baseline.

Come si può osservare, il modello Exp1, basato sull'algoritmo RLM e comprendente 4 predittori, supera del 5% il valore di riferimento (modello Baseline); le prestazioni sono pertanto considerate significative. Il modello Exp2, che include 8 caratteristiche ed è basato su SVM, presenta valori ancora superiori (40,8% di accuratezza). La combinazione di diversi set di funzionalità sembra quindi migliorare le prestazioni dei modelli che si limitano alle singole caratteristiche.
Il metodo migliore risulta essere quello automatico basato su SVM, che presenta un'accuratezza del 49,1% e un coefficiente di correlazione multipla pari a 0,73. Questo modello include tutte e 46 le caratteristiche linguistiche delle 4 famiglie.

6.4.19. Chen et al. 2013

Chen et al. (2013) indagano sull'opportunità di utilizzare l'analisi della coesione lessicale per valutare la leggibilità di testi in cinese. Nello specifico, costruiscono catene lessicali (cioè sequenze di termini semanticamente correlati tra loro, come sinonimi, iperonimi, ecc.) per rappresentare la struttura lessicale coesa dei testi.
La relazione semantica tra le parole è determinata dai legami coesivi, individuati solitamente tramite risorse lessicografiche, come i thesaurus. In questo studio, viene utilizzato HowNet, un database lessicale per il cinese sviluppato da Zhedong Dong e Qiang Dong nel 2000.
Gli studiosi valutano una combinazione di caratteristiche legate alla frequenza delle parole e alle catene lessicali per generare modelli di classificazione della leggibilità per il cinese; i set di funzionalità sono valutati tramite le macchine a vettori di supporto (SVM).
Per l'addestramento, è utilizzato un corpus di 740 testi tratti dai libri di testo di tre materie (mandarino, educazione civica, scienze biologiche) delle scuole elementari di Taiwan. La leggibilità è classificata in tre livelli, in base ai livelli di lettura degli studenti: basso, medio e alto, che corrispondono rispettivamente a primo e secondo anno, terzo e quarto anno, quinto e sesto anno[170].

[170] In Cina esistono tre diversi sistemi scolastici per quanto riguarda la scuola dell'obbligo (la cui durata è 9 anni): quello principale, tipico delle città, prevede 6 anni di scuola primaria + 3 di scuola

Livello di lettura	Livello di istruzione	Mandarino	Educaz. civica	Scienze biologiche	N. testi
Basso	1° anno	42	0	73	115
	2° anno	56	0	55	111
Medio	3° anno	61	53	0	114
	4° anno	67	50	0	117
Alto	5° anno	83	58	0	141
	6° anno	88	54	0	142
Totale		397	215	128	740

Tabella 71. Statistiche del corpus impiegato.

L'algoritmo SVM è usato per costruire classificatori che prevedono i livelli di lettura dei testi, in particolare, classificatori binari specifici per il livello basso e il livello medio.
A differenza della maggior parte degli studi precedenti, che considera i livelli di lettura come classi discrete, in questo lavoro la classificazione è vista come un metodo continuo: l'ipotesi è che se un testo è comprensibile per studenti di un dato livello, allora deve essere comprensibile anche per studenti con un livello superiore; allo stesso modo, se uno studente può capire un testo che presenta un dato livello di difficoltà, deve anche essere in grado di comprendere qualsiasi testo che presenta un livello di difficoltà inferiore. Quindi, nella costruzione di un modello per il livello basso, saranno utilizzati i testi di livello 1 e 2 come dati positivi e gli altri come negativi; nel modello per il livello medio, invece, saranno considerati positivi tutti i documenti di livelli 1-4 e negativi quelli di livello superiore (livello 5).
Per quanto riguarda il set di caratteristiche legate alle catene lessicali, vengono utilizzati 5 parametri: numero di catene lessicali (lc-1), lunghezza media delle catene lessicali (lc-2), estensione media delle catene lessicali (lc-3), numero di catene lessicali con estensione superiore a metà della lunghezza del testo (lc-4) e numero medio di catene attive per parola (lc-5). I risultati sono mostrati nella tabella seguente.

Livello	Caratteristiche	Precisione	Recupero	Punteggio F	Accuratezza
basso	lc-1-2-3-4-5	0,76	0,57	0,65	0,81
medio	lc-1-2-3-4-5	0,70	0,83	0,76	0,68

Tabella 72. Punteggi ottenuti dal classificatore considerando il solo set di caratteristiche delle catene lessicali.

Vengono poi prese in considerazione le funzioni *tf-idf* (*term frequency–inverse document frequency*). L'indice *tf-tdf* è una funzione usata nell'*information retrieval* per ponderare i termini presenti in un documento o in una collezione di documenti, definita dal prodotto di due componenti: *tf* e *tdf*. *Tf* rappresenta la frequenza del termine nel documento e *tdf*

secondaria inferiore; il sistema diffuso nelle zone rurali prevede invece 5 anni di primaria + 4 di secondaria inferiore; esiste poi un sistema unificato, tipico delle zone periferiche, che non prevede divisioni tra cicli scolastici.

l'inverso del rapporto tra la quantità di documenti che contengono quel termine e il totale di documenti della collezione.
Per costruire i classificatori sono utilizzate le caratteristiche *tf-idf* generate dai primi 500 termini. Le Tabelle 73 e 74 mostrano i risultati della classificazione per il livello basso e quello medio.

Caratteristiche	Precisione	Recupero	Punteggio F	Accuratezza
Tf-top-50	0,78	0,87	0,82	0,88
Tf-top-100	0,81	0,86	0,83	0,89
Tf-top-200	0,80	0,89	0,84	0,90
Tf-top-300	0,82	0,89	0,85	0,90
Tf-top-400	0,86	0,89	0,87	0,92
Tf-top-500	0,84	0,89	0,87	0,92

Tabella 73. Punteggi ottenuti dal classificatore per il livello basso considerando il solo set di caratteristiche *tf-idf*.

Caratteristiche	Precisione	Recupero	Punteggio F	Accuratezza
Tf-top-50	0.81	0.88	0.84	0.79
Tf-top-100	0.81	0.90	0.85	0.81
Tf-top-200	0.83	0.92	0.87	0.83
Tf-top-300	0.86	0.90	0.88	0.84
Tf-top-400	0.82	0.92	0.87	0.83
Tf-top-500	0.82	0.95	0.88	0.84

Tabella 74. Punteggi ottenuti dal classificatore per il livello medio considerando il solo set di caratteristiche *tf-idf*.

Chen et al. considerano infine la combinazione dei due diversi set di caratteristiche (catene lessicali e funzioni *tf-idf*), come illustrato nelle Tabelle 75 e 76.

Caratteristiche	Precisione	Recupero	Punteggio F	Accuratezza
Lc + Tf-top-50	0.85	0.85	0.85	0.91
Lc + Tf-top-100	0.83	0.87	0.85	0.91
Lc + Tf-top-200	0.90	0.83	0.86	0.92
Lc + Tf-top-300	0.95	0.91	0.93	0.95
Lc + Tf-top-400	0.93	0.93	0.93	0.96
Lc + Tf-top-500	0.93	0.89	0.91	0.95

Tabella 75. Punteggi ottenuti dalla combinazione dei due set di caratteristiche per il livello basso.

Caratteristiche	Precisione	Recupero	Punteggio F	Accuratezza
Lc + Tf-top-50	0.82	0.87	0.84	0.80
Lc + Tf-top-100	0.84	0.89	0.86	0.82
Lc + Tf-top-200	0.87	0.88	0.88	0.84
Lc + Tf-top-300	0.89	0.87	0.88	0.85
Lc + Tf-top-400	0.83	0.93	0.88	0.84
Lc + Tf-top-500	0.83	0.93	0.88	0.84

Tabella 76. Punteggi ottenuti dalla combinazione dei due set di caratteristiche per il livello medio.

Come si può osservare, per entrambi i modelli (livello basso e livello medio) l'uso combinato dei due set di caratteristiche consente di migliorare le prestazioni. In particolare, nel modello costruito per il livello basso, l'uso delle sole caratteristiche legate alle catene lessicali fornisce un punteggio di accuratezza pari a 0,81; le funzioni *tf-idf* riportano valori più alti, compresi tra 0,88 e 0,92. La combinazione delle due funzionalità raggiunge invece gradi di accuratezza che vanno da 0,91 a 0,96.

Nel modello costruito per il livello medio, invece, le caratteristiche delle catene lessicali ottengono un punteggio di 0,68 e le funzioni *tf-idf* valori compresi tra 0,79 e 0,84: anche in questo caso, la combinazione dei due set consente di migliorare le prestazioni, con punteggi di accuratezza che vanno da 0,80 a 0,85.

Una sintesi dei principali lavori sulla valutazione automatica della leggibilità è presentata nella Tabella 77. Dall'analisi di questi studi possiamo trarre alcune conclusioni generali.

In primo luogo, l'approccio della classificazione, che è il metodo più impiegato, sembrerebbe anche essere quello più adatto al compito di valutare la leggibilità dei testi. Il problema della classificazione, ma vale anche per la regressione, è che richiede dati di addestramento già etichettati che potrebbero non essere disponibili, soprattutto per lingue diverse dall'inglese. Anche la costruzione manuale di un corpus di testi annotati presenta qualche complicazione, soprattutto per il fatto che le assegnazioni ai vari livelli di leggibilità da parti di giudici umani, esperti o meno, sono arbitrarie e potrebbero non essere precise. Il metodo del ranking risolve il problema della mancanza di dati di addestramento etichettati e rappresenta un'alternativa valida: i testi devono infatti essere annotati soltanto rispetto a due livelli di leggibilità (facile o difficile). Tuttavia, come notato da Tanaka-Ishii et al. (2010), questo sistema manca di *assolutezza* nel determinare una norma, riportando solo valori relativi di leggibilità.

Per quanto riguarda gli algoritmi di apprendimento, molti lavori mostrano che l'approccio basato sul *Support Vectore Machine* offre una maggiore accuratezza e una precisione in alcuni casi superiore all'80% rispetto ad altri modelli (come i Naïve Bayes e gli alberi decisionali), ma anche rispetto alle classiche formule di leggibilità (come la formula di Flesch-Kincaid). Dallo studio di Al-Kalifa e Amani (2010), per la lingua araba, emerge che il modello Naïve Bayes risulta più accurato quando viene impiegata una singola caratteristica, mentre il classificatore SVM è più accurato quando si considera una combinazione delle variabili. Alle stesse conclusioni giunge Wang (2006) per l'inglese (australiano), il quale esamina le prestazioni del metodo SVM utilizzando vari set di caratteristiche: le prestazioni sono superiori quando i set sono usati in combinazione tra loro. Sarebbe interessante verificare se questi risultati valgono anche per altre lingue, come l'italiano.

In ogni caso, osserva Collins-Thompson (2014), la scelta delle caratteristiche linguistiche da estrarre dai dati sembra avere un peso maggiore nel determinare le prestazioni del modello di apprendimento rispetto alla selezione del *contesto* di apprendimento, come la scelta del tipo di approccio o dell'algoritmo. Questa ipotesi sarebbe confermata dagli studi di Kate et al. (2010).

In molte ricerche la migliore funzionalità risulta la lunghezza media della frase, che ottiene alti valori di correlazione con i livelli di lettura, per varie tecniche di misurazione: da notare il fatto che si tratta di una delle metriche tradizionali di leggibilità. Al secondo posto si colloca il modello statistico del linguaggio. Come possiamo osservare dalla Tabella 78, che offre un riepilogo delle principali caratteristiche considerate nei vari studi, queste due funzionalità sono presenti nella quasi totalità dei lavori. L'utilizzo di modelli statistici specifici di genere per l'addestramento del sistema contribuisce a migliorare le previsioni delle leggibilità.

In tutti i casi, la combinazione di diversi tipi di funzionalità risulta essere l'approccio più efficace, raggiungendo gradi di accuratezza del 70% - 80%.

Sembra, infine, che ci sia uno spostamento rispetto ai destinatari di riferimento. Le tradizionali formule di leggibilità nascono infatti come supporto agli insegnanti, per la selezione dei materiali di lettura appropriati da sottoporre agli studenti; successivamente, le ricerche ampliano il loro campo di applicazione e iniziano a rivolgersi a tutte quelle figure che si occupano di produrre testi di vario genere, come quotidiani e riviste, pubblicità,

pubblicazioni amministrative, manuali tecnici, libri di testo e materiale per l'educazione degli adulti, ecc. Negli studi più recenti il target di riferimento non sembrerebbe più essere soltanto chi produce i testi, ma anche i destinatari stessi dei materiali di lettura, come gli studenti (di L1 o L2) o più genericamente, gli utenti web. La maggior parte dei sistemi di valutazione automatica della leggibilità non si propone, infatti, come strumento di supporto alla produzione di documenti, ma come ausilio per l'utente nell'identificazione dei testi adeguati al proprio livello di istruzione, in particolar modo nella fruizione dei contenuti web.

	Nome	Approccio	Algoritmo	Caratteristiche	Cosa valutano	Campo di applicazione	Destinatari	Lingua
Si e Callan 2001		Classificazione	Algoritmo EM	caratt. linguistiche + modello statistico unigram	Livelli di lettura di pagine web	Pagine web didattiche	Studenti	Inglese
Inui e Yamamoto 2001		Ranking	Regole di classificazione + SVM	morfosintattiche	Leggibilità dei testi	Assistenza alla lettura	Studenti	Giapponese
Liu et al. 2004		Classificazione	SVM	semantiche + sintattiche	Livelli di lettura delle ricerche degli utenti	Motori di ricerca	Studenti non udenti	Inglese
Collins-Thomson e Callan 2004	Smoothed Unigram	Classificazione	Naïve Bayes	semantiche (unigram di parole)	Livello lettura di pagine web	Pagine web + testi brevi	Utenti web	Inglese (+ francese)
Schwarm e Ostendorf 2005		Classificazione	SVM	caratt. linguistiche + modelli statistici	Livello di lettura dei testi	Inglese L2	Studenti	Inglese
Larsson 2006		Classificazione	SVM	semantiche + sintattiche + unigram	Leggibilità dei testi	Motori di ricerca	Studenti di L2	Svedese
Wang 2006		Classificazione	SVM + Naïve Bayes + Alberi decisionali	caratt. linguistiche + modello statistico unigram	Leggibilità info sanitarie dei siti web	Siti web di assistenza sanitaria	Utenti motori di ricerca + insegnanti	Inglese (Australia)
Heilman et al. 2007		Classificazione	Naive bayes + k-NN	sintattiche + modello statistico unigram	Livello di lettura di testi scritti e web	Utenti web	Utenti web	Inglese L1 e L2
Miltsakaki e Trout 2007	READ X	Classificazione	Classificatore automatico	sintattiche	Leggibilità + Contenuto tematico testi web	Motori di ricerca	Studenti e insegnanti	Inglese
Pitler e Nenkova 2008		Ranking	SVM	lessicali + sintattiche + parti del discorso	Leggibilità dei testi	Vari campi di applicazione	Studenti, pubblico adulto istruito	Inglese

	Nome	Approccio	Algoritmo	Caratteristiche	Cosa valutano	Campo di applicazione	Destinatari	Lingua
Peterson e Ostendorf 2009		Classificazione e regressione	SVM	caratt. linguistiche + modelli linguistici	Livello di lettura dei testi	Inglese L2	Studenti di L2	Inglese
Kate et al. 2010		Regressione	Algoritmi di regressione	sintattiche + lessicali + modelli ling. per generi testuali	Leggibilità dei testi	Vari campi di applicazione	Utenti generici	Inglese
Kanungo e Orr 2009		Regressione	alberi decisionali a gradiente stocastico	Caratteristiche sintattiche + lessicali + leggibilità	Leggibilità abstract dei risultati della SERP	Motori di ricerca	Utenti web	Inglese
Al-Kalifa e Amani 2010	Arability	Classificazione	SVM + Naïve Bayes + Alberi decisionali	semantiche + sintattiche	Leggibilità dei testi	Vari campi di applicazione	Vari	Arabo
Tanaka-Ishii et al. 2010	Terrace	Ranking	SVM	lessicali	Ordina due testi in base alla leggibilità	Insegnamento delle lingue	Studenti e insegnanti di lingue	Inglese e giapponese
Aluisio et al. 2010		Classificazione + ranking e regressione	SVM (class. e regres.) + meta classificatore	Coh-Metrix-Port + sintattiche + modelli linguistici	Leggibilità testi	Semplificazione dei testi	Utenti web + autori di contenuti	Portoghese (Brasile)
Dell'Orletta 2011	READ-IT	Ranking	SVM	Di base + lessicali + sintattiche + morfosintattiche	Livello leggibilità dei testi	Semplificazione dei testi	Lettori con bassa alfabetizzazione	Italiano
François e Fairon 2012		Classificazione e regressione	SVM + algoritmi di regressione	sintattiche, lessicali, semantiche e specifiche FFL	Leggibilità dei testi	Francese L2	Studenti di francese L2	Francese
Chen et a. 2013		Classificazione	SVM	Catene lessicali + funzioni tf-idf	Livelli di lettura dei testi	Insegnamento	Studenti	Cinese

Tabella 77. Approcci di valutazione automatica della leggibilità.

	Caratteristiche lessicali e semantiche	Caratteristiche sintattiche	Parti del discorso
Si e Callan 2001	unigram di parole	lunghezza della frase	
Inui e Yamamoto 2001		50 variabili morfosintattiche	
Liu et al. 2004	frequenza di n-grammi	lunghezza frase + lunghezza parole	
Collins-Thomson e Callan 2004	unigram di parole, TTR, ecc.		
Schwarm e Ostendorf 2005	trigrammi di parole	lunghezza frase + lunghezza parole	
Larsson 2006	frequenza delle parole (unigram) + parole difficili + quoziente nominale + n. nomi/pronomi + n. attributi + articoli determinativi	profondità sintattica + lunghezza frase + frasi preposizionali + n. congiunzioni subordinati + n. vocali per frase + lunghezza espressione	
Wang 2006	unigram + difficoltà parole	lunghezza frase, lunghezza parole	
Heilman et al. 2007	unigram	uso del passivo, frasi relative, tempi verbali	
Miltsakaki e Trout 2007		numero di frasi + numero di parole + numero di parole lunghe (con sette o più caratteri) + numero di lettere	
Pitler e Nenkova 2008	vocabolario (modello linguistico)	lunghezza delle frasi + lunghezza delle parole + altezza alberi di analisi + n. frasi nominali, n. frasi verbali, n. subordinate	coerenza + coesione (n. pronomi + n. articoli determinativi) + relazioni tra parti del discorso
Peterson e Ostendorf 2009	n-grammi	lunghezza frase, lunghezza parole	
Kate et al. 2010	percentuale di parole OOV + parole funzionali + n. di pronomi + n. di parole conosciute + modelli linguistici specifici per generi testuali	lunghezza delle frasi + frasi nominali + frasi verbali, ecc.	
Al-Kalifa e Amani 2010	modello statistico del linguaggio + frequenza dei termini	lunghezza frase + lunghezza parole + n. sillabe per parola	
Tanaka-Ishii et al. 2010	vocabolario (frequenza e leggibilità delle parole)		
Aluisio et al. 2010	Probabilità e perplessità di unigrammi, bigrammi e trigrammi + parole fuori vocabolario	coordinate, subordinate, passivi, frasi avverbiali, frasi relative, apposizioni + lunghezza frasi e parole	Coesione (connettivi causali, temporale, ecc.) + POS (parole funzionali, prole di contenuto)
Feng et al. 2010	Perplessità di unigrammi, bigrammi e trigrammi	altezza alberi di analisi + n. di subordinate + n. di frasi nominali + di n. frasi verbali + frasi preposizionali +la lunghezza delle frasi	Parti del discorso + carat. discorso (catene lessicali, densità entità, ecc.)
Dell'Orletta 2011	Composizione vocabolario, TTR	Caratt. Alberi sintattici + subordinazione + predicati verbali	Modello statistico POS + densità lessicale

	Caratteristiche lessicali e semantiche	Caratteristiche sintattiche	Parti del discorso
François e Fairon 2012	frequenze lessicali + % di parole fuori lista di riferimento + lunghezza parole + diversità lessicale + vicini ortografici + livello personalizzazione + densità concettuale + modelli n-gram	lunghezza frase + tempi verbali + modi verbali	relazioni parti del discorso + coesione
Chen et a. 2013	Catene lessicali + funzioni TF-IDF		

Tabella 78. Riepilogo delle caratteristiche considerate nei diversi studi di valutazione automatica della leggibilità.

6.5. Tra tradizione e innovazione: altri studi di leggibilità

Accanto alle ricerche sulla valutazione automatica della leggibilità, si collocano una serie di studi che potremmo definire *intermedi*, *di transizione*: sebbene infatti queste ricerche siano ancora in parte collegate alla misurazione tradizionale della leggibilità, costituiscono in un certo modo un superamento di queste, per la metodologia impiegata (l'analisi di una serie di variabili più complesse, valutate tramite strumenti di NLP), o per l'oggetto della valutazione (i documenti e le risorse presenti sul web). Si tratta quindi di strumenti innovativi, che però non rientrano nella categoria della valutazione automatica in quanto non fanno uso di tecniche di apprendimento automatico.

6.5.1. Coh-Metrix, Coh-Metrix Port e Coease

Una sintesi dei progressi negli studi sulla leggibilità è stata la creazione di Coh-Metrix (Graesser et al. 2004), uno strumento informatico, sviluppato presso l'università di Memphis, che analizza i testi e misura in modo automatico oltre 200 parametri linguistici: "Its modules use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other components that are widely used in computational linguistics" (Graesser et al. 2004, p. 193).
Uno dei livelli di analisi su cui si concentra in particolar modo Coh-Metrix è la coesione, aspetto che viene generalmente ignorato dalle tradizionali formule di leggibilità. "Cohesion is the linguistic glue that holds together the events and concepts conveyed within a text. Beyond the words and separate sentences in the text, are the relationships between the sentences and larger units of text. Explicit cues in the text help the reader to process, understand, or infer those relationships. […] Cohesive cues help the reader to understand connections among sentences and paragraphs. This, in turn, facilitates understanding of the words and sentences, and enhances the reader' global understanding of the text. Many studies, across a variety of paradigms and dependent measures, have shown that cohesive cues in text facilitate reading comprehension and help readers construct more coherent mental representations of text content. […] For this reason, the primary bank of indices provided by Coh-Metrix assesses the cohesion of the text" (McNamara e Graesser 2012, p. 4).
Coh-Metrix considera più di 50 aspetti legati alla coesione e più di 200 metriche linguistiche, testuali e legate alla leggibilità. Vediamone alcune.

- Informazioni sulle parole

Vengono fornite informazioni su varie proprietà linguistiche delle parole, come la *familiarità* (la frequenza con cui una parola occorre), la *concretezza*, l'*immaginabilità*, l'*età di acquisizione*, la *significatività in base al corpus Colorado*[171] e la *significatività in base alle norme di Paivio*[172].

[171] Toglia e Battig 1978.
[172] Paivio et al. 1968.

- Frequenza delle parole

La misurazione della frequenza si basa su 4 corpora: il database lessicale CELEX[173], che contiene le frequenze di circa 18 milioni di parole; l'analisi delle frequenze effettuata da Francis e Kucera (1982); la lista delle frequenze compilata da Thorndike e Lorge (1944); il conteggio delle frequenze dell'inglese parlato realizzato da Brown (1984). Il principale parametro valutato è il logaritmo medio delle frequenze delle parole.

- Parti del discorso (POS)

Misura la presenza di una determinata parte del discorso, come le parole funzionali (articoli, preposizioni, ecc.) o le parole di contenuto (nomi, verbi, aggettivi, ecc.). Coh-Metrix considera più di 50 categorie POS.

- Punteggi di densità

Misurano l'incidenza, il rapporto o la proporzione di determinate classi di parole o costituenti nel testo. Un parametro altamente predittivo della difficoltà è la densità dei pronomi, valutata tramite la proporzione di frasi nominali.

- Operatori logici

Viene calcolato un punteggio di incidenza per ciascun tipo di operatore logico e per l'insieme totale di tutti gli operatori impiegati. Se il testo ha un'alta densità di operatori logici, significa che è molto denso e ricco di informazioni.

- Connettivi

I connettivi misurano la coesione testuale. Sono considerate diverse categorie di connettivi: causali (*perché*, *così*), additivi (*e*, *inoltre*), temporali (*dopo*, *prima*, *fino a*), logici (*e*, *o*) e avversativi/contrastivi (*sebbene*, *mentre*). Vi è inoltre una distinzione tra connettivi positivi (*anche*, *inoltre*) e negativi (*tuttavia*, *ma*).

- Rapporto TTR

Misura la diversità lessicale.

- Polisemia e iperonimia

La polisemia misura l'ambiguità delle parole, l'iperonimia il grado di astrattezza.

- Complessità sintattica

Coinvolge una serie di parametri che valutano la difficoltà sintattica delle frasi. Le analisi sintattiche sono effettuate tramite il parser *ApplePie* (Sekine e Grishman 1995) e il tagger POS di Brill (1995).

- Leggibilità

La leggibilità è valutata tramite la formula Reading Ease di Flesch e l'indice Flesch-Kincaid.

- Coesione co-referenziale

Una semplice ma efficace misura della coesione è la sovrapposizione referenziale e semantica di frasi adiacenti, che possono trovarsi nello stesso paragrafo o in due paragrafi

[173] Baayen et al. 1995.

diversi: quando parole o concetti si sovrappongono, quel contenuto forma un collegamento tra le frasi. Coh-Metrix considera diverse forme di coreferenza, tra cui la sovrapposizione di nomi, di argomenti, di radici.

- Coesione causale

Riguarda il riferimento nel testo a eventi o azioni correlati causalmente. La misura della coesione causale è data dal conteggio dei verbi causali, sulla base di WordNet (Fellbaum 1998, Miller et al. 1990)[174]. Le relazioni di coesione causale sono segnalate anche da alcune particelle causali, come congiunzioni, avverbi e altri connettivi.

- LSA

L'*analisi semantica latente* (*Latent Semantic Analysis*, LSA)[175] viene adottata in Coh-Metrix come misura di coesione semantica. Considera la sovrapposizione semantica delle parole tra le frasi o tra i paragrafi. Viene misurata in diversi modi, ad esempio tramite la somiglianza tra frasi adiacenti, la somiglianza tra tutte le possibili coppie di frasi, la somiglianza tra paragrafi, ecc.

Esiste una versione gratuita di questo strumento, Coh-Metrix 2.0, la quale comprende circa 60 variabili linguistiche, da quelle più semplici, come il conteggio delle parole, a quelle più complesse, misurate tramite algoritmi specifici.

Scarton et al. (2009) presentano un adattamento di Coh-Metrix alla lingua portoghese brasiliana, chiamato Coh-Metrix-Port. Il loro studio fa parte di un più ampio progetto di semplificazione testuale, PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital, 'Semplificazione testuale del portoghese per l'inclusione e l'accessibilità digitale')[176], che prevede lo sviluppo di tecnologie per facilitare l'accesso alle informazioni sul web da parte di analfabeti e persone con deficit cognitivi, come afasia e dislessia.

---

[174] WordNet è un database semantico-lessicale, sviluppato dal linguista George Miller presso l'Università di Princeton, che si propone di organizzare, definire e descrivere i concetti rilevanti della lingua inglese. La concettualizzazione del lessico è realizzata attraverso i *synset* (da *synonym set*), insiemi di termini dal significato equivalente, strutturati in nodi e collegati da relazioni di senso, come iperonimia, iponimia, ecc. In ogni synset, le differenze di senso (polisemie) sono distinte, numerate e definite mediante relazioni tassonomiche e associative.
Nell'ambito del progetto EuroWordNet (EWN), finanziato dalla Comunità Europea dal 1996 al 1999, sono stati sviluppati lessici WordNet per i vari linguaggi europei, collegati in un database multilingue attraverso un Inter-Lingual Index (ILI). Il WordNet per l'italiano, ItalWordNet (IWN), è stato sviluppato dall'Istituto di Linguistica Computazionale del CNR di Pisa. Nella sua versione generica, il database è costituito da un wordnet (una rete semantica di concetti) contenente circa 47.000 lemmi, 50.000 synset e 130.000 relazioni semantiche.
Ancora per l'italiano, è sviluppato, presso l'istituto ITC-irst di Trento, il progetto MultiWordNet (Pianta et al. 2002), che mira a creare un Italian WordNet strettamente allineato con Princeton WordNet. MultiWordNet è un database lessicale multilingue che contiene le seguenti informazioni sull'inglese e sull'italiano: relazioni lessicali tra parole, relazioni semantiche tra concetti lessicali, corrispondenze tra concetti lessicali italiani e inglesi, campi semantici (domini).
[175] Cfr. Foltz 1996, Landauer e Dumais 1997, Landauer et al. 1998.
[176] Cfr. Aluisio et al. 2008.

La lingua portoghese non ha a disposizione molte risorse computazionali e l'unico strumento di analisi della leggibilità è una versione adattata dell'indice di Flesch (Martins et al. 1996). Questa formula è implementata in Coh-Metrix-Port.
Per quanto riguarda le altre metriche da analizzare, sono scelte due categorie di variabili dalla prima versione di Coh-Metrix: *Parole generali e informazioni sul testo* e *Indici sintattici*. Della prima classe fanno parte: conteggi di base, analisi delle frequenze, concretezza, iperonimi; della seconda: costituenti, pronomi, rapporto TTR, operatori logici e somiglianza tra le frasi.

Tonelli et al. (2012) propongono Coease, l'adattamento di Coh-Metrix alla lingua italiana. Alla base del loro studio c'è l'idea di *rendere leggibili gli indici di leggibilità*: spesso infatti i sistemi di previsione della leggibilità non riescono a mostrare in modo chiaro e comprensibile quanto sia difficile un testo e soprattutto quali siano gli aspetti che contribuiscono a determinare tale difficoltà. Gli autori presentano un sistema che, dato un testo, non solo fornisce un elenco di indici di leggibilità ispirati a Coh-Metrix, ma anche una rappresentazione grafica della difficoltà del documento rispetto ai tre livelli di istruzione italiana (primaria, secondaria di primo e secondo grado): "We believe that this kind of representation makes readability assessment more intuitive, especially for educators who may not be familiar with readability predictions via supervised classification." (Tonelli et al. 2012, p. 40).
Il corpus usato per lo studio è composto da testi in uso nelle scuole italiane: *Classe 1* comprende 63 testi destinati a bambini delle scuole elementari; in *Classe 2* sono raccolti 55 documenti per le scuole medie; *Classe 3* contiene 62 testi per le scuole superiori.
Il corpus comprende sia testi narrativi che testi espositivi, tratti da varie materie (storia, letteratura, biologia, fisica, chimica, geografia, filosofia); include anche i test di comprensione ufficiali utilizzati nelle prove Invalsi. Come si osserva nella Tabella 79, i documenti presentano un'alta variabilità, in particolare quelli appartenenti alla *Classe 3* (la deviazione standard è 1152).

	Classe 1	**Classe 2**	**Classe 3**
Lunghezza testi (in parole)	530 (±273)	776 (±758)	1085 (±1152)
GULPEASE	55,92 (±6,35)	53,88 (±6,13)	50,54 (±6,98)

Tabella 79. Statistiche del corpus. Il valore tra parentesi è la *deviazione standard*.

Come verifica, viene valutata anche la leggibilità tramite l'indice GULPEASE. Anche se i punteggi sono leggermente più alti per le elementari e scendono via via che sale il livello di istruzione, i valori risultano piuttosto simili tra loro e non sembrano rendere conto della divisione tra le 3 categorie. In particolare, i punteggi ottenuti dai testi della Classe 2 si sovrappongono con le altre classi.
Per quanto riguarda gli indici utilizzati, gli autori cercano di seguire la descrizione degli indici riportati nella seconda versione di Coh-Metrix. Risultano 46 indici, organizzati come segue:

- Indici 1 - 6: Conteggi di base
  Riguardano la lunghezza dei contenuti, in termini di sillabe, parole, frasi e paragrafi. Le sillabe sono calcolate tramite il modulo per la sillabazione Lingua::IT::Hyphenate.

- Indici 7 - 10: Frequenza
  Si focalizzano sulla familiarità delle parole *contenuto* (nomi, verbi, avverbi, aggettivi), valutando la loro frequenza; come corpus di riferimento è utilizzata Wikipedia in italiano.
- Indici 11 - 12: Iperonimia
  Calcolano l'astrattezza di nomi e di verbi, misurando la distanza tra synset WordNet contenente il lemma (senso più frequente) e la radice. Viene poi calcolata la distanza media di tutti i sostantivi e verbi nel testo. Per ottenere questo indice è usato MultiWordNet (Pianta et al. 2002).
- Indici 13 -17: Informazioni sui costituenti
  Misurano la complessità sintattica delle frasi, tramite l'incidenza delle frasi nominali e delle negazioni, il numero di parole prima del verbo principale, ecc.
- Indici 18 - 19: Pronomi, types e tokens
  L'indice 18 riguarda il rapporto tra i pronomi e l'incidenza delle frasi nominali. L'alta densità dei pronomi comporta una leggibilità ridotta poiché rende la coesione referenziale meno esplicita. L'indice 19 è il TTR (type/token ratio).
- Indici 20 - 29: Connettivi
  Rendono conto della coesione delle frasi. I connettivi sono divisi in varie categorie: additivo, causale, logico e temporale; per ciascuna classe sono identificati i connettivi positivi e quelli negativi.
- Indici 30 - 31: Somiglianza sintattica
  Misurano la somiglianza sintattica tra le frasi; si basano sul presupposto che un testo con un'elevata variabilità sintattica sia più difficile da capire.
- Indici 32 - 34: Coreferenza
  L'indice 32 calcola la sovrapposizione degli argomenti, cioè la percentuale di frasi adiacenti che condividono almeno un argomento, espresso da un sostantivo o un pronome; gli altri due indici calcolano la stessa percentuale rispetto a parole *contenuto* e a radici.
- Indici 35 - 40: Dimensioni del modello di situazione
  Esprimono il grado di complessità del modello mentale evocato da un testo; coinvolgono 4 dimensioni: causalità, intenzionalità, tempo e spazio. La coesione causale e quella intenzionale sono misurate dal rapporto tra particelle causali o intenzionali (connettivi e avverbi) e verbi causali o intenzionali. Un testo che presenta molti verbi causali e poche particelle causali sarà poco leggibile in quanto le connessioni tra gli eventi non sono espresse in modo esplicito. La coesione temporale è calcolata tramite il numero medio di ripetizioni di tempo e aspetto nel testo. La coesione spaziale riflette la misura in cui le frasi sono correlate da particelle o relazioni spaziali e corrisponde al rapporto tra il punteggio di posizione e movimento.
- Indici 41 – 46: Caratteristiche non incluse in Coh-Metrix
  Si tratta di indici aggiuntivi, non presenti in Coh-Metrix: gli indici 41 e 42 sono basati sul *Vocabolario di Base* di De Mauro (2000) e misurano la percentuale di tokens e types nel testo. Questa funzionalità è ripresa dallo studio di Dell'Orletta et al. (2011). L'indice 43 corrisponde alla valutazione della leggibilità tramite la formula GULPEASE. Gli ultimi indici (44-46) riguardano la sovrapposizione lessicale per

ciascuna delle tre classi: la sovrapposizione è calcolata confrontando le parole di un dato testo con il lessico usato in un corpus di riferimento, formato da 180 documenti (60 per ogni livello di istruzione).

ID	Caratteristiche
General word and text information	
Basic Count	
1-3	N. of words, sent. and parag. in text
4 *	Mean n. of syllables per content word
5	Mean n. of words per sentence
6	Mean n. of sentences per paragraph
Frequencies	
7	Raw frequency of content words
8	Log of raw frequency of content words
9	Min raw frequency of content words
10	Log min raw frequency of content words
Hypernym	
11	Mean hypernym values of nouns
12	Mean hypernym values of verbs
Syntactic indices	
Constituents information	
13	Noun phrase incidence
14	Mean n. of modifiers per NP
15	Higher level constituents
16	Mean n. of words before main verb
17	Negation incidence
Pronouns, Types, Tokens	
18	Pronoun ratio
19	Type-token ratio
Connectives	
20	Incidence of all connectives
21-22	Incidence of pos./neg. additive connectives
23-24	Incidence of pos./neg. temporal connectives
25-26	Incidence of pos./neg. causal connectives
27-28 *	Incidence of pos./neg. logical connectives
29	Incidence of conditional operators
Syntactic similarity	
30	Tree intersection between adj. sentences
31	Tree intersection between all. sentences
Referential and Semantic Indices	

ID	Caratteristiche
Coreference	
32 *	Adjacent argument overlap
33	Stem overlap between adjacent sentences
34	Content word overlap between adj. sent.
Situation model dimension	
35-36	Causal content and cohesion
37-38 *	Intentional content and cohesion
39-40	Temporal and spatial cohesion
Features not included in Coh-Metrix	
41 *	Lemma overlap with VBI (token-based)
42 *	Lemma overlap with VBI (type-based)
43 *	Gulpease index
44 *	Lexical overlap with Class 1
45 *	Lexical overlap with Class 2
46 *	Lexical overlap with Class 3

Tabella 80. Caratteristiche considerate in Coease. I 10 indici maggiormente correlati sono contrassegnati da (*).

Si noti (Tabella 80) il fatto che 6 dei 10 indici più correlati non fanno parte delle caratteristiche considerate da Coh-Metrix. In generale, le correlazioni risultano moderate, essendo comprese tra 0,3 e 0,6.

È disponibile un'interfaccia web di Coease che consente di valutare un testo tramite gli indici appena descritti[177]. Gli studiosi riportano un esempio della rappresentazione grafica restituita dal sistema dopo aver analizzato un articolo tratto dal mensile *Due Parole* (cfr. 5.2.4). i punteggi ottenuti dal testo sono comparati con i valori medi dei 10 indici più correlati. Come mostrato nella Figura 28, la leggibilità del documento è confrontata con i valori dei tre livelli di istruzione.

[177] La pagina di riferimento è attualmente http://terence.fbk.eu/services/api/computeReadability/v2/.

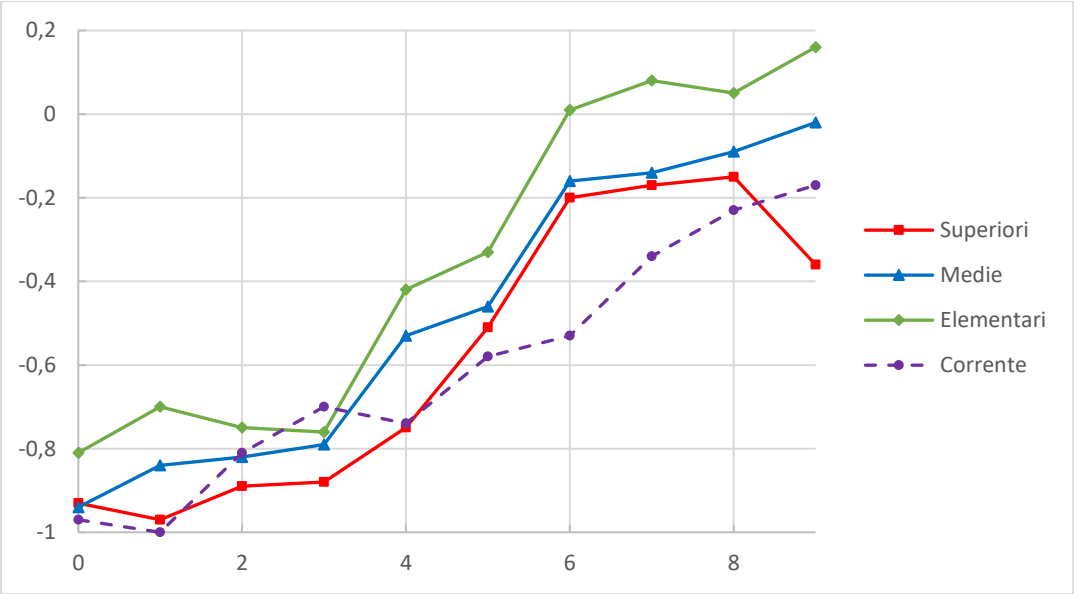


Figura 28. Rappresentazione grafica della leggibilità del testo analizzato (come indicato in Tonelli et al. 2012).

In realtà, come si può vedere nella Figura 29, attualmente l'interfaccia online restituisce una rappresentazione grafica diversa da quella indicata dagli autori. Nonostante l'intenzione dichiarata dagli autori di voler semplificare gli indici di leggibilità, la rappresentazione grafica non risulta molto intuitiva. I valori di riferimento associati ai livelli di istruzione non sono mostrati insieme ma è possibile scegliere un solo livello per volta. Oltre a questo, si pone il problema della lingua. Nonostante il modello sia sviluppato per l'italiano, sia l'esito dell'analisi che la guida relativa ai vari indici sono in lingua inglese: viene quindi chiesto agli utenti non solo di avere una certa padronanza dell'inglese, ma di conoscere (e comprendere) molti termini specialistici.

Figura 29. Rappresentazione grafica della leggibilità fornita attualmente dal sistema Coease.

6.5.2. La leggibilità dei contenuti web

La natura estremamente varia e non tradizionale dei contenuti web, dai commenti dei blog, ai post e ai tweet dei social, alle pagine dei risultati dei motori di ricerca fino alla pubblicità online, porta a nuove sfide per la previsione della leggibilità (Collins-Thompson 2014).

Oltre ad avere contenuti testuali con strutture non tradizionali, le pagine web possono contenere anche immagini, video, audio, tabelle e altri elementi di layout che influenzano la leggibilità del testo. "The ability of a user to understand a document would seem to be a critical aspect of that document's value, and yet a document's reading difficulty is a factor that has typically been ignored in designing access to Web content." (Collins-Thompson 2014, p. 114).

Uno degli aspetti spesso più trascurati è la differenza tra testi cartacei e testi sul web; in particolare, non si tiene conto di come questa diversità influisca sulle strategie di comprensione di lettura dei lettori/utenti. I lettori che leggono contenuti presenti nei siti web e cercano informazioni nei motori di ricerca adottano infatti nuovi tipi di strategie di lettura, necessarie "to learn within this interactive, informationally rich, and relatively new text environment" (Coiro e Dobler 2007).

Si pensi ad esempio agli strumenti di supporto alla navigazione, che si trasformano da sommari o indici con l'indicazione dei numeri di pagina nei testi a stampa a menù di navigazione e mappe del sito: i contenuti effettivi non sono più evidenti ma sono spesso nascosti sotto strati multipli di informazioni che si distribuiscono tramite i collegamenti ipertestuali. Ai lettori è inevitabilmente chiesto di avere un ruolo molto più attivo nel processo di comprensione della lettura: non solo sono chiamati a costruire il significato del testo ma a farlo attraverso scelte flessibili e mirate di collegamenti ipertestuali rilevanti, icone e diagrammi interattivi (Coiro e Dobler 2007). La lettura sul web richiede quindi la capacità di integrare molteplici strutture di conoscenza, adattandosi alle varie situazioni di lettura, ovvero la capacità di riassemblare in modo flessibile le conoscenze esistenti con nuove applicazioni di conoscenza personalizzate per ogni nuova situazione di lettura (Spiro 2004).

Va inoltre considerato il fatto che gli utenti leggono contenuti testuali sul web in modo diverso da quelli cartacei. Non è più una lettura lineare e completa ma una lettura selettiva: l'utente "scansiona" il testo alla ricerca del contenuto rilevante, ignorando tutto il resto, come i contenuti aggiuntivi, gli annunci pubblicitari, ecc.

Coiro e Dobler (2007) sottolineano inoltre che i testi presenti sul web costituiscono una sfida aggiuntiva anche rispetto agli ipertesti informativi. Con ipertesti, gli autori intendono i testi informativi che si trovano nei sistemi ipertestuali chiusi (come una biblioteca online o un'enciclopedia su dvd) e che si differenziano dai testi web che si trovano in un sistema aperto, quello della rete. Un ambiente ipertestuale chiuso è tipicamente un sistema statico: gli utenti accedono generalmente dallo stesso punto di partenza e seguono percorsi sempre all'interno del sistema; anche gli strumenti di ricerca forniscono un insieme finito e costante di risultati. Al contrario, un sistema aperto come il web cambia ogni giorno nella struttura e nei contenuti: gli utenti accedono ai testi da innumerevoli punti di partenza, incontrando spesso strutture testuali incoerenti, elementi di distrazione come annunci pubblicitari, collegamenti ipertestuali non funzionanti e una quantità infinita di informazioni del tutto estranee al loro scopo di lettura.

Cosa si intende quindi per *leggibilità* in un contesto web? Molti studi associano la leggibilità al concetto di *usabilità* dei siti web; altri la considerano come un parametro per il posizionamento (*ranking*) nei motori di ricerca; altri, come una metrica per classificare i risultati delle ricerche sul web in base ai livelli di lettura o per filtrare i documenti in base alle capacità di lettura degli utenti.

Yu e Miller (2010) definiscono la leggibilità dei contenuti web come "a combination of reading comprehension, reading speed and user satisfaction" (p. 2523). Esistono diversi fattori "fisici" che possono influenzare la leggibilità di un testo sul web, come il tipo di carattere impiegato, la sua dimensione, l'interlinea del testo, la suddivisione in paragrafi, la presenza di elementi che mettono in risalto parole o parti del testo più rilevanti, il contrasto dei colori, ecc. I due autori propongono Jenga Format, un metodo per trasformare i contenuti testuali al fine di migliorare la leggibilità delle pagine web. La trasformazione dei contenuti testuali si basa su due elementi che influiscono sulla lettura: la separazione tra le frasi e la spaziatura all'interno di un paragrafo. Lo studio è rivolto in particolare a lettori non madrelingua inglese che si trovano a leggere contenuti sul web principalmente in questa lingua.

Gli studiosi testano Jenga Format su 30 lettori asiatici che hanno una buona padronanza dell'inglese e dimostrano che il loro metodo è in grado di migliorare la comprensione della lettura (e dunque la soddisfazione dell'utente) senza influire negativamente sulla velocità di lettura. Yu e Miller presentano anche Froggy, un'estensione per il browser Firefox che implementa il metodo Jenga e aggiunge due nuove funzionalità: l'identificazione automatica degli elementi di distrazione (come pubblicità, animazioni, loghi, video, immagini) e l'enfatizzazione delle parole chiave (cioè fornisce una sintesi dei contenuti tramite parole chiave).

Ali et al. (2013) valutano gli effetti che il tipo di font può avere sulla leggibilità di un testo web (in lingua malese). In particolare analizzano 4 tipologie di caratteri, alcuni con grazie (*serif*), altri senza (*san serif*): due progettati per la stampa, Times New Roman (*serif*) e Arial (*san serif*), e due progettati specificamente per adattarsi alla lettura tramite lo schermo di un computer, Georgia (*serif*) e Verdana (*sans serif*). Nonostante i font sviluppati per la lettura al computer risultino migliori in un ambiente web, non vi sono sostanziali differenze tra la leggibilità di caratteri con o senza grazie.

Gradišar et al. (2006) invece valutano gli effetti che specifici design dei siti web hanno sulla velocità di lettura: testano 30 diverse combinazioni di colori e dimostrano che queste differenze non influiscono in modo significativo sulla leggibilità del documento.

Oltre a questi studi che si occupano di migliorare l'esperienza di lettura sul web ce ne sono altri che si concentrano sul contenuto stesso delle pagine web: come per i testi cartacei, la leggibilità è associata a quelle caratteristiche linguistiche che influiscono sulla comprensione del materiale di lettura. Come già notato, molte di queste ricerche sono ancora in parte collegate alla misurazione tradizionale della leggibilità: salvo qualche raro caso, si tratta principalmente di studi in cui si applicano le classiche formule di leggibilità ai testi presenti sul web.

In questa sezione cerchiamo di fornire una panoramica di questi lavori, segnalando quegli esempi che sembrano maggiormente rappresentativi sia per quanto riguarda le diverse lingue della ricerca, sia per quanto riguarda i vari campi di applicazione.

Gottron e Martin (2009, 2012) si occupano di valutare la leggibilità di pagine web tramite una combinazione di strumenti tradizionali, come la formula di Flesch e la formula SMOG, e algoritmi che consentono l'estrazione di contenuti. Questo studio è uno dei pochi in cui viene preso in considerazione il problema del *rumore:* nei siti web infatti sono presenti tutta una serie di elementi, come i menù di navigazione, gli elementi del layout, i collegamenti ipertestuali, ecc. che possono alterare il punteggio di leggibilità. La valutazione della leggibilità delle pagine web dovrebbe invece essere limitata alle sole parti di contenuto testuale.
Uno dei metodi per eliminare il rumore è tramite dei filtri che "puliscono" i documenti web: ad esempio, il processo di *Content Extraction* (CE) consente di recuperare da documenti HTML quelle parti che rappresentano il contenuto testuale principale[178].
Per l'estrazione dei contenuti i due studiosi impiegano l'algoritmo *Document Slope Curves* (DSC), sviluppato da Pinto et al. (2002), che identifica le parti del documento che contengono la maggior parte del testo ed esclude altri tag. Utilizzano anche l'approccio del *content code blurring* (CCB), proposto da Gottron (2008), che sfrutta le caratteristiche visive delle diverse tipologie di contenuto: i contenuti aggiuntivi sono in genere altamente formattati e contengono brevi porzioni di testo, mentre i contenuti principali sono formattati in modo omogeneo e sono piuttosto lunghi. È possibile quindi identificare queste tipologie all'interno del documento HTML calcolando i rapporti del codice del contenuto a livello del carattere. Un CCB adattato (ACCB) è anche in grado di ignorare i tag dei collegamenti ipertestuali.
Il corpus su cui si basa lo studio è composto da 1.114 documenti tratti da 5 siti web (BBC News, The Economist, Herald Tribune, MSNBC News e Yahoo News). Gottron e Martin confrontano i valori di leggibilità delle pagine web complete, dei contenuti principali (estratti in modo manuale) e delle versioni dei documenti su cui è effettuata l'estrazione dei contenuti tramite l'algoritmo DSC o il metodo ACCB. La tabella seguente presenta i risultati ottenuti tramite la formula SMOG.

Sito web	N. testi	Testi completi	Contenuti principali	ACCB	DSC
BBC News	337	4,0569	4,8323	4,9360	4,8052
The Economist	53	4,2486	5,0578	5,1433	5,0835
Herald Tribune	300	4,0891	5,0477	5,0650	5,0412
MSNBC News	197	4,4675	4,8949	4,9050	4,8491
Yahoo News	227	4,2063	4,9416	4,7563	4,7670

Tabella 81. Valori di leggibilità secondo la formula SMOG per i diversi siti (completi e dopo l'estrazione di contenuti).

I dati mostrano che entrambi gli approcci di CE riportano valori di leggibilità più accurati rispetto a un indice calcolato in modo automatico sull'intera pagina web. Come valori di riferimento si considerano i punteggi dei documenti estratti a mano.

[178] Una valutazione approfondita dei diversi algoritmi di estrazione dei contenuti (CE) è fornita da Gottron 2007.

Uitdenbogerd (2006), dell'università di Melbourne in Australia, valuta la leggibilità di testi web e testi cartacei, mettendo a confronto le caratteristiche linguistiche che possono influire sulla difficoltà. Scopo del progetto è sviluppare un metodo che consenta all'utente di migliorare le proprie competenze linguistiche attraverso l'accesso a materiali di lettura appropriati provenienti dal web; i principali destinatari sono gli studenti che studiano l'inglese come seconda lingua.
La leggibilità è misurata tramite alcuni indici tradizionali: la formula di Flesch, la formula di Flesch-Kincaid, la formula ARI, la formula di Coleman e Liau, l'indice Fog, la formula SMOG e la formula Lix.
I risultati mostrano che spesso le formule sottostimano il livello di leggibilità delle pagine web e che dunque andrebbero modificate per adattarsi ai contesti web.
Nonostante le ricerche più recenti, ad esempio quelle che si basano su modelli statistici del linguaggio, abbiano consentito migliori risultati nella valutazione dei livelli di lettura delle pagine web rispetto alle formule tradizionali, l'autrice non ritiene che questi strumenti siano adatti nel caso delle lingue straniere. Non esclude, tuttavia, che alcuni approcci, come quello di Schwarm e Ostendorf (2005), possano esser applicabili all'apprendimento delle lingue.

Lau e King (2006) propongono uno schema di valutazione bilingue (inglese e cinese) per la leggibilità delle pagine web basato sulle caratteristiche testuali. La leggibilità è misurata tramite una nuova formula, creata dalla combinazione dell'indice di Flesch (per la lingua inglese) e l'indice di Yang (per il cinese)[179].

$$r_p = \begin{cases} -84{,}6X_{E1} - 1{,}015X_{E2} + 206{,}835 \text{ per l'inglese} \\ 2 \times \begin{Bmatrix} 13{,}90963 + 1{,}54461X_{C1} + 39{,}01497X_{C2} - 2{,}52206X_{C3} - \\ 0{,}29809X_{C4} + 0{,}36192X_{C5} + 0{,}99363X_{C6} - 1{,}64671X_{C7} \end{Bmatrix} \text{ per il cinese} \end{cases}$$

Dove:
X_{E1} = numero di sillabe per parola;
X_{E2} = lunghezza media della frase;
X_{C1} = proporzione di frasi complete;
X_{C2} = proporzione di parole che si trovano nella lista di base del cinese;
X_{C3} = numero medio di battute per carattere;
X_{C4} = percentuale di caratteri in gruppi di 5 battute (su 100 caratteri);
X_{C5} = percentuale di caratteri in gruppi di 12 battute (su 100 caratteri);
X_{C6} = percentuale di caratteri in gruppi di 22 battute (su 100 caratteri);
X_{C7} = percentuale di caratteri in gruppi di 23 battute (su 100 caratteri);
I punteggi della formula vanno da 0 a 100, dove un valore più basso indica una maggiore difficoltà del testo.

Ancora legato ad un contesto bilingue è lo studio di Hussain et al. (2011).
Il Pakistan ha due lingue ufficiali: l'inglese e l'urdu. L'urdu è la lingua nazionale, parlata da oltre il 75% dei pakistani; tuttavia rappresenta soltanto l'8% della lingua principale della

[179] Cfr. 4.1.5

popolazione. L'inglese viene insegnato come lingua straniera ma è utilizzato solo in ambiti ristretti, ad esempio negli atti governativi; soltanto il 10% della popolazione conosce la lingua inglese. Oltre a queste, si parlano più di 60 lingue, tra cui alcune lingue provinciali (le principali sono: Punjabi, Sindhi, Pashto e Balochi) e due lingue regionali (dialetto Saraiki e Kashmir). Risulta evidente il fatto che la comprensione di pagine web scritte in inglese sia un grande problema per parte dei lettori pakistani.
Nel loro studio, Hussain et al. propongono una doppia versione di siti web: una originale in inglese e una versione modificata e semplificata (in *plain language*) in lingua locale (urdu), utilizzando però l'alfabeto inglese. L'efficacia del modello è testata su un campione di 40 soggetti con diversi livelli di padronanza dell'inglese. I risultati mostrano che la modifica dei contenuti contribuisce a migliorare la leggibilità dei siti web, aumentando il livello di comprensione da parte dei lettori.

Garais (2011) conduce uno studio per valutare la leggibilità dei contenuti presenti sul sito web di un'agenzia di stampa rumena[180]. L'autore misura inizialmente la leggibilità tramite l'indice di Flesch ma i risultati restituiscono valori incoerenti; dopo altri esperimenti, risulta che le uniche formule in grado di valutare la lingua rumena sono quella italiana e quella rumena, più vicine dal punto di vista lessicale e sintattico. In particolare, Garais utilizza per l'italiano la versione di Flesch adattata da Vacca (1972) e per lo spagnolo l'adattamento di Huerta (1952). La formula di Kandel e Moles (1958), che costituisce l'adattamento dell'indice di Flesch al francese, non risulta invece adatta a valutare il rumeno.

Al-Badi et al. (2012) presentano una rassegna delle ricerche che riguardano l'usabilità e l'accessibilità dei siti web, analizzando gli strumenti esistenti e le linee guida fornite dai vari autori. Una delle sezioni di studio riguarda propriamente la leggibilità: in particolare, gli autori si concentrano sulle linee guida per scrivere sul web e sulle limitazioni delle attuali formule di leggibilità.

Guo, Zhang e Zhai (2011) propongono di integrare un indice di leggibilità nel motore di ricerca di Twitter. Il sistema consentirebbe, tramite una specifica applicazione, di calcolare, secondo l'indice di Flesch e l'indice di Flesch-Kincaid, il livello di leggibilità dei vari *tweet*, cioè i messaggi di testo scritti dagli utenti del social network[181].
Il loro studio ha finalità principalmente didattiche: un sistema di recupero di tweet etichettati consentirebbe infatti agli studenti di *seguire* quegli utenti che presentano un valore di leggibilità adeguato al loro livello di lettura, aumentandone così la comprensione. Sarebbe tuttavia possibile estendere il campo di applicazione alle varie tipologie di materiale online.
"The idea behind the project may have some potential for not only educational but also academic or commercial purposes, as it is practicable to retrieve a variety of data through web services in the age of Web 2.0. The readability index can be integrated into search

[180] Il sito web di riferimento è www.amosnews.ro.
[181] I tweet presentano diverse specificità; la caratteristica più immediata è la loro brevità: fino al 2017 la lunghezza massima prevista per i tweet era infatti di 140 caratteri; dalla fine del 2017 la lunghezza è aumentata a 280, ad esclusione dei paesi asiatici il cui il sistema di scrittura richiede meno caratteri.

results of webpages, blogs, archives, databases, e-books, book reviews, forum postings, online newspapers and magazines, Wikipedia, government or law document, online textbook, Really Simple Syndication feed, or Yahoo Answers. The possibilities are almost endless." (Guo et al. 2001, p. 104).

Concludiamo questa rassegna con un interessante studio che riguarda la leggibilità dei risultati delle ricerche effettuate su Google.
Nel 2010 Google introduce, per la lingua inglese, una nuova funzionalità nella ricerca avanzata, ovvero la possibilità di visualizzare (ed eventualmente filtrare) i siti web recuperati per livello di lettura. Le pagine web sono classificate in 3 livelli: base (livello elementare), intermedio e avanzato (il livello di lettura che si trova in Google Scholar). Nella pagina di documentazione relativa al *Reading level* si legge: "Sometimes you may want to limit your search results to a specific reading level. For instance, a junior high school teacher looking for content for her students or a second-language learner might want web pages written at a basic reading level. A scientist searching for the latest findings from the experts may want to limit results to those at advanced reading levels"[182].

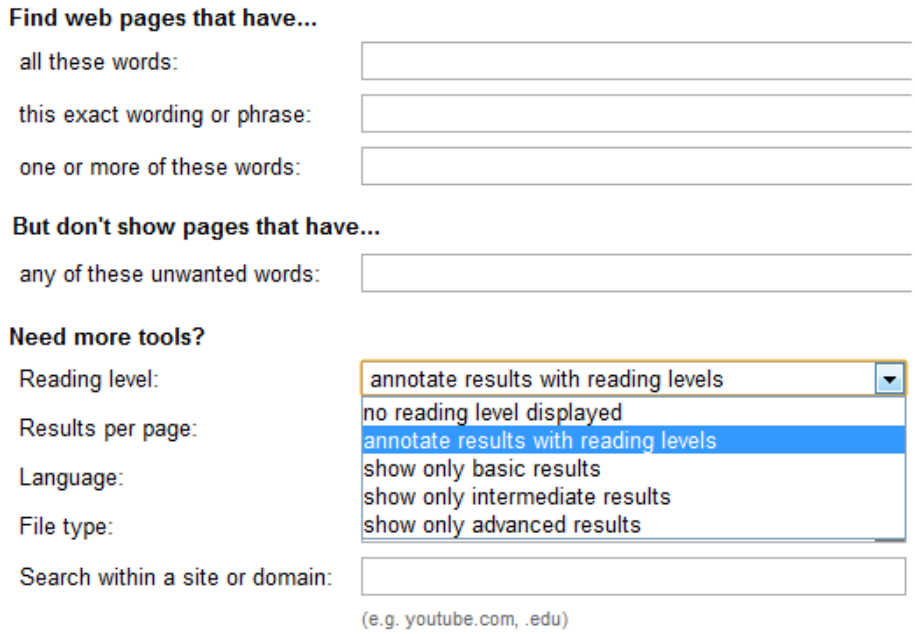


Figura 30. Ricerca avanzata di Google nel 2010.

La funzionalità si basa su modelli statistici del linguaggio costruiti sulle classificazioni effettuate da vari insegnanti: le parole presenti in una pagina web sono confrontate con le parole dei modelli linguistici così da classificare la pagina in uno dei tre livelli di lettura.

[182] La pagina non è più attiva ma può essere recuperata tramite Wayback Machine, un archivio digitale del web creato da Internet Archive; il servizio consente di visualizzare versioni archiviate di pagine web, raccolte periodicamente, conservate e rese disponibili, sotto forma di immagini, in ordine cronologico di acquisizione.

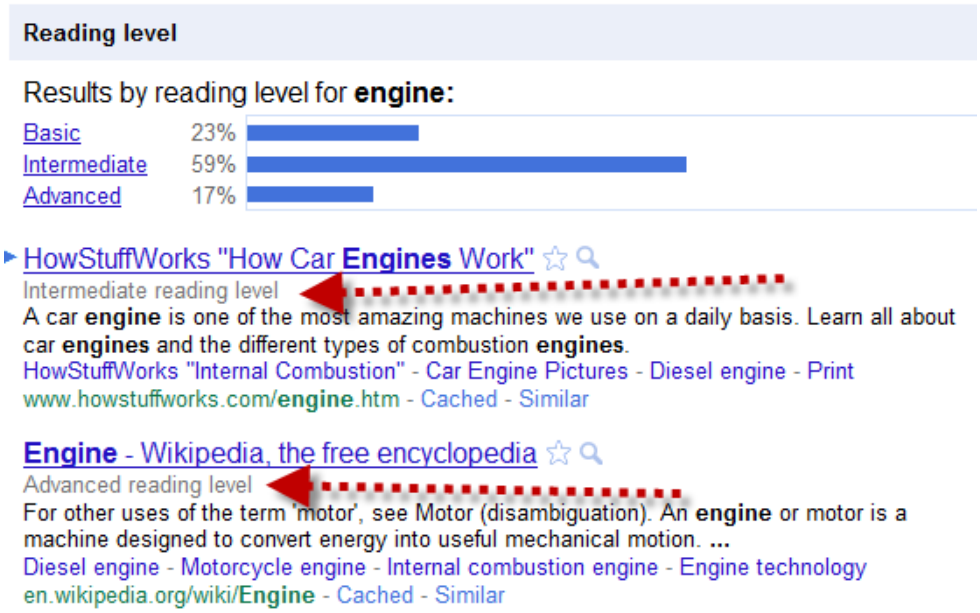


Figura 31. Risultati della ricerca classificati in livelli di lettura.

Come si vede nella Figura 31, nella parte superiore della pagina che restituisce i risultati della ricerca, viene mostrato un filtro, affiancato da un grafico che mostra le percentuali di pagine etichettate da Google per quel dato livello di lettura. Sotto ogni risultato viene visualizzato (in grigio) il livello di lettura corrispondente. È possibile filtrare i risultati, visualizzando solo quelli annotati in uno specifico livello, ad esempio quello *base* (Figura 32).

Figura 32. Risultati della ricerca filtrati per il solo livello *base*.

La funzionalità viene rimossa nel 2015 e dunque non è più possibile effettuare ricerche per valutare il modello di classificazione impiegato da Google. L'unico studio rintracciato è quello di Bilal (2013), in cui si confronta il *Reading level* di Google con due indici di leggibilità tradizionali (la formula di Flesch e quella di Flesch-Kincaid).

Nello studio vengono utilizzate 15 query di ricerca formulate da bambini delle scuole medie, 5 corrispondenti ad una sola parola, 5 composte da due parole e 5 scritte in linguaggio naturale. Per quanto riguarda il primo gruppo, ai bambini è chiesto di trovare informazioni sul pellegrinaggio Hajj, sui giochi minoritari in Gran Bretagna, sul modo di trovare la direzione usando il sole e un bastone, sull'alta pressione sanguigna e su un argomento a scelta libera collegato al medioevo. Le parole scelte per le interrogazioni dagli studenti sono: *Hajj*, *Rugby*, *compass*, *hypertension*, *plague* ('peste').
Per le query del secondo gruppo viene chiesto ai bambini di scegliere un argomento a piacere o, in alternativa, di cercare informazioni su uno dei temi proposti (Clint Dempsey, habitat delle tarantole, social network, ferrovie sotterranee, riserve di petrolio). Le formulazioni dei bambini sono: *Clint Dempsey*, *tarantula habitat*, *social networking*, *underground railroad*, *oil reserves*.
Per quanto riguarda il terzo gruppo, si sottopongono ai bambini alcune domande: *come funziona il cuore?*; *È necessario conoscere le medicine che possono aiutare le persone a smettere di fumare; usando Internet trova tre di queste medicine*; *In quale anno è stato introdotto il pattinaggio di velocità nelle Olimpiadi moderne?*; *In che modo i tipi di sogni e i sogni ad occhi aperti influenzano il modo in cui dormiamo?*; *Gli ambientalisti sono preoccupati del fatto che lo strato di ozono si sta esaurendo. Scopri come la mancanza di ozono nell'atmosfera terrestre sta influenzando le nostre foreste*. Le corrispondenti interrogazioni da parte degli studenti sono: *How does heart work*; *medicine stop smoking*; *dreams affecting sleep*; *what year speed skating Olympics*; *ozone affecting forests.*
Per queste 15 query, Google recupera 300 risultati ed assegna loro il livello di lettura corrispondente (Tabella 82). Come si può osservare, a un'alta percentuale di risultati non viene assegnato alcun livello di lettura.

Query	Avanzato	Intermedio	Base	Non assegnato
Una parola (5)	32%	34%	14%	22%
Due parole (5)	8%	50%	32%	10%
Linguaggio naturale (5)	38%	20%	38%	4%

Tabella 82. Percentuali dei risultati recuperati da Google in base alle 15 query.

I livelli attribuiti da Google sono confrontati con i punteggi di leggibilità ottenuti dall'applicazione delle formule di Flesch e Flesch-Kincaid (Tabella 83). La formula Flesch-Kincaid è misurata in gradi di istruzione (ad esempio 9 = 9° grado); la scala di difficoltà della formula di Flesch è così definita:

- 0-29 = molto difficile
- 30-49 = difficile
- 50-59 = abbastanza difficile
- 60-69 = standard
- 70-79 = abbastanza facile
- 80-89 = facile
- 90-100 = molto facile

Query	Link/Frammenti		Testi pagine web		Google			
	Flesch	F-K	Flesch	F-K	A	I	B	n. d.
Una parola								
Hajj	29,95	11,59	43,73	11,03	20%	60%	10%	10%
Rugby	55,71	9,07	52,43	11,23	40%		30%	30%
Compass	37,90	10,81	30,06	14,31	10%	60%	30%	
Hypertension	35,26	10,62	33,69	17,18	80%	10%		10%
Plague	30,26	11,40	34,93	13,64	50%	40%		10%
Due parole								
Clint Dempsey	58,86	8,50	48,47	12,64		40%	30%	30%
Oil reserves	43,35	10,34	32,28	14,79	30%	60%		10%
Social Network	29,15	11,35	25,84	15,28	10%	60%	30%	
Tarantula habitat	29,37	11,11	45,00	10,76		60%	30%	10%
Undergrond railroad	38,87	10,11	36,01	14,64		30%	70%	
Linguaggio naturale								
Dreams affecting sleep	35,80	12,23	51,44	10,48	30%	40%	20%	10%
Speed skating Olympics	38,17	11,36	48,91	12,15		20%	80%	
Ozone affecting forest	41,18	11,42	28,63	14,90	70%		30%	
How does heart work	54,48	9,45	67,39	7,26	50%	20%	20%	10%
Medicine stop smoking	27,35	13,16	54,13	9,46	40%	20%	40%	

Tabella 83. Confronto tra i livelli di lettura assegnati da Google e i punteggi di leggibilità ottenuti con le due formule.

Se si considerano le query formate da una parola sola, vediamo che i punteggi di leggibilità ottenuti tramite la formula di Flesch sono simili per entrambe le categorie (link/frammenti di testo e contenuti delle pagine web) e si collocano tra *difficile* e *abbastanza difficile.* La formula Flesch-Kincaid fornisce per la prima tipologia valori corrispondenti ai gradi 9-12 e per la seconda valori tra il grado 11 e il 17 (livello universitario). Google assegna la maggior parte dei risultati al livello avanzato o intermedio (A: 40%, I: 34%, B:14%, n.d.:12 %).

Per quanto riguarda le query composte da due parole, l'indice di Flesch valuta la classe link/frammenti come *difficile* e *abbastanza difficile*, mentre inserisce i testi tra *difficile* e *molto difficile*. L'indice di Flesck-Kincaid riporta rispettivamente valori compresi tra i gradi 9-11 e tra i gradi 10-15. Rispetto a questi punteggi, Google sembra sottostimare il livello di difficoltà, assegnando la maggior parte dei risultati al livello intermedio e a quello base (A: 8%, I: 50%, B:32%, n.d.:10 %).

Le query basate su formulazioni scritte in linguaggio naturale mostrano una maggiore variabilità: l'indice di Flesch restituisce punteggi che vanno da 27 a 54 (*abbastanza difficile - molto difficile*) per i link/frammenti e da 28 a 67 (*molto difficile - standard*) per i testi.

L'indice Flesch-Kincaid riporta valori corrispondenti ai gradi 9-13 per la prima tipologia e 7-15 per la seconda. Anche in questo caso, vi è discrepanza tra le formule di leggibilità e il *Reading level*: Google recupera un 38% di risultati associati al livello avanzato ma anche un 38% associati al livello base, seguiti dal 20% di risultati a livello intermedio e un 4% a cui non è assegnato alcun grado di lettura.
Dal confronto emerge quindi che i punteggi di leggibilità ottenuti con le formule tradizionali non corrispondono ai livelli di lettura attribuiti dal modello di Google. "The mismatch between Google's assigned readability to text and the F-K Grade Level found in this study suggest that either the engine's Reading Level algorithm is inadequate, or that the Flesch formulae may not be valid for predicting the readability of Web text (links, snippets, and corresponding pages) since it was originally developed for measuring the readability of print instead of Web text" (Bilal 2013, p. 8).

Nonostante la funzionalità *Reading level* sia stata dismessa, c'è chi sostiene che Google continui a calcolare i livelli di leggibilità delle pagine web e che usi questo parametro come fattore di ranking. In un interessante articolo rintracciato in rete, *Leggibilità e motori di ricerca*, l'autore si chiede appunto se la leggibilità sia un fattore che aiuta il posizionamento[183].
Ogni anno l'azienda Searchmetrics fornisce un'analisi dettagliata dei fattori di posizionamento impiegati dai motori di ricerca, in particolare Google. Nello studio condotto nel 2015, *Ranking Factors 2015*, viene analizzata, tra i fattori relativi al *contenuto*, la correlazione tra il posizionamento in Google e il punteggio di leggibilità dei siti ottenuto tramite la Formula di Flesch.
"When it comes to search rankings, the importance of good quality, relevant content cannot be understated. Once again this year we have carried out detailed analyses of key content ranking factors including word count and Flesch readability. The aim is to give a clearer insight into which aspects of content in particular can improve the overall ranking of your site. As the trend away from keywords and towards relevant content continues, high-ranking sites are shifting their focus from using keywords based on search queries to trying to understand the user's intention as a whole" (Tober et al. 2015, p. 39). Come possiamo notare dalla figura, i siti web con un più alto grado di posizionamento presentano anche valori di leggibilità più alti; la tendenza è inoltre cresciuta dal 2014 al 2015. Dal 2016 la *rilevanza del contenuto* viene inserita tra i principali fattori di ranking.

[183] Federico Sasso, *Leggibilità e motori di ricerca,* 20 settembre 2016. L'articolo è reperibile al seguente link: https://visual-seo.com/it/SEO-Blog/Readability-and-Search-Engines#readability-and-search-engines.

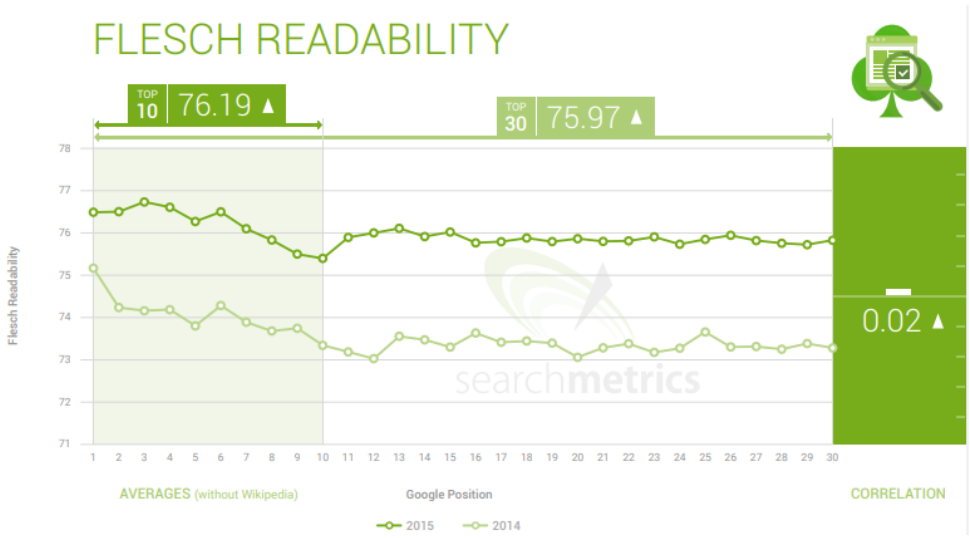


Figura 33. La correlazione tra il posizionamento e il valore di leggibilità dei siti web (l'immagine è tratta da *Ranking Factors 2015*).

6.5.2.1. Il corpus PAISÀ

In contesto italiano, vale la pena segnalare il corpus PAISÀ (Piattaforma per l'Apprendimento dell'Italiano Su Corpora Annotati)[184], sviluppato nell'ambito dell'omonimo progetto (2009-2012)[185]. Nonostante il progetto preveda l'utilizzo di metodi di apprendimento automatico per la classificazione dei testi web per quanto riguarda l'argomento, l'intenzione comunicativa e il genere, la valutazione della leggibilità resta ancora legata alle metriche tradizionali. Per questo motivo, sembra più appropriato inserire tale studio in questa sezione.

PAISÀ è un *web corpus*, cioè un corpus di testi in italiano contemporaneo scaricati dal web, ideato principalmente con finalità glottodidattiche, come supporto all'apprendimento e insegnamento dell'italiano come lingua straniera, ma reso inoltre disponibile per scopi di ricerca, anche di tipo statistico-quantitativo[186].

La novità, rispetto ad altri web corpora di italiano esistenti, è l'utilizzo di documenti non soggetti al vincolo di copyright, ma distribuiti con licenze *Creative Commons* (CC), che permettono una maggiore libertà e flessibilità nella gestione della ridistribuzione dell'opera.

Va precisato che la scelta di utilizzare esclusivamente le licenze CC influisce sulla composizione del corpus e sulla presenza delle tipologie testuali: ad esempio, il 63% dei documenti scaricati tramite il motore di ricerca proviene da blog personali o istituzionali che trattano tipicamente argomenti connessi al sociale, alla politica, all'informatica e al mondo dei motori. Sono, inoltre, inevitabilmente escluse dal corpus molte forme di comunicazione elettronica, come la posta elettronica, chat, forum, social network, ecc. "In

[184] Cfr. Borghetti et al. 2011, Lyding et al. 2014.

[185] Al progetto PAISÀ (2009-2012), cofinanziato dal Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) tramite il Fondo per gli Investimenti della Ricerca di Base (FIRB), collaborano quattro partner: Università di Bologna, CNR di Pisa, Accademia Europea di Bolzano e Università di Trento.

[186] Il corpus è reso accessibile e interrogabile tramite un'interfaccia specificamente concepita per studenti e insegnanti di italiano sul sito web del progetto http://www.corpusitaliano.it; è inoltre possibile scaricare gratuitamente l'intero corpus, sia nella versione annotata che non annotata.

altre parole, la presenza della licenza CC restringe considerevolmente la varietà dei documenti in termini di genere e argomento, tanto che risulta impossibile affermare che PAISÀ sia rappresentativo della comunicazione elettronica nel suo complesso" (Borghetti et al. 2011, p. 151).

I testi del corpus sono stati selezionati utilizzando due criteri. Il primo, ispirato al progetto WaCky (Baroni et al. 2009)[187], prevede che si identifichino le URL dei documenti da scaricare effettuando ricerche per combinazioni casuali di parole su un motore di ricerca (in questo caso Yahoo!). Per PAISÀ, le parole utilizzate sono state tratte dal *Vocabolario di Base* della Lingua Italiana (De Mauro 1989), organizzate in una lista di 50.000 coppie. La ricerca è limitata alle pagine in lingua italiana con licenza *Creative Commons* (CC) che prevedono la possibilità di riutilizzo e condivisione all'interno di un'altra opera[188]. Una volta ottenuta la lista delle URL, sono eliminate le pagine erroneamente condotte alle licenze CC, sulla base di una *black list* realizzata manualmente in precedenti versioni del corpus e si procede con lo scaricamento e ripulimento dei documenti con il sistema KrdWrd (Steger e Stemle 2009).

La seconda componente del corpus comprende testi provenienti dalle versioni italiane di alcuni dei progetti web di Wikimedia Foundation (Wikipedia, Wikinews, Wikisource, Wikibooks, Wikiversity, Wikivoyage). Invece di scaricare i siti, sono utilizzati i *dump* ufficiali rilasciati da Wikimedia Foundation[189].

Il corpus completo contiene circa 388.000 documenti da 1.067 siti diversi, per un totale di circa 250 milioni di parole. Di questi, circa 269.000 testi provengono dai progetti di

---

[187] Il progetto WaCky (*Web as Corpus kool ynitiative*) comprende un insieme di corpora linguistici di alcune delle principali lingue europee, ognuno dei quali contiene da 1,5 a 2 miliardi di parole: itWac per l'italiano, ukWaC per l'inglese e il deWac per il tedesco. L'obiettivo era quello di creare corpora di lingua generale (non specialistici) di grandi dimensioni, utilizzando il web come fonte di dati linguistici. I corpora sono stati creati tra il 2005 e il 2007 attraverso il *web crawling*, utilizzando cioè un programma (*crawler*) per la raccolta e lo scaricamento di pagine dal web. Seguendo la metodologia esposta in Baroni et al. (2009), il meccanismo di funzionamento del crawler è il seguente: si identificano differenti URL (cioè di indirizzi Internet) che possano essere considerate rappresentative sia dei contenuti che dei generi testuali. I siti web sono indentificati tramite il motore di ricerca Google, in risposta alle query formulate con 1000 coppie di parole di contenuto. Per l'italiano le coppie sono derivate da una selezione di parole che hanno una frequenza media nel corpus di Repubblica e dal vocabolario di base dell'italiano, da cui sono escluse le parole funzionali. La lista delle URL individuate viene assegnata al programma, che si occupa di scaricare le pagine, tramite BootCat (Baroni e Bernardini 2004). Successivamente, nella fase di *post-crawl*, vengono filtrate le pagine e si effettuano dei passaggi di "pulizia" del testo, rimuovendo il codice HTML e quelle parti che generano rumore, come i menù di navigazione, le intestazioni e altro materiale privo di contenuto testuale. Per ognuno dei corpora, è effettuata la tokenizzazione, la lemmatizzazione e l'annotazione delle parti del discorso (POS tagging). Il corpus itWac (http://wacky.sslmit.unibo.it/doku.php) contiene 2 miliardi di token; per il POS tagging è stato utilizzato il programma TreeTagger e per la lemmatizzazione MORPH-IT!

Altri *web corpora* realizzati tramite *web crawling* sono Webbit, un corpus di 150 milioni di parole, realizzato da Marco Baroni e liberamente accessibile e i TenTen corpora, una serie di risorse multilingui che comprende arabo, cinese, inglese, francese, tedesco, italiano, giapponese, coreano, portoghese, russo e spagnolo. Un metodo alternativo è presentato nell'ambito del progetto RIDIRE (Moneglia e Paladini 2010), che prevede la costruzione di un corpus di italiano scritto derivato dal web tramite la procedura del *crawling mirato*, in cui i siti sono selezionati da esperti dei vari domini.

[188] Sono quindi incluse le licenze: CC-Attribuzione, CC-Attribuzione-Condividi allo stesso modo, CC-Attribuzione-Non commerciale, CC-Attribuzione-Non commerciale-Condividi allo stesso modo. Per approfondimenti si veda http://creativecommons.it/Licenze.

[189] Il *dump* di un database è un file che contiene il contenuto e la struttura del database; viene usato solitamente per il backup.

Wikimedia Foundation (con circa 263.300 pagine da Wikipedia, 1.680 da Wikinews, 410 da Wikisource, 2.380 da Wikibooks, 740 da Wikiversity, 390 da Wikivoyage). I documenti sono raccolti tra settembre e ottobre 2010.

I testi hanno un doppio livello di annotazione: una prima annotazione linguistica e un'annotazione tramite metadati che riguardano l'argomento, l'intenzione comunicativa e il genere testuale.

L'annotazione linguistica comprende la divisione in frasi, la tokenizzazione, la lemmatizzazione, l'annotazione morfosintattica (tramite il POS tagger utilizzato in Dell'Orletta 2009) e l'analisi delle dipendenze tramite il parser DeSR (Attardi et al. 2009). Il POS tagger, che rappresenta lo stato dell'arte per la lingua italiana, con un'accuratezza che raggiunge il 97,10%, ha dimostrato alla sua prima applicazione al corpus PAISÀ un'accuratezza del 95.10%. Per adattarlo ai testi del web, sono stati effettuati due cicli di correzione manuale di 20.000 token ciascuno[190] e due conseguenti fasi di riaddestramento, riuscendo infine ad ottenere un'accuratezza del 96,03%, difficilmente incrementabile.

L'inserimento dei metadati relativi a genere, argomento e l'intenzione comunicativa mira alla costruzione di un classificatore automatico che sia in grado di associare in modo affidabile questi parametri ai testi. Per ciascun parametro sono definite un insieme di classi e sono condotti 4 cicli di annotazione manuale su campioni casuali di testi. Per verificare l'appropriatezza delle tassonomie, sono confrontate le categorie con i dati risultanti dall'applicazione di metodi non supervisionati di classificazione dei testi.

Per quanto riguarda l'argomento, si deve considerare il fatto che le pagine web tendono ad avere la tendenza alla sovrapposizione di argomenti diversi: per risolvere il problema, è stato stabilito di attenersi a criteri quantitativi e annotare il documento in base all'argomento prevalente.

Le 8 categorie relative all'argomento, individuate a partire dalla proposta di Sharoff (2004), sono le seguenti.

Argomento	**Descrizione**
Business	Economia, commercio, finanza, lavoro, ecc.
Arti	Arti visive, letteratura, architettura, cinema, musica, ecc.
Hi-tech	Informatica, computer, web, telefonia, elettronica, ecc.
MSP	Medicina, Salute, Psicologia, ecc.
Leisure	Tv, moda, astrologia, sport, videogiochi, viaggi, cucina, ecc.
FLERS	Filosofia, Lingue, Educazione/formazione, Religione, Sociologia, ecc.
S. Naturali	Scienze naturali: meteorologia, astronomia, biologia, fisica, chimica, matematica, geografia, ecc.
PSS	Politica, Società, Storia: istituzioni, amministrazione (trasporti, esercito, ecc.), legge, geopolitica, ecologia, etica, ecc.

Tabella 84. La tassonomia degli argomenti del corpus PAISÀ.

[190] Il primo ciclo è effettuato su un campione casuale, il secondo su testi selezionati in base alle difficoltà riscontrate nelle fasi precedenti dell'annotazione. I cicli hanno messo in evidenza errori di segmentazione del testo, abbreviazioni/sigle precedentemente non considerate, emoticon e altre sequenze 'anomale' di caratteri.

La validità della tassonomia è sperimentata tramite un modello di clustering (Sharoff 2010) che raggruppa i testi del corpus in 20 argomenti (*cluster*), in base alla presenza di parole chiave tipiche dei diversi argomenti. Dal confronto emerge che 7 delle 8 categorie sono confermate: solo la classe dedicata alle *scienze naturali* non trova riscontri nei risultati del clustering, probabilmente perché nel corpus ci sono pochi testi appartenenti a tale dominio.
Anche la classificazione delle intenzioni comunicative è ispirata al lavoro di Sharoff (2004).

Intenzione	**Descrizione**
Raccomandare	Raccomandare, consigliare, convincere, persuadere, ecc.
Informare	Informare, descrivere, presentare e raccontare, esprimere se stessi/raccontarsi, ecc.
Argomentare	Argomentare, discutere, commentare e valutare
Intrattenere	Intrattenere, divertire
Istruire	Dare istruzioni, insegnare

Tabella 85. La tassonomia delle intenzioni comunicative del corpus PAISÀ.

Questo parametro riporta i peggiori dati relativi all'accordo degli annotatori: se infatti, durante il quarto ciclo di annotazione, l'annotazione è risultata omogenea per quanto riguarda l'86,7 % dei testi in relazione all'argomento e il 78.7% in relazione al genere, sulle intenzioni comunicative non si è superato il 73.3%. Il disaccordo è dovuto principalmente all'ambiguità tra testo informativo e testo argomentativo.
Per la definizione delle categorie dei generi testuali, gli studiosi fano riferimento alla ricca bibliografia in merito dedicata ai corpora in lingua inglese, in particolare Santini (2005, 2011), Lee (2001), Rehm et al. (2008). La tassonomia dei generi possiede una struttura gerarchica a due livelli, come mostrato nella tabella seguente.

Blog	**Genere - 1° livello**	**Genere - 2° livello**
	Fiction	Prosa – Poesia - Sceneggiatura
✓	Guida	Tutorial – FAQ – Turismo - Ricetta
✓	Giornalismo	Cronaca – Editoriale – Intervista – Reportage - Recensione
	Accademia	Prosa – Lezione - Abstract
	Doc. ufficiale	Legge- Relazione - Contratto
	Scheda	Prodotto – Curriculum Vitae – About page
	Annuncio	
	Commento	
	Lemma	

Tabella 86. La tassonomia dei generi del corpus PAISÀ.

Nella classificazione, il blog assume un ruolo particolare: non è considerato né un genere di primo livello né di secondo livello, ma viene trattato come un formato, un contenitore in cui possono essere pubblicati tutti i generi individuati. Viene quindi incluso nella tassonomia come attributo opzionale da aggiungere alla classificazione di genere.
Per quanto riguarda invece la valutazione della leggibilità, vengono calcolate alcune statistiche:

- La lunghezza della frase misurata in parole;
- Il numero di frasi per testo;
- Il rapporto type/token;
- La percentuale di parole non appartenenti al *Vocabolario di Base* (De Mauro 1991);
- Il punteggio di leggibilità secondo l'indice GULPEASE.

Questi valori sono codificati come metadati e possono essere utilizzati per filtrare i documenti e creare sottocorpora.

Il lavoro futuro sarà concentrato sulla classificazione automatica di generi e intenzioni comunicative. In una prima fase, il classificatore sarà addestrato esclusivamente su caratteristiche linguistiche, come la lunghezza media delle frasi, il rapporto tra *content* e *function words*, la punteggiatura, combinazioni di parti del discorso, frequenze di parole, ecc. In un secondo momento, potranno essere combinate anche informazioni estratte dal layout delle pagine web (immagini, formattazione, link, ecc.) e dal codice HTML.

6.6. READ-IT: uno strumento italiano

Per quanto riguarda la lingua italiana, il primo – e attualmente unico – strumento di valutazione automatica della leggibilità è rappresentato da READ-IT (Dell'Orletta et al. 2011b), realizzato dall'*Italian Natural Language Processing Laboratory* (ItaliaNLP Lab) dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa[191]. READ-IT nasce come supporto al processo di semplificazione dei testi e pertanto si rivolge a un pubblico di destinatari specifico, cioè lettori caratterizzati da una bassa alfabetizzazione o da lieve deficit cognitivo.
Uno degli aspetti innovativi di READ-IT è che la valutazione della leggibilità è effettuata su due livelli: il documento e la singola frase. "La valutazione rispetto alla frase rappresenta un'importante novità dell'approccio sottostante a READ-IT, che riveste un ruolo centrale quando la valutazione della leggibilità sia finalizzata alla semplificazione del testo. Attraverso l'identificazione dei luoghi di complessità del testo (in termini di frasi) che necessitano di revisione e semplificazione, accompagnata da una classificazione semantica del tipo di difficoltà riscontrata (di natura lessicale vs grammaticale), READ-IT può anche essere utilizzato come ausilio per la semplificazione del testo"[192].

[191] Una descrizione completa dell'approccio sottostante a READ-IT è presentata anche in Dell'Orletta et al. 2014a; per la valutazione della leggibilità a livello della frase si veda Dell'orletta et al. 2014b. Sul rapporto tra leggibilità e genere testuale cfr. Dell'Orletta et al. 2012 e 2013. Per le applicazioni di READ-IT si veda: nell'ambito medico, Dell'Orletta et al. 2016, Dell'Orletta et al. 2017, Venturi et al. 2015; nell'ambito giuridico e amministrativo, Brunato e Venturi 2014, Brunato e Venturi 2016 e Brunato 2014.
[192] READ-IT *Documentazione Demo online*, ItaliaNLP Lab, novembre 2012, p. 2.

In linea con gli approcci più recenti, la misurazione della leggibilità è considerata un compito di classificazione, nello specifico una classificazione binaria (ranking) che distingue tra due livelli di lettura (*facile* e *difficile*). Per la creazione del modello viene impiegato il metodo SVM (usando LIBSVM); l'annotazione morfosintattica è effettuata con il PoS tagging descritto in Dell'orletta (2009) e l'analisi delle dipendenze tramite il parser DeSR (Attardi et al. 2009)[193].
Le caratteristiche considerate nell'approccio sono organizzate in quattro categorie principali:

- Caratteristiche di base (si riferiscono al testo non elaborato)
    - Lunghezza della frase
    - Lunghezza delle parole
- Caratteristiche lessicali
    - Composizione del vocabolario
        - Percentuale di parole appartenenti al VdB
        - Ripartizione dei lemmi in FO, AU, AD
    - Rapporto type/token (TTR)
    - Densità lessicale
- Caratteristiche morfosintattiche
    - Modello statistico delle parti del discorso
    - Modi verbali
- Caratteristiche sintattiche
    - Probabilità incondizionata dei tipi di dipendenza
    - Caratteristiche della profondità dell'albero sintattico
        - Profondità dell'albero di analisi
        - Profondità media di strutture nominali complesse
        - Profondità media delle catene di subordinazione
    - Caratteristiche di subordinazione
        - Distribuzione delle proposizioni subordinate rispetto alle principali
        - Posizione delle subordinate rispetto alla principale
    - Caratteristiche dei predicati verbali
        - Numero di radici verbali (arità verbale)
        - Numero di dipendenti per testa verbale
    - Lunghezza delle relazioni di dipendenza

Le caratteristiche di base corrispondono a quelle tipicamente usate nelle tradizionali formule di leggibilità, ovvero la lunghezza della frase (numero medio di parole per frase) e

[193] Tali strumenti rappresentano lo stato dell'arte per quanto riguarda la lingua italiana: sono infatti risultati gli strumenti più precisi e affidabili nell'ambito della campagna di valutazione di strumenti per l'analisi automatica dell'italiano, EVALITA-2009. In particolare, il modulo di annotazione morfo-sintattica ha dimostrato un'accuratezza del 96,34% nell'identificazione simultanea della categoria grammaticale e dei tratti morfologici associati; il modulo di annotazione sintattica a dipendenze realizzato dal parser DeSR raggiunge livelli di LAS e UAS in linea con lo stato dell'arte dell'analisi a dipendenze, pari a 83,38% e 87,71%. LAS (Labelled Attachment Score) è una metrica che indica la proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda sia la testa sintattica sia la dipendenza che le lega. UAS (Unlabelled Attachment Score) è una metrica che indica la proporzione di parole del testo che hanno ricevuto un'assegnazione corretta per quanto riguarda l'identificazione della testa sintattica.

la lunghezza delle parole (numero medio di caratteri per parola). Sono i due parametri utilizzati anche nell'indice GULPEASE.
Come caratteristiche lessicali sono misurate la composizione del vocabolario, il rapporto type/token e la densità lessicale (calcolata come la proporzione di parole piene, o parole contenuto, sul totale delle occorrenze). Per quanto riguarda il vocabolario dei testi, viene preso come riferimento il *Grande Dizionario Italiano dell'uso* (GRADIT)[194] e sono calcolate la percentuale di parole appartenenti al *Vocabolario di Base* (VdB) e la distribuzione dei lemmi rispetto ai 3 repertori d'uso: fondamentale (FO), alto uso (AU) e alta disponibilità (AD). Il rapporto type/token rappresenta una misura della ricchezza lessicale di un testo; essendo un indice sensibile alla lunghezza del testo, viene calcolato su campioni con la stessa dimensione (le prime 100 parole del testo).
Tra le caratteristiche morfosintattiche sono prese in considerazione: il modello statistico del linguaggio basato sulle probabilità degli unigrammi delle parti del discorso (cioè la distribuzione delle categorie grammaticali nel testo) e la distribuzione dei modi verbali.
Per quanto riguarda le variabili sintattiche, sono misurate: la probabilità incondizionata dei diversi tipi di dipendenze sintattiche (ad esempio *soggetto*, *oggetto diretto*, *modificatore*, ecc.), le caratteristiche della profondità dell'albero sintattico, le caratteristiche della subordinazione (% delle proposizioni subordinate rispetto alle principali, posizione delle subordinate rispetto alla principale), le caratteristiche dei predicati verbali (numero di radici verbali, numero di dipendenti per testa verbale, ecc.), la lunghezza delle relazioni di dipendenza (calcolata come la distanza in parole tra la testa e il dipendente).
La misura della profondità dell'albero sintattico è un parametro rilevante per valutare la complessità di un testo e riguarda i livelli di incassamento gerarchico (cioè se esistono diverse proposizioni subordinate all'interno dello stesso periodo e se siano incassate l'una dentro l'altra). Include diverse misure: profondità dell'albero di analisi, calcolata come la distanza massima che intercorre tra la radice dell'albero e una foglia; profondità media di strutture nominali complesse, cioè costituite da una testa nominale e da modificatori aggettivali e/o complementi preposizionali; profondità media delle catene di subordinazione, ovvero la distribuzione delle probabilità di catene di proposizioni subordinate incassate.
Per l'addestramento sono impiegati due campioni: un corpus di giornali, i cui testi sono tratti dal quotidiano *La Repubblica* (Rep) e un corpus di giornali di facile lettura, i cui testi sono presi da Due Parole (2Par)[195]. La scelta di selezionare due corpora appartenenti allo stesso settore consente di evitare interferenze sulla misurazione delle leggibilità dovute alla variazione del genere testuale. A differenza degli studi di Schwarm e Ostendorf (2005) e Peterson e Ostendorf (2009), in cui è impiegata una doppia versione dell'Enciclopedia Britannica (versione completa e versione semplificata, rivolta ai bambini), in questo caso non vi è una correlazione diretta tra gli articoli dei due corpora: la comparabilità è garantita dall'appartenenza allo stesso genere testuale (prosa giornalistica).
Le prestazioni del modello sono testate in tre diversi set di esperimenti: classificazione della leggibilità a livello di documento, classificazione della leggibilità a livello della frase, identificazione delle frasi facili all'interno dei testi difficili. Al livello delle frasi, si pone infatti il problema che non tutte le frasi che si trovano nei testi complessi sono a loro volta difficili

[194] De Mauro 2000.
[195] Cfr. 5.2.4

(mentre è quasi sempre vero il contrario, cioè che nei testi semplificati le frasi possono essere considerate di facile lettura). Per risolvere il problema, gli studiosi introducono la nozione di *distanza* rispetto alle frasi di facile lettura.
Per quanto riguarda la classificazione della leggibilità dei documenti, è utilizzato un corpus composto da 638 testi, di cui 319 tratti da 2Par (che rappresenta la classe *facile*) e 319 da Rep (che rappresenta la classe *difficile*). Viene effettuata una validazione incrociata di 5 volte. Per la classificazione a livello della frase, viene usato un corpus di addestramento formato da circa 6.000 frasi (3.000 da 2Par e 3.000 da Rep) e un test set di 1.000 frasi (500 da 2Par e 500 da Rep). Per il terzo esperimento viene impiegato un corpus di dimensioni maggiori (2,5 milioni di parole), i cui testi sono tratti dal quotidiano La Repubblica (Rep 2.5), per un totale di 123.171 frasi.
L'analisi della leggibilità è effettuata in base a 4 diversi modelli di analisi

- Modello base: sono valutate soltanto le caratteristiche di base (questo modello può essere considerato come un'approssimazione dell'indice GULPEASE);
- Modello lessicale: valuta una combinazione di caratteristiche di base e caratteristiche lessicali;
- Modello morfosintattico: valuta una combinazione di caratteristiche di base, lessicali e morfosintattiche;
- Modello sintattico: combina tutti i tipi di caratteristiche: di base, lessicali, morfosintattiche e sintattiche.

Le caratteristiche utilizzate per la valutazione della leggibilità a livello di documento differiscono da quelle utilizzate a livello di frase. In particolare, nel modello lessicale e in quello sintattico, per quanto riguarda la valutazione delle frasi, non vengono conteggiati: il rapporto type/token, la profondità media delle catene di subordinazione e le caratteristiche dei predicati verbali.
La Tabella 87 mostra i punteggi ottenuti dai diversi modelli nella classificazione a livello dei documenti.

Modello	**Accurat.**	**2Par (prec.)**	**2Par (recup.)**	**Rep (prec.)**	**Rep (recup.)**
Base	76,65	74,71	80,56	78,91	72,73
Lessicale	95,45	95,60	95,30	95,31	95,61
Morfosintattico	**98,12**	**98,12**	**98,12**	**98,12**	**98,12**
Sintattico	97,02	97,17	96,87	96,88	97,18

Tabella 87. Risultati della classificazione dei documenti.

Come si può osservare, il modello morfosintattico risulta essere il migliore. Il modello base restituisce le prestazioni più basse ma i valori di accuratezza aumentano considerevolmente (da 76,65% a 95,45%) nel modello lessicale, cioè quando si considerano sia le caratteristiche di base che quelle lessicali. I risultati della classificazione a livello della frase sono mostrati nella Tabella 88.

Modello	Accurat.	2Par (prec.)	2Par (recup.)	Rep (prec.)	Rep (recup.)
Base	59,6	55,6	95,0	82,9	24,2
Lessicale	61,6	57,3	91,0	78,1	32,2
Morfosintattico	76,1	72,8	83,4	80,6	68,8
Sintattico	**78,2**	**75,1**	**84,4**	**82,2**	**72,0**

Tabella 88. Risultati della classificazione delle frasi.

In questo caso, le prestazioni migliori sono riportate dal modello sintattico, cioè quello che considera tutti i set di caratteristiche. Si nota che, rispetto alla classificazione dei documenti, la differenza tra i punteggi ottenuti dal modello lessicale, quello morfosintattico e quello sintattico è maggiore: se nella classificazione dei documenti i valori oscillano di un 2,6%, in quella a livello della frase variano di un 17%.
La Tabella 89 illustra i risultati del terzo esperimento: sono valutate le prestazioni del modello sintattico (che è risultato essere il migliore per la valutazione delle frasi di Rep 2.5) nel test set Rep.

	Distanza
Corrette	52,072
Test set Rep	45,361
Sbagliate	37,843

Tabella 89. Distanza tra 2par e Rep valutata con il modello sintattico.

Come già accennato, viene introdotta una nuova metodologia di valutazione, basata sulla *distanza euclidea* tra i vettori di caratteristiche. Ogni vettore di caratteristiche rappresenta un insieme di frasi: due vettori con distanza 0 rappresentano lo stesso insieme di frasi, cioè quelle frasi che condividono gli stessi valori per le caratteristiche linguistiche misurate; maggiore è la distanza tra due vettori, più distanti saranno i gruppi di frasi rispetto alle caratteristiche monitorate.
La nozione di *distanza* viene utilizzata per distinguere le classificazioni errate da quelle corrette, cioè per valutare quelle frasi che sono classificate erroneamente come appartenenti a 2Par (perché di facile lettura). Sono calcolate:

- La distanza tra 2Par e le 140 frasi classificate erroneamente come appartenenti a 2Par (37,843);
- La distanza tra 2Par e le 360 frasi classificate correttamente come appartenenti a Rep (52,072);
- La distanza tra 2Par e l'intero set Rep.

Come si può osservare, la distanza tra 2Par e le frasi erroneamente classificate è molto più bassa rispetto a quella tra 2Par e le frasi classificate correttamente: questo perché le frasi classificate come 2Par (e quindi di facile lettura) sono più facili rispetto a quelle classificate correttamente (come appartenenti a Rep).

6.6.1. Leggibilità e generi testuali

Diversi studi (Kate et al. 2010, Štajner et al. 2012) hanno dimostrato che i metodi di valutazione della leggibilità sono più accurati se si utilizzano moduli specifici per ogni genere testuale. La distribuzione delle varie caratteristiche linguistiche cambia infatti a seconda del genere testuale preso in esame. Alcune caratteristiche possono essere comuni in alcune varietà ed essere rare in altre; si pensi ad esempio alla distribuzione di nomi e pronomi nei testi letterari o nei manuali scolastici (nei primi si ha un uso maggiore di pronomi, nei secondi si usano meno pronomi ma si ripetono spesso i nomi). "This suggests that textual genre and readability do not represent orthogonal dimensions of classification, but intertwined notions whose complex interplay needs to be further investigated in order to envisage solutions which could be successfully exploited in real educational applications" (Dell'Orletta et al. 2012, p. 92).

Dell'Orletta et al. (2012) conducono uno studio sull'italiano per verificare se gli strumenti di leggibilità sviluppati per un uso generale risultano affidabili se applicati a testi appartenenti a diversi generi testuali. Il lavoro è strutturato in due parti: in primo luogo, gli autori confrontano i risultati ottenuti dalla classificazione di documenti, utilizzando sia un modello generale che modelli specifici di genere e dimostrano che la valutazione della leggibilità è dipendente dal genere. In secondo luogo, propongono un approccio alternativo ai modelli di classificazione specifici per genere, che essi ritengono troppo impegnativi, cioè un metodo di ranking basato sulla nozione di distanza per la valutazione della leggibilità, ma che possa essere utilizzato anche per la costruzione automatica di corpora di addestramento specifici di genere.

Il training corpus utilizzato nella ricerca è formato da 4 diversi generi testuali: giornalismo, letteratura, materiale didattico e prosa scientifica[196]. Ogni genere è poi suddiviso in due sottoclassi, in base alla tipologia di destinatario presa come riferimento: il giornalismo comprende un corpus di giornali i cui testi sono tratti da La Repubblica (Rep) e un corpus di giornali di facile lettura i cui testi sono tratti da Due Parole (2Par); la scrittura educativa e la letteratura sono divise in testi che si rivolgono ad adulti (AdEdu e AdLit) e testi rivolti a bambini (ChildEdu e ChildLit); la prosa scientifica comprende articoli scientifici (ScientArt) e articoli tratti da Wikipedia (Wiki). Il corpus di letteratura per adulti fa parte del corpus italiano PAROLE (Marinelli et al. 2003) e comprende 44 romanzi pubblicati tra il 1974 e il 1989; anche il corpus di letteratura per bambini fa parte di un corpus più ampio (Marconi et al. 1994) e include romanzi rivolti a bambini della scuola primaria. La composizione del corpus è mostrata nella tabella seguente.

Nome	Corpus	Genere	N. testi	N. parole
Rep	La Repubblica (Marinelli et al. 2003)	giornalismo	321	232.908
2Par	Due Parole (Piemontese 1996)	giornalismo	322	73.314
ChildLit	Letteratura per bambini (Marconi et al. 1994)	letteratura	101	19.370
AdLit	Letteratura per adulti (Marinelli et al. 2003)	letteratura	327	471.421

[196] Uno studio approfondito sul solo genere letterario (in particolare sulla prosa narrativa italiana) è condotto in Dell'Orletta et al. 2013.

Nome	Corpus	Genere	N. testi	N. parole
ChildEdu	Materiali per la scuola primaria (Dell'Orletta et al. 2011a)	educazione	127	48.036
AdEdu	Materiali per scuola secondaria (Dell'Orletta et al. 2011a)	educazione	70	48.103
Wiki	Wikipedia (articoli tratti da "Ecologia e ambiente")	prosa scientifica	293	205.071
ScientArt	Articoli scientifici	prosa scientifica	84	471.969

Tabella 90. Statistiche dei corpora.

Come algoritmo di apprendimento è impiegato SVM (usando LIBSVM); l'annotazione morfosintattica è effettuata con il PoS tagging descritto in Dell'Orletta (2009) e l'analisi delle dipendenze tramite il parser DeSR (Attardi et al. 2009). Per la valutazione della leggibilità è utilizzato lo strumento READ-IT.

Per dimostrare che la misurazione della leggibilità è dipendente dal genere sono condotte due serie di esperimenti, volte a dividere i documenti in due classi, i testi di facile lettura e quelli difficili: nella prima è utilizzato un modello di addestramento generale, nella seconda sono impiegati modelli specifici di genere. Nel primo set di esperimenti sono testati tre modelli, che si distinguono in base ai dati di addestramento: un corpus di testi appartenenti allo stesso genere (giornalismo), un corpus di testi appartenenti a due generi diversi (2Par è scelto come rappresentante della classe di facile lettura, ScientArt della classe di testi difficili) e un corpus costruito combinando tutti i testi facili e tutti i testi difficili per ciascun genere.
La Tabella 91 mostra i punteggi ottenuti da ciascun genere testuale in base ai tre modelli descritti (2Par/Rep Model, 2Par/ScientArt Model e All Easy/All Difficult Model) e in base a modelli specifici di genere.

	2Par/Rep Model			2Par/ScientArt Model			All Easy/All Difficult Model			Modelli specifici di genere		
Genere	**Prec.**	**Rec.**	**Punt. F**	**Prec.**	**Rec.**	**Punt. F**	**Prec.**	**Rec.**	**Punt. F**	**Prec.**	**Rec.**	**Punt. F**
2Par	100	96.67	98.30	50.85	100	67.41	93.55	96.67	95.08	100	96.67	98.30
Rep	96.78	100	98.36	100	3.33	6.45	96.55	93.33	94.91	96.77	100	98.36
	Accuratezza: 98.33			Accuratezza: 51.67			Accuratezza: 95			Accuratezza: 98.33		
ChildLit	0	0	0	46.81	73.33	57.14	100	46.67	63.63	84.61	73.33	78.57
AdLit	50	100	66.67	38.46	16.67	23.25	65.22	100	78.95	76.47	86.67	81.25
	Accuratezza: 50			Accuratezza: 45			Accuratezza: 73,33			Accuratezza: 80		
ChildEdu	90	31.03	46.15	49.15	100	65.91	56.67	58.62	57.63	78.79	89.65	83.87
AdEdu	59.18	96.67	73.42	0	0	0	58.62	56.67	57.63	88.46	76.67	82.14
	Accuratezza: 64.41			Accuratezza: 49.15			Accuratezza: 57.63			Accuratezza: 83.05		
Wiki	100	20	33.33	81.25	86.67	83.87	47.17	83.33	60.24	53.57	100	69.77
ScientArt	55.55	100	71.43	85.71	80	82.76	28.57	6.67	10.81	100	13.33	23.53
	Accuratezza: 60			Accuratezza: 83.33			Accuratezza: 45			Accuratezza: 56.67		

	2Par/Rep Model			2Par/ScientArt Model			All Easy/All Difficult Model			Modelli specifici di genere		
TOT Facili	97.78	36.97	53.66	54.31	89.91	67.72	66.40	71.43	68.82	74.30	89.91	81.37
TOT Difficili	61.34	99.17	75.80	71.43	25	37.04	69.34	64.17	66.67	87.37	69.17	77.21
	Accuratezza: 68.20			Accuratezza: 57.32			Accuratezza: 67.78			Accuratezza: 79.51		

Tabella 91. Risultati della classificazione in base ai diversi modelli.

Per quanto riguarda i modelli generali, i dati mostrano che i metodi di classificazione sono accurati solo quando analizzano documenti appartenenti allo stesso genere di quelli usati nell'addestramento, come si può vedere dai punteggi ottenuti dal modello 2Par/Rep (98,33%). Negli altri casi, i valori risultano piuttosto bassi.
Per quanto riguarda i modelli specifici per generi testuali, si nota che complessivamente l'accuratezza risulta maggiore. L'unica eccezione è rappresentata dalla classificazione dei documenti di prosa scientifica, la cui accuratezza sembra piuttosto bassa (56,67%). Ciò dipende probabilmente dal fatto che i testi che compongono il corpus Wiki sono eterogenei; è possibile che il set non includa solo documenti che risultano di facile lettura rispetto alla classe ScientArt, ma anche testi più tecnici e dunque appartenenti alla classe difficile.
In base a risultati, sembra comunque confermata l'ipotesi che la leggibilità sia strettamente correlata al genere testuale: le prestazioni sono migliori se viene costruito un modello specifico per ciascun genere.

Nella seconda parte del loro lavoro, Dell'Orletta et al. (2012) presentano un approccio alternativo alla risoluzione di questo problema, ovvero un metodo di ranking in grado di assegnare un punteggio di leggibilità ai documenti senza richiedere corpus di addestramento specifici per genere. Questo metodo si basa sulla nozione di distanza tra vettori di caratteristiche, già illustrata in Dell'Orletta et al. (2011b). Il punteggio è calcolato come la combinazione lineare tra la distanza di un documento (d) e due vettori n-dimensionali che rappresentano la classe facile (EV) e difficile (DV):

$$readability(d) = CosineDistance(d, EV) - CosineDistance(d, DV)$$

Il punteggio va da 1 (difficile) a -1 (facile). La classe facile è rappresentata da 2Par e quella difficile da ScientArt; il resto dei vettori è così ordinato:

2Par < ChildEdu < ChildLit < Rep < Wiki < AdLit < AdEdu < ScientArt

La Tabella 92 mostra i risultati della classificazione dei documenti in base a questa metodologia. Si nota che, per ogni genere, il numero di documenti facili è più alto nei primi gruppi di documenti.

	Giornalismo		Letteratura		Educazione		Prosa Scientifica	
	2Par	Rep	ChildLit	AdLit	ChildEdu	AdEdu	Wiki	ScientArt
0 - 30	15	0	4	0	8	0	3	0

	Giornalismo		Letteratura		Educazione		Prosa Scientifica	
31 – 60	6	1	11	0	9	0	3	0
91 – 90	4	6	7	6	3	1	1	0
91 – 120	1	5	1	12	2	5	4	0
121 – 150	2	3	2	7	5	6	4	1
151 – 180	1	1	2	3	2	11	4	6
181 - 210	1	8	2	2	1	3	5	8
211 - 240	0	6	1	0	0	4	4	15

Tabella 92. Risultati della valutazione dei documenti tramite ranking.

Per testare l'efficacia del sistema di classificazione gli studiosi si concentrano sul solo genere di prosa scientifica, per il quale hanno ottenuto i risultati peggiori. Il corpus Wiki viene rivisto, in modo da utilizzare soltanto i documenti che risultano facili da leggere. I risultati sono illustrati nella tabella seguente.

Genere	Precisione	Recupero	Punteggio F
Wiki	72,97	90	80,60
ScientArt	86,96	66,67	75,47
	Accuratezza: 78,33		

Tabella 93. Risultati della classificazione (ranking) della prosa scientifica.

Come si può osservare, rispetto al metodo precedente (cfr. Tabella 91), il ranking risulta avere una migliore accuratezza (da 56,67% a 78,33%). Ciò significa che questo nuovo approccio può essere sfruttato anche per costruire modelli di addestramento specifici per genere testuale.

6.6.2. Applicazioni di READ-IT

Come abbiamo visto, questo nuovo modello di valutazione automatica della leggibilità è sperimentato su diverse tipologie di testi (giornalismo, letteratura, prosa scientifica e materiale educativo); oltre a questi, READ-IT è utilizzato anche in altri studi, come metodo per valutare la leggibilità di documenti appartenenti all'ambito medico, giuridico o burocratico. "Within an information society, where everyone should be able to access all available information, improving access to written language is becoming more and more a central issue. This is the case, for instance, of administrative and governmental information which should be accessible to all members of the society, including people who have reading difficulties for different reasons: because of a low education level or because of the fact that the language in question is not their mother tongue, or because of language disabilities. Health related information represents another crucial domain which should be accessible to a large and heterogeneous target group" (Dell'Orletta et al. 2011, p. 73).
In tutti questi lavori, l'annotazione linguistica è effettuata con gli strumenti software integrati nella piattaforma *LinguA* (Linguistic Annotation pipeline), una catena di strumenti

statistici di NLP sviluppati dall'ILC del CNR di Pisa e dall'Università di Pisa[197]. Questi strumenti comprendono il PoS tagging descritto in Dell'orletta (2009) e il parser DeSR (Attardi et al. 2009) per l'analisi delle dipendenze.

Brunato (2014) conduce uno studio sulle peculiarità della prosa burocratica e la leggibilità dei testi amministrativi. L'obiettivo è mostrare come le tecnologie linguistico-computazionali possano contribuire all'identificazione delle caratteristiche linguistiche che influiscono sulla complessità per questo genere testuale, "permettendo di discriminare in maniera automatica le caratteristiche di complessità "necessaria" da quelli che invece appaiono come inutili artifici del "burocratese" (p.2). Per la ricerca, viene selezionato un 'corpus parallelo monolingue' di 89 documenti amministrativi italiani: il corpus comprende sia le versioni originali (che rappresentano la classe di testi difficili) che le relative riscritture, semplificate da esperti linguisti (classe di testi facili). Si tratta di documenti che appartengono a diverse tipologie (lettere al cittadino, modulistica, bandi di concorso, ecc.) ma che sono accumunati dallo stesso tipo destinatario esterno all'amministrazione: il cittadino comune.

Il corpus è stato comparato non solo internamente, ma anche rispetto ad altri 5 corpora rappresentativi di altrettanti generi testuali (prosa letteraria, linguaggio giornalistico, materiali didattici, linguaggio scientifico, linguaggio legislativo). La Tabella 94 mostra la composizione dei corpora utilizzati nello studio.

Nome	**Corpus**	**Genere**	**N. testi**	**N. parole**
Rep	La Repubblica (Marinelli et al. 2003)	giornalismo	321	232.908
2Par	Due Parole (Piemontese 1996)	giornalismo	322	73.314
Narr_child	Narrativa per bambini (Marconi et al. 1994)	letteratura	101	19.370
Narr_adult	Narrativa per adulti (Marinelli et al. 2003)	letteratura	327	471.421
Edu_child	Materiali scuola primaria (Dell'Orletta et al. 2011a)	educazione	127	48.036
Edu_adult	Materiali scuola secondaria (Dell'Orletta et al. 2011a)	educazione	70	48.103
Wiki	Wikipedia (articoli tratti da "Ecologia e ambiente")	prosa scientifica	293	205.071
Scient_art	Articoli scientifici	prosa scientifica	84	471.969
Norm_acts	Atti legislativi in materia ambientale.	ling. legislativo	553	1.309.866
It_Const	Costituzione italiana	ling. legislativo	1	10.487
Bur_orig.	Testi burocratici originali	ling. burocratico	89	61.208
Bur_simpl.	Testi burocratici semplificati	ling. burocratico	89	43.780

Tabella 94. Statistiche dei vari corpora.

Un esempio di valutazione della leggibilità di alcuni testi estratti dal corpus di Brunato (2014) è riportato in Brunato e Venturi (2014). Le due ricercatrici presentano due esperimenti in cui READ-IT è impiegato per l'analisi di un testo legislativo particolare, la *Costituzione italiana* e per l'analisi di documenti amministrativi. "L'intento è quello di

[197] Una demo di LinguA è disponibile alla pagina http://linguistic-annotation-tool.italianlp.it/.

dimostrare come READ-IT possa essere usato con successo per calcolare il livello di leggibilità di testi giuridici e per valutare l'efficacia della comunicazione legislatore e/o amministratore-cittadino, allo scopo di semplificare e migliorare i processi di comunicazione tra istituzioni e cittadini" (p. 117).
La Tabella 95 mostra i risultati dell'esperimenti condotti sulla *Costituzione italiana*, nella sua versione originaria del 1947.

Livello di leggibilità	**Difficoltà**
base	21,9%
lessicale	87,3%
sintattico	46,9%
globale	99,4%
GULPEASE	54,9

Tabella 95. Analisi della leggibilità della Costituzione Italiana.

Come si può notare, il livello globale restituisce un valore di difficoltà del 99,4%, per cui il testo risulta di difficile lettura. A differenza del punteggio di leggibilità secondo l'indice GULPEASE (54,9), READ-IT fornisce un valore diverso a seconda del modello considerato: rispetto al livello base e al livello sintattico la Costituzione si rivela semplice o comunque non particolarmente complessa; a livello lessicale, invece, la difficoltà sale fino all'87,3%.
Nel secondo esperimento viene effettuato il confronto tra un testo amministrativo nella sua versione originale e nella sua versione semplificata. Il testo di esempio è tratto dal corpus di Brunato 2014: si tratta di una lettera inviata da un'amministrazione comunale ad un privato cittadino, in cui viene comunicata la necessità di richiedere un sopralluogo tecnico come condizione preliminare per dichiarare la condizione di inabitabilità del proprio immobile[198].
La Tabella 96 mostra i risultati della valutazione della leggibilità dei due documenti.

Livello di leggibilità	**Testo originale**	**Testo semplificato**
base	97,2%	54,2%
lessicale	69,3%	68,5%
sintattico	100%	75,5%
globale	100%	87,9%
GULPEASE	44,5	49,4

Tabella 96. Risultato della misurazione della leggibilità sul testo originale e sul testo semplificato.

Come si può osservare, il modello base restituisce per la versione semplificata un punteggio di leggibilità quasi raddoppiato rispetto al testo originale, mentre il modello lessicale restituisce valori pressoché invariati. La difficoltà a livello sintattico tra le due versioni diminuisce di quasi 25 punti percentuali: questo risultato suggerisce che le caratteristiche

[198] Per le due versioni del documento si veda Brunato e Venturi 2014.

linguistiche monitorate dal modello sintattico siano "effettivamente buone spie per tradurre in una metrica computazionale tipologie diverse di interventi di semplificazione sintattica, quali ad esempio lo scioglimento delle nominalizzazioni o la riduzione dei fenomeni di marcatezza (es. frasi passive o impersonali)" (Brunato e Venturi 2014, p. 138).
Un altro caso di studio è descritto in Brunato e Venturi (2016). Il confronto stavolta è effettuato tra le due versioni della legge provinciale della Provincia autonoma di Bolzano n. 7 del 14 luglio 2015, che promuove la partecipazione e l'inclusione sociale delle persone con disabilità. L'attività di riscrittura è promossa dall'amministrazione stessa che, il 25 agosto 2015, approva una versione semplificata della legge. I risultati della valutazione sono mostrati nella tabella seguente.

Livello di leggibilità	**Testo originale**	**Testo semplificato**
base	26,8%	2,0%
lessicale	2,8%	0,4%
sintattico	98,3%	18,9%
globale	99,6%	29,1%
GULPEASE	46	71,8

Tabella 97. Risultato della valutazione della leggibilità sul testo originale e sul testo semplificato

Risulta evidente che il processo di riscrittura della legge ha permesso di migliorare decisamente livello di leggibilità globale (da 99,6% a 29,1%). La considerevole diminuzione della difficoltà interessa tutti i modelli, ma è soprattutto a livello delle strutture sintattiche che i revisori della legge sono intervenuti maggiormente: la leggibilità varia infatti da 98,3% a 18,9%.
In definitiva, gli esperimenti mostrano che, nonostante la metodologia adottata sia stata progettata per l'analisi di testi rappresentativi della lingua comune, essa riesca comunque ad intercettare le difficoltà di altre tipologie linguistiche, come la lingua del diritto, evidenziandone gli specifici luoghi di complessità.

Per quanto riguarda l'ambito medico, gli studi si sono occupati di valutare la leggibilità di quelle tipologie di testi considerate rappresentative della comunicazione medico-paziente, come ad esempio i foglietti illustrativi (bugiardini) dei farmaci senza obbligo di prescrizione medica (Dell'Orletta et al. 2016) e le informative di consenso per le procedure diagnostico-terapeutiche (Venturi et al. 2015, Dell'Orletta et al. 2017).
Dell'Orletta et al. (2016) hanno costituito un corpus di 7335 bugiardini, estratti dal portale di informazione sanitaria e farmaceutica http://www.torrinomedica.it. Tra questi, hanno selezionato 100 foglietti illustrativi di farmaci vendibili senza obbligo di prescrizione medica, tra i più diffusi in commercio. Per valutare la difficoltà del corpus è stata prima effettuata in modo automatico l'annotazione morfo-sintattica e sintattica e successivamente si sono

confrontati i dati con quelli di altri corpora di riferimento: Due parole (2Par), Repubblica (Rep), materiali didattici per la scuola primaria (MDE) e per la scuola secondaria (MDS)[199].
Dal confronto risulta che le frasi del corpus dei bugiardini risultano essere piuttosto brevi (numero di *token* per frase pari a 11,18), anche in riferimento ai due corpora di controllo ritenuti facili (12,14 per 2Par e 18,36 per MDE). Tuttavia, se si valuta il numero medio di token per clausola verbale la situazione si capovolge: il corpus di bugiardini presenta un numero medio di 17,36 token, valore molto alto se si considerano i punteggi ottenuti dai corpora di testi difficili (10,12 per Rep e 9,2 per MDS). Per quanto riguarda la distribuzione del lessico, risulta che la percentuale di parole appartenenti al VdB è di solo 41,12% (67,09% nel corpus Rep): il lessico risulta quindi molto difficile. Ciò è dovuto probabilmente al fatto che molti dei termini presenti nel corpus sono tecnicismi che non sono contenuti nel vocabolario di riferimento.
Il software READ-IT viene impiegato per valutare i risultati della semplificazione dei testi dei bugiardini, in particolare del foglietto illustrativo del medicinale VIVIN C®, scelto perché vendibile senza obbligo di prescrizione medica. La Tabella 98 mostra i risultati della valutazione della difficoltà dei testi originali:

Livello	**Difficoltà**
globale	100%
base	43,8%
lessicale	99,7%
sintattico	97%

Tabella 98. Valutazione della complessità del bugiardino di VIVIN C®.

Come si osserva, il corpus mostra una difficoltà globale del 100%[200]. Interessante il fatto che, se si considera soltanto il livello base, il livello di complessità è di solo 43,8%, mentre è molto più alto a livello lessicale (99,7%) e sintattico (97%).
La semplificazione del testo avviene rispetto alle singole frasi e "prevede interventi linguistici volti alla semplificazione della lingua a livello lessicale, con l'eliminazione di tecnicismi, nomi astratti e deverbali, e morfo-sintattico, con la preferenza per l'allocutivo voi e l'eliminazione di frasi nominali, impersonali e passive a vantaggio di strutture transitive in cui tanto il recupero dell'agente quanto l'esplicitazione delle relazioni sintattiche tra gli elementi costitutivi della frase risultino facili e accessibili a tutte le categorie di utenti" (Dell'Orletta et al. 2016, p. 226).
Come esempio dell'esito della semplificazione, gli autori riportano la prima frase del testo del bugiardino: 1. *Mal di testa e di denti, nevralgie, dolori mestruali, dolori reumatici e muscolari*. La stima della complessità a livello globale, che risultava essere 89,4%, scende dopo la riscrittura al 10%.

[199] I corpora utilizzati per il confronto sono quelli utilizzati nei precedenti studi, in particolare il corpus dei testi giornalistici (Piemontese 1996 e Marinelli et al. 2003) e dei materiali educativi (Dell'Orletta et al. 2011a).
[200] Il livello globale risulta dalla combinazione delle caratteristiche degli altri tre modelli.

Il Centro gestione rischio clinico e sicurezza del paziente della Regione Toscana (GRC), in collaborazione con l'Istituto di Linguistica computazionale del CNR di Pisa (ILC), conduce uno studio sulla leggibilità dei consensi informati impiegati nelle Aziende sanitarie toscane[201].

Il corpus di testi sanitari, raccolti dal Centro GRC nel 2015, comprende 583 documenti rappresentativi dei diversi tipi di comunicazione scritta medico-paziente (informative di consenso, lettere di accompagnamento, fogli informativi, moduli di assenso/dissenso, ecc.), in uso nelle 16 Aziende ospedaliere del Servizio sanitario della Toscana (4 Ospedali universitari e 12 Aziende sanitarie locali). I testi coprono 29 specialità e sono organizzati in 4 macro aree: chirurgica, medica, prevenzione, servizi. A queste si aggiungono altre 3 specialità: generici, pediatria, riabilitazione e rieducazione funzionale). Il corpus di consensi informati è confrontato con i due set di riferimento, ovvero 2Par, che rappresenta la classe dei testi facili da comprendere, e Rep, che rappresenta la classe difficile. I risultati sono mostrati nella Tabella 99.

Area	**Base**	**Lessicale**	**Sintattico**
Chirurgica	63,59	95,72	78,19
Medica	54,26	96,85	78,24
Prevenzione	55,76	87,13	73,56
Servizi	57,38	96,05	78,02

Tabella 99. Valutazione della leggibilità dei documenti del corpus per area medica.
Per i risultati delle singole specialità cfr. Dell'Orletta et al. 2017.

I valori variano su una scala che va da 0 (facile da leggere) a 100 (difficile da leggere).

Come possiamo osservare, il corpus è caratterizzato da un basso livello di leggibilità. Rispetto al modello Base, l'area medica risulta la più semplice (54,26) e quella chirurgica la più difficile (63,59); a livello lessicale tutte le aree (e le specialità) risultano più difficili, in particolare quella medica, al contrario del modello precedente. È proprio a livello lessicale e sintattico che le specialità mediche mostrano una maggiore variabilità, con valori di leggibilità che vanno da un minimo di 65,14 (screening) ad un massimo di 100 (chirurgia colo-rettale, diabetologia e vaccini).

La maggiore difficoltà registrata a livello lessicale è sicuramente collegata al fatto che molti dei termini impiegati nel corpus appartengono al dominio medico e non sono presenti nel *Vocabolario di Base*. Sarebbe quindi utile poter integrare nel vocabolario dello strumento READ-IT alcuni vocabolari specialistici (o una selezione di termini appartenenti a specifici ambiti) per non penalizzare i punteggi di leggibilità. La specializzazione del lessico di riferimento rispetto al quale valutare la complessità lessicale è una delle attività di ricerca su cui i ricercatori dell'ILC intendono concentrarsi in futuro.

[201] Cfr. Venturi et al. 2015, Dell'Orletta et al. 2017.

PARTE SECONDA

Proposta di un metodo di valutazione automatica della leggibilità di pagine web in lingua italiana

Che dite? Come? Non comprendo: vi dispiacerebbe ripetere? Comprendo ancora meno. Finalmente indovino: volete dirmi, Acis, che fa freddo; perché non avete detto: «Fa freddo?». Volete farmi sapere che piove o nevica; dite: «Piove, nevica». Trovate che ho una buona cera e volete rallegrarvene con me; dite: «Trovo che avete una buona cera». - Ma così, ribattete voi, è troppo piatto e chiaro; e d'altronde chi non saprebbe dire altrettanto? - Che importa, Acis? È forse un male così grande essere capiti quando si parla, e parlare come fanno tutti? Vi manca una cosa, Acis, a voi e ai vostri simili, parlatori lambiccati; non ne avete nemmeno il sospetto e vi farò sbalordire: una cosa vi manca, l'arguzia. Non è tutto: in voi c'è una cosa di troppo, il pregiudizio di averne più degli altri; ecco l'origine dei vostri ampollosi sproloqui, delle vostre frasi intricate e delle vostre parolone che non significano nulla. Vi avvicinate a quest'uomo o entrate in questo salotto; vi tiro per l'abito e vi sussurro all'orecchio: «Non vi passi per la mente di fare dello spirito, siatene del tutto sprovvisto, è questo il vostro ruolo; abbiate, se potete, un linguaggio semplice e simile a quello di coloro che ritenete del tutto privi di arguzia: forse allora si crederà che ne abbiate voi».

(Jean de La Bruyére, *I caratteri*)

7. Proposta di un metodo di valutazione automatica della leggibilità di pagine web in lingua italiana

La seconda parte del mio lavoro è rivolta alla costruzione di un metodo di valutazione automatica della leggibilità di siti web in lingua italiana.
La linea di ricerca che abbiamo scelto prevede l'abbandono dei metodi tradizionali di costruzione delle formule di leggibilità per percorrere una strada alternativa, quella della valutazione automatica basata su tecniche di *machine learning*, che risulta, almeno per quanto riguarda la lingua inglese, ampiamente sperimentata per la classificazione di pagine web.
Nonostante i tradizionali indici di leggibilità siano stati applicati con successo per molti anni, sono molte le obiezioni che sono state rivolte a tali tecniche. Come abbiamo visto, le principali critiche riguardano il fatto che le formule non tengono conto di diversi fattori che influenzano il processo di comprensione, come il livello culturale e la preparazione del lettore, il vocabolario impiegato, la correttezza ortografica, grammaticale e sintattica del testo, la struttura logica, ecc. Inoltre, la maggior parte delle formule è stata creata prima della diffusione del web: essendo progettati esclusivamente per i testi scritti, gli indici non prendono in considerazione le caratteristiche tipiche dei testi e delle pagine web. L'applicazione delle tecniche di apprendimento automatico alla predizione della difficoltà dei testi consente di superare tali limitazioni: i nuovi approcci consentono infatti di esplorare una più ampia varietà di caratteristiche linguistiche e sperimentare variabili più complesse; i modelli, inoltre, hanno la possibilità di essere riadattati facilmente in base a nuovi dati e a diverse applicazioni.
Abbiamo inoltre deciso di non adeguare gli strumenti già esistenti per l'italiano alle peculiarità del web, ma di sviluppare un metodo tarato specificatamente su singole varietà linguistiche presenti sul web. Non esiste infatti una "varietà monolitica" (Biffi 2014) che possa essere definita "lingua del web", ma una molteplicità di livelli, ognuno dei quali presenta proprie caratteristiche linguistiche.
La lingua del web si muove all'interno di tutto spazio linguistico, non soltanto lungo l'asse diamesico, ma rispetto anche agli altri assi di variazione. Si pensi, in particolare, alla profondità diafasica della lingua in rete: la variazione diafasica, determinata dalla situazione comunicativa, è presente sul web in tutta la sua ampiezza, dalle conversazioni informali nelle chat o nei social, al registro formale del testo burocratico o dell'articolo scientifico.
Questa variabilità, propria della realtà del web italiano, ma comune anche ad altre lingue, si riflette inevitabilmente sulle specificità linguistiche dei testi, come il lessico e la sintassi. D'accordo con Tavosanis (2011), si nota infatti che "parte della variazione linguistica del web si può ricondurre alla variazione di genere testuale, cioè alla presenza di tipi di testo differenti con caratteristiche linguistiche diverse".
Non è dunque possibile sviluppare un indice di leggibilità generale per la lingua del web (o adattarne uno), ma è necessario definire un modello in funzione dello spazio linguistico, in grado cioè di rendere conto della variabilità della lingua in rete e della molteplicità dei generi testuali.
Ciò è possibile soltanto ricorrendo a un approccio di valutazione basato sull'apprendimento automatico, che permette di costruire un modello capace di riadattarsi a seconda della

varietà linguistica considerata. Per l'addestramento del modello è sufficiente scegliere un genere testuale di esempio rappresentativo di una varietà presente sul web. Una volta addestrato, il sistema sarà poi in grado di riadattarsi ad eventuali altri generi testuali.

Ma quali sono i generi testuali individuabili sul web?

Definire i generi specifici del web è altrettanto difficile che definire quelli tradizionali. Numerosi sono i tentativi di inquadrare il genere sul web, ma manca ancora una classificazione sistematica ed esaustiva[202].

In prima approssimazione, si può stimare che tutti i generi testuali appartenenti alla carta stampata abbiano oggi una rappresentazione sul web (Tavosanis 2011). A questi si aggiungono nuovi generi originali della rete, in continua trasformazione. Un aspetto importante nell'individuazione dei generi è dato dalla granularità della classificazione (Santini 2008): un'etichetta di genere può infatti essere assegnata ad un sito nel suo complesso oppure a una singola pagina o ai contenuti stessi presenti nelle pagine. Per quanto riguarda le pagine web, è possibile individuare: pagine personali, home page aziendali, commercio elettronico, pagine di ricerca, FAQ, schermate iniziali, annunci, home page delle istituzioni, pagine di istruzioni e manuali online, mappe dei siti, link utili, *about* page, ecc.

Per quanto riguarda i contenuti, possiamo includere blog, forum, aggiornamenti di stato sui social, wiki, guide, manuali, tutorial, FAQ, ricette, guide turistiche, articoli di giornale (cronaca, editoriali, interviste, reportage), recensioni, articoli scientifici, abstract, lezioni, questionari, relazioni, curriculum vitae, prosa, poesia, sceneggiature, modulistica, schede dei prodotti, testi istituzionali e documenti ufficiali (leggi, contratti), annunci, commenti, pubblicità, ecc.

Spesso i siti contengono diversi generi testuali. Si pensi ad esempio ai siti istituzionali, in cui è possibile trovare informazioni istituzionali, ma anche notizie, comunicati stampa, modulistica, linee guida, istruzioni, ecc. Viceversa, la stessa tipologia testuale è presente in due tipi diversi di siti.

Ciò che accumuna i diversi generi testuali è il loro profilo linguistico; i testi appartenenti allo stesso genere condividono infatti alcune caratteristiche, non solo linguistiche, ma riguardanti anche il contenuto e lo scopo.

Si tratta di un aspetto che deve essere tenuto in considerazione nello sviluppo di un metodo di valutazione della leggibilità che si basa proprio sull'analisi dei tratti linguistici dei testi. Diversi studi (Kate et al. 2010, Štajner et al. 2012) hanno infatti dimostrato che i metodi di valutazione della leggibilità sono più accurati se si utilizzano moduli specifici per ogni genere testuale. Questa ipotesi è confermata per l'italiano dagli studi di Dell'Orletta et al. (2012, 2013).

In particolare, i dati mostrano che i metodi che usano modelli generali (che comprendono diversi generi testuali senza però alcuna distinzione a livello della classificazione) sono accurati solo quando analizzano documenti appartenenti allo stesso genere di quelli usati nell'addestramento. Negli altri casi i valori risultano piuttosto bassi. Invece, la costruzione di modelli specifici per generi testuali per l'addestramento del sistema contribuisce a migliorare le previsioni delle leggibilità.

[202] Sui generi testuali del web e la loro identificazione cfr. Santini (2005, 2006, 2007, 2008, 2011), Santini et al. (2009), Mehler et al. (2010), Rehm et al. (2008), Tavosanis (2011).

Il lavoro di addestramento del classificatore per moduli specifici è però piuttosto dispendioso e, soprattutto, richiede che il corpus sia già etichettato in base ai generi testuali.
Dell'Orletta et al. (2012) hanno sviluppato un metodo di ranking in grado di assegnare un punteggio di leggibilità ai documenti senza richiedere corpus di addestramento specifici per genere. Nonostante la metodologia sia stata progettata per l'analisi di testi rappresentativi della lingua comune, essa riesce ad intercettare le difficoltà di altre tipologie linguistiche, come ad esempio la lingua del diritto o testi di ambito medico. L'approccio è basato sulla nozione di distanza euclidea tra vettori di caratteristiche, già illustrata in Dell'Orletta et al. (2011b). Ogni vettore di caratteristiche rappresenta un insieme di frasi: due vettori con distanza 0 rappresentano lo stesso insieme di frasi, cioè quelle frasi che condividono gli stessi valori per le caratteristiche linguistiche misurate; maggiore è la distanza tra due vettori, più distanti saranno i gruppi di frasi rispetto alle caratteristiche monitorate.
Il punteggio di leggibilità è calcolato come la combinazione lineare tra la distanza di un documento e due vettori n-dimensionali che rappresentano la classe facile e la classe difficile.

Il metodo di valutazione proposto prevede la costruzione di un modello che permetta di classificare in modo automatico un insieme di documenti testuali in base al loro livello di leggibilità. Il processo comprende diverse fasi:

- definizione di un corpus di apprendimento;
- selezione delle caratteristiche linguistiche da analizzare;
- estrazione automatica delle caratteristiche dai dati;
- selezione dell'algoritmo di apprendimento;
- creazione del modello;
- validazione del modello.

Per prima cosa, è necessario costruire un corpus di addestramento, rappresentativo di quell'aspetto che si intende valutare (un dato genere testuale, una certa varietà linguistica, ecc.). A ogni testo del corpus deve essere assegnato un livello di leggibilità di riferimento (*gold standard*): la misurazione dei livelli di difficoltà dei testi è un aspetto essenziale, in quanto è proprio su tali standard che si baserà il modello.
La seconda fase prevede la selezione di quelle caratteristiche linguistiche che dovranno essere analizzate in ciascun testo. La scelta dovrebbe essere indirizzata su quelle caratteristiche che potrebbero essere dei buoni predittori della leggibilità.
Una volta effettuata la selezione, si procede con l'estrazione automatica delle caratteristiche: si trasforma ogni testo in un vettore di caratteristiche numeriche che serve da input per l'algoritmo di apprendimento. L'algoritmo crea quindi il modello: impara cioè, in base agli esempi forniti, ad associare ogni vettore di caratteristiche che rappresenta un testo al livello di leggibilità assegnato a quel dato testo.
L'ultima fase prevede la validazione del modello su un nuovo set di dati. Il modello ottimizzato viene applicato a un nuovo corpus per stimare la sua capacità di predizione, cioè per valutare se il sistema è in grado di prevedere correttamente il livello di leggibilità dei nuovi testi. Generalmente i risultati sono valutati tramite alcune misure, l'*accuratezza* (*accuracy*), la *precisione* (*precision*) e il recupero (*recall*), detto anche *copertura* o *richiamo*.

In questo capitolo, illustreremo in modo dettagliato le diverse fasi del progetto; per ciascuna, esporremo la metodologia che abbiamo scelto di seguire e gli eventuali approcci alternativi; cercheremo inoltre di affrontare le diverse problematiche e le varie questioni che possono emergere.

7.1. Il corpus di apprendimento

La prima fase prevede la definizione di un corpus di pagine web in italiano.
Le due possibili alternative sono la costruzione di un nuovo corpus o l'impiego di un corpus già definito (ed eventualmente già annotato). Per quanto riguarda la lingua italiana, i principali web corpora attualmente disponibili sono il ItWac, Paisà e RIDIRE.
Il corpus ItWac fa parte del progetto WaCky (*Web as Corpus kool ynitiative*), sviluppato presso l'Università di Bologna. WaCky comprende un insieme di corpora linguistici di alcune delle principali lingue europee, ognuno dei quali contiene da 1,5 a 2 miliardi di parole: itWac per l'italiano, ukWaC per l'inglese e il deWac per il tedesco. I corpora sono stati creati tra il 2005 e il 2007 attraverso il *web crawling*, utilizzando cioè un programma (*crawler*) per la raccolta e lo scaricamento di pagine dal web. Per ognuno dei corpora è effettuata la tokenizzazione, la lemmatizzazione e l'annotazione delle parti del discorso (POS tagging). Il corpus itWac contiene 2 miliardi di token; per il POS tagging è stato utilizzato il programma TreeTagger e per la lemmatizzazione MORPH-IT!
Il corpus Paisà (Piattaforma per l'Apprendimento dell'Italiano Su Corpora Annotati) comprende circa 388.000 testi in italiano contemporaneo raccolti, tra settembre e ottobre 2010, da 1.067 siti web, per un totale di circa 250.000 occorrenze; è stato ideato principalmente con finalità glottodidattiche, come supporto all'apprendimento e insegnamento dell'italiano come lingua straniera. È sviluppato nell'ambito dell'omonimo progetto (2009-2012), a cui hanno collaborato quattro partner: Università di Bologna, CNR di Pisa, Accademia Europea di Bolzano e Università di Trento. Per la costruzione del corpus sono stati utilizzati esclusivamente documenti non soggetti al vincolo di copyright, ma distribuiti con licenze *Creative Commons* (CC); tale scelta ha influito sulla composizione del corpus e sulla presenza delle varie tipologie testuali. I testi hanno un doppio livello di annotazione: una prima annotazione linguistica e un'annotazione tramite metadati che riguardano l'argomento, l'intenzione comunicativa e il genere testuale. L'annotazione linguistica comprende la divisione in frasi, la tokenizzazione, la lemmatizzazione, l'annotazione morfosintattica (tramite il POS tagger utilizzato in Dell'Orletta 2009) e l'analisi delle dipendenze tramite il parser DeSR (Attardi et al. 2009).
Il corpus RIDIRE (RIsorsa Dinamica Italiana di REte), sviluppato presso l'Università di Firenze, contiene circa 1,5 miliardi di token estratti tra il 2009 e il 2013 da siti web in italiano. Il corpus è costruito tramite la procedura del *crawling mirato*, in cui i siti sono selezionati da esperti dei vari domini. I testi sono etichettati in 12 domini semantici e funzionali. È stata effettuata la lemmatizzazione e l'annotazione morfosintattica (POS tagging) tramite TreeTagger.
La scelta di utilizzare un corpus già definito e annotato può essere utile nel caso si abbia poco tempo o poche risorse a disposizione, ma presenta comunque dei problemi. In primo luogo, il corpus dovrebbe essere disponibile online e liberamente scaricabile; per quanto riguarda i tre web corpora, presentano tutti tale caratteristica. A seconda delle analisi che

vogliamo effettuare, il corpus dovrebbe inoltre avere un'annotazione linguistica specifica (sintattica, morfosintattica, semantica): il corpus Paisà, ad esempio, è l'unico ad avere un'annotazione di tipo sintattico.
Paisà presenta anche altre caratteristiche utili: i testi sono già etichettati per livelli di leggibilità in base alla formula GULPEASE e per generi testuali (sia manualmente che tramite la classificazione automatica); questo tipo di annotazione è funzionale alla costruzione di classificatori specifici di genere. Il corpus presenta però dei problemi per quanto riguarda la rappresentatività: le dimensioni sono molto ridotte rispetto agli altri due corpora e, come abbiamo già detto, la scelta di utilizzare esclusivamente documenti distribuiti con licenze CC ha influito sulla varietà di tipologie testuali raccolte, rendendo il corpus non particolarmente rappresentativo della comunicazione in rete.
La soluzione maggiormente preferibile risulta quindi quella di costruire un nuovo corpus, possibilmente bilanciato e quanto più possibile rappresentativo della lingua italiana sul web.
Nella fase di costituzione del corpus di apprendimento è innanzitutto necessario scegliere il tipo di corpus che si vuole realizzare: un corpus più generale, contenente testi appartenenti a generi e domini diversi, un corpus generale ma etichettato per generi testuali, un corpus specifico di una data varietà o genere testuale.
Tenuto conto delle precedenti considerazioni, quest'ultima soluzione ci sembra quella migliore. Si tratta quindi di definire un corpus di testi rappresentativi di un dato genere testuale su cui costruire un modello di apprendimento di esempio; una volta che tale modello sarà validato, sarà possibile impiegare lo stesso metodo di valutazione su un nuovo corpus rappresentativo di un altro genere testuale per costruire un ulteriore modello, specifico di quella varietà. Uno dei vantaggi dei sistemi di apprendimento automatico è infatti la possibilità di adattarsi ai nuovi dati in entrata.
Per lo sviluppo del nostro metodo di valutazione, abbiamo quindi deciso di concentrarci su una specifica varietà linguistica, la lingua istituzionale degli enti sanitari. La ricostruzione del profilo linguistico del corpus potrebbe essere il punto di partenza sia per l'individuazione dei parametri legati alla complessità, sia per l'individuazione delle caratteristiche che identificano quella data varietà testuale.
La comunicazione online in ambito sanitario raccoglie diverse tipologie testuali. Ai fini di creare un sistema maggiormente accurato, abbiamo deciso di analizzare una specifica tipologia, i testi informativi destinati ai cittadini/pazienti. In particolare, il campione di testi è raccolto dai siti web delle Aziende Sanitarie Locali (ASL) italiane.
La creazione di un corpus è suddivisa in due fasi distinte: la fase progettuale, in cui vengono stabiliti i parametri di tipo quantitativo (dimensioni del corpus: numero di testi, numero di occorrenze) e i criteri di selezione dei testi e la fase di acquisizione del materiale.
Alla fase di raccolta segue poi quella di ripulitura dei dati, per eliminare il *rumore* prodotto dal codice HTML e da tutti quegli elementi privi di contenuto, come i menù di navigazione, gli elementi del layout, i collegamenti ipertestuali, che possono alterare il punteggio di leggibilità.
La metodologia seguita per la selezione dei testi e la costruzione del corpus di addestramento è presentata in modo dettagliato nel capitolo 8.

7.1. Livelli di leggibilità

Come già accennato, a ogni testo del corpus deve essere assegnato un livello di leggibilità, il quale sarà poi usato come standard di riferimento (*gold standard*) per l'addestramento dei dati e la costruzione del modello. Gli studi finora analizzati hanno adottato diversi metodi per l'assegnazione dei livelli di leggibilità ai documenti del corpus:

- Annotazione manuale dei livelli di leggibilità da parte di valutatori (esperti o non esperti).
  La costruzione manuale di un corpus di testi annotati presenta qualche complicazione, soprattutto per il fatto che le assegnazioni in varie classi di leggibilità da parti di giudici umani, esperti o meno, sono arbitrarie e potrebbero non essere precise.
- Utilizzo di corpora già annotati per classi di leggibilità / Creazione di un nuovo corpus tramite testi già annotati per livelli di lettura.
  Il problema principale è che non vi è un'effettiva verifica del livello di difficoltà/livello di lettura di questi documenti. Spesso infatti i livelli sono indicati dalla fonte o dagli autori stessi dei testi o dedotti dai ricercatori che li selezionano. Lo stesso problema si ha nel caso di corpus raccolti da libri di testo e manuali scolastici.
- Misurazione della leggibilità tramite gli strumenti disponibili (in genere indici tradizionali, come la formula di Flesch).
  Questo metodo si porta dietro tutta quella serie di critiche rivolte agli indici di leggibilità: il fatto che le formule non tengono conto di diversi fattori che influenzano il processo di comprensione, che siano tarate sulla lingua scritta, che non siano adatte a valutare testi brevi come quelli sul web, ecc.
- Utilizzo di un metodo di ranking: il ricorso a tale metodo è frequente soprattutto nel caso di lingue in cui non esistono corpus già etichettati, come l'italiano.
  Il problema di tale sistema è che restituisce soltanto valori relativi di leggibilità.

L'annotazione manuale non ci sembra una strada percorribile: se infatti i linguisti esperti sono in grado di individuare le difficoltà dei testi, non è però detto che sappiano assegnare punteggi di leggibilità in base ai livelli di istruzione. Viceversa, non è detto che gli insegnanti, che scelgono quotidianamente testi per i loro studenti, siano in grado di individuare le complessità del testo in rapporto alla capacità di lettura dei propri studenti. Lo stesso problema si presenta in caso di costruzione di un corpus da libri di testo e manuali scolastici.
Per quanto riguarda invece gli strumenti di misurazione della leggibilità, attualmente per la lingua italiana sono disponibili tre diversi metodi:

- La formula GULPEASE: restituisce i risultati sia in base a una scala di valori che indicano la difficoltà del testo, sia in base ai livelli di istruzione. I punteggi vanno da 100 (leggibilità massima) a 0 (leggibilità nulla). Le classi relative alla difficoltà del testo sono 5 (molto facile, facile, difficile, molto difficile, quasi incomprensibile); quelle relative al livello di scolarizzazione dei lettori sono 3 (istruzione elementare, media e superiore).
  Il problema di questo indice è che è tarato sull'italiano scritto degli anni '80 e molto probabilmente non rispecchia i livelli di lettura/difficoltà attuali. Inoltre la formula è

tarata soltanto su bambini e ragazzi in età scolare e mancano invece verifiche sistematiche ed estese su gruppi di adulti.

- READ-IT: restituisce i risultati della misurazione della difficoltà dei testi in base a una scala di valori, che variano da 0 (facile da leggere) a 100 (difficile da leggere). I valori non corrispondono a dei livelli ma a una percentuale di difficoltà; è possibile comunque rapportare questi punteggi con la scala di valori di GULPEASE: ad esempio, un valore ottenuto con READ-IT pari a 55% potrebbe rappresentare la classe "difficile".
- Coease: il sistema restituisce i risultati della valutazione della leggibilità sulla base di diversi indici e in rapporto ai tre livelli di istruzione italiani: primaria, secondaria di primo grado e secondaria di secondo grado. Non restituisce un valore di leggibilità globale ma valori relativi a ciascun indice. Il vantaggio rappresentato da questo strumento è che è costruito, oltre che su testi in uso nelle scuole italiane, in base anche ai test di comprensione ufficiali utilizzati nelle prove Invalsi (che vengono costruiti e riadattati in base alla comprensione effettiva da parte degli studenti).

La nostra proposta è quella di effettuare online dei test di comprensione dei documenti del corpus e verificare l'effettiva difficoltà in base ai diversi livelli di istruzione del campione. Le prove di comprensione consentono infatti di avere dati più accurati sui livelli di lettura della popolazione attuale e una taratura specifica per il web.
Nell'assegnazione dei livelli di leggibilità ai testi sono inoltre da definire alcuni parametri, come la scala di misurazione, cioè il numero e la tipologia di classi che saranno scelte come standard di riferimento: è possibile, infatti, considerare come classi di leggibilità sia i vari livelli di difficoltà dei testi che i livelli di scolarizzazione del campione. Nella maggior parte degli studi considerati, i livelli *gold standard* indicano i livelli di comprensione della lettura di una data popolazione e si basano sul sistema scolastico americano che prevede una divisione in 12 gradi (livelli di istruzione).
La scelta influisce direttamente sul tipo di approccio di addestramento che sarà impiegato: ad esempio, se scegliamo come classi di difficoltà le categorie testuali (facile, difficile, molto difficile, ecc.) dobbiamo adottare un modello basato sulla classificazione; se invece scegliamo i valori numerici, dobbiamo adottare il metodo della regressione.
La stabilità e l'affidabilità del modello dipendono dalla quantità di dati utilizzati (Larsson 2006). Un sistema dinamico e flessibile può comunque consentire la creazione di un metodo efficace di valutazione della leggibilità, capace di riadattarsi a nuovi dati di input ed eventualmente a nuove classi di riferimento.

7.2. Caratteristiche linguistiche

L'annotazione automatica multi-livello del testo costituisce il punto di partenza per la metodologia di monitoraggio linguistico (come individuato e descritto in Montemagni 2013a), di cui la valutazione automatica della leggibilità costituisce una delle principali applicazioni.
L'annotazione automatica del testo permette di identificare la struttura linguistica sottostante al testo e di renderla progressivamente esplicita. L'individuazione della struttura linguistica avviene in maniera incrementale, attraverso analisi linguistiche

progressivamente più complesse: la tokenizzazione, ovvero la segmentazione del testo in parole ortografiche (o *tokens*); l'analisi morfo-sintattica e la lemmatizzazione del testo "tokenizzato"; infine l'analisi della struttura sintattica della frase in termini di relazioni di dipendenza (Montemagni 2013a).

I risultati dell'annotazione linguistica possono contribuire alla ricostruzione del profilo linguistico di un testo. La scelta dei tratti da monitorare varia a seconda del fenomeno che si vuole indagare, ad esempio potremmo voler individuare i parametri legati alla complessità o quelle caratteristiche utili a identificare un dato genere e varietà testuale.

La seconda fase del progetto prevede proprio la selezione delle caratteristiche linguistiche che dovranno essere analizzate da ciascun testo.

Come osservano Collins-Thompson (2014) e Kate et al. (2010), la scelta delle caratteristiche linguistiche da estrarre dai dati influenza molto le prestazioni del modello di apprendimento e sembra avere un peso maggiore anche rispetto alla selezione del *contesto* di apprendimento, come la scelta del tipo di approccio o dell'algoritmo.

A seguito dell'analisi dei principali lavori sulla valutazione automatica della leggibilità possiamo fare alcune considerazioni:

- le migliori funzionalità, cioè le caratteristiche che ottengono i più alti valori di correlazione con i livelli di lettura o la difficoltà dei testi, risultano la lunghezza media della frase e il modello statistico del linguaggio (distribuzione della frequenza delle parole). Questi due parametri sono presenti nella quasi totalità dei lavori.
- diverse ricerche[203] hanno dimostrato che la combinazione di diversi tipi di caratteristiche (lessicali, sintattiche, morfosintattiche, semantiche) risulta essere l'approccio più efficace, raggiungendo gradi di accuratezza del 70% - 80% e restituendo valori più alti di precisione e richiamo.

Tenendo conto di queste valutazioni, si propone l'utilizzo di una combinazione di diversi tipi di caratteristiche. Il set completo, come mostrato nella tabella seguente, comprende le caratteristiche linguistiche considerate da Dell'Orletta et al. (2011) per lo sviluppo dello strumento READ-IT e, in aggiunta, alcuni degli indici proposti in Coease (Tonelli et al. 2012), l'adattamento di Coh-Metrix alla lingua italiana. In particolare, sono inclusi gli indici relativi alla coesione, rimasti fuori dalle ricerche collegate a READ-IT, ma che sembrerebbero far parte, secondo lo studio di Tonelli et al. (2012), delle 10 caratteristiche più correlate con la leggibilità[204].

Le caratteristiche di base corrispondono a quelle tipicamente usate nelle tradizionali formule di leggibilità, ovvero la lunghezza media della frase (numero medio di parole per frase) e la lunghezza media delle parole (numero medio di caratteri per parola). La lunghezza dei paragrafi (numero di frasi per paragrafo) non risulta un indice particolarmente correlato alla leggibilità, ma si tratta comunque di un parametro che influisce sull'organizzazione, e dunque sulla chiarezza, dei testi sul web.

Come caratteristiche lessicali sono misurate la composizione del vocabolario (percentuale di parole appartenenti al *Vocabolario di base*), la ricchezza lessicale (rapporto type/token),

[203] Larsson 2006, Wang 2006, Aluisio et al. 2010, Kate et al. 2010.

[204] Oltre agli studi di Coh-Metrix e Coease, la coesione è valutata in Pitler e Nenkova 2008, François e Fairon 2012 e Chen et al. 2013.

la densità lessicale (calcolata come la proporzione di parole piene - nomi, aggettivi, verbi e avverbi - sul totale delle occorrenze) e la frequenza delle parole[205].
Tra le caratteristiche morfosintattiche sono prese in considerazione: il modello statistico del linguaggio basato sulle probabilità degli unigrammi delle parti del discorso (cioè la distribuzione delle categorie grammaticali nel testo: sostantivi, aggettivi, verbi e congiunzioni) e la distribuzione dei modi verbali.
Per quanto riguarda le variabili sintattiche, sono monitorate: la distribuzione dei vari tipi di relazioni di dipendenza (soggetto, oggetto diretto, modificatore, ecc.), le caratteristiche relative alla profondità dell'albero sintattico (altezza media dell'albero, lunghezza delle relazioni di dipendenza), le caratteristiche relative alla subordinazione (distribuzione delle proposizioni subordinate e delle principali, posizione delle subordinate rispetto alla principale, lunghezza media di sequenze consecutive di subordinate), le caratteristiche dei predicati verbali (arità verbale o numero di radici verbali, numero di dipendenti per testa verbale), la profondità media di strutture nominali complesse, cioè costituite da una testa nominale e da modificatori aggettivali e/o complementi preposizionali. Oltre a queste, sono valutate le caratteristiche legate alla coesione e, in particolare, l'incidenza delle diverse categorie di connettivi, il numero di pronomi e la coreferenza (sovrapposizione di argomenti, sovrapposizione lessicale).

Tipo di caratteristiche	**Caratteristica**
Di base	Lunghezza media delle frasi
	Lunghezza media delle parole
	* Lunghezza dei paragrafi (n. di frasi per paragrafo)
Lessicali	Ricchezza lessicale (rapporto type/token)
	Percentuale di lemmi appartenenti al *Vocabolario di Base* del *Grande dizionario italiano dell'uso* (De Mauro, 2000)
	Densità lessicale (% parole piene sul totale delle occorrenze)
	* Frequenza delle parole
Morfosintattiche	Modello statistico delle parti del discorso (distribuzione delle categorie morfosintattiche)
	Distribuzione dei modi verbali

[205] In Coh-Metrix la frequenza delle parole è misurata in base al confronto con 4 corpora: il database lessicale CELEX, che contiene le frequenze di circa 18 milioni di parole; l'analisi delle frequenze effettuata da Francis e Kucera (1982); la lista delle frequenze compilata da Thorndike e Lorge (1944); il conteggio delle frequenze dell'inglese parlato realizzato da Brown (1984). Il principale parametro valutato è il logaritmo medio delle frequenze delle parole. In Coease, gli indici relativi alla frequenza si focalizzano sulle parole contenuto (nomi, verbi, avverbi, aggettivi); come corpus di riferimento è utilizzata Wikipedia in italiano.
I principali studi sulla valutazione automatica delle leggibilità impiegano invece modelli statistici del linguaggio che considerano le probabilità di unigrammi, bigrammi, ecc. Questo è l'approccio seguito da: Collins-Thompson e Callan 2004, Larsson 2006, Wang 2006, Aluisio et al. 2010, Feng et al. 2010, François e Fairon 2012.

Tipo di caratteristiche	Caratteristica
Sintattiche	Distribuzione dei vari tipi di relazioni di dipendenza
	Caratteristiche relative alla struttura dell'albero sintattico analizzato: • altezza media dell'intero albero • lunghezza media della più lunga relazione di dipendenza
	Caratteristiche relative all'uso della subordinazione: • distribuzione di frasi principali vs. subordinate • posizione delle subordinate rispetto alla principale • lunghezza media di sequenze consecutive di subordinate
	Caratteristiche relative alla modificazione nominale: • lunghezza media dei complementi preposizionali dipendenti in sequenza da un nome
	Caratteristiche dei predicati verbali: • arità verbale (numero di radici verbali) • numero di dipendenti per testa verbale
	* Coesione: • incidenza delle diverse categorie di connettivi (positivi/negativi): causali, additivi, temporali. • numero di pronomi • coreferenza: sovrapposizione di argomenti + sovrapposizione lessicale

Tabella 100. Set di caratteristiche da considerare nella ricerca.
I campi contrassegnati con (*) sono gli indici ripresi da Coease.

Sarebbe interessante anche valutare alcune caratteristiche legate più strettamente alla fruizione del testo via web e verificare se esiste una correlazione con i livelli di leggibilità[206].
La maggior parte degli studi sulla leggibilità dei testi web si concentra infatti sul solo contenuto testuale e tratta gli elementi "fisici" propri delle pagine web come un problema

[206] Ad esempio, Ali et al. (2013) hanno valutato gli effetti che la tipologia di font può avere sulla leggibilità (in lingua malese) e non hanno notato sostanziali differenze tra caratteri con o senza grazie.
Gradišar et al. (2006) hanno invece valutato gli effetti che specifici design dei siti web hanno sulla velocità di lettura: hanno testato 30 diverse combinazioni di colori e hanno dimostrato che queste differenze non influiscono in modo significativo sulla leggibilità del documento.
Yu e Miller (2010) hanno proposto il metodo Jenga Format per la trasformazione dei contenuti testuali al fine di migliorare la leggibilità delle pagine web. La trasformazione si basa su due elementi: la separazione tra le frasi e la spaziatura all'interno di un paragrafo. I due studiosi hanno testato il loro metodo su 30 lettori e hanno dimostrato che agire su questi 2 aspetti migliora la comprensione della lettura senza influire negativamente sulla velocità di lettura. Il campione è però piuttosto ridotto.
Kanungo e Orr (2009) hanno studiato la leggibilità degli abstract dei risultati delle ricerche sul web usando tecniche di apprendimento automatico; tra le caratteristiche considerate nello studio vi erano la quantità di lettere maiuscole, la presenza dei puntini di sospensione all'inizio o alla fine delle frasi, la quantità di segni interpuntivi.

di usabilità. Gli esperimenti di Gottron e Martin (2009, 2012) dimostrano inoltre che un approccio che prevede l'estrazione dei contenuti testuali e l'eliminazione del rumore restituisce valori di leggibilità più accurati rispetto a un indice calcolato in modo automatico sull'intera pagina web.
Tuttavia, si deve tenere conto di alcuni aspetti. In primo luogo, la lettura sul web è diversa da quella dei testi cartacei: è una lettura selettiva, in cui l'utente "scansiona" il testo alla ricerca del contenuto rilevante. Anche le strategie di comprensione messe in atto dagli utenti sono diverse: al lettore è richiesto un ruolo più attivo e flessibile nel processo di comprensione della lettura e nella costruzione del significato del testo. Esistono poi alcuni elementi, legati soprattutto all'organizzazione testuale, che influiscono inevitabilmente sulla comprensione dei contenuti:

- il tipo di carattere;
- la dimensione del font;
- l'interlinea del testo;
- la suddivisione dei paragrafi;
- la presenza di elenchi puntati o numerati;
- la presenza di elementi che mettono in risalto parole o parti del testo più rilevanti (ad esempio l'uso del grassetto, del maiuscolo, la sottolineatura, ecc.)
- il contrasto dei colori;
- la presenza di immagini, video, audio;
- la presenza di tabelle;
- la quantità di collegamenti ipertestuali.

La valutazione di questi elementi dovrebbe però essere distinta da quella delle caratteristiche linguistiche: per effettuare l'annotazione automatica è infatti necessario che i testi siano già ripuliti e si presentino in formato esclusivamente testuale. La misurazione di tali parametri andrebbe invece effettuata manualmente, risalendo alle URL dei documenti o eventualmente tramite il ricorso a un algoritmo che misura le variabili all'interno delle pagine.

7.3. Algoritmi e modelli di apprendimento

La nostra proposta è di utilizzare due diverse tipologie di approccio (la classificazione e il ranking) e verificare quale dei due metodi sia più accurato per la valutazione automatica delle pagine web.
L'approccio della classificazione è il metodo generalmente più impiegato nelle ricerche e, in base ai risultati, sembrerebbe anche essere quello più adatto al compito di valutare la leggibilità dei testi web. Aluisio et al. (2010) hanno sviluppato un metodo di valutazione automatica della leggibilità per la lingua portoghese; per la creazione del modello, gli studiosi hanno testato tre diverse tipologie di apprendimento automatico: classificazione, ranking e regressione. In base ai diversi esperimenti, gli autori hanno infine scelto l'approccio della classificazione, che è risultato il più semplice e il più accurato.
Il modello basato sulla classificazione richiede dati di addestramento già etichettati per livelli di leggibilità; il presupposto per l'applicazione di tale metodo è quindi la misurazione della leggibilità e l'assegnazione dei livelli di difficoltà ai testi del corpus.

Il metodo del ranking, l'unico finora utilizzato per i testi in italiano, risolverebbe tale problema, dal momento che i testi dovrebbero essere annotati soltanto rispetto a due livelli di leggibilità (facile o difficile). Tuttavia, come notato da Tanaka-Ishii et al. (2010), questo sistema manca di *assolutezza* nel determinare una norma, riportando solo valori relativi di leggibilità.
Anche nel caso del ranking, è necessario scegliere le classi che rappresentano gli standard di riferimento. Esistono varie possibilità: individuare all'interno del corpus gruppi di testi che possono corrispondere a questi due poli, anche se, per la classe *facile*, l'identificazione non è così immediata. Oppure, è possibile selezionare le due classi in base ai punteggi di leggibilità; tale soluzione richiede però che i dati siano etichettati per livelli di difficoltà. In alternativa, è possibile prendere come riferimento le due classi impiegate nello sviluppo dello strumento READ-IT: il corpus di Repubblica (Rep), che identifica la classe dei testi difficili e il corpus di due parole (2Par), che identifica i testi di facile lettura.
Per quanto riguarda gli algoritmi di apprendimento, molti lavori hanno mostrato che l'approccio basato sul *Support Vectore Machine* offre una maggiore accuratezza e una precisione in alcuni casi superiore all'80% rispetto ad altri modelli (come i Naïve Bayes e gli alberi decisionali).
In particolare, il classificatore SVM risulta più accurato quando si considera una combinazione delle diverse tipologie di caratteristiche (cfr. Al-Kalifa e Amani 2010, Wang 2006).
Una delle possibili scelte è il software LIBSVM (Chang e Lin 2001), che implementa l'algoritmo SMO per SVM. LIBSVM è il pacchetto software utilizzato anche per la definizione di READ-IT e consente di: selezionare i parametri da analizzare, effettuare una classificazione multiclasse, effettuare la convalida incrociata.

7.4. Validazione del modello

L'ultima fase prevede la validazione del modello su un nuovo set di dati. Generalmente, prima della costruzione del modello di addestramento, il corpus viene diviso in due parti: un corpus di apprendimento (*training corpus*), che serve per la costruzione del modello e un corpus di verifica (*test corpus*), che serve per stimare la sua capacità di predizione, cioè per valutare se il sistema è in grado di prevedere correttamente il livello di leggibilità dei nuovi testi.
La qualità del modello dipende da una serie di fattori coinvolti nel processo: la scelta del set di dati di allenamento, la scelta di un algoritmo di apprendimento efficiente e la selezione delle caratteristiche linguistiche da estrarre dai dati.
La validazione del modello è spesso effettuata tramite la tecnica della validazione incrociata. I risultati sono valutati tramite alcune misure, l'*accuratezza* (*accuracy*), la *precisione* (*precision*) e il recupero (*recall*).

8. Il corpus delle Aziende Sanitarie Locali (ASL) italiane

Tra le varie tipologie di comunicazione sanitaria in rete, abbiamo concentrato la nostra analisi sugli enti istituzionali e, in particolare, sui siti web della Aziende Sanitarie Locali (ASL) italiane.

Il modo in cui le aziende sanitarie comunicano attraverso i loro siti rappresenta ancora un aspetto critico: spesso infatti i testi di ambito medico-sanitario risultano troppo complessi ed espressi in un linguaggio tecnico e scientifico che determina difficoltà nella lettura e nella comprensione da parte di persone meno istruite. Una mancata comprensione delle informazioni sanitarie non solo lede il diritto del cittadino di essere informato ma influisce anche sui comportamenti e le azioni che questo intraprenderà nei confronti del sistema sanitario. Una comunicazione inefficacie porta il cittadino a contattare maggiormente gli uffici per avere chiarimenti, a lunghe code presso gli uffici URP; porta a un uso inappropriato dei servizi sanitari e a un numero maggiore di prestazioni e ospedalizzazioni non necessarie; e, in alcuni casi, porta l'utente a intraprendere azioni legali nei confronti dell'ente sanitario.

L'accesso ai contenuti sanitari e la comprensione di questi devono essere garantiti a tutti, indipendentemente dal livello di istruzione. È quindi fondamentale il ruolo svolto dalle amministrazioni sanitarie nella diffusione di informazioni chiare e di materiali ad alta comprensibilità. In questa prospettiva, Il sito web diviene uno strumento molto potente di comunicazione e può diventare un punto di riferimento per le informazioni e i servizi sanitari.

Questo capitolo è dedicato alla costruzione del corpus delle Aziende Sanitarie italiane: dopo una breve introduzione sulla comunicazione sanitaria in rete e la valutazione della qualità e della leggibilità delle informazioni sanitarie dei siti web, saranno presentati i criteri di selezione dei testi e la composizione del corpus. I testi inclusi nel campione andranno a formare il corpus su cui si baserà il modello di apprendimento.

8.1. La comunicazione sanitaria in rete

"La comunicazione in ambito sanitario fa parte di quelle aree 'trasversali' della comunicazione pubblica destinata ad assumere un ruolo centrale e strategico nelle relazioni tra Stato e cittadino. Questa 'centralità' è nata non solo grazie all'obbligo delle Amministrazioni sanitarie di rispondere al diritto del cittadino di essere informato ma, soprattutto, per la dimensione più articolata e complessa che termini come 'salute' e 'cura' hanno assunto nella società contemporanea. […] Dalla fine del secolo scorso, il progresso delle tecnologie e lo sviluppo e la diffusione di Internet hanno reso possibile l'implementazione di canali web in grado di fornire ai cittadini informazioni di carattere sanitario, con l'obiettivo principale di aumentare l'accesso della popolazione ad informazioni sulla salute di alta qualità, coinvolgendo di più i cittadini/pazienti, rendendoli il più possibile responsabili rispetto alla propria condizione di salute e consapevoli riguardo

alle malattie e le cure ed i trattamenti sanitari a cui devono essere sottoposti" (*Linee guida per la comunicazione on line in tema di tutela e promozione della salute*, p. 3)[207].

L'importanza del web come mezzo per la diffusione di informazioni sanitarie è aumentata in modo esponenziale negli ultimi anni. Lo sviluppo di nuovi strumenti e nuove strategie di informazione ha reso possibile il miglioramento della comunicazione da parte della sanità pubblica. Ciò ha garantito una maggiore accessibilità alle notizie di tipo sanitario, migliorando la conoscenza degli utenti sulle problematiche relative alla salute e creando maggiore consapevolezza nelle scelte legate al proprio benessere.

Allo stesso tempo, è aumentato l'uso di Internet e, in particolare, il numero di utenti che cercano in rete informazioni sanitarie. Secondo i dati emersi dalla ricerca Censis Assosalute 2017, sono 15 milioni gli italiani che, in caso di piccoli disturbi (mal di testa, raffreddore, ecc.), cercano informazioni sul web. Di questi, il 17% consulta siti web generici sulla salute, il 6% siti istituzionali, il 2,4% i social network[208]. Anche dalla ricerca *Health Information Journey*[209], realizzata da GfK Eurisko nel 2015, emerge che, su un campione di 2.066 italiani (maggiorenni), un italiano su due ricerca informazioni di tipo sanitario. Il medico di base rimane il principale riferimento (43%), seguito dallo specialista (34%), anche se ben il 24,5% degli intervistati utilizza il web per le proprie esigenze informative sanitarie. Ricerche comparative a livello internazionale hanno mostrato che, nei paesi in cui l'uso di Internet è particolarmente diffuso, il numero di persone che cerca informazioni sanitaria in rete raggiunge, e in alcuni casi supera, il numero di persone che richiedono assistenza medica.

Una maggiore disponibilità di risorse e informazioni sul web si scontra però con il problema della qualità e dell'affidabilità dei testi pubblicati: chiunque può diffondere e pubblicare online materiali di qualsiasi genere, senza però avere competenze in materia. Talvolta sono proprio le istituzioni pubbliche e gli enti sanitari a pubblicare, in buona fede, informazioni inaccurate[210]. A questo si aggiungono i casi in cui, spesso in modo anonimo ma

---

[207] *Linee guida per la comunicazione on line in tema di tutela e promozione della salute*, a cura del Ministero della Salute e dell'Università La Sapienza di Roma, Roma, 2011. Il documento è consultabile all'indirizzo:
http://www.salute.gov.it/portale/documentazione/p6_2_2_1.jsp?lingua=italiano&id=1473

[208] Si tratta della ricerca del Censis «Il valore socio-economico dell'automedicazione», realizzata in collaborazione con Assosalute (Associazione nazionale farmaci di automedicazione) nel 2017. Alcuni dei risultati sono presentati nel comunicato stampa *Allarme fake news in sanità: 8,8 milioni di italiani hanno trovato sul web informazioni mediche sbagliate*: http://www.censis.it/7?shadow_comunicato_stampa=121153.

[209] Si tratta di una ricerca condotta annualmente da GfK Eurisko su un campione di 2.000 individui, rappresentativo della popolazione italiana adulta; a questa si affianca *Total Single Source Panel*, che viene invece condotta su un campione di 10.000 individui, rappresentativo della popolazione italiana 14+.

[210] Esistono diversi studi che confermano questa tesi. McClung et al. (1998) hanno valutato la qualità delle informazioni reperibili in rete relativamente al trattamento della diarrea acuta infantile, confrontandone la conformità con le raccomandazioni pubblicate dall'American Academy of Pediatrics (AAP). Di 60 articoli pubblicati da autorevoli associazioni scientifiche e istituti accademici, solo 12 (20%) erano conformi alle linee guida AAP per il trattamento della diarrea acuta dei bambini. Uno studio analogo è stato condotto da alcuni ricercatori italiani. Impicciatore et. al. (1997) hanno valutato l'affidabilità delle informazioni sanitarie presenti su 41 siti web per quanto riguarda i consigli sulla gestione della condizione febbrile nei bambini. L'affidabilità delle informazioni è stata verificata confrontando i materiali con le linee guida pubblicate. Solo 4 siti presentavano contenuti conformi alle raccomandazioni; ben 28 pagine fornivano indicazioni sbagliate circa la temperatura minima che dovrebbe essere considerata come stato febbrile nel bambino.

consapevole, sono diffuse *fake news* per influenzare l'opinione pubblica; si pensi ad esempio a tutte le notizie che circolano sul web relative alle vaccinazioni nei bambini. Dalla ricerca Censis Assosalute 2017 è emerso che ben 8,8 milioni di italiani sono state vittime di *fake news* nel corso del 2017 e, in particolare, 3,5 milioni di genitori si sono imbattuti in indicazioni mediche sbagliate.

Il mancato controllo sull'informazione in rete è un problema che coinvolge tutti i settori della comunicazione e in ambito sanitario ciò diviene particolarmente critico. Nel reperire informazioni tramite motori di ricerca, l'utente non sempre è in grado di valutarne la qualità, l'attendibilità, l'accuratezza e la veridicità. La diffusione e circolazione di informazioni sbagliate o inaffidabili non solo è dannoso per la salute dei pazienti ma influenza anche la percezione dell'utente e il suo modo di rapportarsi alla sanità. Uno studio condotto nel 2010 negli USA ha dimostrato come il web possa influenzare la decisione dei pazienti di recarsi dal medico, di chiedere spiegazioni su specifiche problematiche o di effettuare ulteriori consulti[211].

Il digitale non sta cambiando soltanto il modo di fruizione dei servizi sanitari ma anche il rapporto tra cittadino e sistema sanitario: *l'utente/paziente digitale* si aspetta di trovare informazioni dettagliate e precise sui servizi sanitari delle pubbliche amministrazioni e se non è soddisfatto si rivolge ad altre fonti; è un utente che sempre più spesso ricorre al *Doctor Google* per ottenere risposte in merito alle domande sul proprio stato di salute, effettuando purtroppo anche auto-diagnosi. È un utente che vuole essere messo al centro, che vuole un ruolo più attivo. Sempre più spesso si parla di *empowerment* del paziente, definito dall'Organizzazione Mondiale della Sanità (OMS, o in inglese World Health Organization, WHO) come il processo attraverso il quale le persone acquisiscono un maggiore controllo sulle decisioni e le azioni che riguardano la propria salute[212]. L'*empowerment* si ottiene principalmente tramite l'alfabetizzazione sanitaria (*health literacy*) della popolazione, cioè quell'insieme di competenze personali, cognitive e sociali che determinano la capacità di comprendere e usare le informazioni per promuovere e mantenere corretti stili di vita e un buono stato di salute. Attraverso l'educazione sanitaria e la promozione della salute si forniscono alle persone gli strumenti critici con cui consentire l'esercizio dei propri diritti di essere informati e con cui compiere in modo consapevole e responsabile le scelte relative alla propria salute. Un paziente maggiormente consapevole è infatti in grado di riconoscere e di riferire il proprio stato di salute; ha maggiore fiducia nel sistema sanitario ed è più propenso a seguire le prescrizioni del medico. Il paziente diviene così protagonista del proprio benessere.

"L'obiettivo del sito di un ente sanitario quindi deve essere quello di offrire una piattaforma telematica in grado di incontrare in maniera efficace le esigenze informative, le aspettative e le priorità dei cittadini-pazienti e di favorire in loro lo sviluppo dell'apprendimento di comportamenti di promozione della salute e di prevenzione della malattia, contribuendo anche ad un uso più appropriato dei servizi sanitari. In particolare, grazie anche alle potenzialità offerte da Internet nei processi di ricerca dell'informazione sanitaria, questo canale on line dovrebbe garantire una diffusione delle conoscenze medico-scientifiche tra

---

[211] *Safety and security on the Internet. Challenges and advances in member states*, Global observatory for eHealth, volume 4, 2011: http://www.who.int/goe/publications/ehealth_series_vol4/en/index.html.

[212] *The WHO Health Promotion Glossary*: http://www.who.int/healthpromotion/about/HPG/en/.

gli utenti-pazienti, allo scopo di supportare il cittadino nell'assunzione di decisioni attive ed informate nei confronti delle proprie condizioni di salute" (Id., p. 4).

Esistono diversi sistemi per la valutazione della qualità dell'informazione sanitaria dei siti web istituzionali. Generalmente le valutazioni si basano su criteri come l'affidabilità delle fonti, l'aggiornamento dei contenuti, le politiche di privacy, la qualità dei contenuti. Le iniziative volte al miglioramento della qualità dei siti web sanitari possono essere distinte in tre tipologie[213]:

- Codici di condotta/etici/di autoregolamentazione;
- Certificazioni o valutazioni di terze parti;
- Strumenti di valutazione

I codici di condotta sono basati su principi etici e su una serie di criteri di qualità a cui gli enti sanitari possono aderire. Tra i più importanti e più conosciuti vi sono l'*eHealth Code of Ethics*, una serie di raccomandazioni nate in seguito all'*Health Ethics Summit* (Washington DC, 2000), organizzato dall'Internet Healthcare Coalition (IHC) e a cui parteciparono esperti da tutto il mondo. I principi si cui si basa il Codice Etico sono: sincerità, onestà, qualità, rispetto del consenso informato, privacy, professionalità nell'assistenza sanitaria online, partenariato responsabile e responsabilità (cioè dare la possibilità agli utenti di fornire un feedback al sito e monitorare l'aderenza al codice)[214]. Un altro esempio è l'*HON Code of Conduct*, un codice di linee guida per la certificazione di qualità dell'informazione medico scientifica on line elaborato dalla Health On the Net Foundation (HON), un'organizzazione senza fini di lucro con sede a Ginevra. Il codice HON è composto da diversi principi, che riguardano: autori, complementarietà (le informazioni diffuse dal sito sono destinate ad incoraggiare, e non a sostituire, le relazioni esistenti tra paziente e medico), privacy, provenienza delle informazioni, trasparenza e linea di condotta adottata per il reperimento dei fondi[215]. La Fondazione HON ha prodotto anche un certificato che è indice dell'aderenza e della conformità del sito a tali criteri.

In alternativa, esistono dei sistemi di certificazioni di qualità gestite da terze parti e rilasciati in seguito alla validazione di aderenza a un insieme di standard ben definiti (e generalmente dietro pagamento di una quota). Un esempio è costituito dal progetto europeo MEDCERTAIN (MedPICS Certification and Rating of Trustworthy Health Information on the Net), e dal catalogo OMNI (Organising medical network information), entrambi non più attivi. Un'altra iniziativa di certificazione è quella promossa da URAC, un'organizzazione senza scopo di lucro con sede a Washington DC, che si occupa di promuovere la qualità dell'assistenza sanitaria tramite programmi di accreditamento, formativi e di misura.

Gli strumenti di valutazione, infine, si basano su questionari rivolti direttamente ai cittadini e misurano la qualità in base al punteggio ottenuto nelle domande. A questa tipologia appartengono DISCERN, un progetto creato dalla Division of Public Health and Primary Care

[213] Sulla valutazione della qualità dell'informazione sanitaria in rete cfr. Masoni et al. 2014.

[214] I criteri previsti dall'*eHealth Code of Ethics* sono consultabili all'indirizzo: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1761853/

[215] Sul sito dell'HON Code è possibile trovare gli otto principi tradotti in lingua italiana: https://www.hon.ch/HONcode/Italian/

dell'Università di Oxford nel 1998 e, in Italia, il questionario *Misurasiti*, disponibile sul portale del progetto PartecipaSalute[216].
Un'altra iniziativa italiana interessante è *Dottoremaeveroche*, il portale italiano realizzato dalla FNOMCeO (Federazione Nazionale degli Ordini dei Medici Chirurghi e Odontoiatri) contro le *fake news* in tema di salute, in particolare diffuse attraverso il web. Il sito nasce "come porto sicuro nel mare in tempesta della disinformazione in ambito sanitario", con lo scopo di fornire "un'informazione seria, solida e trasparente, corredata da tutti i dovuti riferimenti bibliografici, al tempo stesso resa immediata e accessibile a tutti gli utenti, grazie alla collaborazione di persone che lavorano da anni nella comunicazione in ambito scientifico"[217]. All'interno del portale vi è anche una sezione dedicata alla valutazione dell'informazione sanitaria online[218]: sono proposti 5 criteri per una navigazione consapevole (autorevolezza della fonte, analisi dei contenuti, aggiornamento dei contenuti, trasparenza e tutela della privacy) e una scheda di valutazione della qualità delle informazioni con domande suddivise in base ai 5 criteri.

Nel 2011 il Ministero della Salute, in collaborazione con l'Università La Sapienza di Roma, pubblica le *Linee guida sulla comunicazione on line in tema di tutela e promozione della salute*, una raccolta di criteri e raccomandazioni per progettare una comunicazione sanitaria online di qualità[219].
Il documento, frutto dell'Accordo di collaborazione del dicembre 2009 tra la Direzione generale della Comunicazione e Relazioni istituzionali del Ministero della Salute e Sapienza Università di Roma, è realizzato nell'ambito del progetto "Potenziamento della comunicazione on line del Ministero della Salute e del SSN e progettazione di un canale telematico per i cittadini"[220].
Le *Linee guida* si compongono di due parti. La prima illustra la metodologia utilizzata per la stesura del documento; in particolare sono presentati:

- i risultati delle attività di ricerche condotte sui bisogni informativi di salute tramite il web;
- i risultati delle attività di ricerche condotte sulle esigenze dei cittadini sull'informazione online in tema di salute;
- l'analisi del quadro epidemiologico italiano;
- la rassegna di letteratura scientifica (in lingua inglese) sull'efficacia degli interventi di promozione e prevenzione della salute realizzati via web;

[216] Il questionario è disponibile all'indirizzo https://www.partecipasalute.it/informati-bene/misura-siti-002.php. Oltra al *Misurasiti* sono disponibili il *Misuratesti*, per valutare la qualità delle informazioni contenute in un articolo, il *Misura associazioni*, per misurare la qualità, l'attività e la trasparenza delle varie associazioni e il *Misura campagna*, per valutare una campagna di sensibilizzazione.
[217] Dal sito https://dottoremaeveroche.it/.
[218] La sezione è curata da Maria Rensa Guelfi e Marco Masoni, coordinatori scientifici dell'Unità di Ricerca "Innovazione Didattica ed Educazione COntinua in Medicina" (IDECOM) del Dipartimento di Medicina Sperimentale e Clinica dell'Università di Firenze e autori di numerosi articoli sulla qualità e la comprensibilità delle informazioni sanitarie in rete.
[219] Cfr. nota 207.
[220] Il portale di riferimento indicato dal documento non è più attivo e rimanda al sito più aggiornato https://www.agid.gov.it/it/design-servizi/accessibilita-siti-web.

- la valutazione della qualità dell'informazione dei siti web degli enti istituzionali del Servizio Sanitario Nazionale (Regioni e ASL)[221].

Nella seconda parte sono proposti i criteri per progettare una comunicazione sanitaria online di qualità; le raccomandazioni riguardano in particolare:

- i contenuti informativi e la tipologia di interventi sanitari in Internet;
- le strategie di comunicazione e i criteri redazionali;
- l'impiego delle tecnologie del dialogo e il web 2.0.

Dalle indagini esplorative sui bisogni informativi in tema di salute e dalla revisione sistematica della letteratura scientifica emergono interessanti spunti di riflessione. "Un suggerimento importante che proviene da alcuni studi sull'utilizzo di Internet da parte dei pazienti è che la rete non viene usata soltanto per acquisire informazioni o conoscenze scientifiche: attraverso la rete passa, assai più che in passato, un processo di costruzione sociale della malattia, di elaborazione di senso e di condivisione della propria condizione di malato, processo nel quale il singolo paziente riconquista margini di soggettività e di autonomia, specialmente quando la malattia in questione è soggetta ad una forte valutazione sociale. La tendenza dei cittadini ad utilizzare la rete non soltanto a fini informativi ma anche come mezzo di partecipazione pone con forza il problema della valutazione di efficacia delle diverse strategie di comunicazione via Internet in tema di salute" (*Linee guida,* p. 16).

Le motivazioni che spingono la popolazione a ricercare informazioni in rete sono legate alla fruibilità del mezzo (possono essere ricercate in qualunque momento), alla garanzia di privacy e di anonimato (caratteristiche determinanti soprattutto se i temi ricercati sono delicati), alla facilità di accedere ad informazioni chiare e comprensibili, alla rapidità nella ricerca e alla quantità di informazioni presenti sul web. Le ricerche sono effettuate con diverse modalità ma generalmente sono impiegati i motori di ricerca più conosciuti, soffermandosi nella maggior parte dei casi ai primi siti che vengono consigliati. Gli utenti tendono a non prestare una grande attenzione alla qualità e agli autori dei siti consultati. La tipologia di contenuti più frequentemente ricercata è quella relativa a specifiche malattie o trattamenti sanitari ma è in aumento anche la quota di utenti che ricerca informazioni relative alla promozione della salute, alla prevenzione delle malattie e all'accesso ai diversi servizi sanitari.

Per quanto riguarda i dati più circoscritti al contesto italiano, i risultati mostrano che il 62% dei cittadini che hanno partecipato all'indagine usano Internet soprattutto per la ricerca di informazioni di carattere generale in tema di salute. Poco meno del 30% si rivolge in prima battuta alla rete per un problema di salute, percentuale che risulta relativamente omogenea nelle varie fasce di età; il 58% del campione preferisce cercare on line informazioni sulla tutela della salute piuttosto che rivolgersi direttamente al medico di base per la rapidità con cui è possibile ottenere informazioni.

[221] La rilevazione è stata condotta su tutte le regioni italiane (19 + 2 Province autonome) e su un campione rappresentativo di ASL (84 ASL così ripartite: 37 al Nord, 20 al Centro. 27 al Sud). Per i siti delle regioni il periodo di riferimento è stato dal 1 al 30 giugno 2010; per i siti delle ASL, invece, la rilevazione si è svolta in due fasi: tra il 30 aprile 2010 e il 30 giugno 2010 sono stati analizzati i siti di 6 Regioni rappresentative delle 3 aree geografiche: 11 al Nord (Lombardia e Emilia Romagna), 10 al Centro (Toscana e Lazio), 10 al Sud (Campania e Puglia); tra il 1 luglio 2010 e il 31 luglio 2010 è stata realizzata la seconda fase che ha completato la rilevazione sulle restanti 54 ASL del campione.

Il percorso seguito dalla quasi totalità dei rispondenti è quello dei motori di ricerca (Google, Yahoo, ecc.). I siti più frequentemente consultati sono il Ministero della Salute (24%), Wikipedia (20%), siti di associazioni di pazienti con specifiche patologie (17%). Le informazioni e i servizi a carattere sanitario prevalentemente cercati in rete riguardano specifiche malattie e trattamenti, ospedali e altre strutture mediche, le condizioni per un corretto stile di vita.

L'analisi della qualità dell'informazione contenuta nei siti web degli enti istituzionali del Servizio Sanitario Nazionale è stata progettata cercando di rispondere ad alcune domande: qual è l'offerta informativa online (obiettivi del sito, struttura, contenuto, servizi offerti, tecnologie impiegate, ecc.) proposta dagli enti del SSN (Regioni e ASL)? I siti delle strutture del SSN rispondono adeguatamente alle esigenze informative dei cittadini in termini di assistenza sanitaria, promozione della salute e accesso ai servizi? I siti rispettano i requisiti di usabilità e le principali norme di e-government?

È stata quindi costruita una batteria di indicatori per la misurazione della qualità, articolati in 4 aree tematiche:

1. Caratterizzazione istituzionale e relazionalità;
2. Trasparenza amministrativa;
3. Disponibilità e qualità dei servizi on line;
4. Utilizzabilità e qualità tecnologica.

Questi parametri sono stati scelti tenendo presente anche le indicazioni contenute nelle *Linee guida per i siti web della PA*, art. 4 della Direttiva 8/2009 del Ministro per la Pubblica Amministrazione e l'Innovazione[222], in particolare quelle che illustrano i contenuti minimi dei siti web istituzionali (capitolo 4.2) e i requisiti necessari per l'accessibilità e l'usabilità (4.4), quelle che sottolineano l'importanza dell'aggiornamento e della visibilità dei contenuti (4.3) e della esplicitazione della policy relativa alle caratteristiche generali dei contenuti del sito e alle modalità di trattamento dei dati eventualmente resi disponibili dall'utente (4.7).

Alla luce dei risultati emersi dalle varie ricerche, vengono proposte alcune raccomandazioni per la progettazione di una comunicazione online di qualità in ambito sanitario. Per quanto riguarda i contenuti informativi e le tipologie di interventi sanitari, i punti essenziali sono i seguenti:

- Nei siti deve essere presente materiale informativo/educativo relativo alla fisiopatologia del corpo umano, alle principali malattie (con informazioni su sintomi, diagnosi, trattamenti e prevenzione), ai comportamenti a rischio, agli interventi di prevenzione primaria (ad esempio i vaccini) e secondaria (ad esempio gli screening dei vari tumori). Il materiale informativo deve contenere anche una serie di link utili per ulteriori approfondimenti.
- Devono essere presenti anche interventi volti a promuovere comportamenti finalizzati alla promozione della salute ed alla prevenzione delle malattie e a favorire anche l'adesione ai programmi di prevenzione secondaria di provata

[222] *Direttiva n. 8/09 relativa alla riduzione dei siti web delle P.A. e per il miglioramento della qualità dei servizi e delle informazioni on line al cittadino*: http://www.funzionepubblica.gov.it/articolo/dipartimento/26-11-2009/direttiva-n-8-2009. Nel 2017 sono state pubblicate le nuove *Linee guida per il design dei servizi digitali della pubblica amministrazione*, disponibili sul sito https://design-italia.readthedocs.io/it/stable/index.html.

efficacia (screening mammografico, pap-test, screening per il carcinoma colon-rettale, etc.).

- I siti devono fornire informazioni per garantire il diritto di accesso all'assistenza sanitaria. Le informazioni devono riguardare sia le varie strutture di cura (medici di base, ambulatori specialistici, consultori, strutture per l'assistenza agli anziani e ai malati terminali, dipartimenti di emergenza e pronto soccorso, presidi ospedalieri, aziende ospedaliere, ecc.) che i livelli essenziali di assistenza e dunque i *diritti* garantiti ai cittadini in tema di salute e di assistenza sanitaria. "La centralità di questo duplice bacino di informazioni si spiega con la necessità per il Ministero della Salute di razionalizzare il ricorso dei cittadini-pazienti alle strutture sanitarie: è importante infatti tanto governare e ridistribuire sul territorio la domanda di servizi sanitari, indirizzandola verso strutture con liste d'attesa meno estese, quanto favorire l'utilizzo appropriato delle diverse strutture sanitarie (ad esempio, evitando il ricorso alle strutture di emergenza ospedaliera in caso di problematiche di salute che possono essere appannaggio del medico di base o di strutture sanitarie territoriali)" (p. 45).
- Un contenuto topico della comunicazione sanitaria è infine quello che riguarda la comunicazione dell'emergenza, sia in riferimento al presentarsi di vere o presunte pandemie sia in relazione all'emergenza sanitaria che sempre si accompagna a calamità naturali quali terremoti, alluvioni e disastri ambientali. "Poiché nella società globalizzata i media sono sempre più attori centrali nella costruzione degli eventi (anche quelli di emergenza) è importante che, in questi contesti di emergenza e rischio, le istituzioni sanitarie (e il Ministero in primis) concentrino tutti gli sforzi per garantire la rapidità e la correttezza tecnica e scientifica della comunicazione. Un'informazione non corretta o non tempestiva potrebbe, invece, amplificare il problema creando situazioni di panico nell'opinione pubblica e determinando sprechi di risorse (ad esempio, vaccini influenzali rimasti in giacenza). È opportuno, pertanto, adottare un approccio sistemico alla comunicazione in situazioni di emergenza sanitaria che coinvolga non solo le istituzioni sanitarie (centrali e locali) ma anche i mass-media e le autorità scientifiche, come fonti autorevoli e credibili di informazioni e notizie" (p. 49).

Per quanto riguarda invece la struttura e la comprensibilità del contenuto, le *Linee guida* raccomandano di:

- Curare la semplificazione della struttura dell'informazione, possibilmente adottando la regola della 'Piramide rovesciata', che prevede di esplicitare subito l'informazione primaria, cominciando dalle conclusioni e illustrando via via i dettagli del contenuto. In particolare si dovrebbero adottare i seguenti accorgimenti:
    - evidenziare il contenuto della pagina con titoli e sottotitoli chiari;
    - suddividere l'informazione visivamente: un concetto per ogni paragrafo;
    - iniziare ogni paragrafo con una breve descrizione del concetto;
    - usare il grassetto per evidenziare le parole chiave;
    - fare uso di liste numerate ed elenchi;

    - allineare i testi a sinistra (formattazione più funzionale alla riproduzione delle pagine web sui diversi supporti informatici: video, pc portatili, palmari, ecc.);
    - dotare il testo di strumenti per la navigazione interna, come un indice ipertestuale oppure il link 'Torna su'.
- Facilitare la comprensibilità del contenuto. In particolare, si dovrebbero:
    - rendere espliciti gli acronimi e le abbreviazioni;
    - evitare introduzioni arabescate e attacchi brillanti;
    - evitare titoli ironici e paradossali;
    - usare frasi semplici, senza eccesso di figure retoriche;
- Fornire informazioni e documenti aggiornati, verificando periodicamente i materiali pubblicati.
- Semplificare il linguaggio. "Tale scelta è necessaria non solo per essere in linea con il generale processo di innovazione che mette al centro dell'azione del Servizio Sanitario Nazionale il cittadino-paziente, ma anche per rispettare e valorizzare le potenzialità del mezzo comunicativo adottato: Internet" (p. 56-57). Gli accorgimenti proposti sono i seguenti:
    - evitare errori di grammatica o di scrittura;
    - costruire il messaggio cercando di rispondere alle 5W del giornalismo anglosassone (*who*, *what*, *when*, *where*, *why*);
    - preferire la forma attiva del verbo;
    - esplicitare il soggetto dell'azione, non solo come elemento di chiarezza ma anche come diretta assunzione/attribuzione di responsabilità;
    - preferire forme affermative alla doppia negazione;
    - privilegiare l'utilizzo di termini di uso comune anziché il gergo burocratico o l'eccessivo tecnicismo.
- Prediligere la pubblicazione di testi sintetici, anche attraverso il ricorso all'ipertesto e ai documenti in download.
- Prevedere strumenti per la consultazione e per la ricerca delle informazioni, come il menu di navigazione, il motore di ricerca interno, la mappa del sito, l'indice degli argomenti A-Z, ecc.

8.2. La leggibilità delle informazioni sanitarie in rete

Come abbiamo visto, tra i vari criteri proposti per la produzione di una comunicazione sanitaria online di qualità, si fa riferimento alla chiarezza e alla comprensibilità del contenuto, rese possibili tramite la semplificazione del linguaggio adottato.
Numerosi studi, soprattutto in lingua inglese, si sono occupati di valutare la leggibilità e la comprensibilità dei materiali sanitari presenti sul web. La quasi totalità delle ricerche ha mostrato che i contenuti informativi prodotti da professionisti o enti istituzionali possiedono un livello di difficoltà superiore alla capacità di comprensione dei cittadini[223]. Ne presentiamo alcuni dei principali.

[223] Secondo l'American Medical Association il livello medio di lettura dell'adulto americano varia tra il settimo e l'ottavo grado (secondo e terzo anno della scuola secondaria di primo grado). In realtà, la

Estrada et al. (2000) hanno utilizzato la formula SMOG per valutare la leggibilità degli opuscoli informativi degli anticoagulanti orali rivolti ai pazienti. Il livello di leggibilità media dei materiali è risultato essere estremamente elevato (10,7 grado), considerato soprattutto che il livello massimo consigliato è il sesto grado.
Berland et al. (2001) hanno valutato l'accessibilità, la qualità e la leggibilità di alcune informazioni sanitarie in rete (cancro al seno, depressione, obesità e asma infantile), rese disponibili attraverso motori di ricerca in lingua inglese e spagnola. La ricerca è stata condotta da luglio a dicembre 2000; per la leggibilità è stato impiegato il grafico di Fry. Solo il 20% dei motori di ricerca in inglese e il 12% in spagnolo hanno dato dei risultati pertinenti alla ricerca delle informazioni; tutti i risultati inglesi e l'86% di quelli spagnoli richiedevano un livello di lettura molto elevato.
D'Alessandro et al. (2001) hanno stimato la leggibilità di materiali web informativi di tipo pediatrico. Il campione era costituito da 89 documenti tratti da 100 siti diversi e raccolti nella biblioteca digitale GeneralPediatrics.com. La leggibilità è stata misurata tramite 4 indici: Flesch, Flesch-Kincaid, SMOG e il metodo Fry. Tutti i testi hanno riportato un livello di difficoltà superiore a quello consigliato.
Lo stesso gruppo di ricerca (Graber et al. 2002) si è occupato di determinare il livello di difficoltà delle politiche di privacy disponibili in un'ampia selezione di siti web sanitari. La leggibilità è stata calcolata in base alle formule di Flesch, Fry e SMOG. Degli 80 siti esaminati, il 30% non aveva pubblicato alcuna politica di privacy; i testi presenti nei siti rimanenti richiedevano almeno 2 anni di istruzione universitaria per essere compresi.
Anche lo studio di Ermakova et al. (2015), in Germania, è dedicato alle politiche di privacy dei siti web. I ricercatori hanno esaminato ben 5.000 documenti appartenenti a siti di ambito sanitario e 1.000 a siti di e-commerce. I risultati hanno mostrato che entrambi i gruppi di testi richiedevano almeno il livello universitario per essere compresi.
Paasche-Orlow et al. (2003) hanno analizzato la leggibilità dei documenti disponibili in 114 siti web delle scuole di medicina statunitensi tramite la formula di Flesch-Kincaid. Il punteggio medio ottenuto dal campione era 10.6, valore che risultava maggiore di 2,8 rispetto agli standard di comprensibilità proposti dalle stesse istituzioni.
Oermann et al. (2003) hanno misurato la leggibilità di siti web che offrono informazioni ai genitori sul tema della gestione del dolore nei bambini. Dei 40 siti, 29 (72,5%) fornivano indicazioni utili e rilevanti ma il livello di lettura medio è risultato comunque troppo alto (10,8) per i destinatari.
Gemoets et al. (2004) hanno condotto uno studio esplorativo in cui confrontano due diversi metodi di valutazione della leggibilità dei materiali di ambito sanitario, sia cartacei che online: la procedura cloze e gli indici di leggibilità. Per le formule di leggibilità hanno usato il loro programma *Readability Analyzer.*
Friedman et al. (2004) hanno utilizzato alcune formule (Flesch, Flesch-Kincaid e SMOG) per valutare il livello di leggibilità di 55 siti web che si occupano di vari tipi di tumori: il livello medio risultava maggiore del grado 13, che corrisponde a un'istruzione universitaria.

capacità di comprensione di un documento da parte del cittadino medio è di 2-3 gradi inferiore al grado di istruzione. Per questo motivo, l'American Medical Association e il National Institutes of Health raccomandano che i materiali di tipo sanitario siano scritti a un livello di istruzione tra il 4° e il 6° grado (cfr. Huang et al. 2015, Eltorai et al. 2014), che in Italia corrispondono alla quarta classe della scuola primaria fino al primo anno della secondaria di primo grado.

Ancora sui tumori è la ricerca di Misra et al. (2012). Gli autori hanno analizzato i materiali informativi sui tumori della base del cranio ottenuti mediante una ricerca con il motore Google; delle prime 25 risorse web recuperate, 18 erano dedicate ai pazienti. I documenti, analizzati con dieci diversi indici, sono risultati essere tutti scritti ad un livello minimo di leggibilità molto elevato (11° grado). L'anno successivo, i ricercatori (Misra et al. 2013) si sono occupati dei materiali di educazione alla salute pubblicati sul sito dell'American Academy of Facial Plastic and Reconstructive Surgery (AAFPRS). La leggibilità è stata misurata tramite 10 diversi indici (Flesch, Flesch-Kincaid, SMOG, Coleman-Liau, Fog Index, New Fog Count, New Dale-Chall, FORCAST, Raygor Readability Estimate e il grafico di Fry): il livello di istruzione medio dei documenti è risultato essere il 12° grado. Lo stesso procedimento (e con risultati simili) è stato impiegato da Huang et al. (2015) per la valutazione dei materiali informativi online tratti dalle principali associazioni oftalmologiche.

Schutten e Mc Farland (2009) hanno valutato la leggibilità di 105 pagine di siti web che si occupano di salute, i cui contenuti sono relativi ai 6 fattori di rischio per i giovani, individuati dai Centri per il controllo e la prevenzione delle malattie (CDC) degli Stati Uniti (attività sessuali, lesioni intenzionali e non intenzionali, inattività fisica, alimentazione, consumo di alcool e droghe, tabacco). L'analisi ha mostrato che il livello di difficoltà dei testi era pari all'undicesimo grado e dunque non era adeguato al target di riferimento (adulto medio o studenti delle scuole secondarie).

Badarudeen e Sabharwal (2010) hanno cercato di dimostrare che la maggior parte dei contenuti informativi online relativi all'ortopedia sono scritti ad un livello di lettura che non può essere compreso da una parte sostanziale della popolazione.

Tulbert et al. (2011) si sono occupati della leggibilità dei materiali informativi online che trattano di dermatologia. Hanno confrontato i contenuti presenti su WebMD.com, Wikipedia.org e MedicineOnline.com con quelli prodotti dalla American Academy of Dermatology. Nessuno dei documenti possedeva caratteristiche ottimali per quanto riguarda la leggibilità, la lunghezza dei testi e la presenza di illustrazioni esemplificative.

Eltorai et al. (2014) hanno valutato se i materiali disponibili sul sito dell'American Association for Surgery of Trauma rispettavano i criteri di leggibilità previsti dalle linee guida nazionali. Tutti gli articoli, tranne uno, non rispettano il livello minimo di istruzione previsto (6° grado).

Hansberry et al. (2014) hanno scaricato 138 articoli sull'educazione sanitaria da RadiologyInfo.org, il sito web sponsorizzato dall'American College of Radiology e dalla Radiological Society of North America. I testi risultavano scritti in media tra il 10° e il 14° grado, ben al di sopra delle raccomandazioni dell'American Medical Association e del National Institutes of Health. Qualche anno più tardi (Hansberry et al. 2017) gli studiosi conducono un'analisi completa della leggibilità dei materiali informativi riguardanti le sotto-specialità della medicina interna (allergia e immunologia, cardiologia, endocrinologia, gastroenterologia, geriatria, ematologia, hospice e cure palliative, malattie infettive, nefrologia, oncologia, pneumologia e terapia intensiva, reumatologia, medicina del sonno e medicina dello sport); i materiali sono tratti da 14 siti di organizzazioni professionali. Lo studio conclude che nessuno dei materiali esaminati è conforme alle linee guida nazionali.

Patel et al. (2015) hanno stimato la leggibilità del materiale informativo online relativo alla chirurgia paratiroidea. Hanno individuato 30 testi tramite il motore di ricerca Google e li

hanno analizzati con 4 diverse formule di leggibilità (Flesch, Flesch-Kincaid, Fog Index e SMOG). I documenti sono risultati tutti scritti ad un livello superiore rispetto allo standard raccomandato (6° grado).

Morony et al. (2015) hanno valutato la leggibilità di 80 materiali informativi destinati a pazienti con malattia renale cronica. Il campione, raccolto nel maggio del 2014, era formato da diversi documenti scritti provenienti dall'Australia e da materiali online presenti in siti web di note organizzazioni internazionali. La leggibilità è stata misurata tramite la formula di Flesch-Kincaid e il sistema Lexile: entrambi i sistemi hanno rilevato che l'età minima necessaria per comprendere il testo era 14-15 anni; soltanto il 5% dei contenuti rispettava il livello di istruzione raccomandato (5° grado)[224].

Roberts et al. (2016) hanno condotto uno studio sui materiali informativi presenti sul sito dell'American Academy of Orthopaedic Surgeons (AAOS), confrontandoli con i risultati di una precedente ricerca nel 2008. La leggibilità è stata valutata tramite 5 indici: Flesch, Flesch-Kincaid, Coleman e Liau, Fog Index e SMOG. Indipendentemente dalla metrica utilizzata, i livelli di leggibilità risultavano più elevati di quelli raccomandati; sebbene la formula di Flesch-Kincaid riportasse dei punteggi inferiori rispetto allo studio del 2008, restava la necessità di migliorare la comprensibilità del materiale informativo presente sul sito.

Beaunoyer et al. (2016) hanno descritto gli strumenti e le strategie disponibili per valutare le caratteristiche delle informazioni sanitarie, come la leggibilità, il contenuto emotivo (analisi del *sentiment*), la comprensibilità e l'usabilità. Hanno esaminato sia il materiale cartaceo che quello in rete.

Diversi studi sono stati condotti sui testi presenti nella versione inglese di *Wikipedia.*

Brigo et al. (2015) hanno valutato la leggibilità di 8 popolari siti web in inglese (tra cui appunto Wikipedia) che forniscono informazioni sull'epilessia. Nello stesso anno, Brigo e Erro (2015) hanno analizzato l'articolo relativo al Parkinson presente nel sito dell'enciclopedia online[225]; questo documento è stato scelto in quanto era il primo risultato che appariva (il 7 novembre 2014, giorno in cui la ricerca è stata condotta) interrogando il motore di ricerca Google, con le parole chiave "Malattia di Parkinson". Watad et al. (2017) hanno misurato la leggibilità delle pagine web di Wikipedia relative a 134 malattie autoimmuni. Le formule utilizzate nelle tre ricerche sono state le stesse: Flesch, Flesch-Kincaid, SMOG, ARI, Coleman e Liau e l'indice Fog. I risultati sono stati simili: i testi presentavano in generale un basso livello di leggibilità e per essere compresi richiedevano, nel caso dei primi due studi, almeno un'istruzione superiore e, nel terzo, un'istruzione universitaria.

Nel Regno Unito, Edmunds et al. (2013) hanno misurato la leggibilità di materiali online correlati alle diagnosi oftalmiche, riportando deludenti risultati e Vivekanantham et al. (2017) hanno studiato la leggibilità e la qualità delle informazioni sanitarie online relative alla polimialgia reumatica (PMR). Hanno individuato 50 siti web tramite tre motori di ricerca (Google, Yahoo e Bing) e misurato i testi con le formule SMOG e Flesch: la maggior parte dei testi richiedeva un livello di istruzione superiore a quello raccomandato. Infine, Flinton et

[224] Ancora in contesto australiano, Tieman e Bradley (2013) hanno condotto una revisione sistematica dei vari tipi metodi e approcci alla valutazione dell'efficacia dei siti web di ambito sanitario.

[225] https://en.wikipedia.org/wiki/Parkinson%27s_disease.

al. (2018) si sono occupati dei contenuti informativi online destinati ai pazienti in radioterapia. I documenti sono risultati essere scritti a un livello di lettura troppo alto, anche se inferiore rispetto a uno studio simile condotto nel 2006.

Per quanto riguarda l'Italia, le ricerche sulla leggibilità dei testi in ambito sanitario sono pochissime. Abbiamo già citato alcuni studi che si sono occupati di valutare la leggibilità di quelle tipologie di testi considerate rappresentative della comunicazione medico-paziente, come i foglietti illustrativi (bugiardini) dei farmaci senza obbligo di prescrizione medica (Dell'Orletta et al. 2016) e le informative di consenso per le procedure diagnostico-terapeutiche (Venturi et al. 2015, Dell'Orletta et al. 2017).
Anche la ricerca di Terranova et al. (2012) riguarda i moduli di consenso informato utilizzati in cardiologia. Lo scopo era valutare se i 7 moduli utilizzati in quel periodo, sia nella verisone inglese che in quella italiana, erano conformi alle raccomandazioni degli standard diriferimento e se, eventualmente, era possibile migliorarne la qualità. I testi in inglese sono stati analizzati con l'indice di Flesch-Kincaid, l'indice Fog, le formule SMOG, ARI e Coleman e Liau; quelli in italiano tramite la formula GULPEASE e lo strumento READ-IT. La qualità e la leggibilità complessive risultavano scarse sia nella versione inglese che in quella italiana; è stato ottenuto però un miglioramento sostanziale con la revisione e la semplificazione linguistica dei moduli.
Un altro lavoro interessante è quello di Cavallo et al. (2001) sulla leggibilità dei referti radiologici[226]. Da un corpus di 400 referti, ne sono stati selezionati 40 in modo casuale, 10 per ciascuna procedura diagnostica (RT, ECO, TC RM). Il campione è stato poi sottoposto a un'analisi quantitativa, utilizzando un software dedicato e a un'analisi qualitativa che ha tenuto conto di aspetti formali, sintattici e lessicali. Sulla base dei risultati ottenuti, i referti sono stati modificati e poi sottoposti a una nuova analisi: la comprensibilità e la leggibilità sono notevolmente migliorate, in particolare per le metodiche di ecografia e radiologia tradizionale. Il miglioramento è stato meno evidente per la tomografia computerizzata e la risonanza magnetica a causa della maggiore presenza di tecnicismi.
Più recente lo studio di Brugnolli et al. (2014) relativo alla leggibilità[227] e comprensibilità delle linee guida sull'igiene delle mani. Gli studiosi hanno confrontato le linee guida pubblicate nel 2009 dall'Organizzazione Mondiale della Sanità (OMS)[228] con quelle divulgate dai Centres for Disease Control (CDC) nel 2002[229], per identificare le discrepanze e le novità

[226] Una ricerca simile era già stata condotta in lingua inglese da Sierra et al. (1992). I ricercatori avevano analizzato la leggibilità di alcuni referti radiologici tramite la formula di Flesch-Kincaid. Le modalità radiografiche analizzate erano 4: radiografia generale, mammografia, ecografia e risonanza magnetica. L'analisi aveva dimostrato che la leggibilità era diversa a seconda della procedura diagnostica considerata e che risultava maggiore nel caso di mammografia e risonanza magnetica.
[227] In realtà nello studio non si fa ricorso alle formule di leggibilità; si tratta piuttosto di uno studio delle caratteristiche linguistiche dei testi.
[228] World Health Organization, *WHO Guidelines on Hand Hygiene in Health Care*, Geneva, Switzerland: World Health Organization Press, 2009.
[229] J. M. Boyce, D. Pittet, Healthcare Infection Control Practices Advisory Committee, Society for Healthcare Epidemiology of America, Association for Professionals in Infection Control, Infectious Diseases Society of America, Hand Hygiene Task Force, *Guideline for Hand Hygiene in Health-Care Settings: recommendations of the Healthcare Infection Control Practices Advisory Committee and the HICPAC/SHEA/APIC/IDSA Hand Hygiene Task Force*, Morbidity and Mortality Weekly Report, 51, pp. 1-44.

introdotte; le linee guida del 2009, infatti, integrano e aggiornano le precedenti. Data la rilevanza del lavaggio delle mani per la prevenzione delle infezioni, queste norme sono tra le più diffuse e utilizzate in ambito sanitario. Dall'analisi è emerso che, sebbene risultino comparabili per molte raccomandazioni, le due versioni usano una diversa terminologia, rendendo ambigua l'interpretazione: ad esempio, nelle linee guida del 2002 per *handwashing* ('lavaggio delle mani') si intende il lavaggio con acqua e sapone semplice, in quelle del 2009 è compresa anche l'opzione del sapone antimicrobico.

Più propriamente dedicata ai materiali presenti sul web è la ricerca di Dini et al. (2017).

Gli studiosi hanno analizzato l'affidabilità e la leggibilità dei contenuti dei siti web italiani sulla silicosi[230]. Utilizzando il termine chiave *silicosi,* hanno effettuato un'interrogazione nei principali motori di ricerca (Google, Yahoo, Bing, Ask, Libero Arianna)[231] e hanno preso in considerazione i risultati delle prime 3 pagine (per un totale di 30 risultati per motore di ricerca). Una volta esclusi dalla raccolta i duplicati, i siti non pertinenti alla ricerca, quelli che non contenevano sufficienti informazioni da analizzare, i video di YouTube (in quanto avrebbero dovuto essere trascritti) e le presentazioni di congressi o convegni, i ricercatori hanno ottenuto un campione finale composto da 70 siti. Di questi, il 26% sono siti istituzionali, il 21% siti che si occupano di salute.

L'affidabilità è stata valutata in base alla presenza della certificazione *HONCode*, che viene rilasciata solo nel caso in cui il sito soddisfi gli standard richiesti in termini di qualità dell'informazione sanitaria. Solo l'1,4% è risultato essere conforme allo standard HONCode.

La leggibilità è stata invece misurata tramite la formula GULPEASE e lo strumento READ-IT. È emersa una forte variabilità nei punteggi: in particolare, l'indice GULPEASE e il Modello Lessicale di READ-IT hanno restituito valori diversi tra le varie tipologie di siti (istituzionali, accademici, commerciali, ecc.). In generale, i materiali risultano difficili da comprendere in tutti i tipi di siti e, in particolare, in quelli degli enti istituzionali.

Un'iniziativa interessante è il *Sistema di Valutazione delle Performance dei Sistemi Sanitari Regionali*, promosso dal Laboratorio MeS (Management e Sanità) della Scuola Superiore Sant'Anna di Pisa.

Il progetto è nato con l'obiettivo di fornire un'analisi dei sistemi sanitari regionali e delle aziende operanti nelle diverse aree territoriali, attraverso il confronto di un set di indicatori condivisi. Viene attivato nel 2008, attraverso la collaborazione di quattro regioni (Toscana, Liguria, Umbria e Piemonte, che uscirà nel 2010); nel 2010 si aggiungono la Valle d'Aosta (che uscirà nel 2012), la Provincia Autonoma di Trento, la Provincia Autonoma di Bolzano e le Marche; nel 2011 la Basilicata, nel 2012 il Veneto e nel 2014 l'Emilia Romagna e il Friuli Venezia Giulia. Nel corso del 2015, aderiscono anche la Calabria, il Lazio, la Lombardia, la Puglia e la Sardegna.

[230] La silicosi è una malattia professionale causata dall'inalazione cronica e dalla penetrazione nei polmoni di polveri di silicio biossido di silicio. Recentemente la mattia ha acquisito un rinnovato interesse, a causa del fatto che sono emersi nuovi fattori di rischio, come la sabbiatura dei jeans o la fratturazione idraulica.

[231] Si tratta dei primi 5 motori di ricerca più frequentemente utilizzati dagli italiani per navigare sul web.

Il processo di condivisione tra regioni ha portato alla selezione di circa 300 indicatori, di cui 150 di valutazione e 150 di osservazione, volti a descrivere e confrontare 6 diverse dimensioni della performance del sistema sanitario:

- lo stato di salute della popolazione (dimensione A);
- la capacità di perseguire le strategie regionali (dimensione B);
- la valutazione sociosanitaria (dimensione C);
- la valutazione della soddisfazione e dell'esperienza degli utenti (dimensione D);
- la valutazione da parte dei dipendenti (dimensione E);
- la valutazione della dinamica economico-finanziaria e dell'efficienza operativa (dimensione F).

Ogni regione ha inoltre la possibilità di inserire indicatori specifici, volti a esplorare particolari aspetti che siano rilevanti per il perseguimento di strategie territoriali e non necessariamente di interesse per gli altri soggetti coinvolti.

A partire dalla fine del 2015, il Laboratorio MeS ha condotto uno studio sulla comunicazione online, effettuando una rilevazione su 167 siti web delle aziende sanitarie delle 13 Regioni italiane aderenti al Sistema di Valutazione[232]. Lo scopo era offrire una panoramica sulle modalità con cui le aziende sanitarie comunicano con i cittadini attraverso i siti web e, in particolare, sul livello di digitalizzazione della prenotazione dei servizi sanitari. Dai dati presentati nella ricerca sono stati elaborati alcuni indicatori di valutazione, inseriti nella dimensione B (B31 - *Comunicazione e prenotazione web*).

La ricerca ha fornito anche alcune informazioni sul tema della riorganizzazione e integrazione della comunicazione online in caso di riorganizzazione del sistema sanitario regionale. Nel biennio 2014-2015, infatti, molte aziende sanitarie territoriali si sono accorpate in unità territoriali più vaste; si è trattato di una riorganizzazione aziendale finalizzata a migliorare l'efficienza dei servizi e a ottimizzare le risorse senza però influire sulla qualità dell'offerta per il cittadino.

Tenendo presente le *Linee guida* pubblicate dal Ministero, sono state individuate alcune caratteristiche essenziali su cui effettuare la rilevazione:

- presenza della funzione di ricerca e suo corretto funzionamento;
- responsività del sito (cioè l'adattamento del sito a seconda del dispositivo impiegato per visualizzarlo);
- offerta del servizio di prenotazione delle visite specialistiche nelle varie declinazioni (CUP telefonico, prenotazione via web, APP dedicate), corretta descrizione e spiegazione del servizio, verifica del funzionamento[233];
- leggibilità dei testi (valutazione della leggibilità tramite l'indice GULPEASE e confronto del lessico con il *Vocabolario di Base*).

Per la misurazione della leggibilità è stato utilizzato il software *Corrige!* di Èulogos. Sono stati analizzati, per ogni azienda, i testi presenti nelle pagine web dedicate alla spiegazione

[232] I risultati sono stati raccolti nel Quaderno "Comunicare Sanità. Strumenti online per i servizi ai cittadini" pubblicato nel 2016 e in parte sono confluiti nel sistema di valutazione delle performance sanitarie delle regioni. Ogni anno viene pubblicato un report con i risultati; dal 2010 il report è pubblico.

[233] La funzione di prenotazione è stata scelta in quanto è trasversale rispetto a tutti i tipi di azienda sanitaria.

del servizio di prenotazione dei servizi sanitari. Nei casi in cui tale contenuto non era presente, il valore era considerato nullo.
Dai risultati è emersa una grande variabilità sia tra le varie regioni, sia all'interno della stessa regione, che può presentare buoni risultati rispetto a un ambito e risultati meno positivi su altri aspetti. Ad esempio, la Provincia Autonoma di Bolzano risulta avere il punteggio più alto per l'indice GULPEASE ma tra i più bassi per quanto riguarda l'uso di parole appartenenti al VdB. Questo tipo di risultati può essere spiegato dalla presenza nei testi di numerosi elenchi di prestazioni o malattie, che influiscono sul parametro della formula relativo alla lunghezza dei testi.
In generale, la valutazione della leggibilità mostra che la maggior parte dei testi può essere compresa solo da chi possiede almeno la licenza della scuola superiore. Solo i materiali presenti nel sito della Provincia Autonoma di Bolzano hanno un livello di difficoltà più basso, corrispondente alla scuola secondaria di primo grado; tuttavia, come abbiamo detto, tali documenti presentano percentuali molto basse di parole appartenenti al lessico di base. Veneto, Calabria, Liguria, Emilia-Romagna e Lombardia presentano valori medio-alti per entrambi gli indicatori. In media, la percentuale di parole appartenenti al *Vocabolario di Base* è inferiore al 60%, con risultati che vanno da un minimo del 24% a un massimo del 72%.

8.3. Definizione e costruzione del corpus

Il Servizio Sanitario Nazionale (SNN) è un insieme di enti e organi strutturati su tre livelli: il primo livello è composto dal Ministero della Salute, che coordina il piano sanitario nazionale; al secondo livello si trovano le 20 regioni e le province autonome di Trento e Bolzano; al terzo livello si trovano le aziende sanitarie locali (ASL), le aziende ospedaliere (AO) e le aziende ospedaliere universitarie (AOU).
Il SSN è stato istituito con la riforma sanitaria del 1978 (Legge 23 dicembre 1978); la riforma soppresse il sistema mutualistico e decretò che i servizi sanitari divenissero a carico statale e di competenza delle Unità Sanitarie Locali (USL), anch'esse appena istituite.
Con la legge 502 del 1992 le USL acquisiscono una propria autonoma (organizzativa, patrimoniale, gestionale, ecc.) e si staccano dal governo centrale per divenire parte dei servizi sanitari regionali (cioè dipendenti dalle regioni); durante questo passaggio il nome USL si trasforma in ASL.
Perché si possa identificare l'oggetto della nostra indagine, è necessaria una premessa: ciascuna regione è ormai libera di scegliere una propria denominazione per le aziende sanitarie. Quello che chiamiamo corpus delle ASL comprende quindi tutte le varie aziende sanitarie, indipendentemente dal nome che hanno assunto.
Abruzzo, Campania, Lazio, Liguria, Piemonte, Puglia sono le uniche regioni che hanno mantenuto la denominazione ASL. In Basilicata, Calabria e Sicilia le aziende sono state rinominate ASP (Azienda Sanitaria Provinciale); in Toscana e Umbria c'è stato il passaggio alla denominazione AUSL (Azienda Unità sanitaria Locale) e successivamente a USL (Unità Sanitaria Locale). Anche l'Emilia Romagna e la Valle d'Aosta hanno optato per il nome AUSL. Le altre regioni hanno invece ulteriori denominazioni: nel Friuli l'azienda sanitaria si chiama ASUI (Azienda Sanitaria Universitaria Integrata) o AAS (Azienda per l'Assistenza Sanitaria), a seconda della zona; in Lombardia ATS (Agenzia di Tutela della Salute), in Veneto ULSS (Unità

Locale Socio Sanitaria), nelle Marche ASUR (Azienda Sanitaria Unica Regionale), nel Molise ASREM (Azienda Sanitaria Regionale del Molise), in Sardegna l'ATS (Azienda per la Tutela della Salute) comprende diverse ASSL (Azienda Socio Sanitarie Locali). Per quanto riguarda le province autonome, l'azienda sanitaria di Trento è denominata APSS (Azienda Provinciale per i Servizi Sanitari) e quella di Bolzano è ASDAA (Azienda Sanitaria dell'Alto Adige).

8.3.1. Criteri di selezione dei testi

Prima di selezionare i testi per la composizione del corpus, abbiamo effettuato un'indagine esplorativa sui vari siti delle ASL, al fine di indagare quale fosse l'offerta informativa presente.

Nei siti web degli enti sanitari è infatti possibile individuare tipologie diverse di comunicazione. Una prima classificazione può essere fatta in base al destinatario dei testi: esiste una comunicazione "interna", rivolta a operatori e professionisti sanitari, che comprende bandi, concorsi, selezioni del personale, mobilità, informazioni, avvisi e news; nei casi in cui si tratti di aziende universitarie, esiste anche una comunicazione interna rivolta a studenti, professori e a chiunque sia coinvolto nella formazione. C'è poi una comunicazione rivolta verso "l'esterno", che a sua volta si divide in due categorie: i testi rivolti ai cittadini/utenti dei servizi sanitari e quelli rivolti a fornitori, partner, associazioni di volontariato, scuole, ecc.

Se invece consideriamo le diverse tipologie testuali, è possibile distinguere tra testi di ambito legislativo-giuridico (atti amministrativi, bandi, concorsi, albo pretorio), testi che comprendono informazioni istituzionali (informazioni su orari, luoghi, personale, servizi e istituzioni, carta dei servizi, avvisi e notizie), testi regolativi (istruzioni, regolamenti, linee guida, direttive), testi informativi (opuscoli informativi, guide, FAQ), testi prodotti dagli uffici stampa (rassegna stampa, comunicati stampa, interviste, eventi, ecc.), modulistica varia.

Esiste anche una terza modalità di classificazione della comunicazione in ambito sanitario: alcuni studiosi (Del Vecchio e Rappini 2009, Panini e Fiorini 2014) distinguono infatti tra comunicazione per la salute, comunicazione sanitaria, comunicazione istituzionale e comunicazione interpersonale (medico-paziente)[234]. Si tratta di processi comunicativi distinti, ma con profonde aree d'integrazione e di sinergia.

La comunicazione alla salute rappresenta lo strumento tramite il quale l'azienda opera per l'educazione e la promozione alla salute; comprende diversi tipi di contenuti, come le campagne di prevenzione primaria e quelle di promozione relative agli stili di vita corretti (alimentazione, attività fisica, dipendenze) e si rivolge all'intera comunità.

La comunicazione sanitaria è quella tramite cui vengono presentati i servizi e prodotti offerti dalle aziende sanitarie, in tutti i livelli dell'assistenza (prevenzione, diagnosi, cura e riabilitazione); questo tipo di contenuti mira a guidare le scelte dell'utente verso i servizi offerti dall'azienda, sviluppandone l'*empowerment*. All'interno della comunicazione sanitaria hanno particolare rilievo la comunicazione di crisi ed emergenza e l'ambito della prevenzione secondaria.

[234] La distinzione tra comunicazione per la salute e comunicazione sanitaria è presente anche nel *Documento di indirizzo sulla comunicazione pubblica in sanità*, redatto dalla Commissione Sanità e Salute dell'Associazione Italiana dei Comunicatori Pubblici nel 2006.

La comunicazione istituzionale favorisce la legittimazione dell'immagine dell'azienda nel rapporto con i diversi pubblici di riferimento e con i vari portatori di interesse (*stakeholders*): l'immagine percepita dall'esterno deve essere quanto più possibile vicina a quella realmente offerta. Fanno parte della comunicazione istituzionale tutti quei contenuti informativi utili a fare conoscere le attività dell'azienda, i servizi offerti, gli investimenti, ecc. La comunicazione interpersonale, infine, è quella che riguarda i rapporti tra operatori sanitari e pazienti (ma anche i rapporti tra operatori stessi). "L'indagine della relazione medico-paziente è da sempre un tema scottante e allo stesso tempo affascinante in ambito medico-sanitario. La parola comunicazione, che spesso sembra esserne sinonimo, è un termine opaco e impreciso dal punto di vista semantico per definire un'area di interazione tra medico e paziente molto articolata e complessa e raccoglie sotto un'etichetta linguistica codificata, diversi approcci, studi e metodi di analisi" (Revellino 2017, p. 149).

Alla luce di questa analisi, abbiamo focalizzato la nostra attenzione su alcuni contenuti informativi che fanno parte della comunicazione sanitaria rivolta ai cittadini/utenti. In particolare, abbiamo scelto quattro argomenti:

- il servizio di emergenza-urgenza;
- lo screening oncologico;
- l'assistenza sanitaria agli stranieri;
- l'assistenza domiciliare.

Si tratta di tematiche per le quali la chiarezza e la comprensibilità dell'informazione sono estremamente essenziali; sono inoltre tra i contenuti che, secondo le *Linee guida* del Ministero della Salute per la progettazione di una comunicazione sanitaria online di qualità, dovrebbero sempre essere presenti nei siti web degli enti istituzionali.
Per quanto riguarda il servizio di emergenza sanitaria, ci siamo concentrati su quei testi che riguardano il funzionamento del 118. Questo tipo di informazioni non deve soltanto rispondere a requisiti di chiarezza e correttezza, ma deve essere anche facilmente riconoscibile: in caso di necessità, l'utente deve poter individuare in modo immediato le informazioni principali. Si pensi al numero stesso dell'emergenza sanitaria: in caso di bisogno, il cittadino deve subito poter capire se in quella data regione è attivo il numero unico per le emergenze (112) o se invece si deve comporre il 118.
In alcuni siti, le sezioni che riguardano i servizi di emergenza e urgenza integrano sia le informazioni relative al 118, sia quelle riguardanti il pronto soccorso e il triage. A meno che non facessero parte dello stesso testo, abbiamo deciso di escludere le pagine relative al pronto soccorso.
Gli screening oncologici fanno parte degli interventi di prevenzione secondaria. Tali interventi hanno come obiettivo fondamentale la riduzione della mortalità e dell'incidenza dei tumori. Dal 2001 gli screening sono inclusi nei LEA (Livelli Essenziali di Assistenza)[235]. Nel

[235] I Livelli Essenziali di Assistenza (LEA) sono le prestazioni e i servizi che il Servizio Sanitario Nazionale è tenuto a fornire a tutti i cittadini, gratuitamente o dietro pagamento di una quota di partecipazione (ticket). Sono determinati dalla legislazione statale ma spetta alla legislazione regionale il compito di organizzare ed erogare tali prestazioni. I LEA sono organizzati in tre settori: prevenzione collettiva e sanità pubblica, assistenza distrettuale, assistenza ospedaliera. La prima comprende tutte le attività di prevenzione rivolte ai singoli o alla collettività (prevenzione e controllo delle malattie infettive e parassitarie, tutela della salute e sicurezza sui luoghi di lavoro, sicurezza alimentare, programmi di screening, ecc.). L'assistenza distrettuale comprende tutte le attività e i

corpus abbiamo incluso tutti i testi relativi ai 3 programmi di prevenzione dei tumori: screening mammografici, pap test o screening della cervice uterina, screening del colon-retto. La comprensibilità di tali contenuti, oltre a garantire una corretta informazione sulle varie prestazioni, è funzionale anche per favorire e incrementare l'adesione della popolazione ai programmi di prevenzione.
La terza tematica riguarda l'assistenza sanitaria agli stranieri, che comprende diversi contenuti: le modalità di accesso al SSN da parte degli stranieri e i diritti che vengono loro garantiti, informazioni su strutture, ambulatori e sportelli dedicati, servizi specifici come la mediazione culturale, ecc. Dal momento che spesso tali contenuti sono presenti nella stessa pagina (o comunque sono riuniti sotto la stessa sezione), abbiamo scelto di includerli tutti. Sono stati raccolti anche i testi in cui si parla di *migranti* o di *immigrati*, oltre che di (o invece di) *stranieri*. La chiarezza di questo tipo di testi deve essere massima, non solo per favorire la comprensione anche a coloro che hanno competenze in italiano solo come seconda lingua, ma per garantire loro il diritto stesso all'assistenza sanitaria.
L'assistenza domiciliare (o Cure domiciliari) è un servizio dedicato a anziani, disabili e persone autosufficienti; è interamente a carico del SSN, in quanto è inserito nei LEA. Alle cure domiciliari si accede tramite una valutazione multidimensionale della condizione socio-sanitaria dell'assistito; in base ai bisogni di questo, si distinguono diverse tipologie: l'Assistenza Domiciliare Programmata (ADP) e Assistenza Domiciliare Integrata (ADI). La prima eroga prestazioni sanitarie mediche, infermieristiche e/o riabilitative, limitate all'episodio di malattia in atto ed è attivata dal medico di base o dai servizi distrettuali delle ASL. La seconda riguarda prestazioni sanitarie (mediche, infermieristiche, riabilitative) e sociosanitarie (cura della persona), erogate a domicilio in modo coordinato e continuativo. Per la costruzione del corpus, si è scelto di includere, se presenti, i testi inerenti a entrambe le modalità.

8.3.2. Composizione del corpus

Una volta definiti i criteri di selezione dei testi, ci siamo occupati di stabilire quali siti web delle ASL includere nel campione. Inizialmente, avevamo deciso di valutare soltanto i siti web della ASL dei capoluoghi di regione; successivamente però, abbiamo deciso di ampliare il campione e includere altre città principali. Tale decisione è dipesa da una duplice necessità: da una parte, perché molti dei siti dei capoluoghi sono ancora provvisori o in fase di rinnovamento (alcuni anche da diversi anni) e molti dei contenuti non sono presenti, o si trovano divisi tra il vecchio sito e quello nuovo. È il caso ad esempio di quelle regioni, come la Toscana, in cui c'è stato un processo di fusione delle diverse aziende sanitarie, con la conseguente nascita di un nuovo sito web specifico per l'area vasta.
L'altro motivo è legato ai contenuti scelti come riferimento: abbiamo verificato l'effettiva presenza di tali contenuti nelle pagine web delle ASL ed è emerso che, in alcuni siti, alcuni di questi non sono disponibili. Abbiamo dunque scelto di includere i siti di ASL di altre città, nelle quali sono invece presenti.

---

servizi sanitari e socio-sanitari diffusi sul territorio (assistenza sanitaria di base, emergenza sanitaria territoriale, assistenza farmaceutica, ambulatoriale, domiciliare, residenziale, ecc.). L'assistenza ospedaliera coinvolge le attività di pronto soccorso, day hospital, day surgery, ricovero ordinario, riabilitazione, ecc.

La rilevazione è stata effettuata su tutte le 19 regioni italiane e sulle due province autonome (Trento e Bolzano)[236]; abbiamo selezionato un campione di 30 siti web delle aziende sanitarie, così ripartiti: 13 ASL del Nord, 5 del Centro, 9 del Sud e 3 delle isole. La lista definitiva delle ASL è riportata nella Tabella 101.
Per ciascun sito, abbiamo raccolto i testi riguardanti le 4 tematiche scelte (servizio di emergenza, screening oncologici, assistenza sanitaria agli stranieri e assistenza domiciliare). In alcuni casi, all'argomento corrispondeva un solo testo; in altri, il contenuto era strutturato su più pagine. Abbiamo considerato ogni singola pagina web come un testo. Abbiamo invece deciso di non includere nel campione gli eventuali allegati presenti nelle pagine web, come opuscoli e materiale informativo.
Il corpus è composto da 248 documenti, per un totale di 122.793 occorrenze. I testi sono così ripartiti: 32 riguardano il servizio di emergenza (16.857 occorrenze), 85 gli screening oncologici (44.377 occ.), 83 l'assistenza sanitaria agli stranieri (42.193 occ.) e 48 l'assistenza domiciliare (19.366 occ.).

[236] Le regioni italiane sono 20 ma in realtà il Trentino Alto Adige è una regione autonoma a statuto speciale costituita da due province autonome, Trento e Bolzano. Esistono altre regioni italiane a statuto speciale: Valle d'Aosta, Friuli Venezia Giulia, Sicilia e Sardegna. La Valle d'Aosta è l'unica regione a non essere suddivisa in province ma in comuni. Tutte le altre regioni sono organizzate in enti di area vasta, che comprende le province e le città metropolitane. Attualmente, in Italia esistono 14 città metropolitane; al 1° gennaio 2017 il numero delle province era 92, ma è sceso nel corso del 2017 e del 2018.

Regione	Città	Azienda Sanitaria	URL
Italia settentrionale			
Emilia Romagna	Bologna	AUSL di Bologna	http://www.ausl.bologna.it/
	Parma	AUSL di Parma	https://www.ausl.pr.it/default.aspx
Friuli Venezia Giulia *	Trieste	ASUITS di Trieste	http://www.asuits.sanita.fvg.it/it/index.html
Liguria	Genova	ASL 3 Genovese	http://www.asl3.liguria.it/
Lombardia	Milano	ATS di Milano	https://www.ats-milano.it/portale
	Bergamo	ATS Bergamo	http://www.ats-bg.it/servizi/notizie/notizie_homepage.aspx
Piemonte	Torino	ASL TO1	http://www.asl102.to.it/#
	Torino	ASL TO2	http://www.aslto2.piemonte.it/_front/front.php
	Cuneo	ASL CN1	http://www.aslcn1.it/
Trentino *	Trento	APSS Provincia Autonoma di Trento	https://www.apss.tn.it/
	Bolzano	Azienda sanitaria dell'Alto Adige	http://www.asdaa.it/it/default.asp
Valle d'Aosta *	Aosta	Azienda USL Valle d'Aosta	http://www.ausl.vda.it/homepage.asp?l=1
Veneto	Venezia	ULSS3 Serenissima	https://www.aulss3.veneto.it/
Italia centrale			
Lazio	Roma	ASL Roma 1	http://www.aslroma1.it/
Marche	Ancona	ASUR Area Vasta 2	http://www.asurzona7.marche.it/home.asp
Toscana	Firenze, Empoli, Prato, Pistoia	USL Toscana Centro	http://www.uslcentro.toscana.it/
	Arezzo, Grosseto, Siena	USL Toscana Sud Est	http://www.uslsudest.toscana.it/
Umbria	Perugia	USL Umbria 1	http://www.uslumbria1.gov.it/
Italia meridionale			
Abruzzo	Aquila, Avezzano, Sulmona	ASL 1 Abruzzo	http://www.asl1abruzzo.it/
	Pescara	ASL 3 Pescara	http://www.ausl.pe.it/index.jsp
	Chieti, Lanciano, Vasto	ASL 2 Abruzzo	http://www.asl2abruzzo.it/
Basilicata	Potenza	ASP di Potenza	http://www.aspbasilicata.it/
Calabria	Catanzaro	ASP Catanzaro	http://www.asp.cz.it/
Campania	Napoli	ASL Napoli 1 Centro	http://www.aslnapoli1centro.it/
Molise	Campobasso	ASREM Campobasso	https://www.asrem.gov.it/
Puglia	Bari	ASL Bari	https://www.sanita.puglia.it/web/asl-bari
	Brindisi	ASL Brindisi	https://www.sanita.puglia.it/web/asl-brindisi
Isole			
Sardegna *	Cagliari	ASSL Cagliari	https://www.aslcagliari.it/

Regione	Città	Azienda Sanitaria	URL
Sicilia *	Palermo	ASP Palermo	http://www.asppalermo.org/
	Agrigento	ASP Agrigento	http://www.aspag.it/

Tabella 101.Lista delle Aziende Sanitarie include nel campione.
Le regioni contrassegnate con (*) sono quelle a statuto speciale.

8.3.3. I siti delle ASL

Durante la fase di selezione e raccolta dei testi per la composizione del corpus, abbiamo effettuato uno studio critico dei diversi siti della ASL. Abbiamo cercato di valutare alcuni aspetti legati alla reperibilità dei 4 specifici contenuti e, più in generale, alla navigabilità del sito.

In particolare, l'analisi ha riguardato le seguenti questioni:

- Individuazione del sito web tramite il motore di ricerca;
- Presenza sulla home page del sito della funzionalità di ricerca interna al sito;
- Presenza nel sito dei contenuti cercati:
    - Il contenuto è presente?
    - Si trova in home page o in una pagina interna?
    - Quanti click sono necessari per raggiungere il contenuto?
    - Il contenuto è disponibile in una delle voci del menu (principale o secondario) o va cercato tramite il motore di ricerca interno al sito?
    - È possibile visualizzare il percorso di navigazione relativo a quella specifica pagina?
- Visualizzazione del sito tramite diversi browser;
- Responsività del sito;
- Accessibilità del sito.

Per quanto riguarda i contenuti specifici, ci siamo invece concentrati sui seguenti aspetti:

- Presenza di una pagina di spiegazione del servizio di emergenza;
- Indicazione del numero di emergenza da chiamare;
- Presenza delle norme da seguire in caso di emergenza;
- Presenza di una pagina dedicata agli screening oncologici;
- Presenza di una pagina specifica per ciascun programma di screening;
- Presenza di numeri di telefono; indicazione di giorni, orari, strutture.
- Presenza di opuscoli informativi e approfondimenti;
- Presenza di una sezione dedicata agli stranieri;
- Presenza di informazioni relative alla modalità di accesso al SSN (per cittadini comunitari, extracomunitari, senza permesso di soggiorno)
- Presenza di altri servizi e informazioni (ambulatori, sportelli dedicati, mediazione culturale, ecc.)
- Presenza di una sezione dedicata agli anziani o ai disabili;
- Presenza di una pagina informativa relativa all'assistenza domiciliare (chi è il destinatario, come fare per richiederla, quali tipologie esistono, ecc.);
- Presenza di una pagina specifica per ciascuna prestazione (ADI o ADP).

Abbiamo inoltre considerato i seguenti aspetti:

- L'organizzazione testuale (divisione in paragrafi, presenza di titoli esplicativi, evidenziazione delle parole chiave, ecc.);
- Colore dei testi e dimensione del font;

- Contenuto testuale (è presente un contenuto informativo o si trovano soltanto contatti telefonici? Le informazioni sono presenti nella pagina dedicata al servizio o è necessario scaricare dei documenti?);
- Livelli su cui si trovano le informazioni (se sono presenti delle sottopagine);
- Presenza di allegati o materiali di approfondimento;
- Presenza di menu contestuali che aiutano nella navigazione dei contenuti;
- Riconoscibilità dei collegamenti ipertestuali (sono in qualche modo segnalati? Viene indicato quando i link rimandano a siti esterni?).

A livello generale possiamo fare alcune considerazioni.
La quasi totalità dei siti web analizzati ha la funzione di ricerca in home page. L'unico sito che non offre questo servizio è quello dell'Asl di Torino 1. Nella maggior parte dei portali la casella di ricerca è posizionata in alto a destra, sopra al menu; scelgono la posizione sotto al menu i siti dell'ASL di Roma, l'Asl di Napoli 1 Centro e l'ASP di Potenza. La ricerca si trova invece circa a metà pagina nei siti di ASP Palermo e ASSL di Cagliari.
Soltanto in due casi la casella è posizionata a sinistra (ULSS3 Serenissima di Venezia e Azienda sanitaria dell'Alto Adige di Bolzano). Caso isolato quello dell'ASP di Agrigento, dove la ricerca è inserita direttamente nel *footer*, rendendola così poco visibile. Di difficile individuazione è anche il servizio di ricerca nel sito dell'ASL di Chieti-Lanciano-Vasto: vi si accede infatti dal menu di accesso rapido, cliccando sul form *Cerca nel sito – Servizi* e selezionando la funzione *Cerca nel sito*. Si apre così una nuova pagina con la casella di ricerca.
Per quanto riguarda l'individuazione dei contenuti, spesso non è risultata semplice. In alcuni casi, si è dovuto ricorrere agli strumenti progettati per aiutare la navigazione, come le sezioni *Come fare per* (organizzata per obiettivi) e *Dedicato a* (organizzata per tipologia di utenti), oppure all'elenco dei servizi strutturato alfabeticamente o per argomento. In altri, l'unico modo per reperire un'informazione è tramite la funzione di ricerca.
Per quanto riguarda l'accessibilità[237], le pagine dovrebbero essere conformi ai requisiti tecnici indicati nella Legge Stanca (Legge 4/2004 - *Disposizioni per favorire l'accesso dei soggetti disabili agli strumenti informatici*) e alle linee guida del W3C.

[237] L'articolo 2 della Legge 4/2004 (detta Legge Stanca) definisce l'accessibilità come "la capacità dei sistemi informatici, ivi inclusi i siti web e le applicazioni mobili, nelle forme e nei limiti consentiti dalle conoscenze tecnologiche, di erogare servizi e fornire informazioni fruibili, senza discriminazioni, anche da parte di coloro che a causa di disabilità necessitano di tecnologie assistive o configurazioni particolari".
I siti web delle pubbliche amministrazioni devono rispettare i requisiti tecnici di accessibilità riportati nell'Allegato A del Decreto Ministeriale 8 luglio 2005; tali requisiti sono stati definiti sulla base delle raccomandazioni che il Web Accessibility Initiative (WAI) del World Wide Web Consortium (W3C) ha pubblicato l'11 dicembre 2008, le quali contengono le Web Content Accessibility Guidelines 2.0 (WCAG 2.0).
La prima versione delle linee guida di riferimento (WCAG 1.0), resa pubblica dal W3C nel 1999, è stata invece la base per la redazione dei requisiti tecnici di accessibilità e usabilità definiti nel complesso normativo della Legge 4/2004 (*Disposizioni per favorire l'accesso dei soggetti disabili agli strumenti informatici*). Alla Legge Stanca è seguito il Decreto legislativo n. 82 del 7 marzo 2005, detto *Codice dell'Amministrazione Digitale*.
Il 5 giugno 2018 il W3C ha rilasciato un aggiornamento delle linee guida (WCAG 2.1); le raccomandazioni sono applicabili a tutti i tipi di dispositivi (desktop, laptop, tablet e mobili). Come previsto dalla Direttiva (UE) 2016/2102 del Parlamento europeo e del Consiglio relativa

Soltanto la metà dei siti presenta una pagina dedicata all'accessibilità: di questi, dieci siti dichiarano di essere conformi ai requisiti della Legge Stanca e due di essere conformi al Decreto Ministeriale *Requisiti tecnici e i diversi livelli per l'accessibilità agli strumenti informatici* (DM 8 luglio 2005). Nel sito dell'ASUR di Ancona la voce *accessibilità* è presente ma rimanda alla home page; nel sito dell'ASL di Bari la pagina dell'accessibilità è in realtà dedicata soltanto al catalogo dati; nel sito dell'Asl di Brindisi, invece, sono elencati soltanto gli obiettivi di accessibilità da completare.
Sono esempi positivi i siti dell'ASUITS di Trieste e dell'ASL di Pescara: nel primo è possibile trovare una serie di approfondimenti riguardanti la normativa relativa all'accessibilità; nel secondo, si trova anche una sezione dedicata alla leggibilità dei testi.
La qualità del codice è piuttosto scarsa; soltanto nove siti presentano i loghi che attestano la conformità alle direttive del W3C. In quattro siti il codice è valido rispetto alle specifiche WCAG 1.0 e soltanto in due (ULSS3 Serenissima di Venezia e USL Toscana Centro) è conforme alle più aggiornate WCAG 2.0. Sei siti risultano realizzati in XHTML, due in HTML 5 e otto fanno uso dei fogli di stile (cinque siti ricorrono ai CSS 1, due ai CSS 2.1 e soltanto l'ATS di Bergamo impiega i CSS3).
Per quanto riguarda invece la compatibilità dei siti con i diversi browser, soltanto due siti presentano il logo dedicato e quattro riportano, nelle pagine dell'accessibilità, l'elenco dei browser supportati. Abbiamo effettuato delle prove, utilizzando i principali browser (Google Chrome, Internet Explorer, Mozilla Firefox): la quasi totalità dei siti risulta fruibile indipendentemente dal browser utilizzato. Il sito dell'ASP di Potenza presenta dei problemi nell'esecuzione di Adobe Flash con la visualizzazione tramite Firefox e Chrome; il sito dell'ASL dell'Aquila presenta lo stesso problema ma soltanto con Firefox. Quando si accede al sito dell'ASP di Catanzaro tramite Explorer viene mostrato un avviso per i contenuti non sicuri.
Abbiamo valutato anche la responsività dei siti, ovvero la capacità di adattarsi graficamente ai vari dispositivi mobili con cui sono visualizzati (tablet, smartphone). Un sito responsivo adatta in modo automatico la dimensione del font, la disposizione degli elementi e la visualizzazione dei contenuti.
Soltanto 15 siti risultano responsivi. È interessante notare il fatto che la metà di questi non presenta invece una pagina dedicata all'accessibilità.

8.3.3.1. Valle d'Aosta

L'AUSL di Aosta (denominata Azienda USL della Valle d'Aosta) è l'unica azienda sanitaria della regione. Il sito presenta una buona organizzazione dei contenuti: le informazioni principali si trovano nel menu in alto e nelle tre sezioni in basso (*Come fare per*, *Dedicato a*, *Prevenzione, salute e benessere*). Lo strumento di ricerca è visibile, posizionato in alto a destra.

---

all'accessibilità dei siti web e delle applicazioni mobili degli enti pubblici, i requisiti di accessibilità per i paesi europei dovranno essere allineati alle WCAG 2.1.
La direttiva è stata recepita anche in Italia. Con l'entrata in vigore (il 26 settembre 2018) del Decreto legislativo n. 106 del 10 agosto 2018, la disposizione entra nel vivo, aggiornando e modificando la Legge 4/2004.

Figura 34. Le sezioni principali nel sito dell'AUSL di Aosta.

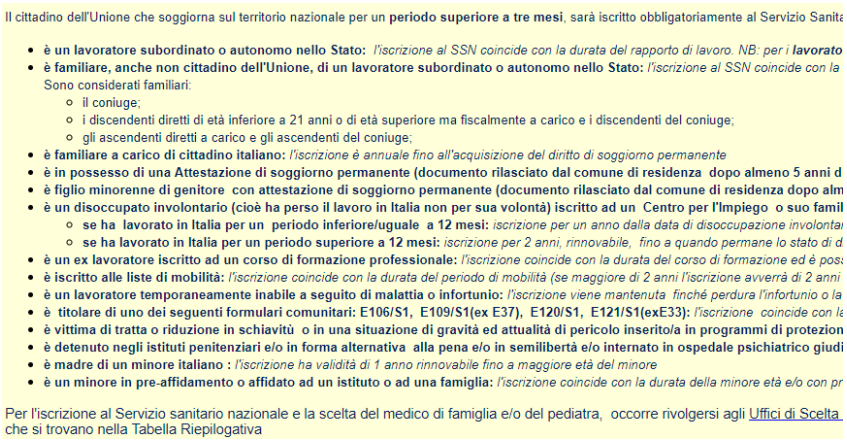
Le informazioni relative all'assistenza agli stranieri e all'assistenza domiciliare sono facilmente individuabili nella sezione *Dedicato a*, dove sono raggruppati i contenuti in base al destinatario di riferimento (in questo caso stranieri e anziani). Anche ai programmi di screening si accede immediatamente dalla sezione *Prevenzione, salute e benessere*.
La pagina relativa all'emergenza sanitaria non è invece facilmente individuabile: il percorso da seguire prevede diversi click (home > chi siamo > ospedale > strutture > emergenza sanitaria) e il modo migliore per arrivare a tale contenuto è attraverso il motore di ricerca interno.
In generale, i percorsi delle varie pagine sono sempre specificati; in ogni pagina sono inoltre presenti dei menu contestuali per navigare all'interno dei vari contenuti.
Anche l'organizzazione testuale è chiara: i paragrafi sono brevi e ben distinti, spesso sono presenti dei titoli in grassetto che introducono i vari argomenti, si fa ricorso a elenchi puntati per organizzare le informazioni, gli allegati sono raccolti in fondo al testo. Pochissimi i link ipertestuali.

8.3.3.2. Piemonte

Dal 1° gennaio 2017 è stata costituita l'ASL Città di Torino tramite l'accorpamento delle ex ASL TO1 e ASL TO2. Il sito risulta ancora provvisorio e soprattutto privo di molti contenuti. Per questo motivo abbiamo deciso di prendere in esame altri siti delle ASL piemontesi: i due siti vecchi delle ASL di Torino (ASL TO1 e ASL TO2) e il sito dell'ASL di Cuneo (ASL CN1)[238].
Nel sito dell'ASL TO1 le informazioni sono organizzate nei menu laterali, mentre nella parte centrale sono disponibili le notizie da parte dell'azienda. La pagina relativa al servizio di emergenza si trova a sinistra, in un box senza titolo; l'assistenza domiciliare e l'assistenza agli stranieri si trovano invece a destra, rispettivamente nelle sezioni *Assist. anziani e disabili* e *Assist. territoriale*. Non è stato invece possibile rintracciare i contenuti relativi agli screening oncologici e, non essendo presente la funzionalità di ricerca, non è possibile verificare se tali pagine siano o meno presenti. Le informazioni relative al servizio di

[238] Oltre alla nuova ASL della Città di Torino, a Torino sono presenti altre aziende sanitarie locali: ASL TO3, che comprende i distretti di Collegno, Giaveno, Orbassano, Pinerolo, Rivoli, Susa, Val Pellice, Valli Chisone e Germanasca, Venaria; ASL TO4, che comprende i distretti Ciriè, Chivasso, Settimo Torinese, San Mauro, Ivrea, Cuorgnè; ASL TO5, che comprende i distretti di Chieri, Carmagnola, Moncalieri e Nichelino.
Dopo Torino, Cuneo è una delle città principali del Piemonte. Ha due aziende sanitarie: l'ASL CN1, che comprende il distretto di Cuneo, Borgo San Dalmazzo-Dronero, Mondovì, Ceva, Savigliano-Fossano, Saluzzo e l'ASL CN2, che comprende i distretti di Alba e Bra.

emergenza si limitano a fornire il numero da chiamare (118) e a specificare che il servizio è attivo 24 su 24.


Il cittadino dell'Unione che soggiorna sul territorio nazionale per un **periodo superiore a tre mesi**, sarà iscritto obbligatoriamente al Servizio Sanita

- **è un lavoratore subordinato o autonomo nello Stato:** *l'iscrizione al SSN coincide con la durata del rapporto di lavoro. NB: per i* ***lavorato***
- **è familiare, anche non cittadino dell'Unione, di un lavoratore subordinato o autonomo nello Stato:** *l'iscrizione al SSN coincide con la*
  Sono considerati familiari:
  - il coniuge;
  - i discendenti diretti di età inferiore a 21 anni o di età superiore ma fiscalmente a carico e i discendenti del coniuge;
  - gli ascendenti diretti a carico e gli ascendenti del coniuge;
- **è familiare a carico di cittadino italiano:** *l'iscrizione è annuale fino all'acquisizione del diritto di soggiorno permanente*
- **è in possesso di una Attestazione di soggiorno permanente (documento rilasciato dal comune di residenza dopo almeno 5 anni d**
- **è figlio minorenne di genitore con attestazione di soggiorno permanente (documento rilasciato dal comune di residenza dopo alm**
- **è un disoccupato involontario (cioè ha perso il lavoro in Italia non per sua volontà) iscritto ad un Centro per l'Impiego o suo famil**
  - **se ha lavorato in Italia per un periodo inferiore/uguale a 12 mesi:** *iscrizione per un anno dalla data di disoccupazione involonta*
  - **se ha lavorato in Italia per un periodo superiore a 12 mesi:** *iscrizione per 2 anni, rinnovabile, fino a quando permane lo stato di d*
- **è un ex lavoratore iscritto ad un corso di formazione professionale:** *l'iscrizione coincide con la durata del corso di formazione ed è pos*
- **è iscritto alle liste di mobilità:** *l'iscrizione coincide con la durata del periodo di mobilità (se maggiore di 2 anni l'iscrizione avverrà di 2 anni*
- **è un lavoratore temporaneamente inabile a seguito di malattia o infortunio:** *l'iscrizione viene mantenuta finché perdura l'infortunio o la*
- **è titolare di uno dei seguenti formulari comunitari: E106/S1, E109/S1(ex E37), E120/S1, E121/S1(exE33):** *l'iscrizione coincide con la*
- **è vittima di tratta o riduzione in schiavitù o in una situazione di gravità ed attualità di pericolo inserito/a in programmi di protezion**
- **è detenuto negli istituti penitenziari e/o in forma alternativa alla pena e/o in semilibertà e/o internato in ospedale psichiatrico giudi**
- **è madre di un minore italiano :** *l'iscrizione ha validità di 1 anno rinnovabile fino a maggiore età del minore*
- **è un minore in pre-affidamento o affidato ad un istituto o ad una famiglia:** *l'iscrizione coincide con la durata della minore età e/o con pr*

Per l'iscrizione al Servizio sanitario nazionale e la scelta del medico di famiglia e/o del pediatra, occorre rivolgersi agli Uffici di Scelta che si trovano nella Tabella Riepilogativa

Figura 35. Un esempio di testo nel sito dell'ASL TO1.

In generale, i testi non hanno una buona organizzazione: sono presenti elenchi puntati e le parti rilevanti sono in grassetto ma il ricorso a questi elementi è talmente frequente che il testo risulta piuttosto condensato. Gli elementi paratestuali (grassetto, corsivo, maiuscolo) dovrebbero servire a mettere in risalto i punti salienti del discorso: se tutto il testo è enfatizzato con uno di questi elementi, non si riesce a individuare l'informazione essenziale (come mostrato nella Figura 35).

Nel sito dell'ASL TO2 le informazioni principali si trovano in parte nel menu in alto e in parte in quello laterale a sinistra. Si tratta di un elenco di voci che riguarda tutti i servizi e le strutture dell'azienda; i contenuti sono posti tutti allo stesso livello e nessuna informazione è messa in risalto. Si trovano qui il servizio di emergenza-urgenza e Prevenzione Serena, il progetto regionale per la prevenzione dei tumori femminili. Assente invece la campagna di prevenzione per il tumore colon-rettale.

Le pagine relative all'assistenza agli stranieri e alle cure domiciliari sono raccolte nella sezione *Come fare per*, che è strutturata a sua volta in un elenco di tutti gli argomenti dalla A alla Z.

A differenza del sito dell'ASL TO1, esiste una sezione relativa all'accessibilità e vi è la possibilità di ingrandire il carattere per una migliore visualizzazione.

Decisamente migliore l'organizzazione del sito dell'ASL di Cuneo. Le informazioni principali si trovano nel menu principale ma sono disposte anche nelle varie sezioni in evidenza: *L'ASL per i cittadini*, *Cosa fare per*, *Dedicato a*. In quest'ultima sono raggruppati i contenuti in base al destinatario di riferimento (bambini, disabili, anziani, donne, stranieri, ecc.) o in base alla tematica (dipendenze, salute mentale, pazienti oncologici, ecc.). Lo strumento di ricerca è posizionato in alto a destra, ben visibile.

Figura 36. La sezione *Dedicato a* presente sul sito dell'ASL di Cuneo.

Il servizio di emergenza si trova in home page, sotto la sezione *L'ASL per i cittadini.* I programmi di prevenzione oncologica si trovano all'interno della voce *Prevenzione* (Prevenzione > Igiene e Sanità pubblica > Screening oncologici) nel menu principale ma vi si accede più facilmente tramite il percorso *Dedicato a* > *Donne*. Anche le informazioni relative all'assistenza agli stranieri e all'assistenza domiciliare sono facilmente individuabili nella sezione *Dedicato a*.

All'interno di tutte le pagine è presente un menu contestuale, chiamato nel sito "Sottomenu di navigazione", che indica il percorso dei vari contenuti.

In generale, si nota una buona organizzazione testuale: i paragrafi sono brevi, ben separati e introdotti da un titolo che ne spiega il contenuto (Come iscriversi, dove iscriversi, Come attivare il servizio); si ricorre spesso a elenchi numerati; le parti rilevanti sono in grassetto; i collegamenti ipertestuali a siti esterni sono segnalati attraverso il ricorso al grassetto e al colore celeste. In alcuni casi, i testi sono affiancati da un box (a destra) in cui sono raccolti gli allegati o la modulistica.

8.3.3.3. Liguria

In Liguria sono presenti 5 aziende sanitarie locali: ASL 1 imperiese (Bussana di Sanremo), ASL 2 savonese (Savona), ASL 3 genovese (Genova), ASL 4 chiavarese (Chiavari), ASL 5 spezzino (La Spezia).

Il sito dell'ASL genovese ha una buona navigabilità. In alto, subito sotto il logo, è presente il menu principale con tutte le informazioni relative all'azienda, ai servizi erogati e alle strutture. Nella parte centrale, oltre alle ultime notizie, sono presenti due sezioni tematiche: *In evidenza*, nella quale sono fornite informazioni su medici e pediatri di scelta libera, vaccinazioni, farmacie, programmi di screening, stili di vita, ecc. e *Per te*, nella quale i contenuti sono organizzati pe target di riferimento (anziano, bambino, donna, diversamente abile, migrante, lavoratore).

Tale chiarezza nell'organizzazione dei contenuti diminuisce nelle pagine interne: alla prevenzione oncologica si arriva con diversi click a partire dalla sezione *In evidenza*, per poi

essere indirizzati in un sito specifico dedicato all'argomento. La pagina relativa all'emergenza si trova tra le voci del menu principale, sotto *Ospedali*, e riunisce in un unico testo informazioni sul 112, sull'accesso al pronto soccorso e sulla continuità assistenziale (Guardia medica). La pagina dedicati ai migranti è in fase di aggiornamento ma rimanda a due testi piuttosto esaustivi circa le modalità di accesso al SSN.
È presente una pagina dedicata all'assistenza domiciliare ma sembra destinata solo a chi ha una disabilità grave o gravissima.
Anche in questo caso si ha una buona organizzazione testuale, con testi brevi, presenza di grassetti o sottolineature per evidenziare le parole chiave, ricorso a elenchi puntati e numerati. Quando previsto, nella parte destra del sito, a fianco dei contenuti, è presente un box *Documenti* in cui è possibile scaricare gli allegati.

8.3.3.4. Lombardia

Dal 1° gennaio 2016 l'ASL di Milano è divenuta ATS Milano – Città Metropolitana (Agenzia di Tutela della Salute della Città Metropolitana di Milano), accorpando i territori di competenza dell'ASL Milano, dell'ASL Milano 1 (Legnano), dell'ASL Milano 2 (Melegnano) e dell'ASL di Lodi.
Nonostante la veste aggiornata, il nuovo sito risulta di difficile navigabilità: l'individuazione dei contenuti non è così immediata e risulta preferibile rintracciarli tramite il motore di ricerca interno al sito. La pagina relativa al servizio di emergenza non è reperibile, neanche ricorrendo alla ricerca.
Gli screening oncologici non sono presenti in home page ma solo sotto la voce *Guida ai servizi* del menu principale. La pagina generale si articola poi nei sotto contenuti corrispondenti ai vari programmi di prevenzione ma il collegamento ipertestuale non è in alcun modo segnalato: la presenza del link si individua solo al passaggio del mouse.
In home page è presente una sezione denominata *Strutture socio-sanitarie e ADI* ma, per le prime, si viene rimandati ad un portale esterno, per le seconde, ad un file in excel con la lista delle strutture. Ai contenuti relativi all'assistenza domiciliare e all'assistenza per gli stranieri si giunge esclusivamente tramite il servizio di ricerca: dal momento che non esistono pagine generiche relative a tali contenuti, si viene reindirizzati ai servizi territoriali delle varie sedi e ai loro canali tematici.

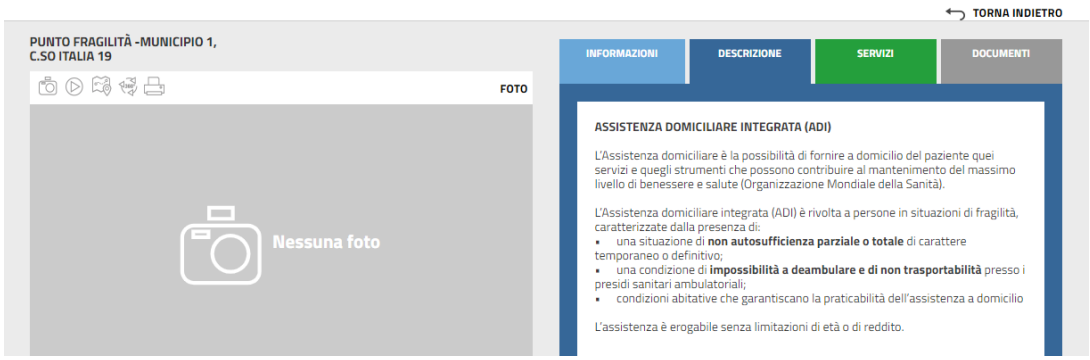


Figura 37. Esempio di pagina dedicata all'ADI.

Figura 38. Esempio di pagina dedicata all'assistenza agli stranieri.

Alla luce di queste problematiche, si è scelto di prendere in esame anche il sito della ATS di Bergamo.

Un aspetto risulta subito evidente: la Regione Lombardia e le ATS puntano sull'immagine coordinata; esiste infatti una certa uniformità, una coerenza grafica tra i vari siti, a partire dai loghi e dal colore principale per i menu, le icone e alcune parti testuali.

Figura 39. Home page del sito dell'ATS di Milano.

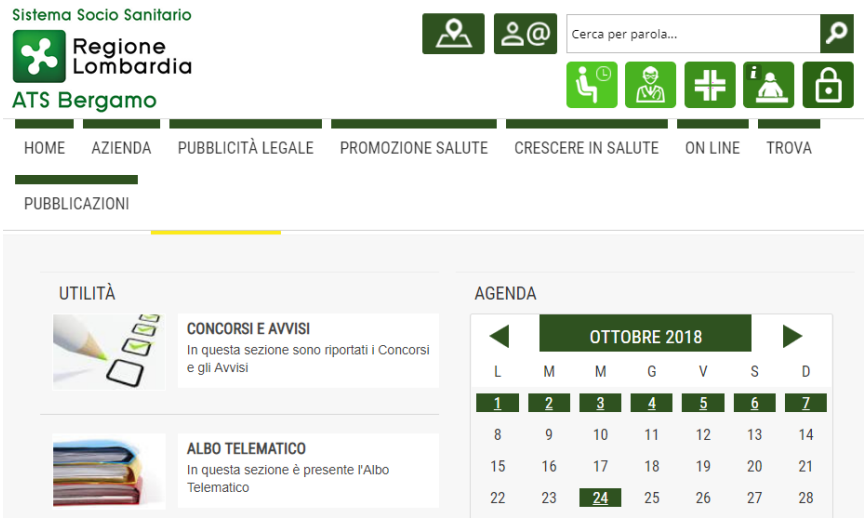


Figura 40. Home page del sito dell'ATS di Bergamo

Nel caso del sito della ATS di Bergamo l'individuazione delle informazioni è più immediata ma non esente da problemi. La pagina relativa alle emergenze non è presente in home page ma si trova nel menu principale in alto, sotto la voce *Trova*. In realtà le informazioni si riferiscono al pronto soccorso più che al funzionamento vero e proprio del servizio di emergenza. Il numero da contattare in caso di emergenza (112) non è particolarmente visibile. Alle campagne di screening si giunge dalla sezione *Approfondimenti*: si trovano informazioni sullo screening mammografico e su quello del colon retto ma manca il programma riguardante il collo dell'utero. Utile la sezione delle FAQ relative all'argomento. Le altre due tematiche si trovano all'interno del menu secondario (che in home è posizionato sotto la vetrina delle immagini o *slider*), sotto la voce *Tempi di attesa* (che poi nell'indicazione del percorso diventa *Prestazioni e tempi di attesa*) e di nuovo sotto la voce *Assistenza esenzioni e ticket*. I contenuti contengono delle pagine di approfondimento (ad esempio i vari tipi di assistenza domiciliare): in entrambi i casi mancano però dei menu contestuali per favorire la navigazione.
Utilissima la sezione *Guida ai servizi dalla A alla Z*, nella quale è possibile trovare tutti i principali argomenti presenti nel sito.

8.3.3.5. Emilia Romagna

Il servizio sanitario dell'Emilia Romagna comprende 8 AUSL (Parma, Piacenza, Reggio Emilia, Modena, Bologna, Imola, Ferrara, Romagna), ciascuna con uno specifico sito. A questi, si aggiungono i siti delle aree vaste: l'Area Vasta Emilia Nord (AVEN) che comprende le AUSL di Parma, Piacenza, Reggio Emilia, Modena; l'Area Vasta Emilia Centrale (AVEC), che comprende le AUSL di Bologna, Imola, Ferrara; l'Azienda USL della Romagna.
Abbiamo preso come riferimento i siti web delle aziende sanitarie di Bologna e di Parma. Come nel caso dei siti delle ATS lombarde, c'è una sorta di continuità grafica tra i due siti (il logo è lo stesso, così come il colore principale), ma le affinità si limitano a questo: il sito dell'AUSL di Parma è sicuramente più recente e organizzato; il sito dell'AUSL di Bologna, oltre ad avere un font piuttosto piccolo, risulta anche più carente nei contenuti.

Figura 41. Header del sito dell'AUSL di Bologna.

Figura 42. Header del sito dell'AUSL di Parma.

Nel sito di Bologna, infatti, non è stato possibile rintracciare né la pagina dedicata all'assistenza agli stranieri, né quella inerente alle cure domiciliari. In home page è presente un box con l'indicazione del 118, ma il link rimanda a un sito esterno e non vengono fornite ulteriori informazioni.

L'unico contenuto presente è quello relativo agli screening oncologici. Si accede dal menu a sinistra (*Informazioni)* che rimanda a una pagina di informazioni generali sugli screening regionali. Tale sezione si articola poi in vari contenuti, corrispondenti ai vari programmi di prevenzione. In ciascuna pagina è presente il percorso di navigazione, talvolta anche piuttosto articolato (Tu sei qui: Portale → Chi siamo → Dipartimenti di produzione territoriale → Dipartimento di Sanità Pubblica → APPS → PI-CS → Screening del tumore del colon-retto).

È presente una pagina dedicata all'accessibilità, nella quale è disponibile la funzione per aumentare la dimensione del font; viene anche fornita una lista con i principali tasti di accesso rapido e le istruzioni per utilizzarli nei diversi browser.

Nel sito dell'AUSL di Parma i 4 contenuti sono invece disponibili. È possibile accedervi tramite il menu a sinistra, che comprende quattro diverse voci: *Per la tua salute*, *Dove curarsi*, *Servizi online*, *Come fare per*. Nella sezione *Per la tua salute* i contenuti sono organizzati in base al destinatario (bambini, giovani, donne, anziani, diversamente abili, migranti, lavoratori), alla tematica (malattie croniche, dipendenze, salute mentale, ecc.) o alle iniziative relative alla prevenzione (vaccinazioni, screening, stili di vita, ecc.). La sezione *Dove curarsi* si articola in altre 9 sotto sezioni, tra cui l'emergenza-urgenza e l'assistenza domiciliare. L'organizzazione testuale è buona, le informazioni sono esaustive, i percorsi all'interno delle pagine sono ben chiari. Utilissima la presenza, sotto ai testi, di alcuni tab con le informazioni e i link utili, i materiali informativi e gli allegati.

Unica pecca l'area dei contenuti: lo sfondo grigio chiaro e il colore del font (grigio scuro) rendono la lettura piuttosto difficoltosa e il testo risulta quasi sfuocato. Non a caso, manca la pagina relativa all'accessibilità.

8.3.3.6. Veneto

Nel Veneto sono presenti 9 ULSS (Unità Locale Socio Sanitaria); quella di Venezia è la ULSS3 Serenissima ed è nata nel 2017 in seguito alla riorganizzazione della sanità, accorpando la ULSS 12 Veneziana, la ULSS 13 Mirano e la ULSS 14 Chioggia. I tre vecchi siti sono ancora attivi ma il sito di riferimento è ormai quello della ULSS3 Serenissima. Nonostante questo, se si cerca il sito dell'azienda sanitaria di Venezia sul motore di ricerca Google, il primo risultato corrisponde ancora al vecchio sito, seguito poi da quello nuovo.

Il sito presenza una buona organizzazione dei contenuti. La maggior parte delle informazioni è raccolta all'interno del menu principale, che conta tra le voci la sezione *Emergenza-urgenza*. Tale sezione si articola poi in sotto menu, tra cui la pagina di spiegazione del servizio e la pagina con le indicazioni per il corretto utilizzo dei servizi di emergenza. Per quanto riguarda gli altri 3 argomenti, non esistono contenuti informativi di carattere generale, ma solo le pagine corrispondenti alle prestazioni erogate nei vari distretti. In questo modo, è necessario un più alto numero di click per raggiungere le informazioni necessarie. Ad esempio, l'assistenza sanitaria e l'assistenza domiciliare hanno il seguente percorso: Home > Distretto del Veneziano > Distretto del Veneziano > Sedi e servizi sul territorio > Assistenza sanitaria di base > Stranieri in Italia (oppure Assistenza domiciliare).

Per quanto riguarda l'organizzazione testuale, in generale i testi sono quasi sempre brevi, divisi in paragrafi spaziati tra loro e introdotti da un titolo esplicativo. Si notano però delle forti differenze a seconda dei contenuti e sembra mancare una certa uniformità: le voci di una lista, ad esempio, sono introdotti graficamente da un puntino nel caso degli screening, da una lineetta nei servizi agli stranieri e da nessun elemento nella pagina delle emergenze. Nelle pagine relative all'ADI, le parole chiave sono a volte in corsivo, a volte in corsivo e grassetto; nelle pagine di assistenza agli stranieri per evidenziare le informazioni importanti si ricorre invece al grassetto o al corsivo sottolineato.

Non è presente una pagina dedicata all'accessibilità del sito, tuttavia sono disponibili alcune funzioni, come la possibilità di ingrandire il testo, di modificare il contrasto, di scegliere la visualizzazione solo testo.

8.3.3.7. Friuli Venezia Giulia

Il sistema sanitario del Friuli è organizzato in due ASUI (Azienda Sanitaria Universitaria Integrata), quella di Trieste e quella di Udine, e in tre AAS (Azienda per l'Assistenza Sanitaria), quella "Bassa Friulana – Isontina", quella "Alto Friuli – Collinare - Medio Friuli" e quella del "Friuli occidentale".

Abbiamo effettuato una valutazione del sito dell'ASUITS (Azienda Sanitaria Universitaria Integrata di Trieste). In home page risulta una buona organizzazione dei contenuti, probabilmente dovuta anche al ricorso ad icone che affiancano le varie voci. Si accede alle informazioni sia dal menu principale sia attraverso le sezioni *Dedicato a / Eventi della vita*, che si articolano in un'ampia gamma di contenuti, organizzati a seconda del destinatario (animali, anziani, bambini, donne, giovani, lavoratori, pubbliche amministrazioni, scuole, imprese, stranieri) o della tematica (avere una casa, vaccinarsi, vita di coppia, viaggiare all'estero, dipendenze, vivere in salute, ecc.).

Figura 43. Le sezioni principali nel sito dell'ASUITS.

Le pagine interne non hanno invece la stessa trasparenza nell'organizzazione delle informazioni: in alcuni casi uno specifico contenuto è trattato in due pagine diverse, talvolta in due sezioni diverse del sito. Questa sovrapposizione influisce sulla chiarezza.

Ad esempio, nella ricerca dei programmi di prevenzione oncologica si viene rimandati alla sezione dedicata alle *Donne*, nella quale troviamo due pagine riguardanti la mammografia e il pap test e una scheda informativa sugli *Screening regionali per la prevenzione dei tumori*, in cui di nuovo si trovano informazioni sulla prevenzione dei tumori femminili e, in aggiunta, anche relative a quello colon rettale.

Per quanto riguarda la sezione dedicata agli stranieri, colpisce il fatto che ci sia una pagina relativa all'iscrizione al SSN per ciascuno dei 4 distretti dell'ASUI. Dal momento che il testo relativo alla procedura di accesso è lo stesso, sarebbe stato più utile raggruppare le 4 pagine in un unico testo, eventualmente con le indicazioni e i numeri utili delle varie strutture di riferimento sul territorio (anche se il punto informativo è unico per tutti e corrisponde al distretto 4).

Il servizio del 118 si trova soltanto ricorrendo al motore di ricerca interno al sito; una volta individuata la pagina corrispondente è possibile visualizzare il percorso relativo (home > Chi siamo > Organigramma > Sistema 118) ma anche andando a ritroso non si riesce a raggiungere tale contenuto. Lo stesso problema si ha con l'assistenza domiciliare: la pagina corrispondente non è rintracciabile partendo dalla home page; effettuando un'interrogazione con la ricerca interna non si trovano risultati né per assistenza domiciliare, né per ADI. La ricerca "cure domiciliari" fornisce 4 risultati uguali, che nel dettaglio risultano essere la pagina *Cure ambulatoriali e domiciliari continuità terapeutica* relativa ad ognuno dei 4 distretti. Per l'analisi abbiamo considerato la pagina del Distretto 1. Si nota che la ricerca rimanda direttamente alla voce *Cosa facciamo* e non alla voce generale *Chi siamo*.

Da sottolineare anche la presenza, nel menu *Contatti,* di una pagina per la valutazione del sito tramite un questionario in forma anonima.

8.3.3.8. Provincia Autonoma di Trento

Nonostante, ai primi di ottobre, il sito sia in fase di aggiornamento in seguito alla riorganizzazione aziendale, il sito dell'APSS (Azienda Provinciale per i Servizi Sanitari) della Provincia Autonoma di Trento rappresenta un esempio molto positivo.

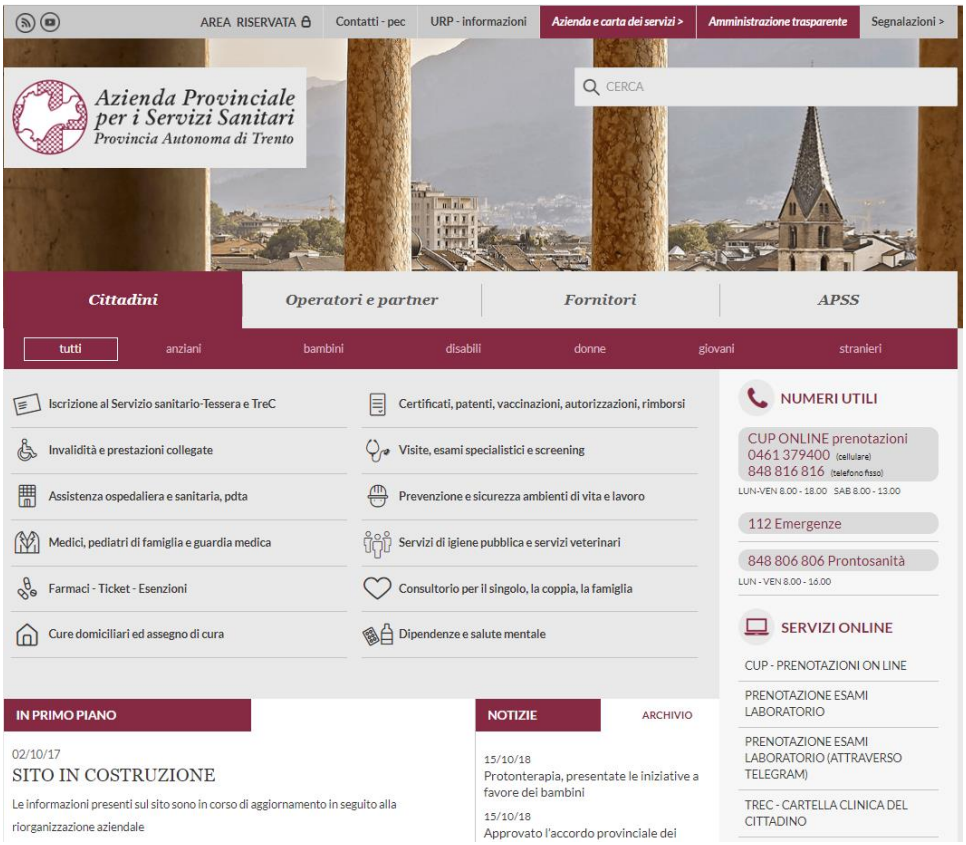


Figura 44. Home page del sito dell'APSS della Provincia Autonoma di Trento.

In home page sono presenti e messe in evidenza tutte le informazioni principali, i numeri utili e i vari servizi online. Tra questi vi è anche il servizio di emergenza (112). Nella parte centrale è possibile visualizzare le informazioni rivolte ai cittadini, filtrandole in base al destinatario (anziani, bambini, disabili, donne, giovani, stranieri). L'individuazione delle informazioni è facile e immediata; i contenuti interni sono a loro volta organizzati in modo chiaro e uniforme. In ogni pagina è presente la data dell'ultimo aggiornamento.

All'interno della sezione *Visite, esami specialistici e screening* sono presenti le pagine dedicate ai tre programmi di prevenzione; ciascuna prevede 4 contenuti che si espandono e si riducono (Descrizione, Destinatari, Strutture aziendali e Documenti, che contiene i vari allegati).

Dalla sezione *Iscrizione al servizio sanitario* si accede alla pagina che riguarda gli stranieri. Anche tale contenuto è articolato in tre parti che si espandono: Descrizione, Destinatari e Servizi o Prestazioni al cittadino, che contiene alcuni link utili. La pagina relativa all'assistenza domiciliare si trova nella sezione *Cure domiciliari*. Prevede diversi contenuti testuali: Descrizione, Destinatari, Costi, Punti di erogazione, Servizi o Prestazioni al cittadino e Documenti.

8.3.3.9. Provincia autonoma di Bolzano

L'azienda sanitaria della Provincia Autonoma di Bolzano è denominata Azienda Sanitaria dell'Alto Adige (ASDAA). Il sito è organizzato come segue: nella parte superiore vi è un menu principale, che illustra l'azienda e i vari servizi; il resto del sito è disposto su tre colonne. La colonna a sinistra contiene informazioni sull'azienda e le varie strutture, quella a destra informazioni generali, contatti e numeri utili; la parte centrale è dedicata ai servizi online e alle notizie in evidenza.
Il servizio di emergenza è immediatamente individuabile nella colonna destra; nella pagina sono presenti i vari contatti e le indicazioni da seguire per un corretto funzionamento del servizio.
Per quanto riguarda l'assistenza sanitaria agli stranieri, vi si accede seguendo il percorso *Servizi* (nel menu principale) > *Servizi di informazione*: sono presenti due pagine, una dedicata ai cittadini dell'Unione Europea e una per gli extracomunitari. Le informazioni sono chiare e sintetiche, rese fruibili anche dalla grafica delle pagine.
Ai programmi di screening si giunge tramite il percorso *Strutture* > *Servizi a livello aziendale* > *Registro tumori - Prevenzione tumori e screening* ma, una volta aperta la pagina corrispondente il percorso di navigazione indicato è diverso (Home > Ospedale di Bolzano > Reparti e servizi > Servizi > Anatomia e Istologia Patologica (aziendale) > Registro tumori - Prevenzione tumori e screening). Nonostante ciò, i contenuti sono chiari e ben organizzati; è presente sia un menu contestuale a sinistra, sia l'indice degli argomenti a destra.

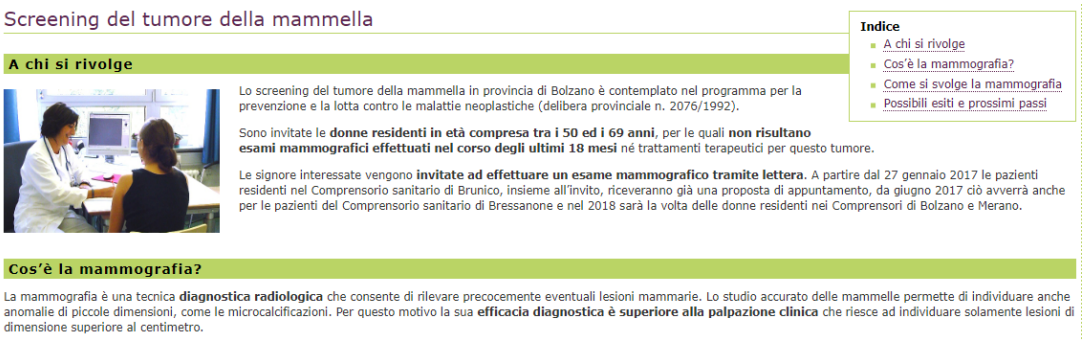


Figura 45. Pagina dedicata allo screening mammografico nel sito dell'ASDAA.

Le cure domiciliari fanno invece parte dei servizi territoriali, erogati a livello dei comprensori sanitari. Non è stato però possibile rintracciare una pagina dedicata allo specifico servizio, neanche tramite la funzione di ricerca.

8.3.3.10. Toscana

Dal 2016 in Toscana vi è stato il passaggio a 3 AUSL, tramite l'accorpamento delle 12 aziende sanitarie in 3 aree vaste: Toscana Nord Ovest (Massa Carrara, Lucca, Viareggio, Pisa), Toscana Centro (Firenze, Empoli Prato, Pistoia), Toscana Sud Est (Arezzo, Siena, Grosseto).
Cercando su Google l'azienda sanitaria di Firenze, il primo risultato visualizzato è ancora il vecchio sito, seguito da quello della AUSL Toscana Centro. Sul sito dell'ASF (Azienda Sanitaria di Firenze) si legge che il portale è in corso di progressiva dismissione e che le informazioni saranno reperibile su quello nuovo. Purtroppo, però, il sito AUSL Toscana

Centro sembrerebbe ancora provvisorio, vista la mancanza di molti dei contenuti principali. Per quanto riguarda le 4 tematiche, siamo riusciti a reperire soltanto la pagina relativa al 118, a cui si accede dalla home page e la pagina relativa all'assistenza agli stranieri. A quest'ultima siamo giunti tramite la ricerca, dato che il percorso per arrivarci è piuttosto curioso: il contenuto si trova infatti nei servizi online (insieme a quello relativo all'emergenza).

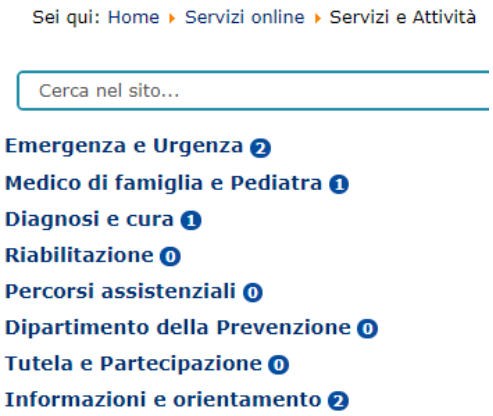


Figura 46. Pagina relativa ai servizi online nel sito dell'AUSL Toscana Centro.

La Figura 46 mostra la quantità di contenuti che è possibile trovare nel sito. L'assistenza domiciliare si trova all'interno della sezione *Diagnosi e cura* ma il testo non è ancora stato inserito.
I programmi di screening oncologici non sono invece rintracciabili in alcun modo.
Più positiva l'esperienza con il sito della AUSL Toscana Sud Est. Dalla sezione *Percorsi assistenziali*, nella home page, si giunge agli screening oncologici. È presente una pagina generale e una pagina informativa per ciascun ambito territoriale (aretino, senese e grossetano). Dalla sezione *Guida ai servizi*, anch'essa in home, si accede alle pagine relative agli altri tre argomenti (sono i primi tre servizi dell'elenco). Anche in questo caso, sono presenti una pagina generale e una per ciascun ambito territoriale; in alcuni casi il rimando all'ambito territoriale corrisponde a un sito esterno, che però non viene segnalato. Nel menu dedicato ai servizi online invece troviamo effettivamente presenti i servizi online.

8.3.3.11. Umbria

L'Umbria è organizzata in 2 AUSL: l'AUSL Locale Umbria1, che comprende 38 comuni e 6 distretti sociosanitari (Perugino, Assisano, Media Valle del Tevere, Trasimeno, Alto Chiascio, e Alto Tevere) e l'AUSL Locale Umbria2, che comprende 54 comuni e 6 distretti sociosanitari (Foligno, Spoleto, Valnerina, Narni e Amelia, Orvieto, Terni). In realtà, nei due siti web dedicati, la denominazione presente nei loghi è USL Umbria1 e USL Umbria2.
Ci siamo concentrati sul sito dell'AUSL Umbria1. Dall'analisi è emersa una buona organizzazione dei contenuti: le informazioni principali si trovano nel menu in alto e nelle tre sezioni in basso (*Come fare per*, *Dedicato a*, *Prevenzione, salute e benessere*).
È possibile accedere ai contenuti desiderati da diverse aree: la voce *Servizi* nel menu principale contiene tutte e quattro le tematiche; dalla sezione *Dedicato a* si arriva alle cure domiciliari (Anziani), agli screening (Donne) e all'assistenza agli stranieri (Stranieri e

migranti); le informazioni sugli screening oncologici sono disponibili anche nella sezione *Prevenzione, salute e benessere*; il servizio di emergenza sanitaria è compreso anche nella sezione *Come fare per*.

Figura 47. Le sezioni principali nel sito dell'AUSL dell'Umbria 1.

In generale, l'organizzazione testuale è buona; i testi risultano brevi, i paragrafi sono ben distinti, le informazioni più importanti sono in grassetto, si ricorre spesso agli elenchi puntati.
Sotto la voce *Per il cittadino* è presente una pagina dedicata al portale contro le *fake news* (Dottore ma è vero che…?).

8.3.3.12. Lazio

A partire dal 1° gennaio 2016 le ASL di Roma sono state riorganizzate ed accorpate; attualmente sono presenti le seguenti aziende sanitarie:

- ASL Roma 1, risultante dalla fusione dell'ASL Roma A e dell'ASL Roma E;
- ASL Roma 2, risultante dalla fusione dell'ASL Roma B e dell'ASL Roma C;
- ASL Roma 3, ridenominazione dell'ASL Roma D;
- ASL Roma 4, ridenominazione dell'ASL Roma F;
- ASL Roma 5, ridenominazione dell'ASL Roma G;
- ASL Roma 6, ridenominazione dell'ASL Roma H.

Per quanto riguarda l'ambito territoriale, la ripartizione è la seguente:

- ASL Roma 1: comprende i Municipi 1, 2, 3, 13, 14, 15 del Comune di Roma;
- ASL Roma 2: comprende i Municipi 4, 5, 6, 7, 8, 9 del Comune di Roma;
- ASL Roma 3: comprende i Municipi 10, 11, 12 del Comune di Roma e il comune di Fiumicino;
- ASL Roma 4, Roma 5 e Roma 6 comprendono vari comuni fuori Roma.

Abbiamo scelto come riferimento per il nostro studio il sito dell'ASL Roma 1.
Nonostante l'ampia mole di informazioni, il sito risulta avere una buona organizzazione. Esistono due menu nella parte superiore del sito ed entrambi rimangono in evidenza anche se si scorre verso la parte inferiore della pagina. Ciascuna voce del menu principale (*Come*

fare per, *Guida ai servizi*, *Servizi online*, *Strutture sanitarie*, *Dedicato a*, ecc.) è poi ripresa nel dettaglio nelle sezioni dedicate, che si trovano via via che si scende nella home page.

Figura 48. La sezione *Come fare per* sul sito dell'ASL Roma1.

Le informazioni relative alle 4 tematiche sono facilmente individuabili: il servizio di emergenza, l'assistenza agli stranieri e l'assistenza domiciliare si trovano nel box centrale *Cosa fare per*. La pagina dedicata agli screening oncologici si trova sotto la sezione *Prevenzione*. A tali contenuti si accede anche dalla sezione *Dedicato a* (Donne, Uomini, Anziani e Migranti). In realtà, la pagina relativa ai migranti è un contenitore più grande, che include anche l'assistenza sanitaria agli stranieri.

La struttura delle pagine si ripete in modo uniforme all'interno del sito: vi è un contenuto testuale con le informazioni di carattere generale, seguito da un box che comprende contenuti aggiuntivi che si espandono/riducono e i link alle sotto pagine relative al quel dato argomento. Le due tipologie di informazioni sono opportunamente segnalate; per ogni pagina è presente anche la data dell'ultimo aggiornamento e la possibilità di stampare il contenuto, inviarlo via mail o condividerlo sui social.

Strano il fatto che non vi sia una pagina dedicata all'accessibilità.

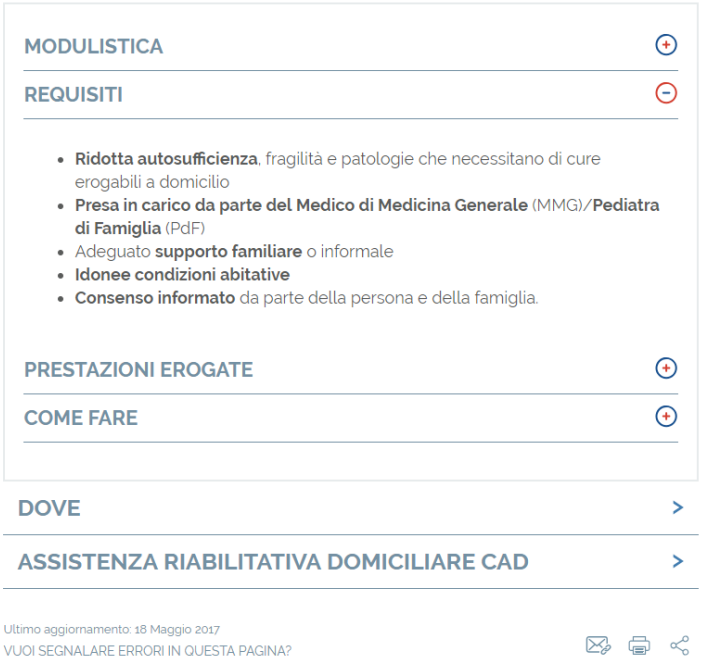


Figura 49. Un esempio di struttura delle pagine informative.

8.3.3.13. Marche

L'Azienda Sanitaria Unica Regionale (ASUR) delle Marche comprende 5 aree vaste:

1. Pesaro, Urbino, Fano;
2. Senigallia, Jesi, Fabriano, Ancona;
3. Civitanova Marche, Macerata, Camerino;
4. Fermo;
5. San Benedetto del Tronto e Ascoli Piceno.

Non è facile individuare i siti web corrispondenti alle varie aziende sanitarie. Attualmente è possibile trovare un sito generale dell'ASUR Marche, alcuni siti specifici per le città e alcuni per le aree vaste.
I siti delle aziende sanitarie di Pesaro, Urbino e Fano reindirizzano al sito unico dell'Area Vasta 1; lo stesso avviene per San Benedetto del Tronto e Ascoli Piceno, che sono riuniti nel sito dell'Area Vasta 5. Le altre ASUR hanno un sito dedicato a ciascuna città.
Dal punto divista grafico, i siti sono tutti uguali. Ci siamo inizialmente concentrati sul sito dell'ASUR di Ancona. I soli contenuti presenti sono quelli relativi all'emergenza sanitaria e all'assistenza domiciliare integrata: l'unico modo per individuarli, e non con poche difficoltà, è però attraverso la ricerca interna al sito. Il resto degli argomenti desiderati va integrato con le informazioni presenti negli altri siti dell'area vasta.
Il dipartimento di prevenzione ha una sotto sezione denominata "screening oncologici" ma il contenuto corrisponde a un file word. In home page è presente anche un box dedicato all'argomento ma rimanda al programma di prevenzione dell'area vasta che si trova sul sito dell'ASUR di Senigallia.
È possibile invece trovare le informazioni relative all'assistenza sanitaria agli stranieri sul sito dell'ASUR di Jesi.

Figura 50. Home page del sito dell'ASUR di Ancona.

Abbiamo provato a prendere in esame anche i siti delle due aree vaste, per verificare se fosse possibile rintracciare tutti gli argomenti in un solo portale, ma sono emerse le stesse problematiche.

8.3.3.14. Abruzzo

L'Abruzzo è organizzato in 4 ASL:

- ASL 1 Avezzano – Sulmona – Aquila;
- ASL 2 Chieti – Lanciano – Vasto;
- ASL 3 Pescara
- ASL 4 Teramo

Abbiamo effettuato una valutazione del sito dell'ASL 1 ma, dal momento che il sito era privo di alcuni contenuti, abbiamo esteso l'analisi anche all'ASL 2 e all'ASL 3. La scelta di includere nel campione entrambe è dipesa dal fatto che si tratta di due esempi abbastanza positivi, anche se piuttosto diversi tra loro.
L'unico argomento presente nel sito dell'ASL 1 è lo screening oncologico. Esiste anche una pagina dedicata al 118 ma contiene soltanto un video in lingua italiana e pdf scaricabili per le altre lingue.
Non è possibile rintracciare gli altri contenuti, neppure ricorrendo al servizio di ricerca (e cercando le possibili varianti: assistenza domiciliare, cure domiciliari, ADI, stranieri, migranti, immigrati, extracomunitari).
Il punto di forza del sito dell'ASL 3 (Pescara) è l'estrema semplicità a livello di navigazione: nella home page è presente, oltre al menu principale e alle ultime notizie, una sezione che organizza i contenuti in base al destinatario (cittadini, professionisti, imprese, area interna). Cliccando sul percorso dedicato al cittadino si apre una pagina con l'elenco di tutti i servizi;

è possibile filtrare in base a un ordine alfabetico o effettuare una ricerca all'interno dei servizi. Tutti i contenuti che cerchiamo sono disponibili in questa sezione.

Figura 51. La sezione *Per il cittadino* sul sito dell'ASL3 di Pescara.

Dal punto di vista grafico, il sito dell'ASL 2 è molto diverso: dalla sobrietà del sito dell'Azienda sanitaria di Pescara (in blu lo sfondo dei menu, in grigio quello dei titoli) si passa ad un portale piuttosto vivace, in cui i colori principali, almeno per quanto riguarda la home page, sono il viola, l'arancione e il rosso acceso.

Figura 52. Home page del sito dell'ASL di Lanciano-Vasto-CHieti

In realtà, nelle pagine interne del sito i tre colori sono presenti solo nel nome e nel logo e il colore dominante è il celeste che fa da sfondo ai menu contestuali.

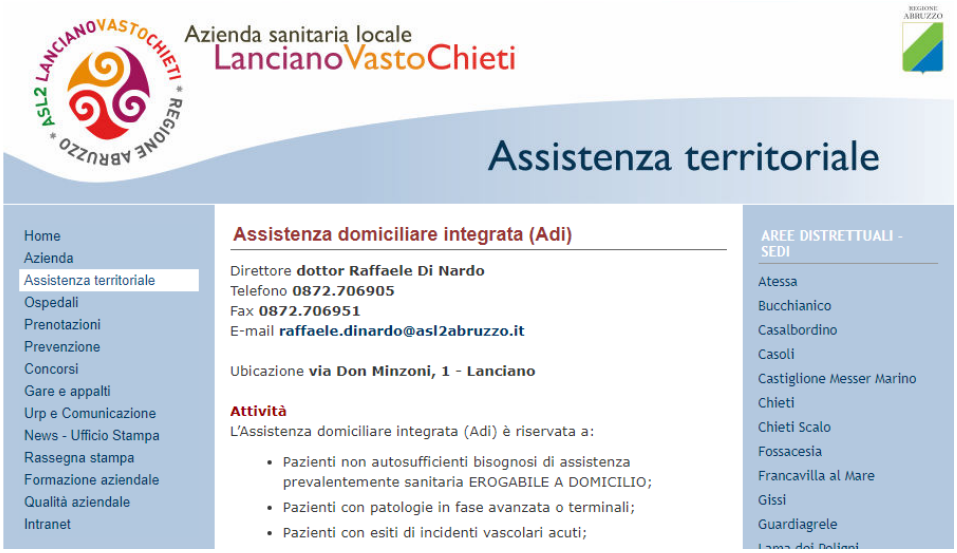


Figura 53. Un esempio di pagina interna.

La voce *Assistenza territoriale* include diversi servizi, tra cui l'assistenza domiciliare e il sevizio di emergenza sanitaria.

Purtroppo non è stato possibile individuare la pagina relativa all'assistenza sanitaria agli stranieri.

Alle campagne di prevenzione oncologica si accede invece da uno dei banner in home page; si viene indirizzati a una pagina che contiene informazioni di carattere generale e che contiene tre collegamenti ai diversi tipi di screening. L'aspetto interessante, almeno dal punto di vista grafico, è che lo sfondo della pagina e il colore dei titoli cambia a seconda del programma di prevenzione, riprendendo i colori del banner dedicato.

Figura 54. La pagina dedicata allo screening dei tumori del colon retto.

Figura 55. La pagina dedicata allo screening della cervice uterina.

8.3.3.15. Campania

Il territorio di Napoli è organizzato in 3 ASL:

- ASL Napoli 1 Centro, articolata in 11 distretti (dal 24 al 33);
- ASL Napoli 2 Nord, articolata in 13 distretti (dal 35 al 47);
- ASL Napoli 3 Sud, 12 distretti (34 + dal 48 al 59).

Ci siamo concentrati sul sito dell'ASL Napoli 1, che nel complesso risulta avere una buona organizzazione. Oltre al menu principale sono infatti presenti diverse sezioni che aiutano la navigazione: *Come fare per*, la rubrica ordinata in base "al problema"; *L'offerta della ASL*, organizzata per argomento; *L'offerta per*, che distingue i contenuti in base al destinatario (qui si trovano ad esempio le pagine dedicate all'assistenza sanitaria per gli stranieri e ad altri servizi per gli immigrati).
Questa diversa possibilità di accesso ai contenuti è spiegata anche nel testo introduttivo alla sezione *L'offerta della ASL*, nel quale si fa ricorso a un registro piuttosto informale.

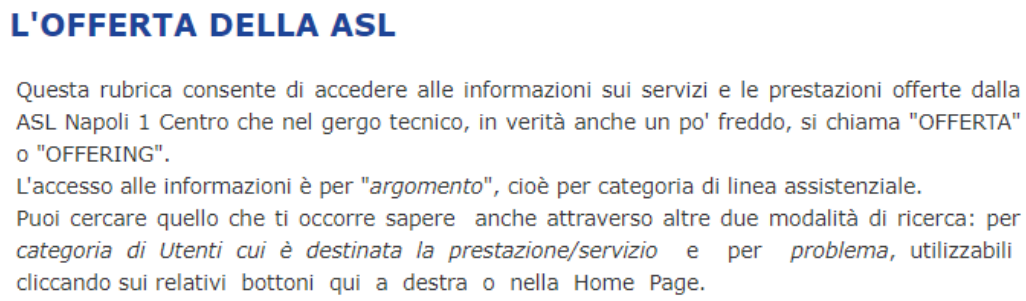


Figura 56. Il testo introduttivo presente nella sezione *L'offerta della ASL* nel sito dell'ASL Napoli 1.

Nella home page è disponibile anche una sezione dedicata al 118 e alla guardia medica e un banner per le cure domiciliari e l'assistenza residenziale.
La parte relativa ai programmi di prevenzione oncologica è invece piuttosto problematica: vi sono infatti diversi punti di accesso alle informazioni ma ciascuno rimanda a un contenuto diverso rispetto agli altri. Non differiscono solo i testi ma anche il percorso di navigazione. Un primo accesso è possibile dal menu *L'offerta della ASL*; ogni tipologia di screening ha una pagina dedicata ma manca un testo introduttivo più generico.

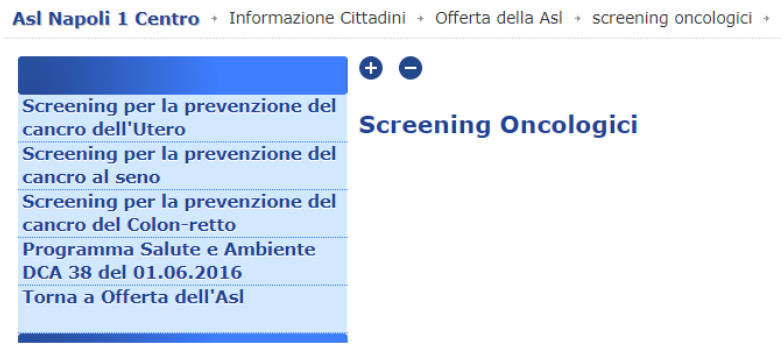


Figura 57. Pagine relative agli screening oncologici (versione 1).

Vi si arriva anche dal menu *L'offerta per* (> Donne): in questo caso, oltre a quelle più specifiche, esiste anche una parte generale.

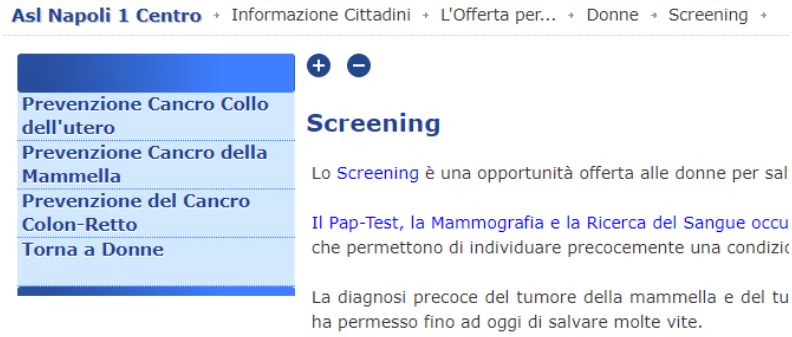


Figura 58. Pagine relative agli screening oncologici (versione 2).

Esiste anche una terza tipologia di contenuti: la pagina sembrerebbe trovarsi all'interno della sezione *Offerta della ASL*, ma ci si arriva invece (per caso) dall'altra rubrica *Offerta della ASL*.

Figura 59. Pagine relative agli screening oncologici (versione 3).

Vi è infine la possibilità accedere agli screening oncologici tramite il banner in home page, subito sotto le notizie in evidenza. Anche in questo caso, si apre una nuova pagina di contenuti.

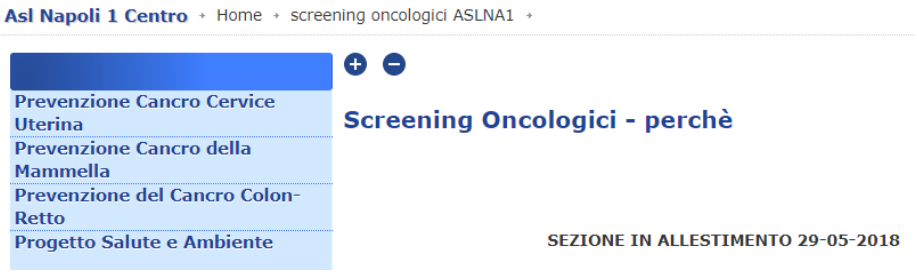


Figura 60. Pagine relative agli screening oncologici (versione 4).

8.3.3.16. Puglia

Il sistema sanitario della Puglia è organizzato in 6 ASL: Bari, Barletta-Andria-Trani, Brindisi, Foggia, Lecce, Taranto. Il sito principale è quello di Puglia Salute, da cui è possibile accedere ai siti delle singole ASL. Attraverso un progetto di immagine coordinata, è stata resa riconoscibile l'identità del sistema sanitario pugliese: il logo di Puglia Salute è presente in tutti i portali; la struttura del sito stesso è uniforme, con la sola differenza del colore predominante.

Per la nostra analisi, abbiamo scelto il sito dell'ASL di Bari. L'organizzazione della home page è piuttosto chiara: il menu principale rimane in evidenza anche quando si scorre la pagina verso il basso; nella parte centrale si trovano tutti i contenuti relativi all'assistenza mentre nella parte destra sono presenti i numeri utili, tra cui il servizio di emergenza. Mancano però delle sezioni che possano aiutare la navigazione, presenti invece, come abbiamo visto, in molti dei portali analizzati (*dedicato a*, *come fare per*, *servizi dalla a alla z*); in realtà, esiste una sezione *Come fare per* ma rimanda al sito generale Puglia Salute.

Un altro problema sono i contenuti: nella pagina relativa agli screening sono presenti soltanto un contatto telefonico e un orario. La pagina dedicata agli stranieri si apre con un titolo, senza nessun contenuto; è necessario un ulteriore click per arrivare finalmente alla pagina desiderata, nella quale, però, è presente soltanto un breve testo: il resto delle informazioni si trova negli allegati che devono essere scaricati.

Figura 61. La pagina dedicata ai cittadini stranieri nel sito dell'ASL di Bari.

Si nota una disomogeneità nell'organizzazione testuale: alcuni testi sono ben strutturati, altri (tra cui le cure domiciliari e l'emergenza sanitaria) sono costituiti da un blocco unitario, senza alcuna separazione tra paragrafi o titoli esplicativi.

Alla luce di tali problematiche, abbiamo deciso di includere nel campione anche il sito dell'ASL di Brindisi, in cui i contenuti sono tutti presenti e facilmente reperibili. Anche in questo caso, il servizio di emergenza si trova nella colonna laterale destra, mentre gli altri contenuti sono raccolti nella sezione *Assistenza.* I testi risultano ben strutturati: i paragrafi sono spaziati, si ricorre agli elenchi numerati, ai titoli e altri elementi (grassetto, sottolineato) per evidenziare le parti importanti.
L'unico problema è di nuovo il contenuto della pagina dedicata agli stranieri: le informazioni si trovano soltanto negli allegati.

Figura 62. La pagina dedicata ai cittadini stranieri nel sito dell'ASL di Brindisi.

8.3.3.17. Basilicata

Fanno parte del servizio sanitario della Basilicata due aziende sanitarie: l'ASP (Azienda Sanitaria di Potenza) e l'ASM (Azienda Sanitaria di Matera).
Il sito dell'ASP non risulta molto facile da navigare: alle informazioni si accede dai menu nella parte alta del sito, sotto l'header. Non vi sono però altre sezioni nella parte centrale che aiutino l'utente nella ricerca delle informazioni. Nella colonna laterale sono collocate in modo disordinato diverse immagini che rimandano a sezioni interne o a campagne promozionali, ma l'effetto è quello del banner pubblicitario.
La reperibilità dei contenuti è comunque relativamente semplice: dalla voce *Servizi per i cittadini* si accede alla pagina dedicata all'emergenza sanitaria e a quella relativa ai servizi per i migranti. Nel menu *Come fare per* è presente l'elenco alfabetico di tutti i servizi: qui si trova il link per accedere alle informazioni sull'assistenza domiciliare programmata. Manca un po' di attenzione alla ridondanza delle informazioni: nella sezione esistono infatti sia i contenuti relativi all'assistenza sanitaria agli stranieri che quelli relativi ai servizi per i migranti. Tali informazioni andrebbero integrate in un'unica pagina.
Non è stato invece possibile individuare le pagine relative ai programmi di prevenzione oncologica. Per scrupolo abbiamo interrogato anche il sito dell'ASM Matera, ma senza risultato. Tale mancanza potrebbe dipendere dal fatto che si tratta di un programma gestito a livello regionale; la regione ha infatti promosso il progetto Basilicata Donna dedicato alla prevenzione dei tumori. Nei siti, tuttavia, non si trova alcun riferimento al progetto.

8.3.3.18. Molise

Nel Molise è presente un'unica ASL regionale, denominata ASREM (Azienda Sanitaria Regionale del Molise), che si articola in 3 distretti sociosanitari (Campobasso, Termoli, Isernia).
A primo impatto si ha un'ottima impressione del sito, che risulta ben organizzato. L'individuazione delle informazioni non è però così immediata. La voce *Servizi per il cittadino*, all'interno del menu principale, raccoglie (come dichiarato) tutte le informazioni utili per gli utenti. Nella sezione dedicata alla prevenzione troviamo in effetti la pagina dedicata agli screening. Le altre 3 tematiche non risultano invece presenti. Per arrivarci, è necessario utilizzare la funzione di ricerca o navigare per un certo tempo all'interno del sito.
Il servizio di emergenza si rintraccia seguendo il percorso *Home > L'Azienda Sanitaria Regionale del Molise > Le reti di ASReM > La rete dell'emergenza*; l'assistenza domiciliare si trova invece in *Home > L'Azienda Sanitaria Regionale del Molise > I Distretti sociosanitari di ASReM > Cure domiciliari*. Non esiste invece una pagina dedicata all'assistenza sanitaria agli stranieri.
È presente infine una pagina per la valutazione del sito, ma non risulta possibile inviare i risultati alla fine del questionario.

8.3.3.19. Calabria

In Calabria esistono 5 Aziende Sanitarie Provinciali (ASP): Catanzaro, Cosenza, Crotone, Reggio Calabria, Vibo Valentia. Nonostante Catanzaro sia il capoluogo di regione, se si interroga Google cercando le aziende sanitarie della regione, tra i primi risultati troviamo le ASP di Reggio Calabria e Cosenza; l'ASP di Catanzaro si trova solo nella quarta pagina (sia per la ricerca "asl calabria" che "asp calabria").
Il sito dell'ASP di Catanzaro presenta una buona organizzazione dei contenuti: le informazioni principali si trovano nel menu in alto e nelle varie sezioni in basso, tra cui *Servizi sanitari*, che raccoglie un elenco di servizi utili al cittadino.

Figura 63. La sezione *Servizi sanitari* presente sul sito dell'ASP di Catanzaro.

In quest'area si trovano i servizi di emergenza, l'assistenza domiciliare, la prevenzione-screening.
La voce relativa all'emergenza rimanda all'unità operativa di Lamezia Terme; le altre aree sono in costruzione. Le informazioni sul 118 sono effettivamente presenti ma per poterle

leggere occorre scorrere molto in basso la pagina. La pagina dedicata allo screening fornisce informazioni generali; per i dettagli dei vari programmi è necessario scaricare un file word. La voce relativa all'assistenza domiciliare rimanda nuovamente al distretto di Lamezia Terme; se si desiderano informazioni su tale servizio in altre città, come Catanzaro, è necessario cliccare sulla pagina del distretto sanitario desiderato.
La sezione dedicata all'*Assistenza sociosanitaria* è in costruzione. È possibile che si tratti del contenitore in cui saranno fornite anche informazioni dedicate agli stranieri. Ricorrendo al motore di ricerca interno al sito si può comunque giungere all'Organismo aziendale immigrazione, dove si trovano indicazioni su servizi, progetti e ambulatori dedicati.

8.3.3.20. Sardegna

Nel 2016 è stata istituita l'ATS (Azienda per la Tutela della Salute) coincidente con l'ambito territoriale della Sardegna. L'ATS comprende 8 ASSL (Azienda Socio Sanitarie Locali): Cagliari, Carbonia, Lanusei, Nuoro, Olbia, Oristano, Sanluri e Sassari. Graficamente i siti sono tutti uguali: il colore dei titoli e dei menu è per tutti il rosso, la struttura è la stessa e riprende quella dell'ATS Sardegna. Il logo dell'ATS è presente in tutti i portali.
Per la valutazione, abbiamo scelto quello dell'ASSL di Cagliari. Il sito non sembrerebbe ottimizzato per monitor medio-grandi. Il font risulta piccolo e il contenuto copre circa là metà dello schermo, risultando un po' "condensato". Si tratta comunque di un portale facile da navigare: le voci del menu principale sono infatti riprese per esteso nella parte centrale, in modo che si possano visualizzare tutti i servizi offerti dall'azienda. Utili le sezioni *Argomenti*, *Servizi al cittadino* e *Servizi sanitari*.

Figura 64. Home page sel sito dell'ASSL di Cagliari.

Le informazioni si individuano facilmente. Il 118 si trova nel menu *Servizi al cittadino*; nella stessa sezione compare la voce *Come fare per*, che fornisce indicazioni su come risolvere alcuni problemi, come *Ricevere assistenza a casa* e "semplici risposte per orientarsi meglio all'interno del Sistema Sanitario Regionale e per sapere a chi rivolgersi quando si hanno dei problemi complessi o per conoscere ed esercitare i propri diritti". Vi si trova anche la voce *Per chi è straniero*.

Dal menu *Argomenti* si arriva agli screening oncologici, trattati in modo esaustivo: vi è una pagina di informazioni, generali, una pagina relativa ai programmi e una pagina con le domande frequenti sui vari tumori.

Figura 65. La pagina dedicata agli screening oncologici.

8.3.3.21. Sicilia

Il servizio sanitario della Sicilia comprende 9 ASP (Azienda Sanitaria Provinciale): Agrigento, Caltanissetta, Catania, Enna, Messina, Palermo, Ragusa, Siracusa e Trapani.
Abbiamo deciso di valutare sia l'ASP di Palermo, in quanto capoluogo, sia l'ASP di Agrigento, per fornire un'alternativa. Il sito dell'ASP di Palermo presenta infatti qualche problematica: in primo luogo, manca un menu principale da cui accedere alle varie sezioni; esistono delle sezioni nelle colonne laterali, ma sono relative principalmente a informazioni istituzionali. La parte centrale raccoglie le ultime notizie e una serie di argomenti vari, accompagnati da un'immagine più o meno esplicativa.

Figura 66.. Home page del sito dell'ASP di Palermo.

Dal menu *Direzione aziendale > Cosa fare per* si accede a vari contenuti, tra cui l'Assistenza Domiciliare Integrata e il servizio di mediazione culturale (da cui è possibile scaricare i vari allegati informativi). Alla mediazione culturale si arriva anche da una delle immagini nella parte centrale; nella stessa area sono presenti anche i contenuti dedicati all'emergenza migranti, all'emergenza urgenza e agli screening. L'immagine relativa al servizio di emergenza rimanda a un sito esterno.

La sezione relativa agli screening oncologici apre una nuova pagina con una diversa URL rispetto al portale (screening.asppalermo.org) e non è ben chiaro se si trova o meno all'interno del sito.

Il sito dell'ASP di Agrigento presenta una maggiore facilità di navigazione. I contenuti sono così organizzati: il menu nella parte superiore contiene informazioni relative all'azienda; il menu laterale comprende una sezione informativa relativa alle strutture e ai servizi sanitari; nella parte centrale sono invece presenti comunicazioni di carattere istituzionale e ultime notizie.

Dalla sezione a sinistra si accede alle pagine dedicate ai cittadini stranieri e al 118.

Per trovare invece le cure domiciliari è necessario cliccare sui servizi offerti dai vari distretti domiciliari. L'individuazione delle pagine relative alle campagne oncologiche non è proprio

immediata: il percorso più veloce è cliccare sul box *New prevenzione*, nella parte centrale, subito sotto alle informazioni istituzionali, poi sulla voce screening. Si apre una pagina con le informazioni sui programmi e la possibilità di scaricare ulteriore materiale di approfondimento. Per rintracciare tali contenuti, abbiamo anche provato con il servizio di ricerca: tra i risultati è presente un riferimento ai programmi ma per arrivare effettivamente ai contenuti sono necessari almeno 3 click. La ricerca, inoltre, non consente di digitare una stringa di oltre 20 caratteri.

9. Il profilo linguistico del corpus delle ASL

Effettuare l'annotazione automatica multi-livello del corpus ci consente di identificare la struttura linguistica sottostante ai testi ed è il prerequisito per la valutazione automatica della leggibilità. L'individuazione della struttura linguistica avviene a partire dal monitoraggio delle caratteristiche linguistiche. I risultati contribuiscono alla ricostruzione del profilo linguistico del corpus, che è il punto di partenza per l'individuazione sia dei parametri legati alla complessità, che delle caratteristiche che identificano quella data varietà testuale.

I tratti linguistici che abbiamo monitorato, riepilogati nella Tabella 102, comprendono caratteristiche di base, lessicali, morfosintattiche e sintattiche. Le caratteristiche di base considerate sono il numero medio di caratteri per parola (lunghezza delle parole) e il numero medio di parole per frase (lunghezza della frase).

Come caratteristiche lessicali sono misurate la ricchezza lessicale (rapporto type/token), la densità lessicale (calcolata come la proporzione di parole piene, o parole contenuto, sul totale delle occorrenze) e la composizione del vocabolario, che considera sia la percentuale di lemmi appartenenti al Vocabolario di Base che la loro distribuzione rispetto ai 3 repertori d'uso: fondamentale (FO), alto uso (AU) e alta disponibilità (AD).

I tratti morfosintattici monitorati sono la distribuzione delle categorie grammaticali nel testo e la distribuzione dei modi verbali. Per quanto riguarda le variabili sintattiche, sono misurate: la struttura dell'albero sintattico (altezza media dell'albero, lunghezza media delle relazioni di dipendenza), le caratteristiche della subordinazione (distribuzione delle proposizioni subordinate rispetto alle principali, posizione delle subordinate rispetto alla principale, incassamento gerarchico, ovvero lunghezza media delle catene di subordinazione), le caratteristiche dei predicati verbali (arità verbale, numero di dipendenti per testa verbale, numero medio di parole per clausola verbale), la modificazione nominale (lunghezza media delle catene preposizionali).

Tipo di caratteristiche	**Caratteristiche monitorate**
Di base	numero medio di caratteri per parola (lunghezza delle parole)
	numero medio di parole per frase (lunghezza della frase)
Lessicali	Ricchezza lessicale (rapporto type/token)
	Percentuale di lemmi appartenenti al *Vocabolario di Base*
	Distribuzione dei lemmi rispetto repertori d'uso (FO, AU, AD)
	Densità lessicale (% parole piene sul totale delle occorrenze)
Morfosintattiche	Modello statistico delle parti del discorso (distribuzione delle categorie morfosintattiche)
	Distribuzione dei modi verbali

Tipo di caratteristiche	Caratteristiche monitorate
Sintattiche	Caratteristiche relative alla struttura dell'albero sintattico: • altezza media dell'albero • lunghezza media delle relazioni di dipendenza • lunghezza media della più lunga relazione di dipendenza
	Caratteristiche relative all'uso della subordinazione: • distribuzione di frasi principali vs. subordinate • posizione delle subordinate rispetto alla principale • lunghezza media di sequenze consecutive di subordinate
	Caratteristiche relative alla modificazione nominale: • lunghezza media dei complementi preposizionali dipendenti in sequenza da una testa nominale
	Caratteristiche dei predicati verbali: • arità verbale • numero di dipendenti per testa verbale • numero medio di parole per clausola verbale

Tabella 102. Caratteristiche linguistiche considerate nel monitoraggio.

L'annotazione linguistica è stata effettuata con gli strumenti software sviluppati dall'Istituto di Linguistica Computazionale (ILC) del CNR di Pisa, che comprendono sia il POS tagging che il parser per l'analisi delle dipendenze. Tali strumenti riescono a monitorare una combinazione molto ricca di caratteristiche linguistiche, ma non includono quegli indici presenti invece in Coease che avevamo proposto di aggiungere al set di tratti da analizzare: la lunghezza dei paragrafi, la frequenza delle parole e la coesione.

La lunghezza dei paragrafi (numero di frasi per paragrafo) non è in realtà un indice particolarmente correlato alla leggibilità, ma è comunque un parametro che influisce sull'organizzazione e sulla chiarezza dei testi sul web. L'annotazione automatica viene effettuata su documenti già ripuliti del codice HTML e che si presentano in formato esclusivamente testuale, per cui, almeno per il momento, tale parametro non viene incluso nel set completo.

Nei principali studi sulla valutazione automatica delle leggibilità, la frequenza delle parole è generalmente valutata tramite modelli statistici del linguaggio, che considerano le probabilità di unigrammi, bigrammi, ecc. Nelle misurazioni tradizionali di leggibilità, la frequenza è misurata invece in base al confronto con corpora di riferimento o database lessicali. Ad esempio, in Coease, gli indici relativi alla frequenza si focalizzano sulla familiarità delle parole piene (nomi, verbi, avverbi, aggettivi), usando come riferimento Wikipedia in italiano.

Anche se non si ha a disposizione un modello statistico che descriva la distribuzione della frequenza delle parole del corpus, è comunque possibile effettuare delle prime osservazioni in base alla composizione del lessico e, in particolare, considerando la distribuzione dei lemmi appartenenti al *Vocabolario di Base* rispetto ai repertori d'uso. Il lessico fondamentale (FO) comprende infatti i primi 2.000 vocaboli del LIF e il lessico di alto uso (AU) i successivi 2.500 - 3.000 lemmi. La percentuale di lemmi appartenenti a tali fasce ci fornisce quindi una prima stima circa la presenza di vocaboli ad alta frequenza nel corpus.

Anche per quanto riguarda la coesione, sebbene non siano stati monitorati indici specifici, è possibile ottenerne una misurazione tramite l'uso di altri tratti considerati: ad esempio, la distribuzione delle parti del discorso ci può fornire dati circa la percentuale di pronomi o di connettivi presenti nel corpus.

Di seguito presentiamo i risultati del monitoraggio linguistico. Per ogni caratteristica valutata, si riporta generalmente il dato medio.
Purtroppo, non avendo ancora effettuato le prove di comprensione sui testi, non disponiamo dei livelli di difficoltà associati a ciascun documento. Non abbiamo quindi a disposizione un range di valori per ciascuna caratteristica da usare come parametro di confronto per la correlazione con la difficoltà. In base ai risultati, possiamo tuttavia riuscire a individuare quali siano i parametri maggiormente legati alla complessità.
Quando possibile, si sono inoltre portati a confronto i dati risultanti dal monitoraggio di corpora rappresentativi di altri generi testuali: giornalismo, letteratura, materiale didattico, prosa scientifica, linguaggio burocratico e linguaggio legislativo[239]. Per quanto riguarda invece l'ambito medico, possiamo confrontare i valori con quelli relativi ad alcune tipologie di testi considerate rappresentative della comunicazione medico-paziente: il corpus dei foglietti illustrativi (bugiardini) dei farmaci senza obbligo di prescrizione medica (Dell'Orletta et. al. 2016) e il corpus che raccoglie le informative di consenso per le procedure diagnostico-terapeutiche impiegate nelle aziende sanitarie toscane (Venturi et al. 2015).
Tali dati non sono però da intendersi come standard di riferimento, ma vanno considerati esclusivamente come termine di paragone puramente indicativo. I corpora appartengono infatti a una varietà diamesica diversa (sono tutti testi scritti) e la maggior parte presenta delle dimensioni nettamente maggiori rispetto al nostro corpus.
La comparazione del profilo linguistico della nostra varietà con i profili di altri corpora ci consente comunque di identificare da una parte, se esistono, tratti comuni a più generi, dall'altra le caratteristiche specifiche dei testi web informativi di ambito sanitario.

Corpus	**Sotto-corpora**	**N. testi**	**N. parole**
Giornalismo	Due Parole (Piemontese 1996)	643	306.222
	La Repubblica (Marinelli et al. 2003)		
Materiali didattici	Materiali didattici per la scuola primaria (Dell'Orletta et al. 2011a)	197	96.139
	Materiali didattici per la scuola secondaria superiore (Dell'Orletta et al. 2011a)		

[239] Si tratta di corpora utilizzati in vari studi dell'ILC. Ciascun genere è formato da due sottoclassi, in base alla tipologia di destinatario presa come riferimento; in particolare, il giornalismo comprende il corpus di testi tratti da La Repubblica (Rep) e il corpus di giornali di facile lettura tratti da Due Parole (2Par); la scrittura educativa e la letteratura sono divise in testi che si rivolgono ad adulti (AdEdu e AdLit) e testi rivolti a bambini (ChildEdu e ChildLit); la prosa scientifica comprende articoli scientifici (ScientArt) e articoli tratti da Wikipedia (Wiki). Casi a parte sono rappresentati dal linguaggio burocratico, in cui le sottoclassi sono rappresentate dalla versione originale dei documenti e dalla loro riscrittura semplificata, e dal linguaggio legislativo, che comprende atti legislativi in materia ambientale e la Costituzione italiana.

Corpus	Sotto-corpora	N. testi	N. parole
Letteratura	Narrativa per bambini (Marconi et al. 1994)	428	490.791
	Narrativa per adulti (Marinelli et al. 2003)		
Prosa scientifica	Articoli italiani di Wikipedia tratti dalla sezione "Ecologia e Ambiente"	377	677.040
	Articoli scientifici di vari settori scientifici (es. cambiamenti climatici, linguistica)		
Ling. legislativo	Atti legislativi in materia ambientale	554	1.320.353
	Costituzione Italiana (1947)		
Ling. burocratico	Testi burocratici originali	178	1.049.88
	Testi burocratici semplificati		
Consensi informati	Consensi informati impiegati nelle ASL toscane	583	607.677
Bugiardini	Foglietti illustrativi di farmaci senza obbligo di prescrizione medica tratti dal portale http://www.torrinomedica.it.	100	189.315
Corpus ASL	Testi relativi al servizio di emergenza-urgenza, allo screening oncologico, all'assistenza sanitaria agli stranieri e all'assistenza domiciliare.	248	122.793

Tabella 103. Corpora di monitoraggio portati a confronto.

9.1. Caratteristiche di base

Come già accennato, le caratteristiche di base monitorate sono il numero medio di caratteri per parola (lunghezza delle parole) e il numero medio di parole per frase (lunghezza della frase). Nella tabella seguente riportiamo i dati relativi sia all'intero corpus, che ai vari sotto-corpora, distinti in base alla tematica affrontata: servizio di emergenza-urgenza, screening oncologico, assistenza sanitaria agli stranieri e assistenza domiciliare.

Corpus	N. di caratteri per parola	N. di parole per frase	N. parole per clausola
Emergenza	5,38	18,18	14,35
Screening	5,56	17,65	12,40
Stranieri	5,86	18,67	11,11
Ass. Dom.	6,14	18,76	15,48
CORPUS ASL	**5,71**	**18,39**	**13,37**

Tabella 104. Risultati del monitoraggio delle caratteristiche di base.

Alcune indicazioni emergono comparando tali valori con i vari corpora di confronto (Tabella 105). Anche per quanto riguarda gli altri corpora, viene fornita la media del genere considerato, senza però evidenziare la distinzione interna tra le due sottoclassi, a meno che questa risulti significativa ai fini della discussione.

Corpus	N. di caratteri per parola	N. di parole per frase
Giornalismo	5,09	22,90
Materiali didattici	5	27,64
Letteratura	4,91	17,61
Prosa scientifica	5,57	28,73
Ling. legislativo	/	20,79
Ling. burocratico	/	23,36
Consensi informati	6,75	16,06
Bugiardini	/	11,18
MEDIA	5,46	21,03

Tabella 105. Caratteristiche di base nei corpora analizzati.

Dal confronto risulta che la lunghezza media delle frasi del nostro corpus (pari a 18,39 parole per frase) si trova al di sotto del valore medio totale dei corpora considerati. Il dato si avvicina maggiormente a quello dei testi di facile lettura (che contengono frasi con una lunghezza media pari a 19 tokens) che non a quello dei testi di difficile lettura (con lunghezza media pari a 27 tokens). La lunghezza delle frasi risulta maggiore rispetto alle altre tipologie testuali dell'ambito sanitario, tuttavia se si considera il numero medio di parole per clausola verbale la situazione si capovolge: ad esempio, il corpus di bugiardini presenta un numero medio pari a 17,36 (più vicino ai testi difficili) mentre il nostro corpus ha un valore di 13,37.
La lunghezza delle parole rientra invece nella media dei corpora; osservando i dati, si nota però che il dato è influenzato molto dal valore riportato dal corpus dei consensi informati, pari a un numero medio di 6,75 caratteri per parola. Analizzando ciascun corpus separatamente si osserva invece che i singoli valori risultano tutti inferiori rispetto al nostro corpus.
Per quanto riguarda la ripartizione interna al corpus, non si notano sostanziali differenze per entrambe le caratteristiche considerate; il sotto-corpora dell'assistenza domiciliare presenta i valori più alti (con lunghezza media delle parole pari a 6,14 caratteri e lunghezza media delle frasi pari a 18,76).

9.2. Caratteristiche lessicali

Se si considera il profilo lessicale, i tratti monitorati sono il rapporto type/token (TTR), la composizione del vocabolario e la densità lessicale.
Il rapporto type/token rappresenta una misura della ricchezza lessicale di un testo. L'indice mette in rapporto il numero di parole diverse e il numero di occorrenze totali di un testo; essendo un parametro sensibile alla lunghezza del testo, viene calcolato su campioni con la stessa dimensione. I valori oscillano tra 0 e 1: quelli vicini allo 0 indicano che il vocabolario del testo è meno vario, quelli vicini a 1 caratterizzano testi particolarmente variegati dal punto di vista lessicale; i valori tendono quindi a crescere all'aumentare della complessità di

un testo. Abbiamo monitorato il TTR rispetto alle prime 100 e 200 parole del testo, sia in relazione ai lemmi che alle forme.
Come mostrato nella Tabella 106, il valore medio del corpus ASL risulta piuttosto alto (0,79) se si considerano le prime 100 forme del testo ma si abbassa a 0,68 se monitorato rispetto ai primi 100 lemmi. I punteggi migliorano quando il TTR è calcolato sulle prime 200 forme (0,69) e lemmi del testo (0,59). Ancora una volta, la differenza interna al corpus è statisticamente irrilevante, con valori che oscillano di 0,3-0,4 punti.

Corpus	**TTR (prime 100 forme)**	**TTR (primi 100 lemmi)**	**TTR (prime 200 forme)**	**TTR (primi 200 lemmi)**
Emergenza	0,79	0,67	0,69	0,58
Screening	0,80	0,69	0,69	0,59
Stranieri	0,78	0,67	0,69	0,59
Ass. Dom.	0,82	0,71	0,72	0,61
CORPUS ASL	**0,79**	**0,68**	**0,69**	**0,59**

Tabella 106. Rapporto type/token (TTR).

La distribuzione del lessico rispetto al *Vocabolario di Base*[240] fornisce un'indicazione qualitativamente più raffinata del tipo di vocabolario usato nei testi. Dai risultati del monitoraggio emerge che il corpus ASL contiene una percentuale di lemmi appartenenti al VdB pari al 59,46%. Sebbene questo valore sia piuttosto basso e sia indicativo di testi di difficile lettura, il parametro relativo alla percentuale di lessico fondamentale (FO) restituisce un punteggio più positivo (71,74%).

Corpus	**% lemmi nel VdB**	**FO**	**AU**	**AD**	**Densità lessicale**
Emergenza	60,52	74,89	18,59	6,52	0,57
Screening	64,72	72,90	20,90	6,20	0,59
Stranieri	58,91	68,99	24,46	6,55	0,60
Ass. Dom.	55,01	70,13	23,90	5,97	0,61
CORPUS ASL	**59,46**	**71,74**	**21,88**	**6,38**	**0,59**

Tabella 107. Risultati del monitoraggio delle caratteristiche lessicali.

Risultano preoccupanti anche i risultati ottenuti dai sotto-corpora dell'assistenza sanitaria agli stranieri e dell'assistenza domiciliare, con soltanto il 59% e il 55% circa di lemmi appartenenti al Vocabolario di Base; nonostante il 69-70% di tali lemmi appartenga al repertorio d'uso del lessico fondamentale, questo valore può essere considerato indicativo della complessità lessicale di tale tipologia di testi, soprattutto considerando il destinatario a cui sono rivolti.

[240] Viene preso come riferimento il *Grande Dizionario Italiano dell'uso* (De Mauro, 2000)

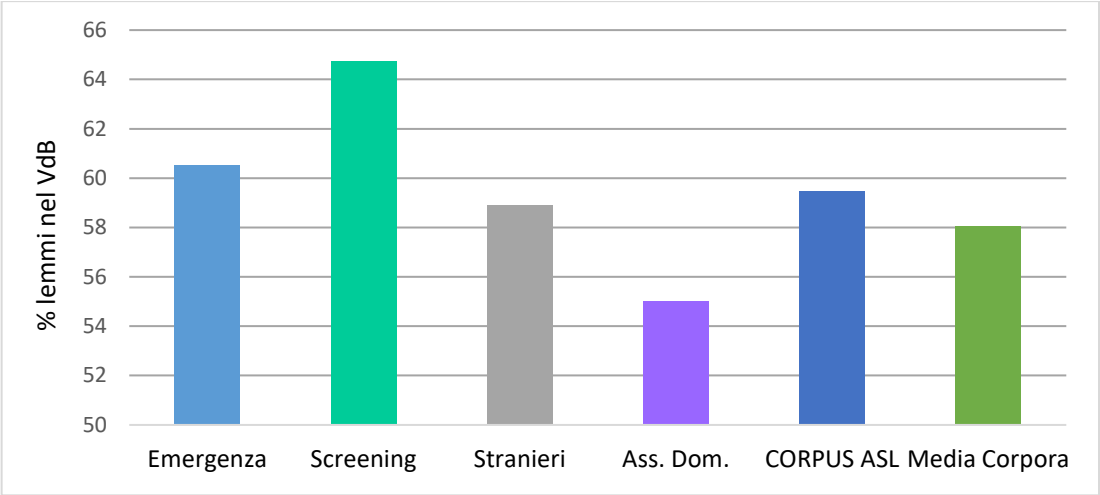


Figura 67. Percentuale di lemmi appartenenti al *Vocabolario di Base*.

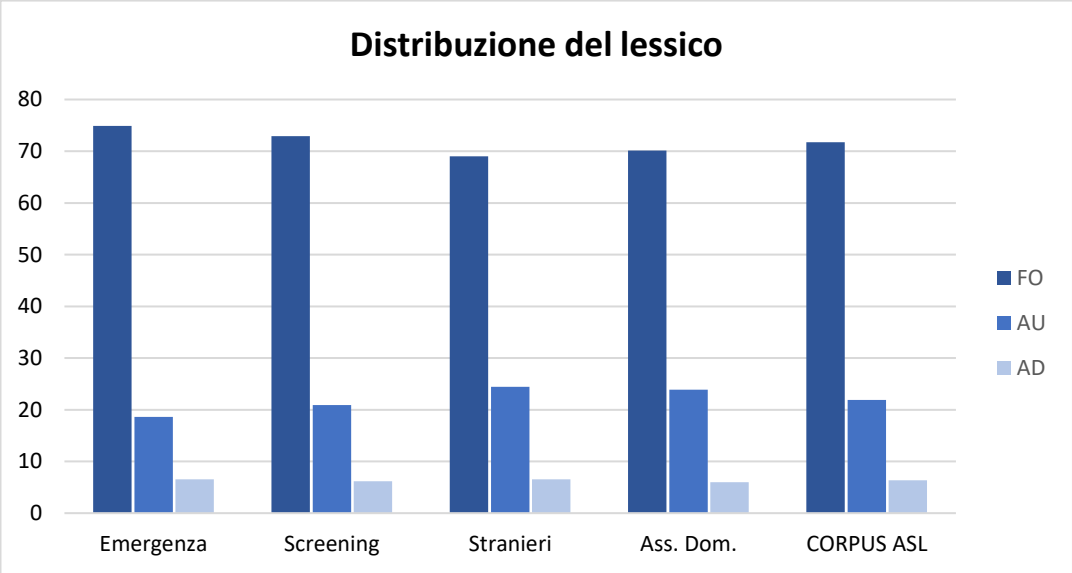


Figura 68. Distribuzione del lessico nei tre repertori d'uso.

Rispetto agli altri corpora analizzati, il corpus ASL presenta una percentuale di lemmi appartenenti al VdB vicina alla media (58,05%), con un punteggio piuttosto simile ai testi burocratici (58,81%) e ai consensi informati (57,24%). Il distacco dai corpora che registrano ricorrenze più alte risulta di almeno 10 punti percentuali. Se si considera invece la percentuale di lessico fondamentale (FO), il corpus ASL mostra una situazione più positiva e più vicina ai corpora che raccolgono testi giornalistici e materiali didattici.

Genere	TTR (prime 100 forme)	% lemmi nel VdB	FO	Densità lessicale
Giornalismo	0,76	70,84	73,53	0,56
Materiali didattici	0,80	73,57	73,15	0,56
Letteratura	0,81	71,76	76,31	0,57
Prosa scientifica	0,78	55,44	67,04	0,58

Genere	TTR (prime 100 forme)	% lemmi nel VdB	FO	Densità lessicale
Ling. legislativo	0,46	35,60	66,69	0,54
Ling. burocratico	0,69	58,81	65,63	0,54
Consensi informati	0,72	57,24	/	0,59
Bugiardini	/	41,12	66,48	/
MEDIA CORPORA	0,72	58,05	69,83	0,56
CORPUS ASL	**0,79**	**59,46**	**71,74**	**0,59**

Tabella 108. Caratteristiche lessicali nei corpora analizzati.

La densità lessicale, calcolata come la proporzione di parole piene, o parole contenuto, sul totale delle occorrenze, è un parametro associato alla ricchezza lessicale di un testo.
Valori superiori di densità lessicale sono solitamente associati a un maggior carico informativo e, quindi, ad una maggiore complessità testuale. I testi informativi, come quelli scientifici, tendono ad essere lessicalmente più densi; l'analisi contrastiva conferma tale ipotesi, con il corpus ASL che ottiene i valori più elevati di densità lessicale (0,59), insieme al corpus dei consensi informati.

9.3. Caratteristiche morfosintattiche

I tratti morfosintattici monitorati sono la distribuzione delle parti del discorso nel testo e la distribuzione dei modi verbali. La Tabella 109 e la Figura 69 mostrano la distribuzione percentuale delle categorie morfosintattiche nei vari corpora considerati.

Genere	Agg.	Avv.	Cong.	Prep.	Pron.	Nomi	Verbi
Giornalismo	6,16	4,18	3,65	15,85	3,04	28,24	13,28
Materiali didattici	7,76	5,79	4,69	14,57	5,36	23,08	13,86
Letteratura	6,15	**5,90**	4,60	**12,21**	**6,51**	**23,02**	**15,39**
Prosa scientifica	8,85	3,98	3,60	17,05	3,03	28,44	10,62
Ling. legislativo	8,28	**1,86**	**4,71**	**19,64**	**2,25**	**30,21**	**10,04**
Ling. burocratico	**5,85**	2,03	**2,92**	19,25	2,96	30,17	11,09
Consensi informati	**9,26**	3,60	4,29	16,19	/	28,51	11,83
MEDIA ALTRI CORPORA	7,47	3,91	4,07	16,39	3,86	27,38	12,30
CORPUS ASL	**8,58**	**2,96**	**3,73**	**17,92**	**2,29**	**32,34**	**9,81**

Tabella 109. Distribuzione delle categorie morfosintattiche nei corpora analizzati
(sono evidenziati in azzurro i valori più bassi e in arancione quelli più alti).

Si osserva che le categorie maggiormente oscillanti tra i diversi generi testuali sono rappresentate da preposizioni, pronomi, nomi e verbi (tutte caratterizzate da una deviazione standard maggiore di 2). Rispetto ai valori medi, il corpus ASL presenta delle differenze maggiori sia nella categoria dei sostantivi che in quella dei verbi.

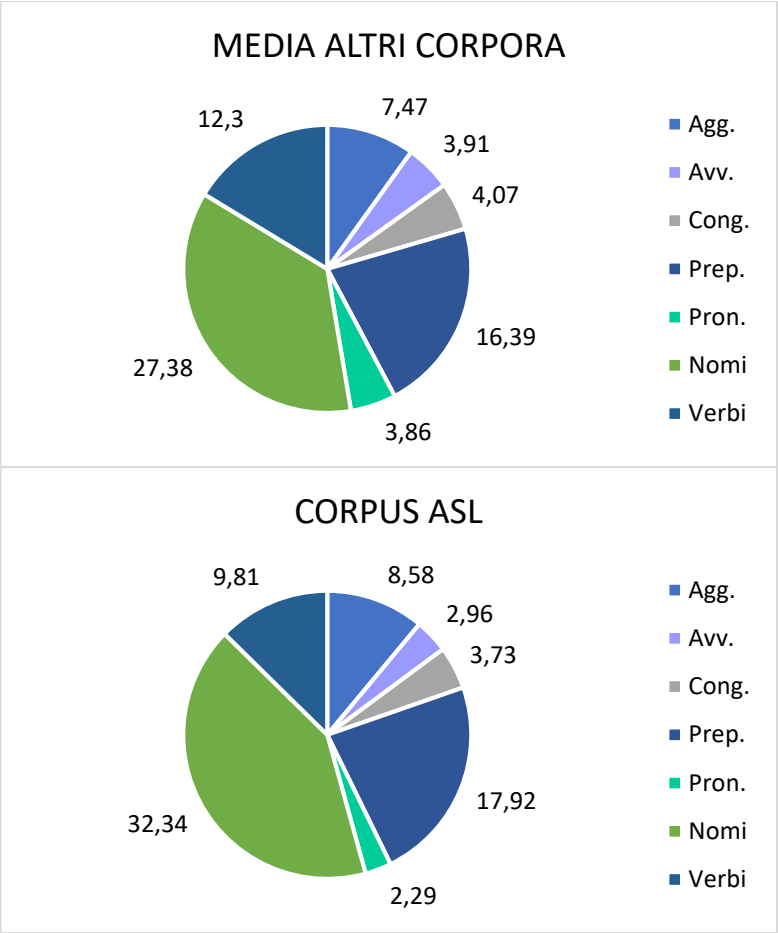


Figura 69. Distribuzione delle categorie grammaticali nei corpora analizzati.

Per quanto riguarda gli aggettivi, il corpus ASL registra una delle percentuali più alte (8,58%), insieme al corpus dei consensi informati (9,26%) e della prosa scientifica (8,85%). Anche nella categoria dei nomi si rileva una maggiore frequenza rispetto agli altri generi (32,34% su una media di 27,38%), con uno scarto di ben 2 punti sul successivo valore superiore (corpus legislativo: 30,21%). Ottiene invece valori più bassi nelle categorie dei verbi (9,81% su una media di 12,30%), dei pronomi (2,29% su una media di 3,86%) e degli avverbi (2,96% su una media di 3,96%). La distribuzione delle congiunzioni si approssima invece al dato medio (3,73% rispetto a una media di 4,07%)[241].

Rispetto alla categoria delle preposizioni, si nota una percentuale abbastanza elevata (17,92%), superiore al dato medio (16,39%). Si tratta di un dato atteso, considerata anche l'alta percentuale di nomi. Come osservato da Biber (1988), e confermato in parte anche dagli studi di Brunato (2014), esiste infatti un'associazione sistematica nella distribuzione di nomi, preposizioni e aggettivi attributivi in testi connotati da una spiccata funzione informativa. Tale ipotesi sembrerebbe avvalorata dai risultati emersi in questo studio: le

[241] La percentuale di pronomi e connettivi ci può fornire una stima circa la coesione testuale. Un'alta densità di pronomi comporta una ridotta leggibilità, poiché rende la coesione referenziale meno esplicita; in questo caso, la percentuale è relativamente bassa (2,29%). Quanto ai connettivi, si registra una bassa presenza di avverbi (2,96%) e congiunzioni (3,73%), ma un'elevata occorrenza di preposizioni (17,92%).

categorie di nomi, aggettivi e preposizioni sono infatti quelle maggiormente rappresentate nel corpus ASL.

A partire dalla distribuzione delle categorie morfosintattiche, è possibile mettere in relazione la frequenza di alcune categorie grammaticali rispetto ad altre, ad esempio il rapporto tra nomi e verbi o, come abbiamo fatto per la densità lessicale, la proporzione delle parole piene (nomi, aggettivi, avverbi e verbi) rispetto al totale delle occorrenze.
Per quanto riguarda il rapporto tra nomi e verbi (Tabella 110), si tratta di una misura generalmente associata alla distinzione sull'asse diamesico, con il parlato che registra valori più bassi nell'andamento di questo parametro rispetto allo scritto, dove invece i nomi predominano sui verbi[242]. "Questa distinzione si è dimostrata significativa per cogliere differenze e similarità tra generi e varietà testuali anche all'interno della stessa dimensione diamesica; focalizzandosi su corpora di produzione scritta annotati automaticamente, Montemagni (2013a) ha osservato che la proporzione tra nomi e verbi varia in maniera tale da discriminare testi informativi, da un lato, e testi di scrittura creativa, dall'altro, che riportano valori quasi comparabili a quelli del parlato" (Brunato 2014, p. 12).
La correlazione tra la predominanza dei nomi e il carattere fortemente informativo del testo risulta confermata anche dai dati qui ottenuti: il corpus ASL riporta infatti un rapporto nomi/verbi molto alto (pari a 3,30), simile a quello del linguaggio legislativo (3,07). Rispetto alla media dei corpora il corpus ASL è superiore di ben 1 punto.
Si osserva una variazione rilevante all'interno dei sotto-corpora ASL, con valori più bassi (e dunque simili al parlato) registrati dai testi relativi all'assistenza sanitaria agli stranieri (il rapporto nomi/verbi risulta 1:1) e una maggiore predominanza di nomi nei testi che riguardano l'assistenza domiciliare (3,67) e gli screening tumorali (3,15).

Genere	**Rapporto nomi/verbi**
Giornalismo	2,13
Materiali didattici	1,67
Letteratura	1,50
Prosa scientifica	2,68
Ling. legislativo	3,07
Ling. burocratico	2,72
Consensi informati	2,41
MEDIA ALTRI CORPORA	**2,31**
Emergenza	2,76
Screening	3,15
Stranieri	1,00
Ass. Dom.	3,67
CORPUS ASL	**3,30**

Tabella 110. Rapporto nomi/verbi nei corpora analizzati.

[242] Cfr. Voghera 2005.

L'analisi delle categorie morfosintattiche può essere condotta anche in relazione alle sotto-classi, come ad esempio le congiunzioni coordinanti e subordinanti (Tabella 111).

Corpus	Cong. Coord.	Cong. Subord.
Emergenza	2,65	0,68
Screening	3,19	0,72
Stranieri	3,57	0,36
Ass. Dom.	3,72	0,25
CORPUS ASL	**3,24**	**0,50**
Ripartizione classe	**86,72%**	**13,28%**

Tabella 111. Ripartizione interna della classe delle congiunzioni.

Se osserviamo la distribuzione delle due sotto-categorie grammaticali, si nota una netta preferenza per le congiunzioni coordinanti (86,72%). Assumendo che un'alta percentuale di congiunzioni subordinanti sia un indicatore della proporzione di costruzioni ipotattiche all'interno del testo, risulta che il corpus ASL fa un limitato ricorso a tali tipi di costruzioni, dato che è associato a una minore difficoltà sintattica.
Altre caratteristiche interessanti da analizzare sono la distribuzione dei tratti flessivi nei verbi e la distribuzione dei pronomi nelle rispettive sotto-classi (personali, dimostrativi, clitici, ecc.).
La Tabella 112 mostra la distribuzione dei tratti flessivi di persona e numero nei verbi principali. Emerge con molta evidenza l'alta frequenza percentuale di verbi alla terza persona singolare e plurale; tale dato è registrato sia dal corpus ASL (3 sing: 56,49%, 3 plu: 21,13%) che dagli altri corpora. L'unica eccezione è presente nel linguaggio burocratico, in cui si ha anche un'alta percentuale di verbi alla prima persona plurale, classe che invece è scarsamente rappresentata negli altri generi.
Questo dato quantitativo non è casuale e conferma la tendenza alla spersonalizzazione che caratterizza spesso i testi istituzionali: la terza persona è infatti associata all'uso di frasi impersonali, passive o al ricorso a soggetti inanimati scelti per rivolgersi al cittadino (l'ufficio, l'azienda, il servizio).

Genere	1 p. sing	2 p. sing	3 p. sing	1 p. plu.	2 p. plu.	3 p. plu.
Giornalismo	2,6	0,5	25,0	2,1	0,1	15,4
Materiali didattici	1,53	0,69	36,59	1,39	0,18	13,97
Letteratura	3,70	1,98	38,82	1,97	0,73	10,18
Prosa scientifica	0,43	0,68	29,54	0,84	0,03	14,37
Ling. legislativo	1,14	0,51	45,28	0,01	0,07	25,10
Ling. burocratico	2,55	0,64	**20,71**	**7,53**	0,26	**4,52**

MEDIA CORPORA	2,0	0,8	**32,6**	2,3	0,2	**13,9**
CORPUS ASL	1,05	3,29	**56,49**	0,47	0,24	**21,13**

Tabella 112. Distribuzione dei tratti flessivi di persona nei verbi all'interno dei corpora analizzati.

Tale tendenza può essere verificata anche misurando l'andamento della distribuzione dei pronomi, in particolare confrontando i pronomi personali e i pronomi clitici (Tabella 113). Un basso uso di pronomi personali è indice di spersonalizzazione, al contrario un uso maggiore riflette una modalità di scrittura più diretta e orientata al destinatario; anche la presenza del clitico può essere interpretabile come segno della presenza di costruzioni impersonali.

Genere	**Pron. personali**	**Pron. clitici**
Giornalismo	0,14	1,28
Materiali didattici	0,61	2,17
Letteratura	0,75	3,38
Prosa scientifica	0,14	2,14
Ling. legislativo	0,11	0,69
Ling. burocratico	0,37	1,48
MEDIA CORPORA	**0,35**	**1,86**
CORPUS ASL	**0,07**	**0,90**

Tabella 113. Distribuzione dei pronomi personali e clitici nei corpora analizzati.

La tabella mostra che i pronomi clitici predominano in ciascun corpora considerato. Il corpus della letteratura presenta le percentuali più elevate in ciascuna sotto classe, seguito da quello dei materiali didattici. Nel corpus ASL la categoria dei pronomi è invece scarsamente rappresentata, con valori percentuali simili al linguaggio legislativo e al di sotto della media degli altri corpora di confronto.
La tendenza alla spersonalizzazione sembrerebbe qui confermata dal basso valore dei pronomi personali (0,07%). Il clitico sembra registrare invece una bassa frequenza, almeno nel confronto con gli altri generi testuali; dal momento che tale pronome può svolgere diversi ruoli sintattici (accusativo, dativo, riflessivo, partitivo) oltre alla funzione impersonale, è preferibile confrontare questo dato con i valori risultanti dal monitoraggio delle relazioni di dipendenza sintattica. In particolare, abbiamo a disposizione i dati relativi alla distribuzione percentuale di tre classi di dipendenza: le dipendenze clitiche, che identificano la relazione tra un pronome clitico e una testa verbale usata in funzione pronominale e le relazioni tra una testa verbale e il clitico in funzione accusativa o dativa.

Genere	Dip. clitiche	Funz. accusativa	Funz. dativa
CORPUS ASL	0,68	2,68	0,08

Tabella 114.Distribuzione percentuale di dipendenze sintattiche nel corpus.

Possiamo utilizzare questi parametri per incrociare i dati relativi alla percentuale di pronomi clitici con quelli relativi alle dipendenze clitiche (Tabella 115). Dal confronto si nota che i corpora che presentano valori simili a quelli del nostro corpus per quanto riguarda la frequenza della categoria dei clitici, presentano una minore preponderanza di dipendenze sintattiche clitiche.

Genere	Pron. clitici	Dip. clitiche
Giornalismo	**1,28**	**0,67**
Materiali didattici	2,17	1,19
Letteratura	3,38	1,36
Prosa scientifica	2,14	0,77
Ling. legislativo	**0,69**	**0,37**
Ling. burocratico	1,48	0,99
CORPUS ASL	**0,90**	**0,68**

Tabella 115. Distribuzione di dipendenze sintattiche di tipo clitico nel corpus.

Per quanto riguarda invece la frequenza dei modi verbali, possiamo considerare ad esempio l'occorrenza dei participi. Come per le congiunzioni, anche l'uso del participio può essere infatti un indicatore della presenza di costruzioni ipotattiche all'interno del testo, in particolare di subordinate implicite, che contribuiscono a una maggiore oscurità e densità informativa.
La Tabella 116 mostra la distribuzione delle forme participiali di verbi principali (V+p) nei vari corpora esaminati. Dai dati relativi alla frequenza del modo verbale sono state precedentemente escluse le forme participiali usate per la formazione di tempi verbali composti.

Corpus	V+p
Giornalismo	10,16
Materiali didattici	10,87
Letteratura	9,07
Prosa scientifica	19,26
Ling. legislativo	29,28
Ling. burocratico	28,81
MEDIA CORPORA	**15,63**
Emergenza	21,16

Corpus	V+p
Screening	16,90
Stranieri	14,95
Ass. Dom.	26,80
CORPUS ASL	**23,06**

Tabella 116. Frequenza dei participi nei corpora analizzati.

Come atteso, l'uso di frasi participiali risulta essere una caratteristica tipica del linguaggio burocratico e legislativo, che raggiungono valori più elevati (rispettivamente 28,81% e 29,28%). Anche nel corpus ASL l'uso del participio è piuttosto elevato (23,06%), con quasi 8 punti di differenza rispetto alla media dei corpora.
In questo caso, si registra una notevole variazione interna al corpus, con percentuali che vanno dal 15-16% nei sotto-corpora degli screening e dell'assistenza agli stranieri al 21,16 % in quello relativo al servizio di emergenza e al 26,80% nell'assistenza domiciliare.

9.4. Caratteristiche sintattiche

Prendiamo adesso in considerazione le caratteristiche selezionate dal livello più profondo dell'analisi automatica, l'annotazione sintattica a dipendenze. I tratti monitorati sono la struttura dell'albero sintattico, le caratteristiche della subordinazione, le caratteristiche dei predicati verbali e la modificazione nominale.
La misura delle caratteristiche strutturali dell'albero sintattico è un parametro rilevante per valutare la complessità di un testo; include diversi parametri:

- la profondità dell'albero di analisi (o media delle altezze massime degli alberi sintattici, per frase), calcolata come la distanza massima che intercorre tra una foglia (rappresentata da parole del testo senza dipendenti) e la radice dell'albero, espressa come numero di archi (ovvero relazioni di dipendenza) attraversati nel cammino foglia-radice. Più alto è il valore di questo parametro, maggiore sarà la complessità sintattica del testo;
- la media della lunghezza delle relazioni di dipendenza (o media della lunghezza dei link), calcolata come la distanza media in parole tra la testa e il dipendente (con esclusione delle relazioni riguardanti la punteggiatura). Dipendenze sintattiche più lunghe sono associate ad una maggior difficoltà di elaborazione da parte del lettore e sono dunque sinonimo di maggiore complessità del testo;
- la media delle lunghezze massime delle relazioni di dipendenza (o media della lunghezza dei link massimi): la lunghezza media (per frase) della più lunga relazione di dipendenza (con esclusione delle relazioni riguardanti la punteggiatura).

La Tabella 117 mostra i dati relativi ai tre parametri per quanto riguarda il corpus ASL e i corpora di confronto.

Corpus	Media altezze massime	Media lunghezza link	Media lunghezza link massimi
Giornalismo	5,90	2,27	9,10
Materiali didattici	6,45	2,39	10,69
Letteratura	4,54	2,33	7,03
Prosa scientifica	7,04	2,47	10,93
Ling. legislativo	5,44	2,68	9,51
Ling. burocratico	6,64	2,36	10,17
Consensi informati	4,86	/	6,43
Bugiardini	3,21	/	4,40
MEDIA CORPORA	**6,00**	**2,42**	**9,57**
Emergenza	5,40	2,14	7,22
Screening	5,25	2,21	7,12
Stranieri	5,34	2,29	7,96
Ass. Dom.	5,52	2,25	7,57
CORPUS ASL	**5,39**	**2,22**	**7,53**

Tabella 117. Caratteristiche dell'albero sintattico nei corpora analizzati.

Rispetto al primo fattore, si nota che la variazione interna al corpus ASL è statisticamente irrilevante. Prendendo a confronto gli altri corpora, si osserva che il corpus presenta uno tra i valori più bassi, avvicinandosi molto al punteggio medio del linguaggio legislativo.
La situazione è simile per la lunghezza media delle relazioni di dipendenza, con una differenza interna pressoché nulla e valori inferiori rispetto agli altri corpora. Per quanto riguarda la lunghezza massima delle relazioni di dipendenza, si nota un maggiore distacco (di circa 2 punti) rispetto ai corpora che registrano i valori più alti; il corpus di letteratura presenta invece un valore simile. Quanto all'ambito sanitario, si osserva che, rispetto ai testi ASL, il corpus dei consensi informati registra un punteggio inferiore di un punto (6,43) e quello dei bugiardini di ben 3 punti (4,40).
Anche la subordinazione è un marcatore di maggiore complessità strutturale. Le caratteristiche della subordinazione monitorate sono:

- la distribuzione delle proposizioni subordinate rispetto alle principali;
- la posizione delle subordinate rispetto alla principale;
- l'incassamento gerarchico, ovvero lunghezza (o profondità) media delle catene di subordinazione.

Consideriamo, in primo luogo, la proporzione tra frasi principali e subordinate (Tabella 118).

Corpus	Frasi principali	Frasi subordinate	Rapporto princ. e sub.
Giornalismo	70,44%	29,25%	0,42
Materiali didattici	66,75%	31,73%	0,48
Letteratura	67,05%	32,31%	0,48

Corpus	Frasi principali	Frasi subordinate	Rapporto princ. e sub.
Prosa scientifica	71,02%	28,22%	0,40
Ling. legislativo	79,73%	20,27%	0,25
Ling. burocratico	62,61%	36,27%	0,58
Consensi informati	74,7%	25,3%	0,34
MEDIA CORPORA	**70,33%**	**29,05%**	**0,41**
Emergenza	74,15%	25,85%	0,35
Screening	70,62%	29,38%	0,42
Stranieri	74,75%	25,25%	0,34
Ass. Dom.	79,40%	20,60%	0,26
CORPUS ASL	**74,16%**	**25,84%**	**0,35**

Tabella 118. Proporzione tra frasi principali e subordinate nei corpora analizzati.

Come si può osservare, il corpus ASL fa un ampio uso delle clausole principali (74,16%), mentre ricorre meno spesso a costruzioni ipotattiche (25,84%). Il rapporto tra frasi principali e subordinate risulta tra i più bassi (0,35) rispetto agli altri corpora di confronto, secondo soltanto al corpus legislativo (0,25) e a quello dei consensi informati (0,34).
La variazione interna al corpus è stavolta più significativa: si va dal sotto-corpus dell'assistenza domiciliare, che ottiene una proporzione media pari a 0,26, fino al sotto-corpus dello screening oncologico che raggiunge un valore medio di 0,42. Si registra una notevole variazione anche all'interno dei vari sotto-corpora (Figure 70-73): ad esempio, nei testi relativi all'emergenza sanitaria i valori oscillano da 0,06 a 2, in quelli dedicati all'assistenza sanitaria agli stranieri da 0,04 a 1,56.

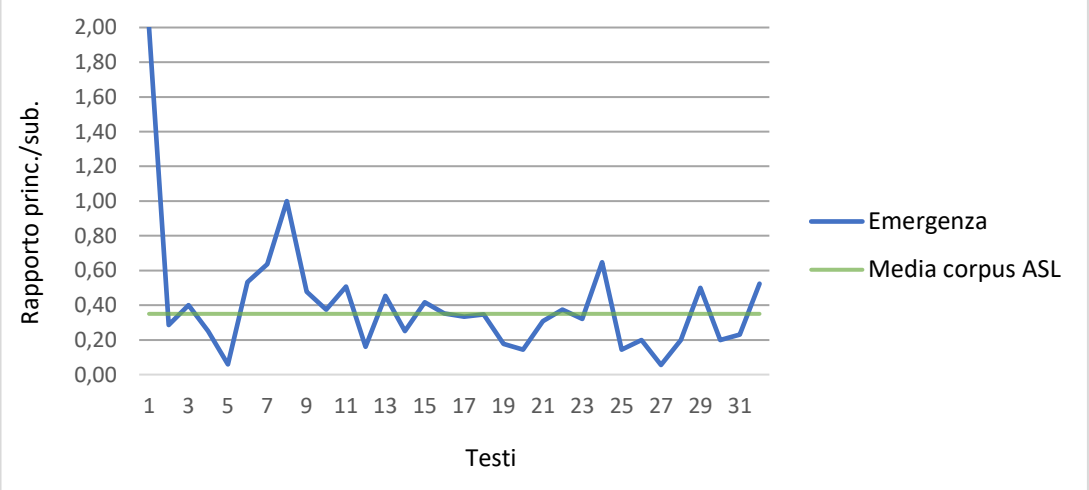


Figura 70. Proporzione tra frasi principali e subordinate nel corpus dell'emergenza sanitaria.

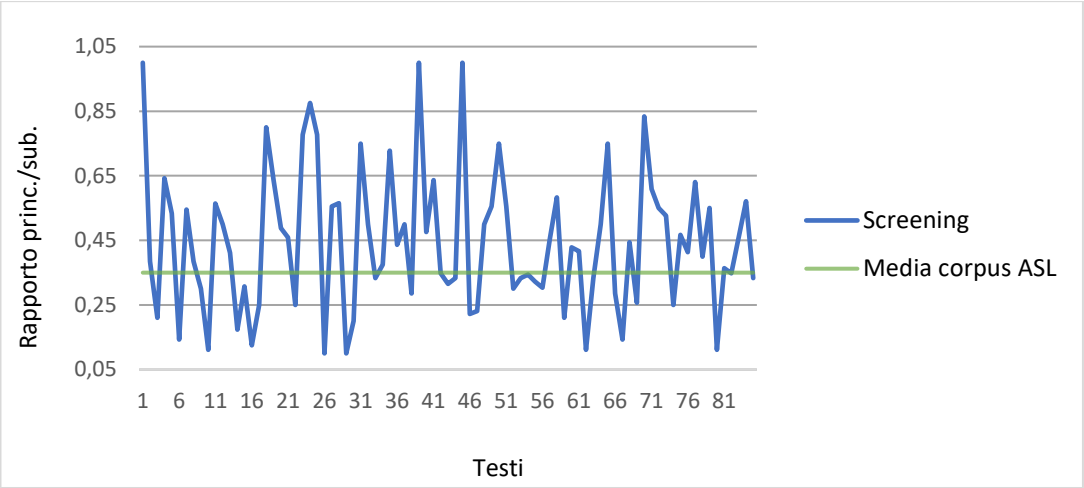


Figura 71. Proporzione tra frasi principali e subordinate nel corpus dello screening oncologico.

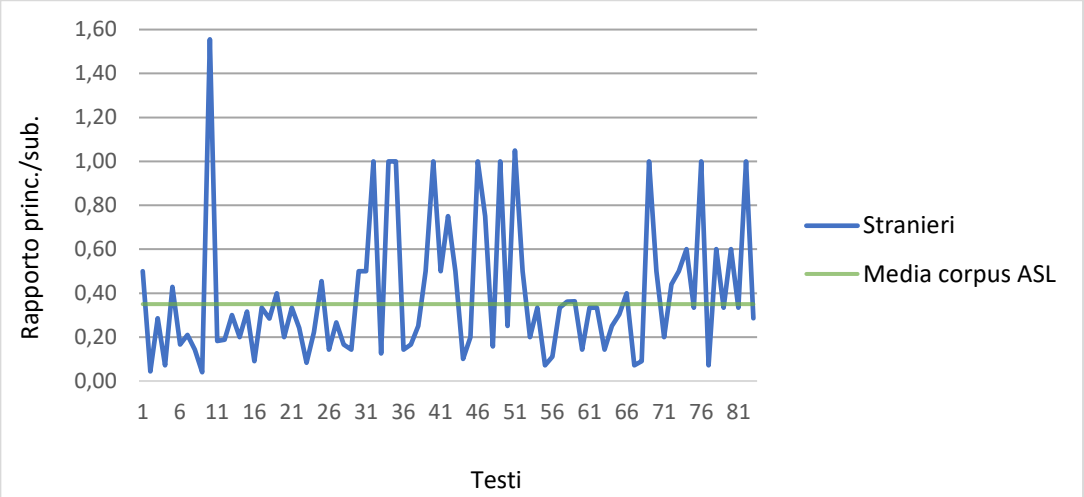


Figura 72. Proporzione tra frasi principali e subordinate nel corpus dell'assistenza sanitaria agli stranieri.

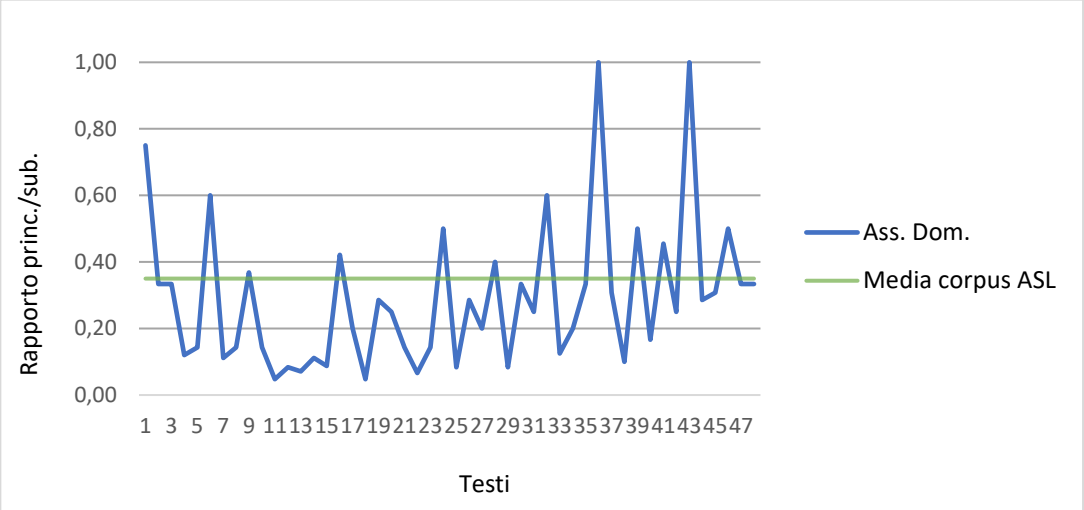


Figura 73. Proporzione tra frasi principali e subordinate nel corpus dell'assistenza domiciliare.

Dato che ciò che costituisce un fattore di complessità non è la sola presenza della subordinazione, ma la combinazione della subordinazione con altri fattori, può essere interessante affinare questi dati, integrandoli con altre informazioni, come l'ordine relativo tra principale e subordinata, il grado di incassamento della subordinata, il tipo di subordinata, ecc.
Per quanto riguarda l'ordine relativo delle subordinate rispetto alla principale (Tabella 119), si nota una prevalenza di subordinate che seguono la principale, ordine che è riconosciuto di più facile elaborazione nella letteratura linguistica. Tuttavia, considerando il singolo dato di subordinate in posizione precedente alla principale, si può osservare che si tratta di una ricorrenza piuttosto alta (la percentuale media si attesta intorno al 18% e si mantiene tale anche nei vari sotto-corpora).

Corpus	**Subordinate pre**	**Subordinate post**
Emergenza	18,10	81,90
Screening	18,11	81,89
Stranieri	18,13	81,87
Ass. Dom.	18,05	81,95
CORPUS ASL	**18,35**	**81,65**

Tabella 119. Ordine relativo delle subordinate rispetto alla principale.

L'altro aspetto rilevante è il grado di incassamento gerarchico, ovvero se la subordinata dipende direttamente dalla radice verbale (frase principale) o se rappresenta una reggenza di secondo o terzo grado; in altre parole, è interessante ricostruire quale rapporto esiste tra le subordinate, cioè se siano ricorsivamente incassate l'una dentro l'altra.

Un grado di incassamento elevato è tipico dei testi complessi. Le misure che si possono ottenere sono la lunghezza (profondità) media delle catene di subordinazione e la distribuzione delle catene di proposizioni subordinate incassate per livello di profondità (con profondità = 1,2,3).

Corpus	**Lunghezza media catene subordinanti**
Giornalismo	1,09
Materiali didattici	1,08
Letteratura	1,16
Prosa scientifica	1,03
Ling. legislativo	1,11
Ling. burocratico	0,95
Consensi informati	1,02
Emergenza	1,04
Screening	1,02

Corpus	Lunghezza media catene subordinanti
Stranieri	0,74
Ass. Dom.	0,84
CORPUS ASL	**0,90**

Tabella 120. Lunghezza media catene subordinanti nei corpora analizzati.

Corpus	Catene subordinanti 1	Catene subordinanti 2	Catene subordinanti 3
Emergenza	87,77%	10,81%	1,42%
Screening	87,77%	10,81%	1,42%
Stranieri	87,70%	10,87%	1,43%
Ass. Dom.	87,64%	10,93%	1,44%
CORPUS ASL	**87,83%**	**10,76%**	**1,41%**

Tabella 121. Distribuzione delle catene di proposizioni subordinate incassate per livello di profondità.

I risultati mostrano che la profondità media delle catene subordinanti è prossima al valore 1, per cui la maggior parte delle subordinate (87,83%) dipendono direttamente dalla radice verbale. Circa il 10% restituisce un livello di incassamento pari a 2, mentre solo l'1,41% rappresenta una reggenza di terzo grado.

Consideriamo il caso della modificazione nominale, ovvero strutture costituite da una testa nominale e da modificatori aggettivali e/o complementi preposizionali. In particolare, possiamo misurare la lunghezza (o profondità) media delle catene preposizionali e la distribuzione delle catene di dipendenza a testa nominale per livello di profondità (con profondità = 1-5).
Il dato relativo alla lunghezza media delle catene preposizionali (Tabella 122) ci mostra l'incidenza di strutture nominali complesse contraddistinte dalla presenza di modificatori (nominali, preposizionali)[243]. Come si può notare, la profondità media è pari a 1,40 (gli altri corpora vanno da una profondità minima di 1,27 a una massima di 1,84); in generale il corpus presenta catene di modificazione poco profonde (69% circa di catene con profondità 1 e 23% circa con profondità 2). Anche i sotto-corpora sono caratterizzati da un andamento analogo.

Corpus	Lunghezza media catene preposizionali
Emergenza	1,40
Screening	1,34
Stranieri	1,41

[243] L'incidenza delle frasi nominali può essere anche un parametro utile a misurare la coesione testuale.

Corpus	Lunghezza media catene preposizionali
Ass. Dom.	1,43
CORPUS ASL	**1,40**

Tabella 122. Lunghezza media delle catene preposizionali

Corpus	Catene prep. 1	Catene prep. 2	Catene prep. 3	Catene prep. 4	Catene prep. 5
Emergenza	69,01	23,44	5,99	1,35	0,16
Screening	68,99	23,42	6,01	1,36	0,17
Stranieri	68,99	23,44	6,00	1,36	0,17
Ass. Dom.	69,01	23,43	5,99	1,35	0,17
ASL	**68,97**	**23,46**	**5,99**	**1,36**	**0,16**

Tabella 123. Distribuzione delle strutture nominali complesse per livello di profondità.

Analizziamo infine le caratteristiche dei predicati verbali. I parametri considerati sono:

- arità verbale (distribuzione di teste verbali per numero di dipendenti istanziati); sono riportate le percentuali di occorrenza di verbi con valenza da 0 a 6;
- media di teste verbali per frase. Questa misura può essere raffinata considerando ulteriori parametri: la percentuale di radici verbali con soggetto esplicito, la percentuale di radici verbali e il numero di token per clausola/proposizione (verbale);
- media di archi entranti in teste verbali (numero di dipendenti per testa verbale, sia argomenti che modificatori).

Corpus	Media di teste verbali per frase	Numero token per clausola	Media di archi entranti in testa verbale
Emergenza	1,52	14,35	1,87
Screening	1,77	12,40	1,88
Stranieri	1,42	11,11	1,88
Ass. Dom.	1,47	15,48	1,88
CORPUS ASL	**1,56**	**13,37**	**1,88**

Tabella 124. Monitoraggio delle caratteristiche dei predicati verbali.

Si considera la distribuzione delle teste verbali per periodo (Tabella 124). Il corpus ASL registra una media di 1,56 teste verbali per frase (possiamo portare a confronto i dati relativi al corpus di Repubblica, che presenta una media di 2,41 e a quello di Due Parole che ottiene una media di 1,64): si tratta quindi di un corpus caratterizzato da una maggiore proporzione di periodi monoclausali. Tale dato può essere poi incrociato con la distribuzione delle diverse clausole nei testi, ovvero la proporzione di principali e subordinate, che abbiamo già analizzato (cfr. Tabella 118).

La misura dei dipendenti per testa verbale restituisce un valore relativamente basso, pari a 1,88; ciò significa che per ogni testa verbale abbiamo quasi 2 dipendenti. Possiamo raffinare ulteriormente questo dato, andando a osservare la distribuzione delle teste verbali per numero di dipendenti istanziati all'interno del testo (Tabella 125 e Figura 74); si nota che il corpus ASL ha un numero significativamente più alto di teste verbali monovalenti (34,12%) e bivalenti (30,93) e numeri più bassi di teste verbali che presentano un numero maggiore di dipendenti (argomenti o modificatori). L'andamento è piuttosto simile a quello degli altri corpora di confronto, che presentano una maggiore incidenza di predicati monoargomentali o biargomentali.

Corpus	Arità 0	Arità 1	Arità 2	Arità 3	Arità 4	Arità 5	Arità 6
Emergenza	7,95	34,13	32,66	16,33	6,50	2,10	0,27
Screening	6,17	34,02	32,91	17,47	5,90	1,95	0,34
Stranieri	8,12	34,14	28,44	18,40	7,51	1,50	0,54
Ass. Dom.	7,75	34,25	30,58	18,92	6,91	1,30	0,28
ASL	**7,36**	**34,12**	**30,93**	**17,92**	**6,71**	**1,70**	**0,39**

Tabella 125. Distribuzione delle teste verbali per numero di dipendenti istanziati.

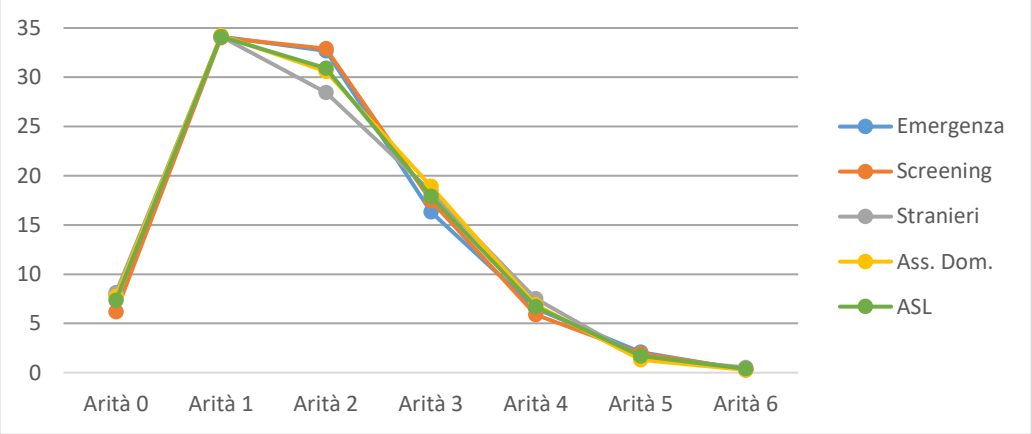


Figura 74. Grafico della distribuzione delle teste verbali per numero di dipendenti istanziati.

9.5. La leggibilità calcolata con READ-IT

Per avere una prima indicazione circa la difficoltà dei testi del corpus ASL, abbiamo effettuato l'analisi della leggibilità tramite lo strumento avanzato READ-IT.
READ-IT consente quattro diversi modelli di analisi:

- Modello base;
- Modello lessicale;
- Modello sintattico;
- Modello globale.

Abbiamo monitorato la complessità del corpus esclusivamente rispetto al livello lessicale e sintattico. Non abbiamo invece preso in considerazione il modello base e quello globale. Il modello base, infatti, fa affidamento unicamente su caratteristiche generali e formali del

testo e può essere considerato come un'approssimazione dell'indice GULPEASE; il modello globale combina i risultati degli altri tre modelli ma non sempre fornisce risultati ottimali.
Per ogni modello, la leggibilità è stimata per l'intero corpus e per ciascun testo singolo; l'analisi è poi svolta per ciascun sotto-corpus e dunque rispetto sia alle quattro tematiche che rispetto alle aziende sanitarie. L'obiettivo è capire se esiste una certa uniformità nella difficoltà dei testi in relazione al tipo di contenuto espresso oppure in relazione alla ASL di appartenenza, ovvero se i testi appartenenti a una data azienda sanitaria risultano più complessi a prescindere dal tipo di contenuto.
La Tabella 126 mostra i valori ottenuti in media dal corpus ASL nel modello lessicale e sintattico; i punteggi vanno da 0 (facile) a 1 (difficile), o in percentuale da 0 a 100.

Corpus	**Lessicale**	**Sintattico**
CORPUS ASL	**0,82**	**0,76**

Tabella 126. Valori di leggibilità ottenuti dal corpus ASL.

Come si può osservare, a livello lessicale il valore ottenuto è piuttosto alto (82%), mentre a livello sintattico può essere considerato medio-alto (76%).
Un tale punteggio elevato rispetto al modello lessicale può essere facilmente spiegato prendendo in esame i dati risultanti dal monitoraggio delle caratteristiche lessicali. Se, per esempio, consideriamo la composizione del vocabolario e la ricchezza lessicale dei testi, ci accorgiamo che sono questi parametri a influenzare negativamente la leggibilità. Il corpus ASL contiene una percentuale di lemmi appartenenti al VdB pari al 59,46%, e dunque molto bassa; è infatti probabile che molti dei termini impiegati nel corpus appartengano al dominio medico o siano comunque termini specialistici, non presenti nel vocabolario di base. Per quanto riguarda la ricchezza lessicale, abbiamo invece un alto rapporto type/token (0,79), indicatore di testi particolarmente variegati dal punto di vista lessicale e un valore elevato di densità lessicale (0,59), spia di una maggiore complessità testuale.
A livello sintattico, la percentuale di difficoltà risulta abbastanza alta ma inferiore rispetto al modello lessicale. Confrontando questo punteggio con i risultati del monitoraggio, si nota una leggera discrepanza: a livello generale, infatti, i dati che emergono dall'analisi sintattica non risultano così preoccupanti, ed anzi, nell'analisi contrastiva con altri corpora testuali, risultano spesso più positivi.
Come si può allora spiegare questo alto livello di complessità nella valutazione della leggibilità sintattica? La risposta più immediata è che READ-IT calcola la difficoltà sintattica dei testi con una stima che si basa sull'intero set di caratteristiche sintattiche e il valore che risulta non è altro che una media di tutti questi parametri. Le ragioni di un alto valore di leggibilità vanno allora cercate all'interno delle singole caratteristiche esaminate.
Un dato che emerge subito è quello relativo alla subordinazione e in particolare alla posizione delle subordinate rispetto alla principale. Il corpus ASL fa un ampio uso delle principali (74,16%), mentre ricorre meno spesso a costruzioni ipotattiche (25,84%); il rapporto tra frasi principali e subordinate risulta piuttosto basso (0,35). Anche il dato che riguarda l'ordine delle subordinate rispetto alla principale è positivo (81,65%). Sono tutti parametri legati a una minore complessità testuale; tuttavia, se andiamo a considerare il singolo valore relativo alle subordinate in posizione precedente alla principale, si nota che si

tratta di una ricorrenza piuttosto alta (18,35%). Dunque, circa il 18% delle subordinate del corpus presenta un ordine delle frasi che risulta di difficile elaborazione e genera complessità sintattica: questo è uno dei parametri che può influenzare la leggibilità sintattica dei testi.
Un'altra ipotesi potrebbe essere che, in questo caso, il modello sintattico READ-IT sovrastimi la difficoltà dei testi rispetto ai risultati derivanti dal monitoraggio, probabilmente perché è addestrato su documenti piuttosto differenti rispetto a quelli del nostro corpus.

Consideriamo adesso la valutazione della leggibilità rispetto a ciascun sotto-corpus (Tabella 127 e Figura 75).

Corpus	**Lessicale**	**Sintattico**
Emergenza	0,75	0,74
Screening	0,74	0,67
Stranieri	0,85	0,83
Ass. Dom.	0,96	0,82
CORPUS ASL	**0,82**	**0,76**

Tabella 127. Valori di leggibilità ottenuti dai sotto-corpora analizzati.

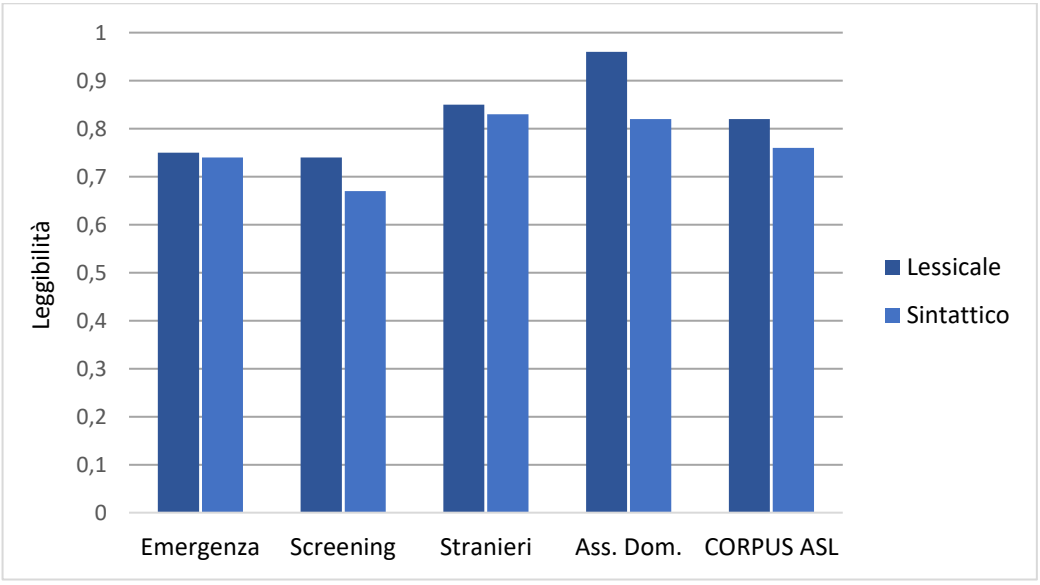


Figura 75. Confronto tra leggibilità lessicale e sintattica dei corpora analizzati.

Si osserva che in generale i valori di leggibilità risultano piuttosto alti e che in tutti i corpora la difficoltà lessicale supera quella sintattica. Trattandosi di tematiche per le quali la chiarezza e la comprensibilità dell'informazione sono estremamente essenziali, tali punteggi non sono molto incoraggianti. Ci si aspetterebbe che il corpus relativo all'emergenza sanitaria fosse quello con una maggiore leggibilità, dato che questo tipo di informazioni non deve soltanto rispondere a requisiti di chiarezza e correttezza, ma anche di immediatezza nella comprensione. Il corpus più semplice, sebbene di poco, risulta invece

quello dello screening oncologico, che registra i valori più bassi sia nel modello lessicale (74%) che in quello sintattico (67%).
Lo stesso vale per il corpus relativo all'assistenza sanitaria agli stranieri: trattandosi di testi rivolti a persone che hanno competenze in italiano solo come seconda lingua, ci si aspetterebbe una maggiore facilità nei contenuti. Anche questo gruppo di documenti presenta invece una percentuale alta di difficoltà per entrambi i livelli di analisi (così come il corpus dell'emergenza sanitaria, anche in questo caso i valori medi sono molto simili).
L'assistenza domiciliare restituisce invece tra i valori più alti in entrambi i modelli di analisi. In particolare, ciò che stupisce è la percentuale relativa al livello lessicale (96%), che sfiora quasi il valore massimo di difficoltà.

Se si confrontano i valori di leggibilità calcolati con il modello lessicale con i risultati del monitoraggio (Tabella 128), si nota che punteggi più bassi di difficoltà registrati dai corpora relativi allo screening oncologico e all'emergenza sanitaria rispecchiano una migliore situazione sul piano delle caratteristiche lessicali, con parole e frasi più brevi, una maggiore percentuale di lemmi appartenenti al Vocabolario di Base e di lessico fondamentale (FO).
Sul piano sintattico si ripresenta invece la stessa discrepanza notata già a livello generale per l'intero corpus ASL: a valori più bassi di leggibilità non corrispondono sempre punteggi maggiormente negativi delle caratteristiche sintattiche e viceversa. Ad esempio, il corpus dello screening tumorale, che ha una leggibilità sintattica pari a 67% e risulta quindi il più facile, registra valori più positivi per quanto riguarda la struttura dell'albero sintattico e la modificazione nominale ma restituisce i punteggi peggiori circa la subordinazione e i parametri relativi ai predicati verbali. Al contrario, il corpus dell'assistenza sanitaria agli stranieri, che risulta il più complesso a livello di leggibilità sintattica (83%), ha valori negativi per quanto riguarda la struttura dell'albero sintattico e la modificazione nominale ma riporta i punteggi migliori per quanto riguarda le caratteristiche della subordinazione e alcuni parametri relativi ai predicati verbali.
È interessante notare che nel monitoraggio delle caratteristiche i due corpora ottengono valori di difficoltà invertiti e dunque che entrambi presentano dei problemi dal punto di vista sintattico. Viene allora da domandarsi quale sia il fattore che comporta una tale differenza rispetto alla valutazione della leggibilità sintattica (pari a 67% nello screening e a 83% negli stranieri): potrebbe dipendere dal diverso peso che assumono le varie caratteristiche nel concorrere alla misurazione della difficoltà; oppure, come già accennato a livello generale per l'intero corpus, potrebbe dipendere dal fatto che READ-IT è addestrato su corpora molto differenti (in termini di caratteristiche linguistiche) da quello della ASL e, in particolare per quanto riguarda il piano sintattico, non riesce a raggiungere livelli elevati di accuratezza.

Caratteristiche	Emergenza	Screening	Stranieri	Ass. Dom
Base				
N. di caratteri per parola	5,38	5,56	5,86	6,14
N. di parole per frase	18,18	17,65	18,67	18,76
Lessicali				
TTR (primi 100 lemmi)	0,67	0,69	0,67	0,71
% Lemmi nel VdB	60,52	64,72	58,91	55,01
FO	74,89	72,90	68,99	70,13
Densità lessicale	0,57	0,59	0,60	0,61
Sintattiche				
Media altezze massime	5,40	**5,25**	**5,34**	5,52
Media lunghezza link	2,14	**2,21**	**2,29**	2,25
Media lunghezza link massimi	7,22	**7,12**	**7,96**	7,57
Rapporto principali e subordinate	0,35	**0,42**	**0,34**	0,26
Lunghezza media catene subordinanti	1,04	**1,02**	**0,74**	0,84
Lunghezza media catene preposizionali	1,40	**1,34**	**1,41**	1,43
Media di teste verbali per frase	1,52	**1,77**	**1,42**	1,47
Media di archi entranti in testa verbale	1,87	**1,88**	**1,88**	1,88

Tabella 128. Risultati del monitoraggio dei sotto-corpora analizzati.

Dal momento che i dati di leggibilità presi in considerazione corrispondono ai dati medi per ciascun sotto-corpus, abbiamo calcolato la deviazione standard. La deviazione standard è una misura che indica quanto i valori si discostano dalla media; in pratica ci consente di sapere se il valore medio è affidabile nel dare una rappresentazione significativa dei dati. Le figure seguenti illustrano la media e la deviazione standard per ogni sotto-corpora, sia a livello lessicale che sintattico.

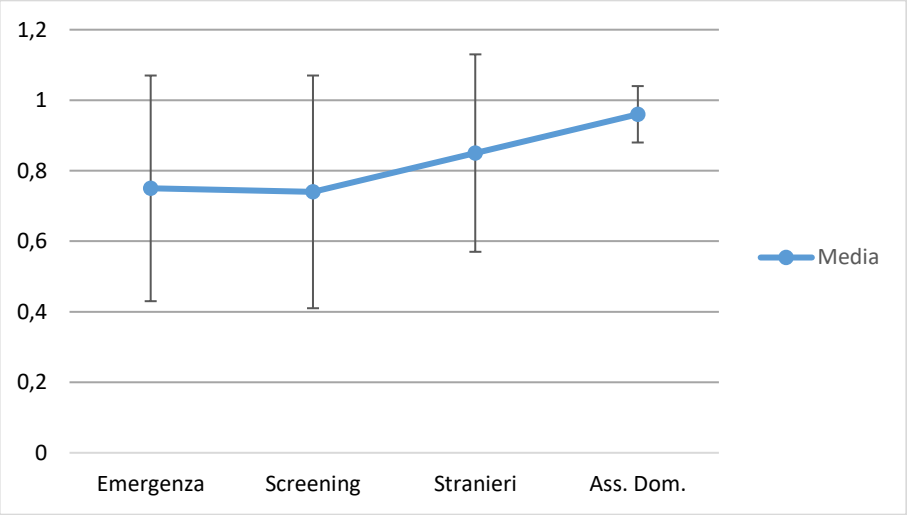


Figura 76. Media e deviazione standard per ciascun sottocorpora a livello lessicale.

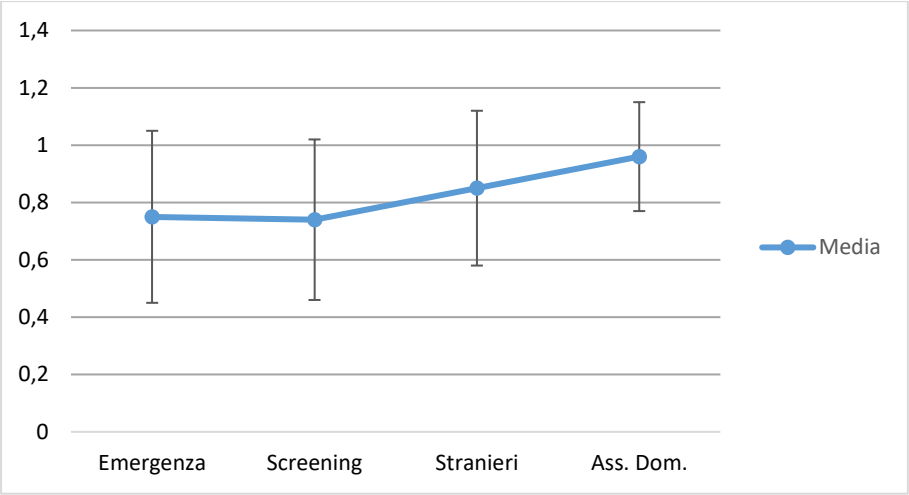
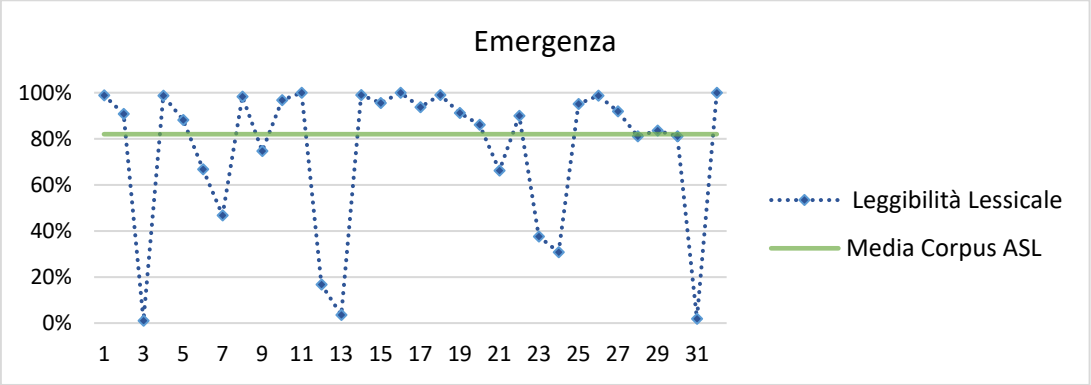


Figura 77. Media e deviazione standard per ciascun sottocorpora a livello sintattico.

Il corpus con il valore più basso di deviazione standard è quello dell'assistenza domiciliare (0,08 a livello lessicale e 0,19 a livello sintattico); gli altri corpora presentano valori simili e più alti (emergenza 0,32 e 0,30; screening 0,33 e 0,28; stranieri 0,28 e 0,27). Ciò significa che soltanto nel corpus dell'assistenza domiciliare i valori sono tutti prossimi tra loro e la media è un valore piuttosto preciso.

Per questo motivo, andiamo ad analizzare nei grafici seguenti l'andamento interno a ciascun sotto-corpora, sia per il modello lessicale che per quello sintattico. L'asse orizzontale rappresenta i vari testi che compongono il corpus; la variazione viene mostrata anche rispetto al livello medio di leggibilità di tutto il corpus ASL (linea verde).

Per quanto riguarda il corpus relativo al servizio di emergenza (Figura 78), si nota un andamento simile in entrambi modelli di analisi, con soltanto 10 testi su 32 che presentano valori di leggibilità al di sotto dell'80%. Anche all'interno di questa percentuale esiste però una notevole variazione, con valori che vanno dall'1% al 75% per il livello lessicale e dal 3% al 77% per quello sintattico.



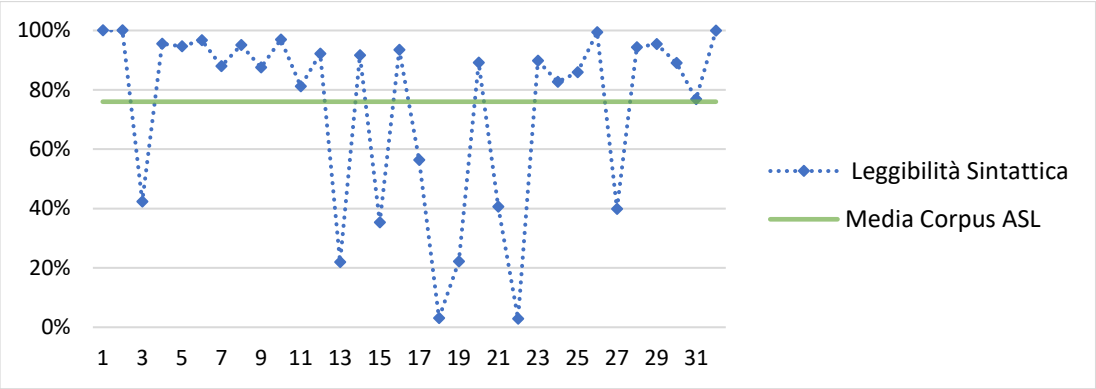


Figura 78. Andamento della leggibilità nel corpus dell'emergenza sanitaria.

Ci si aspetterebbe una certa corrispondenza tra i testi che risultano più facili in un modello con quelli che risultano più semplici nell'altro; in realtà, i testi sono molto diversificati sui due piani: la maggior parte dei documenti che riporta un basso valore a livello lessicale ha invece un valore alto o medio alto in quello sintattico (o viceversa).

Ad esempio, se consideriamo i valori dei 10 testi più semplici sul piano lessicale, vediamo che la quasi totalità registra valori intorno all'80% a livello sintattico; viceversa, prendendo a riferimento i valori dei 10 documenti più facili sul piano sintattico, notiamo che la maggior parte presenta valori massimi di leggibilità lessicale. I testi 1, 2, 3 e 8 (in azzurro) sono gli unici presenti in entrambe le tabelle: si osserva però che, ad eccezione del documento 3, che registra valori piuttosto bassi in entrambi i modelli (4% e 22%) e del documento 8, che riporta valori medi nelle due categorie (41% e 66%), gli altri testi hanno punteggi molto differenti tra loro (1-42%, 2-77%).

Una maggiore corrispondenza tra i testi si ha soltanto nel caso di valori molto alti in entrambi i modelli.

Testo	Lessicale	Sintattico		Testo	Sintattico	Lessicale
testo 1	**1%**	**42%**		testo 16	3%	90%
testo 2	**2%**	**77%**		testo 29	3%	99%
testo 3	**4%**	**22%**		**testo 3**	**22%**	**4%**
testo 4	17%	92%		testo 18	22%	91%
testo 5	31%	83%		testo 22	35%	96%
testo 6	38%	90%		testo 19	40%	92%
testo 7	47%	88%		**testo 8**	**41%**	**66%**
testo 8	66%	41%		**testo 1**	**42%**	**1%**
testo 9	67%	97%		testo 20	56%	94%
testo 10	75%	88%		**testo 2**	**77%**	**2%**

Tabella 129. Valori di leggibilità dei 10 testi più semplici a livello lessicale e sintattico del corpus dell'emergenza sanitaria.

La Figura 79 mostra l'andamento interno del corpus relativo allo screening oncologico (il corpus presenta un numero maggiore di testi rispetto a quello dell'emergenza sanitaria, così come i corpora successivi). Come si può subito notare, esiste una forte variazione tra i testi: a livello lessicale, tale alterazione è maggiormente concentrata sui documenti che presentano valori più bassi, mentre sul piano sintattico, interessa anche i punteggi più alti.

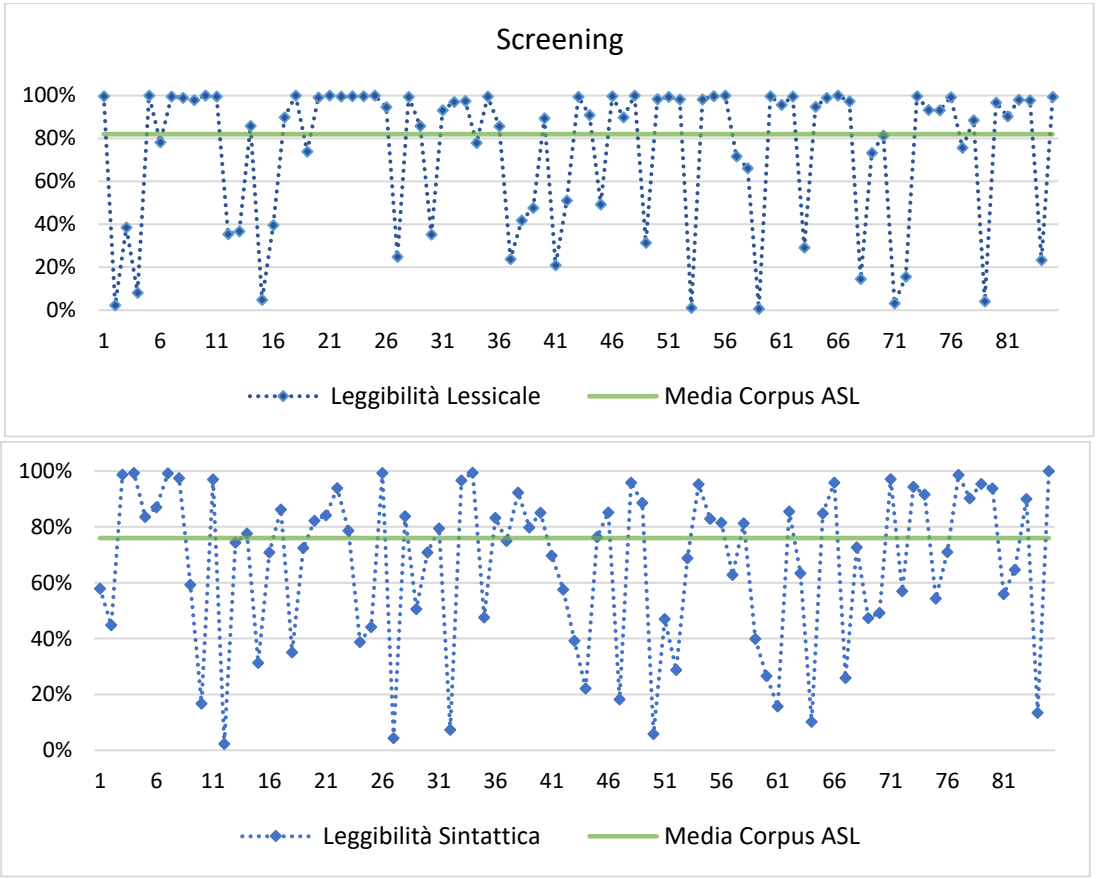


Figura 79. Andamento della leggibilità nel corpus dello screening oncologico.

Per quanto riguarda il corpus relativo all'assistenza sanitaria agli stranieri (Figura 80), si registra un andamento simile in entrambi modelli di analisi, con soltanto 20 testi circa su 83 che presentano valori di leggibilità al di sotto delle medie del corpus (rispettivamente 82% per la leggibilità lessicale e 76% per quella sintattica).

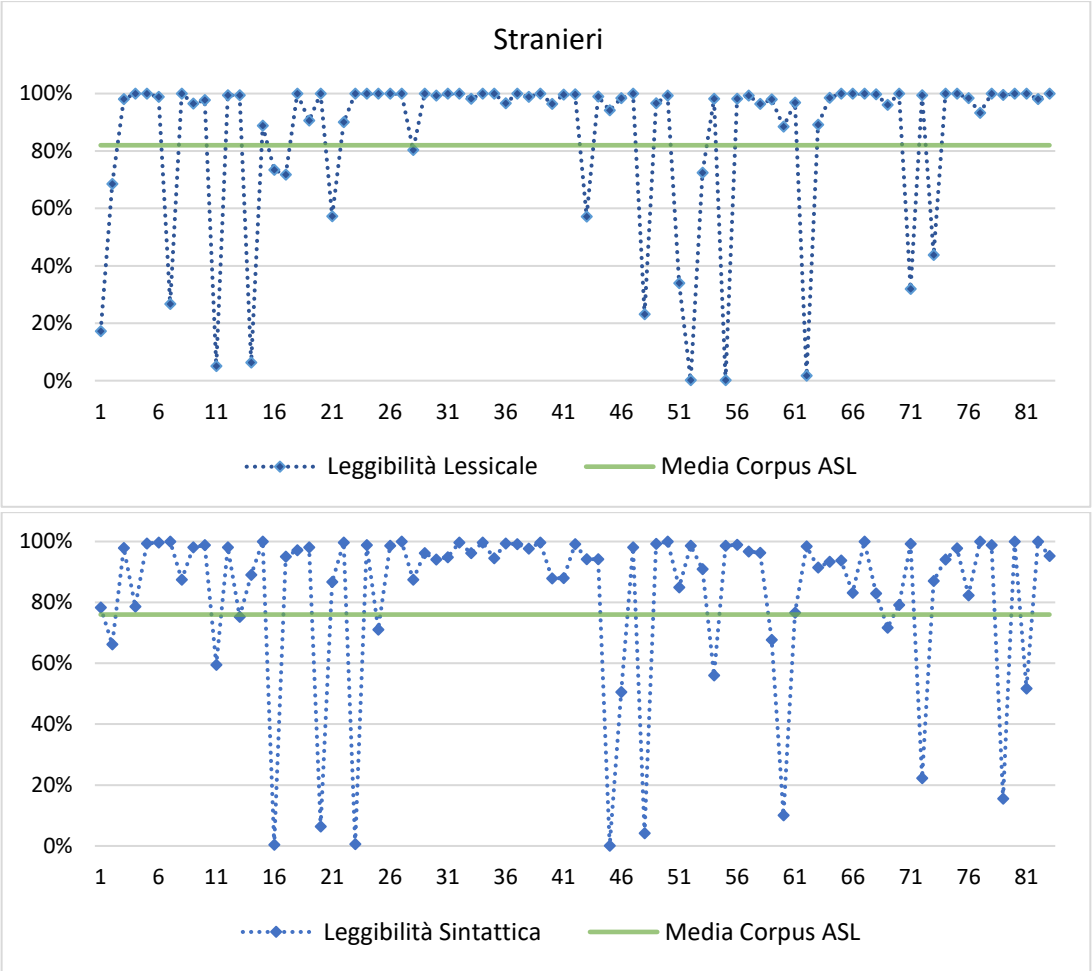


Figura 80. Andamento della leggibilità nel corpus dell'assistenza sanitaria agli stranieri.

Si riscontra anche in questo corpus una notevole variazione interna ai modelli e una mancata corrispondenza tra testi facili nel confronto tra i due livelli di analisi.

Ad esempio, considerando i valori dei 15 testi più semplici sul piano lessicale, vediamo che la quasi totalità registra valori maggiori dell'80% a livello sintattico; viceversa, prendendo a riferimento i valori dei 15 documenti più facili sul piano sintattico, notiamo che la maggior parte si avvicina ai valori massimi di leggibilità lessicale. Gli unici testi presenti in entrambe le liste sono i numeri 4, 7 e 14: il documento 4 registra valori piuttosto bassi in entrambi i modelli (23% e 4%), il documento 14 valori medi (69% e 67%); il testo 7 presenta invece punteggi molto diversi tra loro (5% e 59%).

In questo caso, a differenza del corpus dell'emergenza sanitaria, non vi è corrispondenza tra i testi neanche nel caso di valori elevati di leggibilità.

Testo	Lessicale	Sintattico		Testo	Sintattico	Lessicale
testo 1	0%	99%		testo 25	0%	94%
testo 2	0%	99%		testo 17	0%	73%
testo 3	2%	98%		testo 80	1%	100%
testo 4	**5%**	**59%**		**testo 7**	**4%**	**23%**

Testo	Lessicale	Sintattico		Testo	Sintattico	Lessicale
testo 5	6%	89%		testo 75	6%	100%
testo 6	17%	78%		testo 19	10%	88%
testo 7	**23%**	**4%**		testo 52	16%	99%
testo 8	27%	100%		testo 51	22%	99%
testo 9	32%	99%		testo 40	51%	98%
testo 10	34%	85%		testo 81	52%	100%
testo 11	44%	87%		testo 37	56%	98%
testo 12	57%	94%		**testo 4**	**59%**	**5%**
testo 13	57%	87%		**testo 14**	**66%**	**69%**
testo 14	**69%**	**66%**		testo 34	68%	98%
testo 15	72%	95%		testo 82	71%	100%

Tabella 130. Valori di leggibilità dei 15 testi più semplici a livello lessicale e sintattico del corpus degli stranieri.

Il corpus dell'assistenza domiciliare presenta tra i valori più alti in entrambi i modelli di analisi, come risulta evidente dalla Figura 81. Sul piano delle leggibilità lessicale, si nota che la quasi totalità dei testi (45 su 48) registra una percentuale di difficoltà al di sopra dell'80% e che i tre documenti più "semplici" hanno comunque valori che partono dal 60% circa (59%, 67% e 79%). La variazione interna è ovviamente quasi del tutto assente, con un dato medio pari a 96%. A livello sintattico si osserva una maggiore variazione, anche se i punteggi di leggibilità risultano in generale tutti piuttosto alti (soltanto 15 testi su 48 sono al di sotto dell'80%).

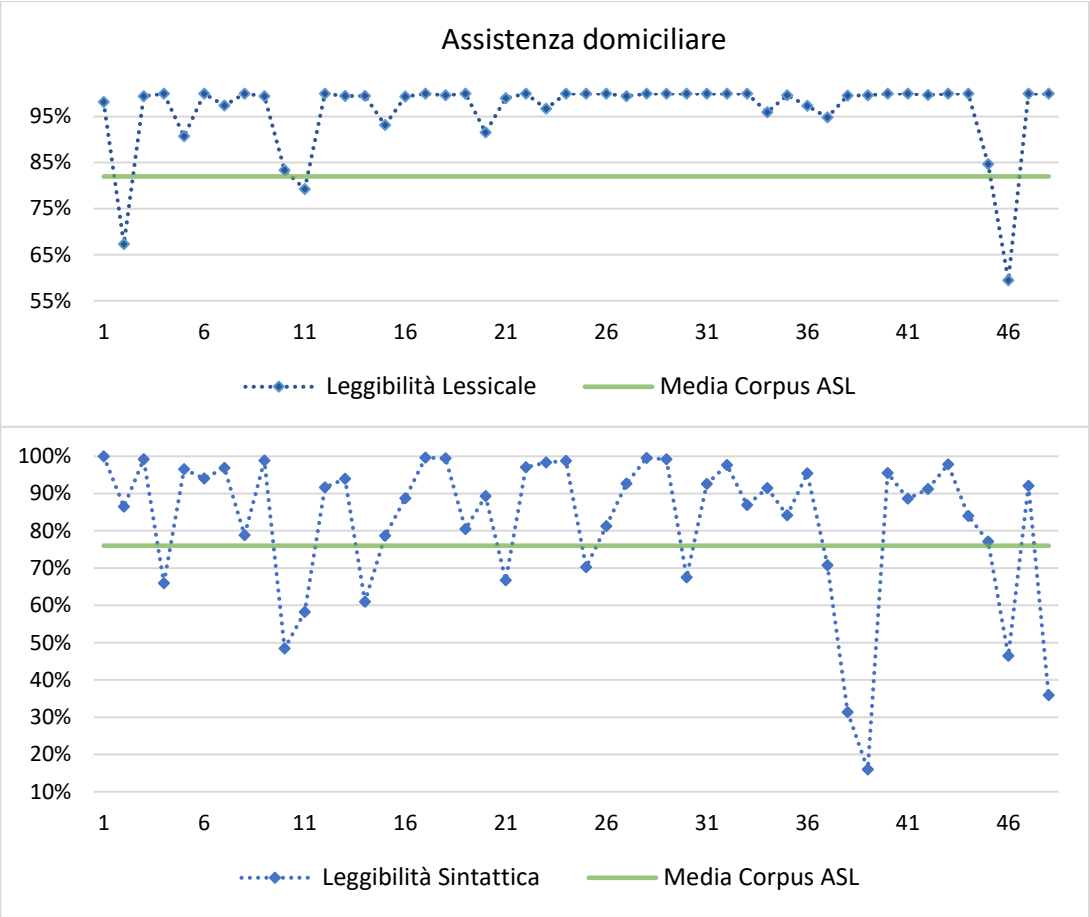


Figura 81. Andamento della leggibilità nel corpus dell'assistenza domiciliare.

In questo caso, i testi che risultano più facili a livello lessicale sono anche quelli che registrano tra i valori più bassi nel modello sintattico: si tratta però di punteggi molto alti in entrambi i livelli, per cui si può considerare una corrispondenza poco indicativa.

Testo	Lessicale	Sintattico		Testo	Sintattico	Lessicale
testo 1	**59%**	**46%**		testo 23	16%	100%
testo 2	67%	86%		testo 22	31%	100%
testo 3	**79%**	**58%**		testo 32	36%	100%
testo 4	**83%**	**48%**		**testo 1**	**46%**	**59%**
testo 5	85%	77%		**testo 4**	**48%**	**83%**
testo 6	91%	97%		**testo 3**	**58%**	**79%**
testo 7	92%	89%		testo 21	61%	100%
testo 8	93%	79%		testo 37	66%	100%
testo 9	95%	71%		testo 15	67%	99%
testo 10	96%	91%		testo 42	68%	100%

Tabella 131. Valori di leggibilità dei 10 testi più semplici a livello lessicale e sintattico del corpus dell'assistenza domiciliare.

Consideriamo adesso i valori di leggibilità rispetto a ciascuna azienda sanitaria. La Tabella 132 mostra i punteggi medi ottenuti da ciascuna ASL nel modello lessicale e in quello sintattico.

Azienda Sanitaria	Lessicale	Sintattico
AUSL di Bologna	0,91	0,93
AUSL di Parma	0,61	0,76
ASUITS di Trieste	0,83	0,86
ASL 3 Genovese	0,97	0,66
ATS di Milano	0,44	0,55
ATS Bergamo	0,94	0,96
ASL TO1	0,53	0,71
ASL TO2	0,84	0,59
ASL CN1	0,97	0,95
APSS Provincia Autonoma di Trento	0,93	0,76
Azienda sanitaria dell'Alto Adige	0,97	0,79
Azienda USL Valle d'Aosta	0,74	0,62
ULSS3 Serenissima - Venezia	0,87	0,71
ASL Roma 1	0,88	0,78
ASUR Area Vasta 2	0,76	0,83
USL Toscana Centro	0,82	0,73
USL Toscana Sud Est	0,68	0,73
USL Umbria 1	0,72	0,83
ASL 1 Abruzzo - Aquila	0,48	0,62
ASL 3 Pescara	0,93	0,92
ASL 2 Abruzzo - Chieti	0,91	0,64
ASP di Potenza	0,82	0,91
ASP Catanzaro	0,96	0,66
ASL Napoli 1 Centro	0,82	0,57
ASREM Campobasso	0,68	0,71
ASL Bari	0,99	0,89
ASL Brindisi	0,92	0,81
ASSL Cagliari	0,72	0,69
ASP Palermo	0,85	0,87
ASP Agrigento	0,97	0,88

Tabella 132. Valori di leggibilità lessicale e sintattica per ciascuna azienda sanitaria.

Si può notare che a livello lessicale i punteggi vanno da un minimo di difficoltà pari al 44% fino al 99%, mentre a livello sintattico si parte da una soglia più alta (55%) per arrivare a un valore massimo di 96%. Rispetto a una distinzione dei testi relativa al contenuto, in cui si registra una difficoltà lessicale sempre maggiore rispetto a quella sintattica, una differenziazione rispetto alla provenienza dei documenti restituisce una situazione di sostanziale equivalenza: in circa metà dei casi i punteggi più alti riguardano il piano lessicale e nell'altra metà il piano sintattico.

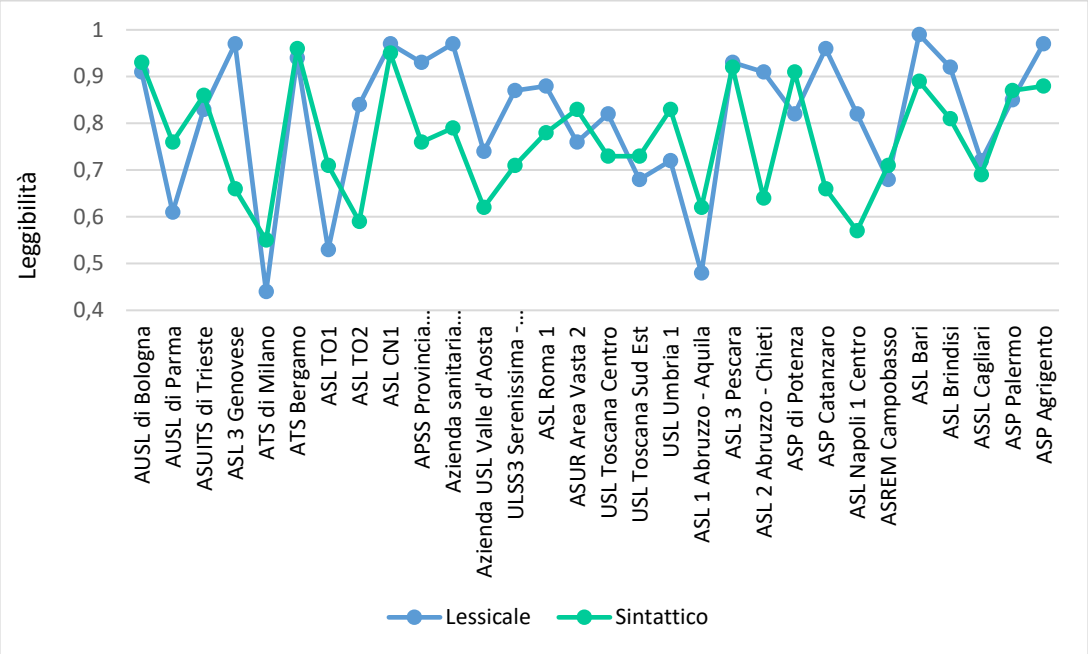


Figura 82. Andamento della leggibilità lessicale e sintattica per azienda sanitaria.

Premesso che i corpora delle varie aziende sanitarie sono costituiti da un numero molto diverso di testi (ad esempio l'ASL TO1 e l'ASL 1 Abruzzo presentano rispettivamente soltanto 3 e 2 documenti, mentre l'AUSL di Parma, l'ASUITS di Trieste e l'ULSS3 Serenissima di Venezia ne hanno ben 15) e che necessariamente ciò influisce sul diverso peso che possono avere, ai fini dell'analisi proviamo a considerarli tutti sullo stesso livello.
Le Figure 83 e 84 mostrano l'andamento del modello lessicale e di quello sintattico rispetto alle varie ASL e in relazione al livello medio di leggibilità di tutto il corpus ASL (linea verde).

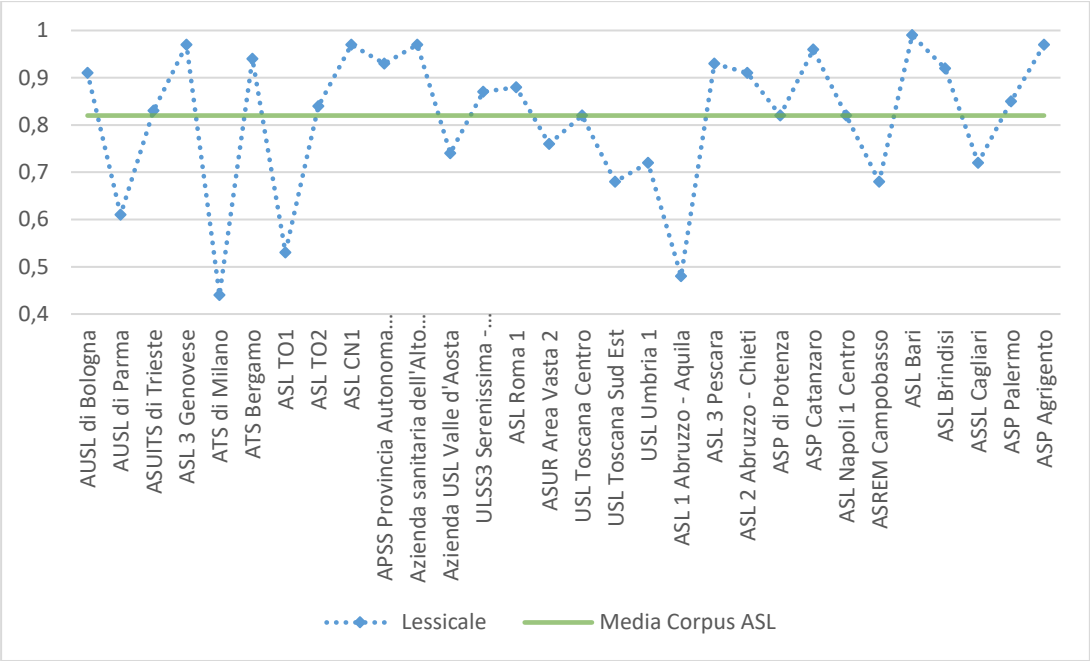


Figura 83. Andamento della leggibilità lessicale per azienda sanitaria.

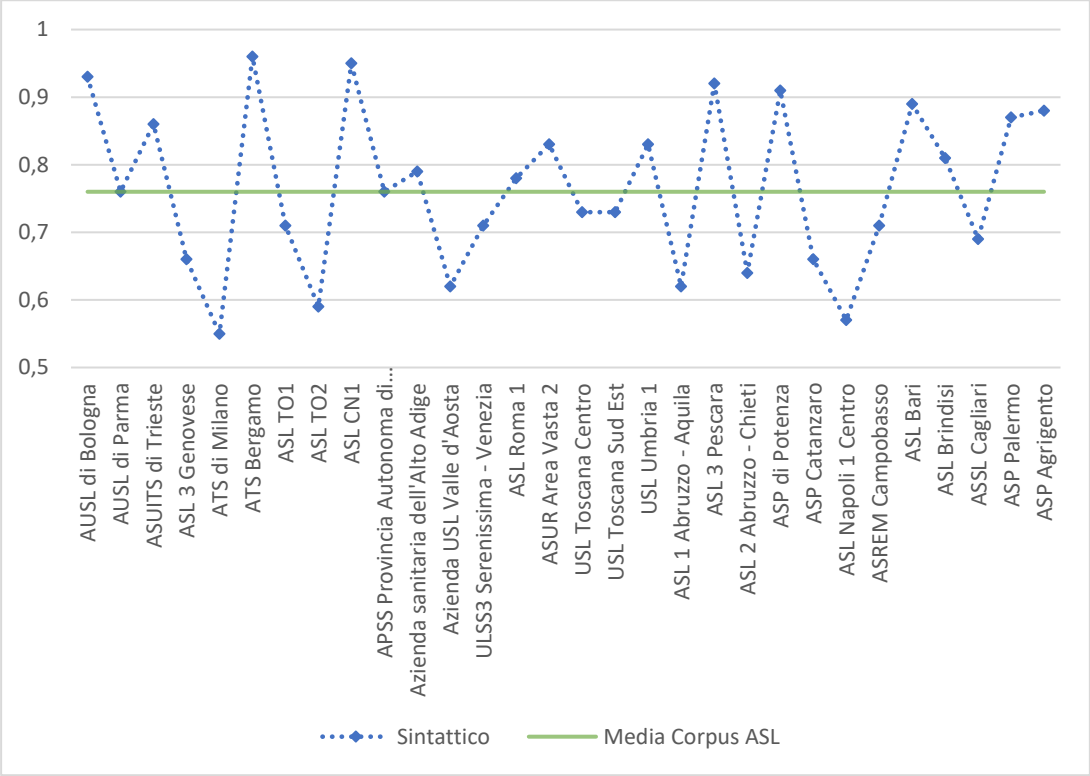


Figura 84. Andamento della leggibilità sintattica per azienda sanitaria.

Come si può osservare, l'ATS di Milano registra i valori più bassi in entrambi i livelli di analisi (44% e 55%), mentre è l'azienda sanitaria di Cuneo (ASL CN1) a riportare due tra i valori più alti nelle due categorie (97% e 95%). Si nota anche una certa corrispondenza tra le aziende che riportano testi più semplici (o difficili) in un livello e quelli che risultano più facili (o complessi) nell'altro, come illustrato nelle tabelle seguenti.

Casi eccezionali (in viola) sono rappresentati dall'ASUR area vasta 2, che occupa il decimo posto sia tra le 10 aziende con testi più semplici a livello lessicale che tra le 10 aziende con testi maggiormente complessi a livello sintattico e dall'ASP di Catanzaro e ASL3 Genovese che, al contrario, rientrano tra le aziende con punteggi più bassi nel modello sintattico (66%) e valori più alti nel modello lessicale (96% e 97%).

ASL	Lessicale	Sintattico		ASL	Sintattico	Lessicale
ATS di Milano	44%	55%		ATS di Milano	55%	44%
ASL 1 Abruzzo	48%	62%		ASL Napoli 1 Centro	57%	82%
ASL TO1	53%	71%		ASL TO2	59%	84%
AUSL di Parma	61%	76%		ASL 1 Abruzzo	62%	48%
ASREM Campobasso	68%	71%		Azienda USL Valle d'Aosta	62%	74%
USL Toscana Sud Est	68%	73%		ASL 2 Abruzzo	64%	91%
ASSL Cagliari	72%	69%		ASP Catanzaro	66%	96%
USL Umbria 1	72%	83%		ASL 3 Genovese	66%	97%
Azienda USL Valle d'Aosta	74%	62%		ASSL Cagliari	69%	72%
ASUR Area Vasta 2	76%	83%		ASL TO1	71%	53%

Tabella 133. Le 10 ASL con i valori più bassi di leggibilità a livello lessicale e sintattico.

ASL	Lessicale	Sintattico		ASL	Sintattico	Lessicale
ASL Bari	99%	89%		ATS Bergamo	96%	94%
ASL CN1	97%	95%		ASL CN1	95%	97%
ASP Agrigento	97%	88%		AUSL di Bologna	93%	91%
Azienda sanitaria dell'Alto Adige	97%	79%		ASL 3 Pescara	92%	93%
ASL 3 Genovese	97%	66%		ASP di Potenza	91%	82%
ASP Catanzaro	96%	66%		ASL Bari	89%	99%
ATS Bergamo	94%	96%		ASP Agrigento	88%	97%
ASL 3 Pescara	93%	92%		ASP Palermo	87%	85%
APSS Provincia Autonoma di Trento	93%	76%		ASUITS di Trieste	86%	83%
ASL Brindisi	92%	81%		ASUR Area Vasta 2	83%	76%

Tabella 134. Le 10 ASL con i valori più alti di leggibilità a livello lessicale e sintattico.

Ci sembra interessante confrontare i dati relativi a quelli che sono risultati essere i due poli di leggibilità, ovvero l'ATS di Milano come esempio di testi più semplice e l'ASL di Cuneo

come esempio di testi maggiormente complessi, soprattutto in riferimento alle caratteristiche linguistiche monitorate nei documenti. Va precisato che il sotto-corpus dell'ATS è costituito da 6 documenti (4 relativi allo screening, 1 all'assistenza domiciliare e 1 agli stranieri; mancano i testi relativi all'emergenza sanitaria), mentre quello di Cuneo è formato da 12 elementi (1 testo relativo allo screening, 6 all'assistenza domiciliare, 4 agli stranieri e 1 all'emergenza sanitaria).
Dalla Figura 85 risulta evidente la differenza nei valori di leggibilità: l'ATS di Milano riporta valori medi piuttosto bassi di difficoltà, ben al di sotto della media generale, con un punteggio più alto nel modello sintattico (55%) rispetto a quello lessicale (44%); l'ASL CN1 registra invece punteggi simili nei due livelli di analisi, entrambi sopra la media e molto vicini ai livelli massimi di complessità (97% e 95%).

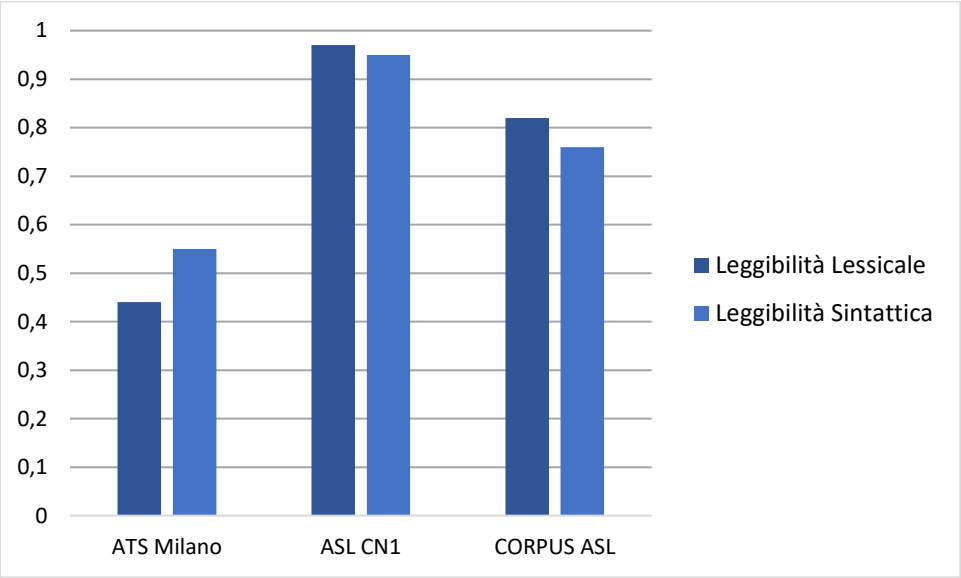


Figura 85. Confronto tra i valori di leggibilià dell'ATS di Milano e dell'ASL CN1.

Consideriamo anche l'andamento interno della leggibilità, sia sul piano lessicale (Figura 86), che su quello sintattico (Figura 87).
Rispetto al modello lessicale, risulta subito evidente l'omogeneità dei testi dell'ASL di Cuneo, che risultano tutti molto difficili e superiori alla media dell'intero corpus (82%); una maggiore variazione interna si ha invece nei documenti dell'ASL di Milano, con valori che in generale tendono verso il basso e dunque verso una maggiore facilità.

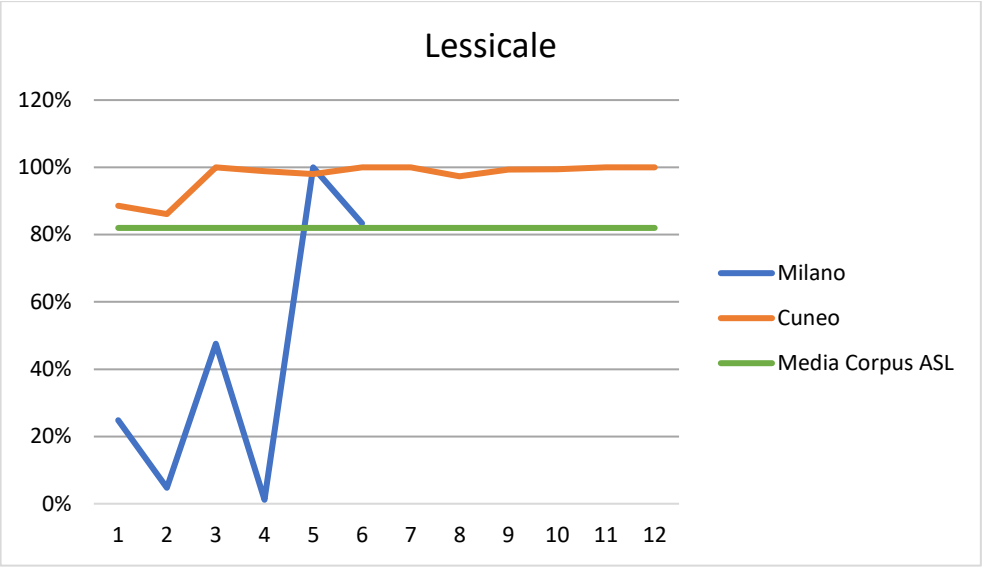


Figura 86. Andamento della leggibilità lessicale nelle ASL analizzate.

La situazione rimane pressoché invariata anche a livello sintattico: i testi dell'ASL di Cuneo risultano tutti molto difficili e superiori alla media dell'intero corpus (76%), anche se si nota una maggiore variazione rispetto a modello lessicale. I documenti dell'ATS di Milano risultano di nuovo piuttosto vari, anche se stavolta un numero più elevato tende maggiormente verso il dato medio e dunque verso una difficoltà più alta.

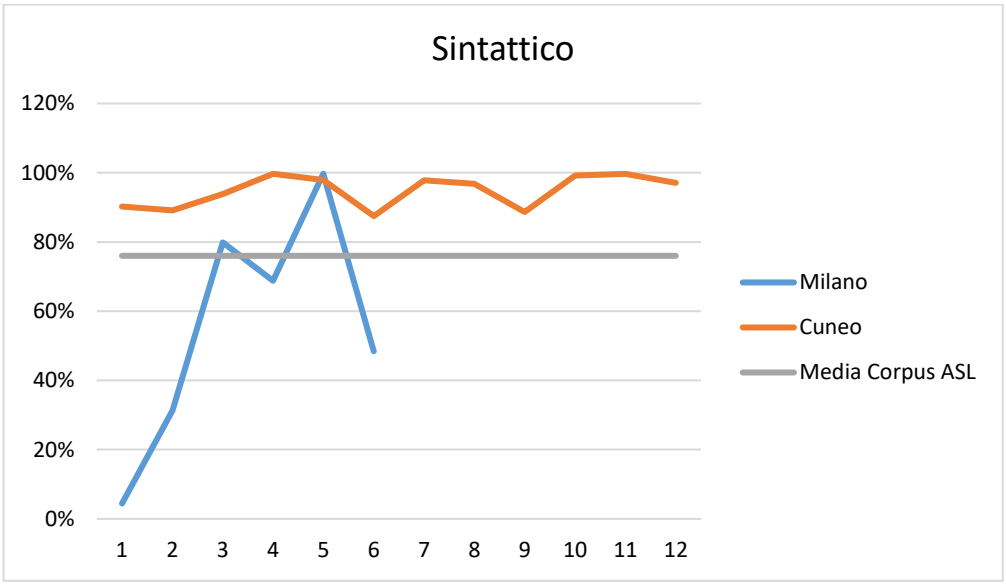


Figura 87. Andamento della leggibilità sintattica nelle ASL analizzate.

Se incrociamo questi valori con i dati medi emersi dal monitoraggio delle caratteristiche linguistiche (Tabella 135), otteniamo però una situazione diversa da quella attesa.

Caratteristiche	ATS Milano	ASL CN1	CORPUS ASL
N. di caratteri per parola	5,64	5,81	5,71
N. di parole per frase	15,56	18,51	18,39
TTR (primi 100 lemmi)	0,66	0,73	0,68
% Lemmi nel VdB	58,43	59,45	59,46
FO	69,73	71,01	71,74
Densità lessicale	0,59	0,59	0,59
Media altezze massime	5,48	5,68	5,39
Media lunghezza link	2,01	2,26	2,22
Media lunghezza link massimi	6,02	7,19	7,53
Rapporto principali e subordinate	0,26	0,29	0,35
Lunghezza media catene subordinanti	1,18	0,93	0,90
Lunghezza media catene preposizionali	1,48	1,35	1,40
Media di teste verbali per frase	1,53	1,44	1,56
Media di archi entranti in testa verbale	1,92	1,73	1,88
Leggibilità Lessicale	0,44	0,97	**0,82**
Leggibilità Sintattica	0,55	0,95	**0,76**

Tabella 135. Risultati del monitoraggio linguistico per le due ASL considerate.

Prendiamo ad esempio l'ASL di Milano: ci si aspetterebbero valori molto positivi in entrambi i livelli, soprattutto nel confronto con il corpus dell'ASL di Cuneo. Notiamo invece che spesso i punteggi risultano peggiori non solo rispetto al dato medio, ma anche rispetto ai testi di Cuneo. Viceversa, con livelli di leggibilità che si attestano intorno ai valori massimi, ci attenderemo punteggi molto più alti della media nel corpus di CN1: invece spesso tali valori risultano piuttosto positivi, sia rispetto al dato medio che addirittura ai testi di Milano.

Consideriamo ad esempio le caratteristiche lessicali che riguardano la composizione del vocabolario (Figura 88), come la percentuale di lemmi appartenenti al *Vocabolario di Base* e al lessico fondamentale (FO). Data la diversità dei punteggi di leggibilità ottenuti dalle aziende sanitarie nel modello lessicale (rispettivamente 44% per Milano e 97% per Cuneo) dovrebbe esserci molto distacco nei dati relativi a tali parametri lessicali; come si può osservare nel grafico, invece, non solo i valori risultano molto simili, ma addirittura Cuneo restituisce punteggi più positivi, con percentuali più alte in entrambe le categorie (59,45 vs 58,43 per i lemmi nel VdB e 71,01 vs 69,73 per FO).

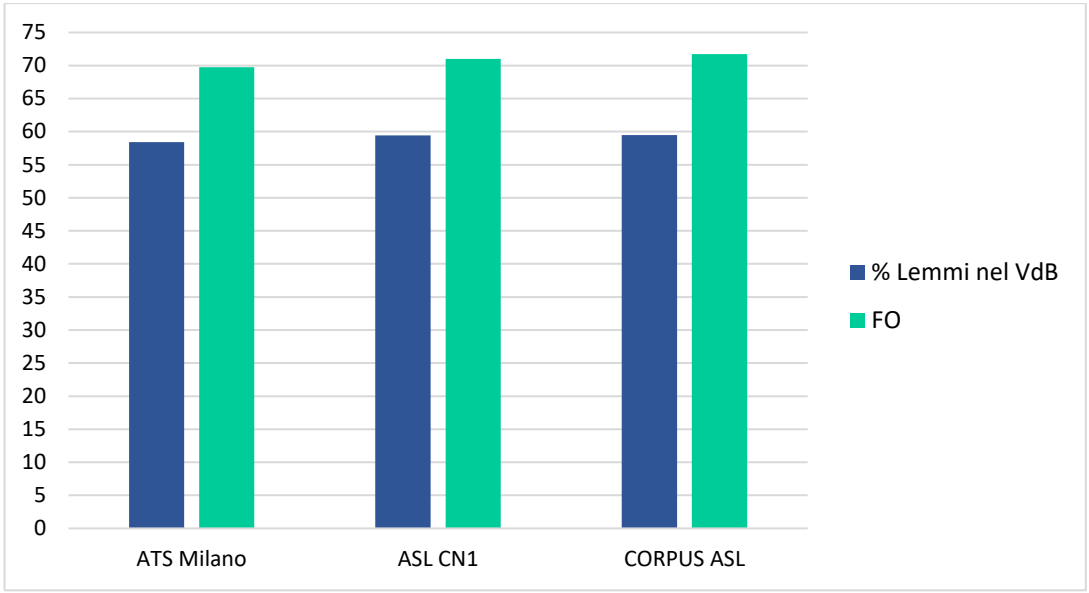


Figura 88. Composizione del vocabolario nei corpora analizzati.

Anche i dati relativi alle altre variabili lessicali non riescono a spiegare questo diverso punteggio di leggibilità: ad esempio, prendendo in esame i parametri che influiscono sulla ricchezza lessicale di un testo, si osserva che, nonostante il rapporto type/token risulti più basso nel corpus di Milano (0,66 vs 0,73), la densità lessicale risulta la stessa in entrambi i corpora (0,59).

La stessa problematica si nota nel livello sintattico. Infatti, come mostrato chiaramente nella Figura 89, non esiste una sostanziale differenza tra i valori assunti dalla maggior parte delle caratteristiche sintattiche nei due diversi corpora, tale da giustificare una così diversa valutazione della leggibilità (55% di difficoltà per l'ATS di Milano vs il 95% dell'ASL CN1). Anzi, se in circa metà dei fattori i testi di Cuneo restituiscono valori peggiori (media delle altezze massime, media della lunghezza dei link, media della lunghezza dei link massimi, rapporto principali/subordinate), nell'altra metà registrano risultati maggiormente positivi (lunghezza media delle catene subordinanti, lunghezza media delle catene preposizionali, media di teste verbali per frase e media di archi entranti in testa verbale).

Dunque, ancora una volta, si osserva una mancata corrispondenza tra i valori di leggibilità calcolati con i modelli READ-IT e i dati emersi dal monitoraggio linguistico.

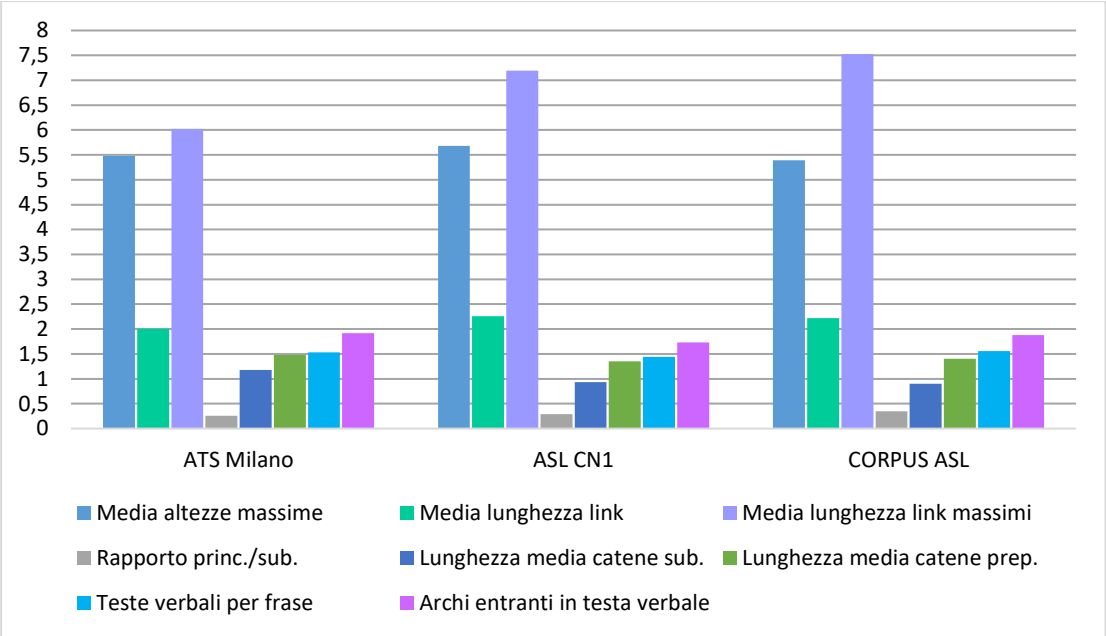


Figura 89. Confronto tral e carattertistiche sintattiche dei corpora considerati.

Facciamo un ulteriore prova andando a confrontare il profilo linguistico del testo che risulta più semplice, per ciascun modello, con quello del testo che risulta più complesso. In questo caso, è preferibile tenere distinti i due livelli di valutazione: non è detto infatti (e del resto lo abbiamo già dimostrato) che un documento che risulta molto semplice a livello lessicale, presenti la stessa facilità a livello sintattico, e viceversa.

Se la nostra ipotesi è corretta, le caratteristiche linguistiche dei due documenti non presenteranno valori tali da giustificare un livello minimo e massimo di leggibilità lessicale e sintattica; in particolare ci aspettiamo che la difficoltà del testo più complesso sia sovrastimata, ovvero che tale testo registri per le varie caratteristiche dei punteggi non affatto superiori (o comunque non negativi) rispetto alla media dell'intero corpus e addirittura anche rispetto al documento considerato semplice.

Il testo che restituisce il valore più basso di leggibilità lessicale (0%) è il numero 143, che appartiene al sotto-corpus dell'assistenza sanitaria agli stranieri e a quello dell'AUSL Umbria1. A titolo indicativo, forniamo anche il valore della leggibilità sintattica, che risulta pari a 99%, dunque all'opposto rispetto a quella lessicale. Riportiamo il testo nella sua versione originale (senza la correzione di eventuali refusi):

Testo n. 143

Stranieri in Italia - Assistenza sanitaria

Assistenza sanitaria agli stranieri in Italia. Iscrizione al SSN obbligatoria o volontaria...

Descrizione del servizio

L'assistenza sanitaria agli stranieri in Italia può essere obbligatoria o volontaria.
La normativa definisce i casi in cui l'iscrizione al Servizio Sanitario Nazionale (SSN) è obbligatoria o volontaria.
Il cittadino straniero iscritto al SSN ha diritto alle stesse prestazioni garantite ai cittadini italiani e partecipa alla spesa sanitaria (ticket) nella misura prevista.

1. L'iscrizione al SSN è OBBLIGATORIA per

- CITTADINI DELL'UNIONE EUROPEA nei casi previsti dalla normativa vigente (Decreto legislativo 3 febbraio 2007, n. 30): che abbiano in corso regolari attività di lavoro subordinato o autonomo, siano iscritti nelle liste di collocamento o in formazione e loro familiari; pensionati; studenti.
- CITTADINI STRANIERI NON APPARTENENTI ALL'UNIONE EUROPEA REGOLARMENTE SOGGIORNANTI: che abbiano in corso regolari attività di lavoro subordinato o autonomo o siano iscritti nelle liste di collocamento; motivi di famiglia e ricongiungimento familiare, asilo politico e umanitario, protezione sussidiaria; acquisto della cittadinanza; attesa adozione e affidamento; motivi di culto con regolare attività lavorativa; gravidanza.

Ai cittadini stranieri non appartenenti all'Unione Europea non è consentita l'iscrizione nel caso siano soggiornanti per periodi inferiori a 3 mesi (visto di ingresso per turismo, visita, affari), o siano titolari di permesso di soggiorno per cure mediche, o ultrasessantacinquenni titolari di permesso di soggiorno per ricongiungimento familiare.

2. L'iscrizione VOLONTARIA al SSN può essere richiesta dagli

- STRANIERI CON PERMESSO DI SOGGIORNO di durata superiore a 3 mesi che non rientrino tra coloro che sono di diritto iscritti al SSN, previo versamento di un contributo. Rientrano in questa fattispecie, per esempio, i motivi di studio, persone collocate alla pari. L'iscrizione volontaria non è consentita per cure mediche, turismo, visite, affari. L'iscrizione è valida per l'anno solare e si estende ai familiari a carico.

DOVE RIVOLGERSI

Presso gli Sportelli Anagrafe Assistibili del Comune di residenza o di domicilio della USL Umbria 1.

COSA DEVE SAPERE LO STRANIERIO CHE ARRIVA IN UMBRIA

Cittadino di un paese della Unione Europea

- in possesso della Tessera Europea di Assicurazione Malattia (TEAM) (EHIC European Health Insurance Card) rilasciata dal tuo paese di residenza
- senza la Tessera Europea di Assicurazione Malattia (TEAM) (EHIC European Health Insurance Card)

Cittadino di un paese NON appartenente all'Unione Europea

- in possesso del permesso di soggiorno
- senza il permesso di soggiorno
- cittadino di Stati con cui l'Italia ha Accordi Bilaterali

Per informazioni nella propria lingua accedere al Sito HFM Health For Migrants elaborato dall'Università degli Studi di Perugia con la Regione Umbria.

Altri link utili: Ministero della Salute

Il testo che restituisce invece il valore più alto di leggibilità lessicale (100%) è il numero 91, che appartiene al sotto-corpus dello screening oncologico e a quello dell'ASUI di Trieste (in questo caso, la leggibilità sintattica è pari all'84%, quindi risulta alta in entrambi i livelli). Il testo originale (senza la correzione di eventuali refusi) è il seguente:

Testo n. 91
Screening regionali per la prevenzione dei tumori

sommario

Pap – test (prevenzione e diagnosi precoce dei tumori del collo dell'utero):
Mammografia (prevenzione e diagnosi precoce dei tumori della mammella)
Sangue occulto (prevenzione e diagnosi precoce dei tumori del colon retto)

Pap – test (prevenzione e diagnosi precoce dei tumori del collo dell'utero):
É un programma organizzato dalla Regione Friuli Venezia Giulia per la prevenzione e la diagnosi precoce dei tumori del collo dell'utero, rivolto alle donne residenti di età compresa tra i 25 e i 65 anni. che non hanno eseguito gratuitamente il pap test negli ultimi tre anni e per le donne di 65 anni che non lo hanno mai eseguito prima.
La convocazione all'esecuzione del pap – test avviene su chiamata attiva da parte dell'Azienda Sanitaria, con l'invio ogni 3 anni di una lettera d'invito ad eseguire gratuitamente un pap-test.
Una volta ricevuto l'invito basta presentarsi all'appuntamento. Per cambiare o disdire l'appuntamento è possibile telefonare

Per cambio appuntamento e rinunce:
Call Center Regionale Unico 800 000 400 dalle 9:00 alle 18:00

Per informazioni, casi complessi e per il secondo livello:
Segreteria Pap-Test / ASUITS - tel. 040 399 7235
- lunedì - martedì - mercoledì dalle 10:00 alle 12:30
- giovedì dalle 13:00 alle 15:30
AVVISO:
ASUITs informa l'utenza che a partire dal 7 gennaio 2019 la segreteria Pap-Test seguirà gli orari sotto elencati:
- lunedì e martedì: dalle 10:00 alle 12:30;
- mercoledì: dalle 14:00 alle 15:00;
- giovedì: dalle 13:00 alle 15:30.

L'esame è gratuito e non occorre l'impegnativa.
Qualora il pap-test venga eseguito al di fuori del programma di screening, la paziente che rientra nella fascia d'età prevista, può autocertificare sotto la propria responsabilità di non aver eseguito gratuitamente l'indagine di screening nei 3 anni precedenti –vedi modulo "Dichiarazione esenzione per screening".
Gli eventuali accertamenti successivi sono gratuiti: l'esenzione deve essere segnalata dal medico prescrivente.

Consultare anche: Sezione dedicata del sito della Regione Friuli Venezia Giulia - Screening per la prevenzione dei tumori (collo dell'utero)

Riferimenti:
Responsabile progetti regionale per ASUITs: Daniela GERIN
Tel 040 3997235
E-mail

Avviso:
Dal 10 dicembre p.v. la Segreteria dello screening mammografico e del colon-retto di Trieste è trasferita dall'attuale sede di via della Pietà 19 all'interno dell'Ospedale Maggiore di Trieste (Primo Piano, area ex BIC, stanze 13 e 14).
Risponde al nuovo numero 040 3992816 - le giornate e gli orari di attività rimarranno invariate.

Mammografia (prevenzione e diagnosi precoce dei tumori della mammella)

Il tumore alla mammella è diventato un problema di sanità pubblica per le implicazioni sociali che ne derivano in fatto di perdite di vite umane, di compromissione della qualità della vita e di impiego di risorse economiche

Se hai tra i 45 e i 49 anni

Se hai tra i 50 e i 69 anni

La Sanità del Friuli Venezia Giulia, seguendo le indicazioni europee ed italiane in fatto di prevenzione, ha avvitato un programma di controlli gratuiti con scadenza biennale.
La convocazione avviene su chiamata attiva da parte dell'Azienda Sanitaria (sempre che non si abbia effettuato una mammografia in esenzione negli ultimi 12 mesi), con l'invio ogni 2 anni di una lettera d'invito ad eseguire gratuitamente una mammografia su un'Unità Mobile (Camper) con indicati il giorno, l'ora e il luogo dell'appuntamento. La lettera viene inviata a tutte le residenti in Friuli Venezia Giulia con un'età compresa tra i 50 e 69.
È sempre possibile cambiare o annullare l'appuntamento telefonando al Call Center – Numero Verde dedicato 800-000 400 dal lunedì al venerdì dalle 9:00 alle 18:00. È possibile inoltre disdire o spostare l'appuntamento contattando anche il Numero Unico – Call Center Regionale 848 - 448 884 o recandosi agli sportelli CUP dell'Ospedale di Cattinara.

Se l'esame è negativo, cioé non evidenzia problemi, la risposta arriva a casa per posta entro un mese.
Se la mammografia evidenzia immagini dubbie, la segreteria del centro di screening contatterà la persona e le comunicherà la necessità di eseguire ulteriori accertamenti. Nella maggior parte dei casi, questo non significa la presenza di un tumore, ma è necessario fare altri esami per esserne certi. Gli eventuali accertamenti e approfondimenti successivi sono gratuiti: l'esenzione deve essere segnalata dal medico prescrivente.
Per le donne che aderiscono al programma di screening mammografico sono inoltre previste due ulteriori chiamate oltre l'età limite, ossia fino ai 74 anni.
Solamente le mammografie eseguite sull'Unità Mobile rientrano nella campagna di screening. Ciò significa che se sei in età di screening (tra i 50 ed i 69 anni) e scegli di fare la mammografia presso una struttura ospedaliera o ambulatoriale dovrai pagare il ticket a meno che tu non sia esente per età e reddito o per patologia.
Dal 1 novembre 2006, infatti, la mammografia finalizzata alla diagnosi precoce dei tumori alla mammella per le donne di età compresa tra i 50 e i 69 anni, non è più in regime di esenzione ticket se effettuata nelle radiologie ospedaliere e ambulatoriali accreditate, ma viene richiesto un versamento di Euro 36,00 a titolo di partecipazione alla spesa. La Regione ha infatti posto a suo carico il finanziamento del programma ed ha scelto di svolgere l'attività su un'unità mobile proprio per non sovraccaricare di lavoro le radiologie ospedaliere ed ambulatoriali.

Riferimenti:
Responsabile progetto regionale per ASUITs: Carla DELLACH
Tel 040 399 2816 - Segreteria Screening mammografico
E-mail: co@asuits.sanita.fvg.it

Ulteriori informazioni:
Sezione dedicata del sito della Regione Friuli Venezia Giulia - Screening per la prevenzione dei tumori (mammella)

Sangue occulto (prevenzione e diagnosi precoce dei tumori del colon retto)
É un programma organizzato dalla Regione Friuli Venezia Giulia per la diagnosi precoce dei tumori del colon retto, rivolto agli uomini e alle donne di età compresa tra i 50 e i 69 anni.
L'Azienda Sanitaria invia una lettera d'invito ad eseguire ogni 2 anni un test semplice e gratuito: la ricerca del sangue occulto nelle feci.
Questo esame e gli eventuali accertamenti successivi sono completamente gratuiti.

Una volta ricevuto l'invito basta ritirare il kit in farmacia e seguire accuratamente le istruzioni.
Il test è semplice e si esegue a casa propria. Bisogna raccogliere un piccolo campione di feci e metterlo nella provetta. Non è necessario seguire una dieta particolare prima dell'esame e basterà un unico campione di feci.
Il kit va riportato al più presto in farmacia e comunque entro 24 ore dall'utilizzo. La provetta verrà successivamente inviata a un laboratorio specializzato.
L'esame è gratuito e non occorre l'impegnativa: basta portare la lettera di invito.
Se l'esame risulterà negativo (vale a dire in assenza di sangue occulto nelle feci) arriverà la risposta a casa, per posta, dopo due settimane circa. Ogni 2 anni la stessa persona sarà invitata a ripetere il test.
Se il test risulterà positivo (presenza di sangue occulto nelle feci), la segreteria del programma contatterà la persona per completare gli accertamenti.
In qualche raro caso il test risulterà inadeguato per motivi tecnici e la persona riceverà un invito a ritirare un altro kit e a ripetere l'esame.

Riferimenti:
Responsabile progetti regionale per ASUITs: Carla DELLACH
Tel 040 399 2813 - Segreteria Screening colon retto
E-mail: co@asuits.sanita.fvg.it

Ulteriori informazioni:

Sezione dedicata del sito della Regione Friuli Venezia Giulia - Screening per la prevenzione dei tumori (colon retto)

La tabella seguente mostra il confronto tra le caratteristiche lessicali emerse dal monitoraggio dei due testi. Come ci aspettavamo, i valori dei parametri non risultano così diversi tra loro e nella maggioranza dei casi il testo più difficile (a livello di leggibilità) registra valori maggiormente positivi rispetto a quello facile. Il testo più semplice ha una percentuale maggiore di lemmi appartenenti al VdB ma lo scarto è di solo 1 punto; in tutti gli altri fattori, è il testo complesso a presentare valori migliori (più bassi per quanto riguarda la ricchezza lessicale e più alti per quanto riguarda la presenza di lemmi appartenenti al lessico fondamentale, con una differenza di ben 9 punti percentuali).
Il problema sembrerebbe ancora una volta una sovrastima della leggibilità che, in particolare nei testi che hanno punteggi elevati, non sembra rispecchiare i valori risultanti dal monitoraggio linguistico. Non è possibile, infatti, che il testo con il più alto valore di leggibilità lessicale del corpus registri valori più bassi in tutte le caratteristiche lessicali rispetto anche alla media stessa del corpus.

Caratteristiche lessicali	**Testo più semplice**	**Testo più complesso**	**Media corpus ASL**
TTR (primi 100 lemmi)	0,60	0,54	0,68
% Lemmi nel VdB	62,05	61,01	59,46
FO	67,76	76,76	71,74
Densità lessicale	0,63	0,56	0,59
Leggibilità Lessicale	0%	100%	82%

Tabella 136. Confronto tra le caratteristiche lessicali dei due testi analizzati.

Consideriamo adesso il modello sintattico.
Il testo che restituisce il valore più basso di leggibilità sintattica (0%) è il numero 67, quello con il punteggio più elevato (100%) è il numero 70. Entrambi appartengono al sotto-corpus dell'assistenza sanitaria agli stranieri (il secondo documento fa parte di una sotto sezione del primo) e all'AUSL di Parma. I rispettivi valori di leggibilità lessicale sono 94% e 27%, dunque le due valutazioni riportano risultati opposti in entrambi casi: a una bassa difficoltà sintattica corrisponde un'alta complessità lessicale e viceversa.
Osservando i testi originali viene però qualche dubbio circa la valutazione della leggibilità; in particolare, sembra strano che il testo n. 70 sia quello che nell'intero corpus presenta il più alto valore di difficoltà sintattica.

Testo n. 67
Per la salute degli STRANIERI
L'Ausl di Parma offre ai cittadini stranieri, provenienti sia da altri paesi europei che da paesi extra-europei, e ai loro famigliari alcuni servizi specifici.

In questa sezione puoi trovare informazioni su:

Spazio salute immigrati

Spazio donne immigrate

Assistenza sanitaria

Guide multilingue

Testo n. 70
Assistenza sanitaria
Iscrizione al Servizio Sanitario Nazionale

Assistenza sanitaria per bambini figli di persone immigrate senza permesso di soggiorno

Iscrizione volontaria di determinate categorie di cittadini comunitari

Assistenza sanitaria per i figli di migranti senza permesso di soggiorno

Ad una prima impressione i testi risultano piuttosto simili: entrambi non sono altro che un elenco di voci che corrispondono a link che rimandano a pagine interne del sito. Proviamo a confrontare le caratteristiche lessicali risultanti dal monitoraggio linguistico (Tabella 137).

Caratteristiche	Testo più semplice	Testo più complesso	Media corpus ASL
Media altezze massime	2,42	4,80	5,39
Media lunghezza link	1,74	1,48	2,22
Media lunghezza link massimi	2,85	3,2	7,53
Rapporto principali e subordinate	0,5	1	0,35
Lunghezza media catene subordinanti	0	0	0,90
Lunghezza media catene preposizionali	1	2	1,40

Caratteristiche	Testo più semplice	Testo più complesso	Media corpus ASL
Media di teste verbali per frase	0,57	0,2	1,56
Media di archi entranti in testa verbale	1,75	1	1,88
Leggibilità Sintattica	0%	100%	0,76

Tabella 137. Tabella 138. Confronto tra le caratteristiche sintattiche dei due testi analizzati.

Stando a tali dati, emerge effettivamente una situazione maggiormente positiva per quanto riguarda i punteggi ottenuti dalle caratteristiche sintattiche nel testo più semplice, con la sola esclusione dei parametri relativi ai predicati verbali. Se però confrontiamo i valori dei due documenti con i dati medi dell'intero corpus, notiamo che tale corrispondenza non risulta sempre valida, in particolare per il testo a elevata complessità sintattica. Rispetto al dato medio, infatti, il documento difficile registra valori più positivi nei parametri relativi alla struttura dell'albero sintattico, nelle catene subordinanti (che in questo caso sono assenti) e nel numero di dipendenti per testa verbale.

In sintesi, dai risultati della valutazione della leggibilità tramite i due modelli di analisi READ-IT emergono alcune considerazioni. La prima è che dobbiamo considerare queste stime della difficoltà come puramente indicative; tali risultati ci sono molto utili a livello di analisi per avere alcune prime indicazioni generali, ma risulta necessario effettuare delle prove di comprensione sul corpus per ottenere dati più reali della difficoltà di questi testi.
READ-IT ha dimostrato in più di un'occasione di essere uno strumento utile e adatto a valutare la leggibilità di testi appartenenti anche a diversi generi testuali. In questo caso, però, abbiamo più volte notato una discrepanza tra la misurazione della leggibilità e i punteggi ottenuti dalle diverse variabili linguistiche, sia lessicali che sintattiche. Il problema non è che READ-IT non funziona o non è un metodo accurato: il problema è che è stato addestrato su corpora molto diversi dal nostro e in base a specifiche caratteristiche linguistiche che evidentemente risultano differenti dal nostro set di parametri. È possibile, ad esempio, che la correlazione tra la difficoltà dei testi e tali variabili risulti molto alta per quanto riguarda i vari corpora di addestramento di READ-IT ma che non lo sia invece per il corpus della ASL. Ne risulta che spesso, nonostante i testi presentino caratteristiche linguistiche con valori più o meno ottimali, siano valutati con punteggi alti di leggibilità, in particolare sul piano sintattico.
Risulta quindi evidente che l'unica strada percorribile sia quella di addestrare un nuovo sistema (o eventualmente READ-IT o un altro modello già esistente) in base alle caratteristiche linguistiche proprie dei testi del corpus ASL e ai loro livelli di difficoltà.
Ci eravamo infine chiesti se potesse esistere una certa uniformità nella difficoltà dei testi in relazione al tipo di contenuto espresso oppure in relazione alla ASL di appartenenza. Per quanto riguarda il contenuto, la risposta è che, almeno per quanto riguarda la leggibilità, tale uniformità non esiste, con l'esclusione forse del corpus dell'assistenza domiciliare, in cui la complessità risulta più o meno sempre costante (si tratta infatti dell'unico corpus che presenta un basso valore di deviazione standard, pari a 0,08 nel livello lessicale e a 0,19 nel livello sintattico). Gli altri corpora presentano invece una notevole variazione interna. Sul

piano lessicale si nota una maggiore tendenza all'uniformità nel caso di punteggi più elevati rispetto al livello sintattico, che rimane invece più vario.
Anche quanto riguarda le aziende sanitarie, si osserva che a livello lessicale esiste una sostanziale omogeneità nel caso di valori più elevati: le ASL che registrano i valori più alti di leggibilità, possiedono valori alti in ogni testo. Invece, le ASL che hanno valori medi più vicini allo 0, e dunque più facili, presentano una maggiore variazione interna. Il piano sintattico risulta invece molto più differenziato in entrambi i livelli di difficoltà.
Rispetto a un raggruppamento per contenuto, quello per ASL comporta una maggiore corrispondenza tra testi che risultano più semplici (o difficili) in un livello e testi che risultano più facili (o complessi) nell'altro: dunque i documenti appartenenti a una data azienda sanitaria risultano più semplici o complessi su entrambi i piani a prescindere dal tipo di contenuto.

Conclusione

Scopo del presente lavoro era la proposta di un metodo per la valutazione della leggibilità dei testi presenti nei siti web in lingua italiana.
La prima parte della tesi era rivolta alla ricostruzione dello stato della ricerca sulle formule di leggibilità. Esistono centinaia di lavori scientifici sull'argomento, soprattutto per quanto riguarda la lingua inglese, ma sono relativamente pochi quelli che propongono una rassegna critica del campo della leggibilità. Il tentativo di ricostruire l'evoluzione delle ricerche che affrontano questo tema ha preso avvio da due presupposti: da una parte, reperire tutti quegli studi che rappresentano i punti di riferimento per chi intende occuparsi dello studio della leggibilità e colmare quindi la mancanza di una bibliografia ragionata in merito, almeno per quanto riguarda la lingua italiana. Dall'altra, effettuare un'analisi comparativa dei criteri che hanno portato alla definizione degli indici di leggibilità, al fine di trovare i parametri statistici maggiormente collegati alla difficoltà dei testi e i migliori strumenti per la misurazione di tali variabili.
Nella parte introduttiva si è tentato di definire il concetto di leggibilità, delimitando il suo campo di applicazione e stabilendo il suo ruolo nel processo di comprensione della lettura. Ci siamo poi occupati delle ricerche che hanno preceduto gli studi sulla leggibilità e che sono stati il punto di riferimento per le indagini successive: lo studio delle frequenze lessicali e le analisi di statistica linguistica.
La ricerca empirica sulla leggibilità ha inizio negli anni Venti, con i primi studi sulla frequenza delle parole. Il primo lessico di frequenza della lingua inglese, *The Teacher's Word Book,* è pubblicato da E. L. Thorndike nel 1921, a cui seguono *A Teacher's Word Book of 20,000 Words nel 1932 e A Teacher's Word Book of 30,000 Words nel 1944 insieme a Irving Lorge.* Iniziano così, prima negli Stati Uniti e poi negli altri paesi, le compilazioni di dizionari fondamentali soprattutto ad opera di pedagogisti e psicologi che vogliono da una parte rendersi conto dell'estensione del vocabolario infantile a diverse età, dall'alta trarre indicazioni sull'apprendimento del lessico. Si compilano e si confrontano liste diverse per tentare di raggiungere risultati più generali, nella convinzione che le parole più frequenti siano anche le più utili.
In Italia i primi spogli di frequenza si hanno solo negli anni Settanta, sulla scia degli studi condotti per l'inglese e il francese. Il *Lessico di frequenza della lingua italiana contemporanea* (LIF) elaborato al Centro Nazionale Universitario di Calcolo elettronico di Pisa nel 1971, rappresenta il primo grande progetto di costruzione di un lessico di frequenza per la lingua italiana. Il LIF apre la strada ad altre opere analogamente basate sulla lingua scritta che si sono susseguite per un ventennio fino a giungere al 1993, anno in cui viene presentato il *Lessico di frequenza dell'italiano parlato* (LIP), che costituisce il primo esempio di analisi statistica in grande scala della lingua italiana parlata. Il LIF è stato la base per la compilazione del *Vocabolario di Base della lingua italiana* (VdB) di Tullio De Mauro, composto da circa 7.000 lemmi, riferimento fondamentale per il controllo del lessico di testi scritti in italiano, per verificarne la comprensibilità e aumentarne la leggibilità.
Accanto agli studi che hanno prodotto le prime liste di frequenza si collocano gli studi di statistica linguistica. Gli indici di frequenza si basano infatti sui procedimenti della statistica lessicale che prevede l'applicazione di metodi statistici all'esame dei fatti linguistici: "Le

unita costitutive di una lingua (fonemi, parole, ecc.) soprattutto considerate sotto il profilo della frequenza con cui appaiono nei testi, costituiscono un tipico insieme di fenomeni di massa e sono perciò suscettibili di indagini statistiche per rilevare le frequenze medie del loro distribuirsi nel discorso e, nel tempo, le eventuali trasformazioni di tali frequenze." (De Mauro 1961). La statistica linguistica mira quindi a individuare le regolarità statistiche delle diverse unità testuali, con particolare attenzione al lessico. In quest'ottica si collocano una serie di studi legati ai nomi dello statunitense George K. Zipf, del polacco Benoit Mandelbrot, dei francesi Pierre Guiraud e Charles Muller, dei praghesi e di Gustave Herdan nell'Europa orientale.
Prima la compilazione delle liste di frequenza del lessico delle lingue e poi l'osservazione, sulla base di queste, delle regolarità statistiche che hanno portato alla definizione di leggi, hanno aperto la strada a nuove direzioni di ricerca sul tema della comprensione della lettura, portando infine allo sviluppo di strumenti oggettivi di misurazione della leggibilità e comprensibilità dei testi: gli indici di leggibilità. Klare (1968), nella sua revisione della ricerca sulla frequenza delle parole, conclude: "Not only do humans tend to use some words much more often than others, they recognize more frequent words more rapidly than less frequent, prefer them, and understand and learn them more readily. It is not surprising, therefore, that this variable has such a central role in the measurement of readability".

Nel secondo capitolo abbiamo preso in considerazione quelli che vengono definiti "studi classici sulla leggibilità", cioè tutte quelle ricerche sulla leggibilità dei testi e le formule matematiche sviluppate per la lingua inglese a partire dagli anni Venti fino agli anni Sessanta.
Le prime ricerche provengono dall'ambito scolastico; da una parte si concentrano sul controllo del vocabolario dei libri di testo, dall'altra sulla comprensione dei materiali di lettura proposti agli studenti. Gli studiosi cercano di ideare dei metodi oggettivi di misurazione della difficoltà dei materiali destinati all'apprendimento.
Sono Lively e Pressey (1923) a proporsi per primi l'obiettivo di ricavare degli indicatori in grado di misurare e predire la leggibilità di un testo; il loro studio ha portato alla definizione della prima formula di leggibilità per bambini, che misura il numero di parole diverse e il numero di termini che non appartengono alla lista di Thorndike (1921) in un campione di 1.000 parole. Lo stesso criterio è adottato da Vogel e Washburne (1928), che individuano nel rapporto tipo-replica un buon indice di misurazione della leggibilità. La loro formula, detta la formula Winnetka, è la prima a prevedere la difficoltà in base al livello scolastico ed è divenuta il prototipo delle moderne formule di leggibilità. Il primo indice basato invece sulla valutazione di materiali specificamente progettati per adulti è quello di Dale e Tyler (1934).
Nel 1935 Gray e Leary pubblicano il loro libro *What Makes a Book Readable*, divenuto punto di riferimento nel campo della ricerca sulla lettura per la metodologia usata. I due autori muovono da una ricerca esplorativa sui fattori che in qualche modo possono contribuire alla leggibilità, individuandone ben 289, raggruppati in quattro categorie: contenuto, stile, formato e organizzazione.
A partire dagli studi di Lorge (1939, 1944) i ricercatori concentrano la loro attenzione soltanto su due criteri, considerati i migliori predittori della difficoltà testuale: la lunghezza

delle parole come indice di difficoltà semantica e la lunghezza delle frasi come indice di complessità sintattica.
Nel 1943 Flesch sviluppa una formula statistica per la misurazione della difficoltà dei materiali di lettura per adulti, basata sul conteggio di tre elementi: lunghezza media delle frasi, numero di affissi e numero di riferimenti personali. Dopo la sua pubblicazione, la formula viene impiegata in molti campi diversi: giornali, pubblicità, pubblicazioni del governo, bollettini e opuscoli, materiale per l'educazione degli adulti, libri per bambini. Il suo uso ne mostrò la validità, ma ne evidenziò anche le insufficienze. Il difetto più grave era il troppo tempo che richiedeva la sua applicazione. Per correggerla e renderla più pratica, nel 1948 Flesch propone alcune varianti. In primo luogo la formula viene sdoppiata: una prima formula misura la facilità di lettura (*Reading Ease Score)*, l'altra valuta il grado di interesse che il materiale suscita nel lettore (*Human Interest*). La previsione della facilità di lettura tiene conto della lunghezza della parola e di quella della frase; il valore dell'interesse umano dipende invece dalle frequenze relative dei riferimenti personali. La formula *Reading Ease* è divenuta la più diffusa tra tutte le formule di leggibilità esistenti, sia per la sua semplicità che per la sua facilità di applicazione.
Anche Dale e Chall (1948) pubblicano una formula a due variabili ampiamente utilizzata, soprattutto negli ambienti scolastici. Come indice di difficoltà sintattica scelgono la lunghezza media della frase, ma per determinare la difficoltà semantica, invece che la lunghezza delle parole scelta come variabile da Flesch, usano una lista di parole familiari. Questa variazione complica la formula, a causa del maggior numero di regole applicative, ma aggiunge anche una piccola quantità di potere predittivo.
L'ultima delle formule classiche di leggibilità è il *Fog Index* di R. Gunning (1952). Gunning è stato tra i primi studiosi ad applicare le nuove ricerche di leggibilità al mondo del lavoro, fondando nel 1944 la Robert Gunning Associates, la prima società di consulenza specializzata sulla leggibilità. La sua formula, divenuta popolare grazie alla sua facilità d'uso, è stata scelta scelta dall'esercito, dalla marina e dall'aeronautica per i loro manuali. Utilizza due variabili, la lunghezza media delle frasi e il numero di parole polisillabiche.

La pubblicazione delle formule di Flesch, Dale-Chall e Gunning segna la fine del primo periodo di studi sulla leggibilità; gli autori hanno il merito di aver portato la questione della leggibilità all'attenzione del pubblico, stimolando l'esigenza di produrre (e leggere) testi in un linguaggio semplice e comprensibile. A questo primo momento segue un periodo di consolidamento e approfondimento delle ricerche, che arriverà fino agli anni Novanta. Il terzo capitolo della tesi era dedicato proprio a questi "nuovi studi di leggibilità".
Gli studiosi si sforzano di migliorare le formule attuali e renderle sempre più di facile applicazione. Un forte impulso alla ricerca è dovuto alla disponibilità di strumenti informatici che consentono di analizzare una grande quantità di testi e considerare un maggior numero di variabili senza però perdere il vantaggio della facilità d'uso. Questo periodo è inoltre caratterizzato dallo sviluppo di formule di leggibilità per lingue diverse dall'inglese e dall'introduzione della procedura *cloze* come criterio per lo sviluppo degli indici.
Danielson e Bryan (1963) sviluppano i primi due programmi informatici per l'applicazione delle formule di leggibilità. La loro è la prima formula creata specificatamente per l'uso automatico. Per facilitare il procedimento, impiegano il conteggio dei caratteri per misurare

sia la lunghezza della frase che quella delle parole. Un altro passo significativo verso la facilità di utilizzo è il grafico sviluppato da Fry (1963); nella sua versione più recente (1977) è uno dei metodi più utilizzati per la valutazione della leggibilità. Nel 1965, in un progetto di ricerca sponsorizzato dalla National Science Foundation, Coleman pubblica quattro formule di leggibilità per uso generale; è il primo studio in cui si utilizza il *cloze* come criterio al posto dei più convenzionali test di lettura a scelta multipla o classificazioni da parte di esperti. Anche Bormuth inizia a sperimentare la procedura *cloze* come nuovo criterio. Il suo primo studio (1966) fornisce una panoramica di tutte quelle variabili, oltre al vocabolario e la lunghezza delle frasi, che possono influire sulla comprensione; il *cloze* gli permette di valutare gli effetti di questi fattori non solo sulla difficoltà di interi brani ma anche su singole parole o frasi. Nel 1969 Bormuth conduce la più ampia analisi di leggibilità che sia stata fatta, fornendo una nuova base empirica per le formule successive. Nello stesso anno McLaughlin pubblica la sua formula SMOG, che egli ritiene essere più veloce, più semplice e più valida rispetto ai metodi precedenti di valutazione della leggibilità. L'autore ritiene necessarie come variabili la lunghezza della frase e quella della parola ma crede che i due valori vadano moltiplicati piuttosto che addizionati.
Nel 1967, Smith e Senter creano una formula destinata ad applicazioni militari, *Automated readability Index* (ARI), che utilizza una macchina da scrivere elettronica modificata con tre microinterruttori collegati a contatori cumulativi per parole e frasi. Anche Caylor, Sticht, Fox, e Ford (1973) sviluppano un indice per conto dell'esercito degli Stati Uniti, la formula *Forcast*. Kincaid et al. (1975) seguono questa tendenza, ricalcolando nuove versioni di vecchie formule per testarle su materiali della Marina. Utilizzando i punteggi di comprensione di manuali formativi militari, ritarano le formule ARI, Flesch e Fog, costruendone versioni specifiche per la Marina, chiamate *Navy Readability Index* (NRI).
Nel 1975, Coleman e Liau presentano una formula molto simile a quella di Bormuth, che impiega le stesse variabili ed è costruita con la stessa tecnica cloze. La loro formula è però tarata su studenti universitari ai primi anni di corso. Anche Dale e Chall (1995) sviluppano una nuova versione della loro formula, *The New Dale-Chall Readability Formula*.
A partire dagli anni Ottanta, anche le grandi aziende commerciali sviluppano, con l'ausilio del computer, nuove e più sofisticate formule di leggibilità. Il sistema di misurazione Lexile Framework (1988), il sistema Degrees of Reading Power (DRP) sviluppato da Touchstone Applied Science Associates (1999) e la formula di leggibilità per libri Advantage Open Standard (ATOS) del 2000 sono tutti strumenti che misurano sia il grado di leggibilità dei testi sia il livello di lettura o di istruzione degli studenti; impiegano variabili tradizionali di lunghezza delle frasi e difficoltà del vocabolario ma, essendo informatizzati, sono in grado di valutare grandi campioni di testi o l'intero contenuto di libri.

Il quarto capitolo era dedicato allo sviluppo di formule di leggibilità per lingue diverse dall'inglese. Nella sua rassegna su questo argomento, Klare (1974, 1984) sottolinea che gran parte delle prime ricerche è condotta negli Stati Uniti a beneficio di studenti di lingua inglese che studiano lingue straniere. Nascono così lo studio di Tharp (1939) sul francese, sette formule per lo spagnolo (Spaulding 1951, Patterson 1972, Thonis 1976, Garcia 1977, Gilliam et al. 1980, Vari-Cartier 1981 e Crawford 1984), strumenti per l'ebraico (Nahshon 1957), il tedesco (Walters 1966, Schwartz 1975), il russo (Rock 1970), il cinese (Yang 1970) e il vietnamita (Nguyen e Henkin 1982).

Le prime misure di leggibilità sviluppate in Europa sono semplicemente adattamenti della formula *Reading Ease* di Flesch: Kandel e Moles (1958) tarano l'indice sulla lingua francese, Huerta (1959) produce una versione spagnola, Douma (1960) e Brouwer (1963) adattano la formula alla lingua olandese, De Landsheere presenta una versione per il francese (1963) e una per il tedesco (1970).
L'attività di ricerca per lo sviluppo di strumenti più originali inizia in Europa alla fine degli anni Sessanta. Si trovano così ricerche sul finlandese (Wiio 1968), francese (De Landsheere 1966, Henry 1973, 1979, Richaudeau 1979), danese (Togeby 1971), svedese (formula Lix di Björnsson 1968 e 1983, Platzach 1974), olandese (van Hauwermeiren 1972, Zondervan, van Steen e Gunneweg 1976, Staphorsius e Krom 1985), tedesco (in Germania: Groeben 1972, Nestler 1977, Dickes e Steiwer 1977; in Austria: Bamberger 1973), spagnolo (in Venezuela: Gutiérrez Polini et al. 1972; in Spagna: Rodriguez 1981, Rodríguez Diéguez 1983).
Più recentemente, sono effettuate ricerche che applicano i nuovi metodi computazionali a diverse lingue, come il cinese (Lau 2006, Chen et al. 2013), il tedesco (Vor Der Brück-Hartrumpf 2007), il francese (François-Fairon 2009), l'arabo (Al-Khalifa-Al-Ajlan 2010), il giapponese (Tanaka-Ishii et al. 2010), il tailandese (Daowadung-Chen 2011) e lo svedese (Sjöholm 2012).

Il quinto capitolo ha riguardato gli studi di leggibilità in Italia.
Mentre negli Stati uniti le ricerche sul rapporto tra leggibilità dei testi e comprensione della lettura si sono sviluppate già a partire dalla fine degli anni Venti, in Italia tali problemi cominciano a imporsi all'attenzione generale solo dalla fine degli anni Sessanta. L'Italia risulta in ritardo anche rispetto agli altri paesi europei: se in questi paesi le prime tarature della formula *Reading Ease* di Flesch si hanno già a partire dalla fine degli anni Cinquanta o dai primi anni Sessanta, il primo adattamento all'italiano risale invece al 1972 da parte di Roberto Vacca. Allo stesso modo, mentre in Europa le prime formule originali sono sviluppate a partire dalla fine degli anni Sessanta, la prima formula tarata sulla lingua italiana si ha soltanto alla fine degli anni Ottanta.
Le occasioni di riflessione sul tema della comprensibilità e della leggibilità dei testi si moltiplicano a partire dagli anni Settanta e soprattutto nel corso degli anni Ottanta. Tra i vari dibattiti e i lavori svolti in quegli anni si segnalano: il dibattito aperto tra il 1976 e il 1978 da alcuni quotidiani e periodici italiani sulla semplificazione della comunicazione; le ricerche che hanno portato alla definizione del *Vocabolario di Base* della lingua italiana e alla nascita della collana dei *Libri di Base* (1980); l'analisi della leggibilità dei *Libri di Base* curata da Tiziana Fiorucci (1982); il XIX Congresso Internazionale della Società di Linguistica Italiana del novembre 1985 dedicato al problema della percezione, comprensione e interpretazione, visto dalla parte del ricevente (De Mauro, Gensini, Piemontese 1985); l'incontro di studio *Leggibilità e comprensione* organizzato dalla cattedra di Filosofia del Linguaggio dell'Università La Sapienza e tenutosi a Roma il 26-27 giugno 1986 (De Mauro, Piemontese, Vedovelli 1986); l'analisi della leggibilità di manuali scolastici da parte di Anna Thornton (1984); le analisi di leggibilità svolte dalla cooperativa Spazio Linguistico (Palombi e Raponi 1984, Palombi 1986); l'analisi di leggibilità di testi giuridici e politici ad opera dell'Istituto di Documentazione Giuridica del CNR di Firenze (Martino, Bianucci 1986); le ricerche condotte nell'ambito di seminari intercattedra di Filosofia del linguaggio e Pedagogia dell'Università La Sapienza sulla leggibilità di testi scolastici e sul livello di

comprensione degli studenti (Lucisano, Piemontese 1986); la formazione (sotto la supervisione di De Mauro e Piemontese) di un gruppo di studenti (Gruppo H) fra il 1984 e il 1987 alla redazione di testi di alta leggibilità, diretti in particolare a persone con deficit intellettivo (Piemontese, Vedovelli 1988); l'analisi da parte del Gruppo H, con il coordinamento di Piemontese e Tiraboschi, di una serie di libretti informativi del sindacato.
Tutti questi lavori suggeriscono che la formula di Vacca poteva fornire indicazioni utili, tuttavia la maggior parte dei ricercatori evidenzia dei problemi nella sua applicazione, come il conteggio automatico delle sillabe, particolarmente difficile per l'italiano. Si notano problemi anche per il computo delle cifre, sigle, abbreviazioni e simboli. I problemi relativi all'uso della formula sono stati affrontati in una serie di esercitazioni di ricerca intercattedra svolte dalle cattedre di Filosofia del Linguaggio (prof. De Mauro) e di Pedagogia (prof.ssa Corda Costa) dell'Università di Roma La Sapienza dal 1987 al 1989, con l'obiettivo di mettere a punto una nuova formula di leggibilità tarata sulla lingua italiana. A queste ricerche ha partecipato un gruppo di lavoro, denominato Gulp (Gruppo Universitario Linguistico Pedagogico), formato da ricercatori, insegnanti, dottorandi di ricerca e studenti. Il progetto si è svolto anche grazie all'intervento dell'IBM Italia che ha messo a disposizione alcuni computer ed un piccolo finanziamento. Risultato di questi studi è stato lo sviluppo dell'indice GULPEASE, la prima formula di leggibilità per l'italiano. Nella formula la frequenza della parola costituisce la variabile semantico-lessicale, mentre la lunghezza media delle frasi è la variabile sintattica; lessico e sintassi sono considerati i due livelli linguistici che rendono conto del grado di difficoltà nella leggibilità di un testo.

Il sesto capitolo era dedicato a uno studio approfondito dei nuovi approcci al tema della leggibilità.
A partire dalla prima metà degli anni 2000, i ricercatori hanno dimostrato un rinnovato interesse per la leggibilità. Il desiderio di superare le limitazioni degli indici tradizionali, insieme ai progressi compiuti nel campo del *machine learning* e lo sviluppo di efficienti tecniche di *Natural Language Processing* (NLP), hanno contribuito alla nascita di nuovi approcci alla valutazione della leggibilità.
Da una parte, l'opportunità di sfruttare metodi computazionali sempre più sofisticati e una crescente disponibilità di nuove fonti di dati hanno consentito ai ricercatori di esplorare una più ampia varietà di caratteristiche linguistiche e sperimentare variabili più complesse; dall'altra, l'uso di modelli di previsione avanzati basati sull'apprendimento automatico ha permesso di costruire nuovi strumenti e algoritmi per la misurazione della leggibilità. Si è quindi registrato un passaggio dalle misure tradizionali a favore dei nuovi algoritmi di valutazione; tali metodi sono rivolti alla costruzione di un modello che permetta di classificare in modo automatico un insieme di documenti testuali in base al loro livello di difficoltà.
È stato necessario, in primo luogo, condurre un'analisi più generica di tutte quelle tecniche che possono essere raggruppate sotto la definizione di *machine learning*; il lavoro di ricostruzione dell'evoluzione delle ricerche in questo ambito è piuttosto complesso, dal momento che presuppone non solo conoscenze avanzate in campo informatico ma anche in campo matematico-statistico.
L'approccio di *machine learning*, in italiano 'apprendimento automatico', inizia a diffondersi a partire dagli anni Novanta, per poi divenire il paradigma dominante. L'approccio prevede

tutta una serie di tecniche rivolte alla costruzione automatica di classificatori di documenti testuali, cioè di programmi informatici in grado di etichettare i documenti scritti in un linguaggio naturale in un insieme di categorie. Le tipologie di apprendimento automatico si distinguono in apprendimento supervisionato e non supervisionato, a seconda che il corpus di apprendimento sia o meno già etichettato. La tecnica più usata è quella della classificazione automatica (o categorizzazione automatica), che appartiene al gruppo delle tecniche supervisionate. Indica l'attività di ordinare automaticamente un insieme di documenti in categorie a partire da un set predefinito. Se ne possono trovare innumerevoli applicazioni, come i filtri anti-spam per la posta elettronica, i metodi di indicizzazione automatica di pagine web, strumenti destinati ai motori di ricerca, al web semantico e alla creazione di ontologie, la disambiguazione automatica, le attribuzioni automatiche di paternità a documenti scritti, ecc.

Dopo una panoramica sul *machine learning* e alcuni algoritmi di apprendimento automatico, abbiamo presentato i metodi più recenti di misurazione della leggibilità, sia per la lingua inglese, che come sempre è la lingua da cui parte l'impulso alla ricerca, sia per le altre lingue, tra cui l'italiano.

Per quanto riguarda i nuovi metodi di valutazione automatica della leggibilità di un testo, è possibile fare una prima distinzione in base al tipo di approccio utilizzato:

- Valutazione della leggibilità come compito di classificazione: assegnazione del documento analizzato ad una specifica classe di leggibilità;
- Valutazione della leggibilità come compito di ranking: assegnazione del documento analizzato ad una posizione all'interno di una scala di leggibilità;
- Valutazione della leggibilità come problema di regressione: i livelli o i punteggi si trovano in un intervallo continuo.

La valutazione della leggibilità come compito di classificazione è l'approccio più utilizzato, ad esempio in studi come quelli di Si e Callan (2001), Liu et al. (2004), Collins-Thomson e Callan (2004), Schwarm e Ostendorf (2005), Heilman et al. (2007), Al-Kalifa e Amani (2010), Aluisio et al. (2010), Chen (2013); il problema principale è rappresentato dal fatto che richiede dati di addestramento che possono non essere disponibili, specialmente per un dominio specifico. La valutazione come compito di ranking rappresenta un'alternativa valida alla precedente in quanto richiede soltanto dati di addestramento rispetto a due livelli di leggibilità (facile o difficile). Questo approccio è utilizzato per esempio da Inui et al. (2001), Tanaka-Ishii et al. (2010). Il modello di regressione è invece utilizzato da Kate et al. (2010) e François e Fairon (2012).

Un'ulteriore distinzione può essere operata a seconda delle caratteristiche linguistiche considerate (lessicali, sintattiche, semantiche e relative alle parti del discorso). Nella maggior parte degli studi è impiegata una combinazione delle diverse caratteristiche: Si e Callan (2001) e Collins-Thompson e Callan (2004) utilizzano modelli statistici del linguaggio di tipo *unigram* combinati con altre caratteristiche, di tipo sintattico o semantico. Liu et al. (2004) e Schwarm e Ostendorf (2005) impiegano l'algoritmo SVM per combinare le caratteristiche sintattiche con quelle semantiche. Heilman et al. (2007) analizzano la struttura sintattica delle frasi tramite la combinazione di modelli statistici n-gram e i più tradizionali alberi sintattici. Kate et al. (2010) usano algoritmi di regressione per combinare caratteristiche sintattiche, lessicali e modelli linguistici specifici per generi testuali. François

e Fairon (2012) considerano ben 46 parametri linguistici diversi (lessicali, sintattici, semantici, oltre a parametri relativi al francese come L2).
I metodi si differenziano tra loro anche in base al campo di applicazione e ai destinatari. Schwarm e Ostendorf (2005), Heilman et al. (2007) e Peterson e Ostendorf (2009) si occupano di classificare il livello di lettura di testi scritti destinati a studenti di L2. Altri studi si concentrano sulla valutazione del livello di lettura di pagine web, come Si e Callan (2001) e Collins-Thompson e Callan (2004). Wang (2006) misura la leggibilità delle informazioni presenti nei siti web di assistenza sanitaria. Liu et al. (2004) determinano il livello di lettura dei risultati delle query dei motori di ricerca. Miltsakaki e Troutt (2007) hanno progettato un'applicazione per valutare la leggibilità dei testi sul web e classificarli in base al loro contenuto tematico. Guo, Zhang e Zhai (2011) hanno sviluppato un indice di leggibilità da integrare nel motore di ricerca di Twitter; Bilal (2013) ha valutato la leggibilità dei risultati delle ricerche di Google.
Parte di questo capitolo riguardava anche una serie di studi che si collocano accanto alle ricerche sulla valutazione automatica della leggibilità, ma che potremmo definire *intermedi*, *di transizione*: sebbene infatti queste ricerche siano ancora in parte collegate alla misurazione tradizionale della leggibilità, costituiscono in un certo modo un superamento di queste, per la metodologia impiegata (l'analisi di una serie di variabili più complesse, valutate tramite strumenti di NLP), o per l'oggetto della valutazione (i documenti e le risorse presenti sul web). Si tratta quindi di strumenti innovativi, che però non rientrano nella categoria della valutazione automatica in quanto non fanno uso di tecniche di apprendimento automatico.
Fanno parte di questa categoria Coh-Metrix, uno strumento informatico sviluppato presso l'università di Memphis, che analizza i testi e misura in modo automatico oltre 200 parametri linguistici e i suoi due adattamenti: Coh-Metrix-Por, per la lingua portoghese brasiliana e Coease per lingua italiana.
Il sesto capitolo si è concluso con la parte relativa a READ-IT, il primo, e attualmente unico, strumento italiano di valutazione automatica della leggibilità. Realizzato dall'*Italian Natural Language Processing Laboratory* (ItaliaNLP Lab) dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa, READ-IT nasce come supporto al processo di semplificazione dei testi e pertanto si rivolge a un pubblico di destinatari specifico, cioè lettori caratterizzati da una bassa alfabetizzazione o da lieve deficit cognitivo. Uno degli aspetti innovativi è che la valutazione della leggibilità è effettuata su due livelli: il documento e la singola frase.
In linea con gli approcci più recenti, la misurazione della leggibilità è considerata un compito di classificazione, nello specifico una classificazione binaria (ranking) che distingue tra due livelli di lettura (*facile* e *difficile*). READ-IT è stato sperimentato su diverse tipologie di testi (giornalismo, letteratura, prosa scientifica e materiale educativo) ed è stato poi utilizzato anche in altri studi, come metodo per valutare la leggibilità di documenti appartenenti all'ambito medico, giuridico o burocratico.

Dalla rassegna critica del campo della leggibilità sono emerse alcune considerazioni generali.

- Fino agli anni Novanta il metodo seguito dai ricercatori per la definizione di una formula di leggibilità è generalmente lo stesso, sia per l'inglese che per le altre

lingue. Questo metodo, che possiamo definire *tradizionale*, prevede che siano effettuate delle misurazioni sui lettori e successivamente delle misurazioni sui testi. Vengono messi a punto dei test di comprensione della lettura (a risposta multipla o *cloze* test) per determinare il grado di facilità o di difficoltà di testi scritti. Se i lettori rispondono in modo rapido e corretto alle domande formulate sul testo in esame, il testo è considerato facile; se invece essi commettono errori, il testo è considerato più o meno difficile. In questo modo si costruisce una scala di leggibilità: i vari campioni di testi vengono ordinati su una scala che va dal più difficile (quello che ha fatto registrare il maggior numero di errori) al più facile (quello che ha fatto registrare il minor numero di errori). Successivamente, si analizzano statisticamente i brani, confrontando le caratteristiche linguistiche dei campioni per individuare se esiste una variabile statistica strettamente correlata con la difficoltà. Sono molti gli indici statistici che possono essere misurati: la percentuale delle parole più frequenti, il numero delle parole difficili, il numero delle parole diverse, il numero dei pronomi, il numero delle frasi complesse, ecc.

Una volta compilata una lista degli indici statistici e stabiliti i punteggi dei vari brani per ciascun indice, è possibile correlare i punteggi con quelli relativi alla leggibilità. Se il coefficiente di correlazione è prossimo a 1 o -1, l'indice statistico misura qualcosa di correlato alla leggibilità. Se la correlazione è prossima allo zero, l'indice statistico non ha rapporto con la leggibilità. In base alle diverse variabili statistiche considerate, i ricercatori costruiscono le varie formule di leggibilità.

- Gli studi classici sulla leggibilità (anni Venti - Sessanta) si sono concentrati principalmente su due aspetti: da una parte la ricerca di un criterio adeguato a determinare la difficoltà di brani specifici, dall'altra l'identificazione di quegli elementi testuali che influenzano tale difficoltà e il modo migliore per misurarli.

  Per quanto riguarda gli indici statistici correlati con la difficoltà, gli studiosi hanno considerato quasi esclusivamente le variabili dello stile, legate alla struttura sintattica del testo o a scelte lessicali. Il vocabolario impiegato risulta avere il più alto valore predittivo. In particolare, nelle formule vengono misurati la diversità (il numero di parole diverse), il numero di parole non comuni, il numero di parole difficili, il numero di tecnicismi, l'interesse umano (ad esempio il numero di pronomi personali), la lunghezza delle parole (in realtà Flesch è l'unico ad impiegarla come variabile semantica). La struttura sintattica viene misurata principalmente tramite la lunghezza della frase (il parametro è introdotto da Grey e Leary nel 1935 ed è mantenuto in tutte le formule successive) ma sono usati anche il numero di frasi semplici e la percentuale di frasi preposizionali.

  Lo *Standard Test Lessons in Reading* di McCall e Crabbs (1926, 1950) sembra essere il miglior criterio per determinare la difficoltà di lettura dei brani. Viene scelto per la prima volta da Lorge (1939) e rimane lo standard per le formule di leggibilità fino agli anni Sessanta. Ogni brano ha un punteggio di difficoltà (già stabilito dagli autori) che può essere usato come parametro di confronto; il test misura la comprensione della lettura per ogni specifico brano. Questo è il suo punto di forza ma anche la sua debolezza di fondo. Le valutazioni sono effettuate tramite un questionario e le domande sono pensate in base ai brani considerati. Ma le

domande variano sia per quanto riguarda il linguaggio impiegato, sia a livello dei concetti considerati; se viene usato un linguaggio troppo difficile il lettore potrebbe non rispondere correttamente anche se ha compreso bene il brano. Il numero di risposte corrette, quindi, dipende anche dal tipo di domande sottoposte al lettore. La misurazione della difficoltà di un testo è inevitabilmente legata alla qualità delle domande che lo valutano. Questa indeterminatezza determina l'ambiguità del criterio (Lorge 1949). L'introduzione del cloze test come nuovo metodo di validazione degli indici apre la strada a misurazioni più precise delle variabili testuali e della comprensione della lettura.

- Tra gli anni Sessanta e gli anni Novanta, gli studi sono rivolti principalmente a migliorare le formule attuali e rendere più semplice la loro applicazione. Una maggiore facilità d'uso è resa possibile grazie alla nascita dei primi programmi informatici in grado di calcolare la leggibilità in modo automatico: l'uso di tali strumenti non solo semplifica il computo dei parametri, che fino ad allora era effettuato in modo manuale, ma consente anche di analizzare una grande quantità di testi e considerare un maggior numero di variabili.
  Questo periodo è inoltre caratterizzato da alcuni aspetti innovativi:
  - Viene introdotto il conteggio dei caratteri come nuovo parametro per la stima della difficoltà delle parole e/o delle frasi. Si tratta di un calcolo più veloce e più semplice rispetto alla misurazione di altre variabili, come ad esempio il numero delle sillabe; inoltre si presta bene al computo automatico.
  - Nascono le prime formule per la comprensione dei testi parlati: la formula di Rogers nel 1962 che utilizza 480 campioni di parlato spontaneo improvvisato di studenti di ogni grado scolastico e la *Easy Listening Formula* (ELF) di Fang nel 1966-67 basata sullo studio comparato di testi provenienti da telegiornali o programmi televisivi.
  - Le ricerche sulla leggibilità iniziano ad essere applicate in campo militare e dunque a materiali tecnici, come manuali, regolamenti, relazioni, documenti di formazione del personale, ecc.: l'*Automated readability Index* (ARI) del 1967 è tarato su testi dell'Air Force (l'Aeronautica militare degli Stati Uniti); la formula FORCAST del 1973 nasce nell'ambito di uno studio condotto per conto dell'esercito degli Stati Uniti sulle abilità di lettura necessarie per i MOB (*Military Occupational Specialties 'Specializzazione Occupazionale Militare'*), cioè sulle particolari esigenze di lettura richieste a seconda del campo di competenza militare; le formule chiamate *Navy Readability Index* (NRI) del 1975 sono un gruppo di nuove versioni di formule di leggibilità esistenti tarate specificatamente per materiali della Marina.

- In generale sono molte le obiezioni che sono rivolte alle tradizionali tecniche di misurazione della leggibilità; non tutti gli studiosi concordano nel ritenere che gli indici costituiscono uno strumento efficace e molti criticano il valore o la validità della misurazione stessa della leggibilità.
  Le principali critiche riguardano il fatto che le formule non tengono conto di diversi fattori che influenzano il processo di comprensione, come il livello culturale e la

preparazione del lettore, il vocabolario impiegato, la correttezza ortografica, grammaticale e sintattica del testo, la struttura logica, l'impaginazione, la dimensione e il tipo di caratteri impiegati, la presenza di tabelle, immagini, grafici o di accorgimenti volti a facilitare la decodifica, come titoli, sottotitoli, sottolineature, grassetti, ecc. Ci sono poi altri fenomeni che influiscono sulla difficoltà dei testi e che andrebbero dunque considerati, ad esempio l'eventuale presenza di nominalizzazione, particolarmente diffusa nell'italiano scritto.
Inoltre, la maggior parte delle formule è stata creata prima della diffusione del web: essendo progettati esclusivamente per i testi scritti, gli indici non prendono in considerazione le caratteristiche tipiche dei testi e delle pagine web.

- Dall'analisi dei principali lavori sulla valutazione automatica della leggibilità possiamo trarre alcune conclusioni. In primo luogo, l'approccio della classificazione, che è il metodo più impiegato, sembrerebbe anche essere quello più adatto al compito di valutare la leggibilità dei testi. Il problema della classificazione, ma vale anche per la regressione, è che richiede dati di addestramento già etichettati che potrebbero non essere disponibili, soprattutto per lingue diverse dall'inglese. Il metodo del ranking risolve il problema della mancanza di dati di addestramento etichettati e rappresenta un'alternativa valida: i testi devono infatti essere annotati soltanto rispetto a due livelli di leggibilità (facile o difficile). Tuttavia, come notato da Tanaka-Ishii et al. (2010), questo sistema manca di *assolutezza* nel determinare una norma, riportando solo valori relativi di leggibilità.
  Per quanto riguarda gli algoritmi di apprendimento, molti lavori mostrano che l'approccio basato sul *Support Vectore Machine* offre una maggiore accuratezza e una precisione in alcuni casi superiore all'80% rispetto ad altri modelli (come i Naïve Bayes e gli alberi decisionali), ma anche rispetto alle classiche formule di leggibilità (come la formula di Flesch-Kincaid).
  La scelta delle caratteristiche linguistiche da estrarre dai dati sembra avere un peso maggiore nel determinare le prestazioni del modello di apprendimento rispetto alla selezione del *contesto* di apprendimento, come la scelta del tipo di approccio o dell'algoritmo. In molte ricerche, la migliore funzionalità risulta la lunghezza media della frase, che ottiene alti valori di correlazione con i livelli di lettura, per varie tecniche di misurazione: da notare il fatto che si tratta di una delle metriche tradizionali di leggibilità. Al secondo posto si colloca il modello statistico del linguaggio. La combinazione di diversi tipi di funzionalità risulta essere l'approccio più efficace, raggiungendo gradi di accuratezza del 70% - 80%.
  Sembra, infine, che ci sia uno spostamento rispetto ai destinatari di riferimento. Le tradizionali formule di leggibilità nascono infatti come supporto agli insegnanti, per la selezione dei materiali di lettura appropriati da sottoporre agli studenti; successivamente, le ricerche ampliano il loro campo di applicazione e iniziano a rivolgersi a tutte quelle figure che si occupano di produrre testi di vario genere, come quotidiani e riviste, pubblicità, pubblicazioni amministrative, manuali tecnici, libri di testo e materiale per l'educazione degli adulti, ecc. Negli studi più recenti il target di riferimento non sembrerebbe più essere soltanto chi produce i testi, ma anche i destinatari stessi dei materiali di lettura, come gli studenti (di L1 o L2) o più

genericamente, gli utenti web. La maggior parte dei sistemi di valutazione automatica della leggibilità non si propone, infatti, come strumento di supporto alla produzione di documenti, ma come ausilio per l'utente nell'identificazione dei testi adeguati al proprio livello di istruzione, in particolar modo nella fruizione dei contenuti web.

La seconda parte del presente del lavoro era rivolta alla costruzione di un metodo di valutazione automatica della leggibilità di siti web in lingua italiana.
In base ai risultati emersi dalla ricostruzione dello stato dell'arte delle ricerche sulla misurazione della leggibilità, è stato necessario modificare la linea di ricerca scelta in precedenza, che prevedeva di realizzare un indice di leggibilità adeguando le formule già esistenti per l'italiano alle caratteristiche specifiche della lingua dei siti web. Si è dunque abbandonato il metodo tradizionale a favore dell'approccio della valutazione automatica basata su tecniche di *machine learning*.
La scelta di non adattare gli strumenti esistenti alle peculiarità del web, ma di sviluppare un metodo tarato specificatamente sulla lingua italiana sul web prende le mosse dal seguente presupposto: non esiste una singola varietà che possa essere definita "lingua del web", ma una molteplicità di varietà, ognuna delle quali presenta proprie caratteristiche linguistiche. La variazione della lingua sul web rispetto alle diverse dimensioni dello spazio linguistico è strettamente collegata al genere testuale.
Non è dunque possibile sviluppare un indice di leggibilità generale per il web (o adattarne uno), ma è necessario sviluppare un modello che renda conto della variabilità della lingua in rete e della molteplicità dei generi testuali. "Anche i concetti di semplificazione, di brevità, e ancora più quello di chiarezza diventano relativi, e vanno di volta in volta bilanciati a seconda del punto preciso dello spazio linguistico del web in cui vogliamo collocarci" (Biffi 2014, p. 97).
Ciò è possibile soltanto ricorrendo a un approccio di valutazione basato sull'apprendimento automatico, che permette di costruire un singolo modello capace di riadattarsi a seconda della varietà linguistica considerata. Per l'addestramento del modello è sufficiente scegliere un genere testuale di esempio rappresentativo di una varietà presente sul web. Una volta addestrato, il sistema sarà poi in grado di riadattarsi ad eventuali altre varietà.
Nel settimo capitolo abbiamo illustrato in modo dettagliato la metodologia proposta e le diverse fasi di realizzazione del progetto. Il metodo di valutazione prevede la costruzione di un modello che permetta di classificare in modo automatico un insieme di documenti testuali in base al loro livello di leggibilità. Il processo comprende diverse fasi: la prima consiste nella definizione di un corpus di apprendimento su cui verrà addestrato il modello; per la costruzione del modello il corpus deve essere precedentemente etichettato: ad ogni testo deve cioè essere assegnato un livello di leggibilità. Tali classi di leggibilità costituiranno lo standard di riferimento.
La fase successiva prevede la selezione di quelle caratteristiche linguistiche che dovranno essere analizzate in ciascun testo. Una volta effettuata la selezione, si procede con l'estrazione automatica delle caratteristiche: si trasforma ogni testo in un vettore di caratteristiche numeriche che serve da input per l'algoritmo di apprendimento. L'algoritmo crea quindi il modello: impara cioè, in base agli esempi forniti, ad associare ogni vettore di caratteristiche che rappresenta un testo al livello di leggibilità assegnato a quel dato testo.

L'ultima fase prevede la validazione del modello su un nuovo set di dati. Il modello ottimizzato viene applicato a un nuovo corpus per stimare la sua capacità di predizione, cioè per valutare se il sistema è in grado di prevedere correttamente il livello di leggibilità dei nuovi testi.

Nel corso del nostro lavoro abbiamo portato a termine le prime fasi. Abbiamo costruito un corpus di lingua istituzionale degli enti sanitari e, in particolare, abbiamo raccolto dai siti web delle Aziende Sanitarie Locali (ASL) italiane testi informativi destinati ai cittadini. Abbiamo focalizzato la nostra attenzione su quattro tipi di contenuti informativi: il servizio di emergenza-urgenza, lo screening oncologico, l'assistenza sanitaria agli stranieri e l'assistenza domiciliare. Si tratta di tematiche per le quali la chiarezza e la comprensibilità dell'informazione sono estremamente essenziali; sono inoltre tra i contenuti che, secondo le *Linee guida* del Ministero della Salute per la progettazione di una comunicazione sanitaria online di qualità, dovrebbero sempre essere presenti nei siti web degli enti istituzionali.
Abbiamo selezionato un campione di 30 siti web di aziende sanitarie, così ripartiti: 13 ASL del Nord, 5 del Centro, 9 del Sud e 3 delle isole. Per ciascun sito, abbiamo raccolto i testi riguardanti le quattro tematiche scelte. Ne risulta un corpus composto da 248 documenti (32 riguardano il servizio di emergenza, 85 gli screening oncologici, 83 l'assistenza sanitaria agli stranieri e 48 l'assistenza domiciliare), per un totale di 122.793 occorrenze. Tale corpus costituisce il modello di esempio su cui, in seguito, potrà essere addestrato il sistema di valutazione.
Abbiamo poi ricostruito il profilo linguistico del corpus tramite il monitoraggio delle caratteristiche lessicali, morfosintattiche e sintattiche dei testi. Il monitoraggio linguistico è il punto di partenza per l'individuazione sia dei parametri legati alla complessità, che delle caratteristiche che identificano quel dato genere o varietà testuale. Abbiamo confrontato i risultati dell'annotazione linguistica con i dati emersi dal monitoraggio di altri corpora rappresentativi di diversi generi testuali (giornalismo, letteratura, materiale didattico, prosa scientifica, linguaggio burocratico, linguaggio legislativo) e, per quanto riguarda invece l'ambito medico, con i valori relativi a due tipologie di testi considerate rappresentative della comunicazione medico-paziente: il corpus dei foglietti illustrativi (bugiardini) dei farmaci senza obbligo di prescrizione medica e il corpus che raccoglie le informative di consenso per le procedure diagnostico-terapeutiche impiegate nelle aziende sanitarie toscane. La comparazione del profilo linguistico della nostra varietà con i profili di tali corpora serviva a identificare l'esistenza di tratti comuni tra i diversi generi ma soprattutto di tratti specifici dei testi web informativi di ambito sanitario.
Ciò che è emerso è che il corpus delle ASL presenta un profilo peculiare che lo rende significativamente diverso dagli altri generi testuali.
Sul piano lessicale, i parametri restituiscono valori che indicano una certa complessità testuale: ad esempio, il corpus mostra un alto rapporto type/token (0,79), indicatore di testi particolarmente variegati dal punto di vista lessicale, un valore elevato di densità lessicale (0,59) e una bassa percentuale di lemmi appartenenti al *Vocabolario di Base* (59,46%), entrambi spie di una maggiore difficoltà. Quest'ultima percentuale dipende sicuramente dal fatto che il corpus è ricco di termini specialistici appartenenti al dominio medico e dunque non presenti nel vocabolario di base: sarebbe quindi auspicabile poter

includere nella valutazione lessicale alcuni vocabolari specialistici che integrino quello di base, oppure, se non possibile, almeno una selezione di termini appartenenti a specifici ambiti. In questo modo i punteggi di leggibilità non sarebbero penalizzati.
Riguardo all'analisi contrastiva, il corpus ASL presenta valori più simili al dato medio generale che a quelli dei singoli corpora. In alcuni fattori, come la densità lessicale, i punteggi risultano più affini, in altri si registra una notevole differenza: ad esempio, considerando la composizione del vocabolario, il corpus ASL presenta una percentuale di lemmi appartenenti al VdB che talvolta si distacca di ben 10 punti rispetto ad altri corpora.
Sul piano morfosintattico e sintattico, il corpus risulta avere invece dati maggiormente positivi, sia in assoluto che in riferimento agli altri generi testuali. In particolare, a livello morfosintattico si nota una certa differenza riguardo alla distribuzione di alcune categorie grammaticali e conseguentemente anche ai rapporti tra tali parti del discorso: il corpus registra percentuali più alte di aggettivi, nomi, preposizioni e più basse di verbi, pronomi e avverbi; la frequenza delle congiunzioni si attesta invece intorno alla media generale. Presenta inoltre punteggi più elevati sia nel rapporto tra nomi/verbi (3,30), che nell'uso del participio (23,06), in entrambi i casi con una notevole variazione interna (i valori del rapporto nomi/verbi vanno da 1 a 3,65, quelli relativi alla frequenza delle forme participiali da 15% a 27% circa).
A livello sintattico, restituiscono valori maggiormente positivi le caratteristiche legate alla struttura dell'albero sintattico (media delle altezze massime, media della lunghezza dei link, media della lunghezza dei link massimi) e alla subordinazione (distribuzione e rapporto principali/subordinate, lunghezza media delle catene subordinanti). Circa l'andamento interno, si osserva una consistente variazione sia rispetto al rapporto tra frasi principali e subordinate (il dato medio è pari a 0,35; i valori dei sotto-corpora vanno da 0,26 dell'assistenza domiciliare a 0,42 dello screening) e alla lunghezza delle catene subordinanti (in questo caso la media è 0,90 e i valori oscillano tra lo 0,74 del corpus degli stranieri e l'1,04 del corpus relativo all'emergenza sanitaria).

Abbiamo infine effettuato l'analisi della leggibilità tramite lo strumento avanzato READ-IT. In particolare, abbiamo monitorato la complessità rispetto a due livelli di analisi, quello lessicale e quello sintattico.
La leggibilità è stimata sia per l'intero corpus che per i vari sotto-corpora, costituiti in base alle quattro tematiche considerate o a seconda delle aziende sanitarie. L'obiettivo era capire se esiste una certa uniformità nella difficoltà dei testi in relazione al tipo di contenuto espresso oppure in relazione alla ASL di appartenenza.
I valori di difficoltà ottenuti dal corpus risultano entrambi piuttosto alti: a livello lessicale il punteggio è pari 82%, a livello sintattico la complessità è inferiore (76%) ma risulta comunque abbastanza alta.
Andando a osservare la ripartizione interna del corpus rispetto ai contenuti, si nota che in generale i valori di leggibilità sono alti e che in tutti i sotto-corpora la difficoltà lessicale supera quella sintattica. Il corpus più semplice risulta quello dello screening oncologico, che registra i valori più bassi sia nel modello lessicale (74%) che in quello sintattico (67%). L'assistenza domiciliare restituisce invece tra i valori più alti in entrambi i modelli di analisi, in particolare nel livello lessicale, che sfiora quasi il valore massimo di difficoltà (96%). Gli altri due corpora presentano valori simili in entrambi i tipi di leggibilità.

Se si considerano invece i testi in base all'azienda sanitaria di appartenenza, si osserva che a livello lessicale i punteggi vanno da un minimo di difficoltà pari al 44% fino a un massimo pari a 99%, mentre a livello sintattico si parte da una soglia più alta (55%) per arrivare a un valore massimo di 96%. Rispetto alla distinzione dei testi relativa al contenuto, in cui si registra una difficoltà lessicale sempre maggiore rispetto a quella sintattica, in questo caso si ha una situazione di sostanziale equivalenza: in circa metà dei casi i punteggi più alti riguardano il piano lessicale e nell'altra metà il piano sintattico. L'ATS di Milano registra i valori più bassi in entrambi i livelli di analisi (44% e 55%), l'azienda sanitaria di Cuneo (ASL CN1) restituisce invece due tra i valori più alti nelle due categorie (97% e 95%).
Circa il fatto che possa esistere una certa omogeneità nella complessità dei testi in relazione al tipo di contenuto o alla ASL di appartenenza, la risposta è senz'altro negativa. Per quanto riguarda il contenuto, i corpora presentano infatti una notevole variazione interna, osservabile soprattutto a livello sintattico; l'unica eccezione è rappresentata dal corpus dell'assistenza domiciliare, in cui la complessità risulta più o meno sempre costante (con una deviazione standard pari a 0,08 per il livello lessicale e a 0,19 per il livello sintattico).
Anche per quanto riguarda le aziende sanitarie, si osserva un'alta variazione interna. A livello lessicale esiste una sostanziale omogeneità nel caso di valori più elevati: le ASL che registrano i valori più alti di leggibilità, possiedono valori alti in ogni testo. Il piano sintattico risulta invece più differenziato.

Sembrerebbe mancare anche una corrispondenza tra facilità/difficoltà nei due piani di analisi, cioè tra i testi che risultano più facili in un modello con quelli che risultano più semplici nell'altro. I documenti, infatti, sono molto diversificati tra loro rispetto ai due modelli: la maggior parte dei testi che riporta un basso valore a livello lessicale ha invece un valore alto o medio alto in quello sintattico, e viceversa. Ciò vale sia per l'intero corpus sia per i sotto-corpora distinti per contenuto; una maggiore correlazione si ha invece a livello di ASL di appartenenza: i testi che fanno parte di una data azienda sanitaria risultano quindi più semplici o complessi in entrambi i livelli a prescindere dal tipo di contenuto.
Il fatto che non vi sia una corrispondenza tra la difficoltà lessicale e quella sintattica può sembrare poco significativo ma porta invece a formulare importanti considerazioni, soprattutto nell'ottica della semplificazione linguistica. Spesso, infatti, si associa il concetto di semplificazione dei testi al solo piano lessicale, agendo per lo più sulla sostituzione di termini complessi e tecnicismi con parole più brevi e familiari.
Il fatto che i testi presentino valori così opposti rispetto ai due piani di analisi, ci porta invece alla conclusione che non è possibile semplificare un documento operando esclusivamente su un livello di analisi; gli interventi devono necessariamente andare in entrambe le direzioni. In particolare, non si può prescindere dalle trasformazioni a livello sintattico, né si devono limitare gli interventi ai soliti due o tre accorgimenti, come la riduzione della lunghezza delle frasi o la sostituzione delle frasi passive con quelle attive. Il monitoraggio delle caratteristiche linguistiche dei testi ci aiuta a individuare i parametri maggiormente legati alla complessità, così che possiamo tradurli in interventi di semplificazione.

I risultati dalla valutazione della leggibilità sono stati comparati anche con i punteggi ottenuti nel monitoraggio dalle diverse variabili linguistiche. Dal confronto è emerso che spesso a livelli alti di leggibilità non corrispondono valori altrettanto negativi dei parametri linguistici e che la difficoltà dei testi, sia essa lessicale o sintattica, è molte volte sovrastimata. Questa discrepanza è stata notata sia a livello generale per l'intero corpus che rispetto ai vari sotto-corpora, ma anche a livello dei singoli documenti, ad esempio mettendo a confronto i dati relativi al testo più semplice del corpus con quelli del testo che registra il più alto livello di complessità.
Ciò non significa che READ-IT non è uno strumento utile e adatto a valutare la leggibilità di testi appartenenti anche a diversi generi testuali ma che semplicemente i modelli di addestramento su cui è costruito presentano un profilo linguistico altamente differente dal nostro corpus. Gli strumenti di valutazione automatica della leggibilità sono infatti costruiti sulla base di un set di caratteristiche linguistiche che rappresentano uno specifico insieme di testi: l'algoritmo di apprendimento impara, in base all'esempio fornito e dunque ai corpora usati come modello, ad associare ogni vettore di caratteristiche che rappresenta un singolo testo a un livello di leggibilità. È dunque possibile che, applicato ad un corpus con un profilo linguistico molto diverso da quelli di addestramento, lo strumento non riesca ad associare in modo accurato i corretti livelli di difficoltà.

Alla luce di queste considerazioni, risulta quindi evidente la necessità, da una parte, di effettuare delle prove di comprensione sul corpus per ottenere dati più reali circa difficoltà di tali testi, dall'altra, di addestrare un nuovo sistema o ri-addestrare un modello già esistente in base alle caratteristiche linguistiche proprie di questi documenti e ai loro livelli di complessità.
Lasciamo la realizzazione di queste fasi del progetto tra i propositi per gli sviluppi futuri.
In particolare, ci auguriamo di poter condurre delle indagini sul campo per verificare l'effettiva comprensione dei testi da parte di un campione rappresentativo della popolazione. Tali prove di comprensione consentirebbero di individuare i livelli di lettura della popolazione attuale e dunque le classi di leggibilità dei testi che costituiranno gli standard di riferimento.
Il confronto con i dati emersi dal monitoraggio ci permetterebbe inoltre di costruire indici di correlazione tra le caratteristiche linguistiche dei testi e il loro livello di difficoltà, in modo da poter verificare quali siano i parametri maggiormente legati alla complessità testuale. Potremmo infatti costruire più modelli di addestramento, selezionando di volta in volta set diversi di caratteristiche fino ad ottenere i migliori punteggi di accuratezza e precisione; le possibilità sono molteplici: potremmo ad esempio costruire un sistema che si basa sull'intero set di caratteristiche, su un unico livello (lessicale, sintattico o morfosintattico) o su una loro combinazione; oppure sarebbe possibile scegliere soltanto le 5 o 10 o 20 caratteristiche che registrano i più alti valori di correlazione con la difficoltà, ecc.
Come abbiamo più volte sottolineato, lo sviluppo di un tale metodo di valutazione, oltre ad essere specificamente tarato su una varietà linguistica presente nei siti web, presenta il vantaggio di potersi riadattare in base a nuovi dati di input (ad esempio corpora appartenenti ad altre varietà testuali, altre tipologie testuali, ecc.) ed eventualmente a nuove classi di riferimento.

Tale tipo di strumento presenta, inoltre, numerosi campi di applicazione: potrebbe costituire un supporto alla produzione di testi in tutti gli ambiti cruciali per la comunicazione (istituzionale, sanitario, giornalistico, aziendale, educativo); potrebbe essere incorporato direttamente nei sistemi di recupero delle informazioni per personalizzare le ricerche degli utenti in base ai loro livelli di lettura o alla difficoltà dei testi. Potrebbe, infine, essere impiegato come ausilio per la semplificazione dei testi e, in particolare, nei sistemi di semplificazione automatica.

Riferimenti bibliografici

- ADAMIC E HUBERMAN 2002 = Lada A. Adamic, Bernardo A. Huberman, *Zipf's law and the Internet, Glottometrics*, 3, 2002, pp.143-150.
- AGGARWAL E ZHAI 2012 = C. C. Aggarwal, C. X. Zhai, *Mining text data*, in Springer Science & Business Media, 2012.
- AGRAWAL ET AL. 1988 = R. Agrawal, T. Imielinski, A. Swami, *Mining Association Rules between Sets of Items,* in Large Databases, SIGMOD, 1993.
- AL-BADI ET AL. 2005 = A. Al-Badi, P. Mayhew, A. Al-Solbi, *Readability formulas and the Web*. In Information Management in Modern Enterprise: Issues and Solutions - Proceedings of the 4th International Business Information Management Association Conference, IBIMA 2005, Vol. 1, pp. 317-322. International Business Information Management Association, IBIMA.
- AL-BADI ET AL. 2012 = Ali Al-Badi, Ali Saqib, Taiseera Al-Balushi, *Ergonomics of usability/accessibility-ready websites: Tools and guidelines*. In Webology, vol. 9, n.2, Dicembre 2012
- ALI ET AL. 2013= Ahmad Zamzuri Mohamad Ali, Rahani Wahid, Khairulanuar Samsudin, Muhammad Zaffwan Idris, *Reading on the Computer Screen: Does Font Type Has Effects on Web Text Readability?*, in International Education Studies, Vol. 6 n. 3, 2013, pp. 26 -35.
- ALPAYDIN 2004 = Ethem Alpaydin, *Introduction to Machine Learning*, MIT Press, 2004.
- ALUISIO ET AL. 2008 = Sandra Maria Aluísio, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick G. Maziero e Renata P. M. Fortes, *Towards Brazilian Portuguese Automatic Text Simplification Systems*, in Em Proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008), São Paulo, Brasil, 2008, pp. 240-248.
- ALUISIO ET AL. 2010 = S. Aluisio, L. Specia, C. Gasperin, C. Scarton, *Readability assessment for text simplification*, in J. Tetreault, J. Burstein & C. Leacock (Eds.), Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 1–9), Los Angeles, California: Association for Computational Linguistics, 2010.
- AMIZZONI E MASTIDORO 1993 = Maurizio Amizzoni, Nicola Mastidoro, *Linguistica applicata alla leggibilità: considerazioni teoriche e applicazioni*, in Bollettino della Società Filosofica Italiana, n. 149, maggio - agosto 1993, pp. 49 – 63.
- AMIZZONI ET AL. 2000 = Maurizio Amizzoni, Nicola Mastidoro, Patrizia Sposetti, *Il lessico dei giovani nella comunicazione in chat*, in Piemontese 2000b, pp. 145-163.
- ANDERSON 1965 = J. Anderson, *Research in Readability for the Classroom Teacher*, in Journal of Reading, vol. 8, n. 6, Marzo 1965, pp. 402-403, 405.
- ANDERSON 1967 = J. Anderson, *A scale to measure the reading difficulty of children's books*, University of Queensland, Faculty of Education, St Lucia, Australia, 1967.

- ANDERSON 1972 = J. Anderson, *The application of cloze procedure to English learned as a foreign language in Papua and New Guinea*, in English Language Teaching, 27, October 1972, pp- 66-72.
- ANDERSON 1983 = J. Anderson, *Lix and Rix*: variations on a little known readability index, in Journal of Reading, 26, 1983, pp.490-496.
- ARMBRUSTER ET AL. 1985 = Bonnie B. Armbruster, Jean H. Osborn, Alice L. Davison, *Readability Formulas May Be Dangerous to Your Textbooks*, in Educational Leadership, vol. 42, Issue 7, aprile 1985, pp. 18 -20.
- ATTARDI 2006 = Giuseppe Attardi, *Experiments with a Multilanguage non-projective dependency parser*. In Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06), New York City, New York, 2006, pp. 166–170.
- ATTARDI ET AL. 1998a = G. Attardi, G. Di Marco, G. Salvi, *Categorization by context*, in Proc. of WebNet, Orlando, USA, e in Journal of Universal Computer Science, 4(9), 1998, pp. 719–736.
- ATTARDI ET AL. 1998b = Giuseppe Attardi, Sergio Di Marco, David F. Salvi, Fabrizio Sebastiani, *Categorization by context,* in David Schwartz, Monica Divitini e Terje Brasethvik (eds.), Proceedings of the 1st International Workshop on Innovative Internet Information Systems (IIIS 1998*),* Pisa, Italia, 1998, pp. 1-13.
- ATTARDI ET AL. 1999 = Giuseppe Attardi, Antonio Gullì and Fabrizio Sebastiani, *Automatic Web page categorization by link and context analysis,* in Chris Hutchison and Gaetano Lanzarone (eds.), Proceedings of the 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence (THAI 1999), Varese, Italia, pp. 105-119.
- ATTARDI ET AL. 2009 = G. Attardi, F. Dell'orletta, M. Simi, J. Turian, *Accurate Dependency Parsing with a Stacked Multilayer Perceptron*, in "Proceedings of Evalita'09 (Evaluation of NLP and Speech Tools for Italian)", Reggio Emilia, 2009.
- AULA 2004 = A. Aula. *Enhancing the readability of search result summaries*. In Proc. of HCI, 2004.
- BAAYEN ET AL. 1995 = R. H. Baayen, R. Piepenbrock, L. Gulikers, *The CELEX lexical database* (CD-ROM). Philadelphia: University of Pennsylvania, Linguistic Data Consortium, 1995.
- BADARUDEEN E SABHARWAL 2010 = S. Badarudeen, S. Sabharwal, *Assessing Readability of Patient Education Materials: Current Role in Orthopaedics*, in Clinical Orthopaedics and Related Research, 2010, 468(10), pp. 2572-2580.
- BALDINI 2004 = M. Baldini, *Elogio dell'oscurità e della chiarezza*, Roma, Armando Editore, 2004.
- BAMBERGER 1973 = R. Bamberger, *Lese-Erziehung,* Vienna, Jugend und Volk, Austria, 1973.
- BAMBERGER E RABIN 1984 = R. Bamberger., A.T. Rabin, *New approaches to readability. Austrian research*, in The Reading Teacher, 37, 1984, pp. 512-519.
- BAMBERGER E VANECEK 1982 = R. Bamberger, E. Vanecek, *Die Lesbarkeit oder die Schwierigkeitsstufen von Texten in deutscher Sprache*, in Unpublished research study, Vienna, Austria, 1982.

- BAMBERGER E VANECEK 1984 = R. Bamberger, E. Vanecek, *Lesen-Verstehen-Lernen-Schreiben*, in Die Schwierigkeitsstufen von Texten in deutscher Sprache, Jugend und Volk, Vienna, Austria, 1984.
- BARBAGLI ET AL. 2014 = A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi, *Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado*, Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it), 9-10, Pisa, Italia, dicembre 2014.
- BARNI E PECCIANTI 1989 = Monica Barni, Maria Cristina Peccianti, *Il progetto «La lingua italiana per il made in Italy*», in Italiano lingua seconda: modelli e strategie per l'insegnamento, atti della Giornata di studi del Centro interfacoltà di ricerca sulla didattica delle lingue straniere moderne, Pavia, 15 dicembre 1989, a cura di Marco Mazzoleni e Maria Pavesi, 1991, pp. 175-186.
- BARONI E BERNARDINI 2004 = marco baroni, Silvia bernardini, *BootCaT: Bootstrapping Corpora and Terms from the Web*. In: LREC. 2004. p. 1313.
- BARONI ET AL. 2009 = M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora*, in "Language Resources and Evaluation" 43(3), 2009, pp. 209-226.
- BARZILAY E ELHADAD 2003 = R. Barzilay and N. Elhadad, *Sentence alignment for monolingual comparable corpora*, in Proc. of EMNLP, 2003, pp. 25-32.
- BARZILAY E LAPATA 2008 = Regina Barzilay, Mirella Lapata, *Modeling local coherence: An entity-based approach*. Computational Linguistics, 34(1), 2008, pp.1–34.
- BASILI E MOSCHITTI 2005 = Roberto Basili, Alessandro Moschitti, *Automatic Text Categorization: from Information Retrieval to Support Vector Learning*, ARACNE Editore, 2005.
- BASILI ET AL. 2006 = Roberto Basili, Marco Cammisa and Alessandro Moschitti, *A Semantic Kernel to Classify Texts with Very Few Training Examples*, in Informatica, vol. 30, n. 2, 2006, pp. 163–172.
- BEAUNOYER ET AL. 2016 = E. Beaunoyer, M. Arsenault, A. M. Lomanowska, M. J. Guitton *Understanding online health information: evaluation, tools, and strategies*, in Patient Education and Counseling, Vol. 100, Issue 2, pp. 183-189.
- BERARDI ET AL. 2015 = Giacomo Berardi, Andrea Esuli and Fabrizio Sebastiani, *Utility-Theoretic Ranking for Semiautomated Text Classification*, *in* ACM Transactions on Knowledge Discovery from Data, 10(1), articolo 6, 2015.
- BERIHUETE ET AL. 1992-1993 = Moro Berihuete, Pilar, Cabero Pérez, Mariví, Rodríguez Diéguez, José Luis, *Ecuaciones de predicción de lecturabilidad*, in Enseñanza: anuario interuniversitario de didáctica, 10-11, 1992-1993, pp. 47-64.
- BERLAND ET AL. 2001 = G. K. Berland, M. N. Elliott, L. S. Morales et al., *Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish*, in JAMA, 285(20), 2001, pp. 2612-2621.
- BHAGOLIWAL 1961 = B. S. Bhagoliwal, *Readability formulae: Their reliability, validity, and applicability in Hindi*, in Journal of Education and Psychology, 19, aprile 1961, pp. 13-26.
- BIAGIOLI ET AL. 1984 = Carlo Biagioli. Pietro Mercatali. Daniela Tiscornia Ghelli, *Banche di dati e divulgazione del diritto: modelli per analisi quantitative del*

linguaggio giuridico, in Informatica e diritto, vol. X, n. 2, maggio-agosto 1984, pp. 257-305.

- BIBER 1988 = Douglas Biber, *Variation across speech and writing*, Cambridge & New York, Cambridge University Press. 1988
- BIFFI 2014 = Marco Biffi, *Scrivere sul Web: qualche considerazione generale in prospettiva linguistica*, in A. Anichini, *Digital writing. Nel laboratorio della scrittura*, pp. 91-104, Sant'Arcangelo di Romagna, Maggioli, 2014.
- BILAL 2013 = Dania Bilal, *Comparing Google's Readability of Search Results to the Flesch Readability Formulae: A Preliminary Analysis on Children's Search Queries*, in Proceedings of the 76th ASIS&T Annual Meeting. Montreal, Canada: Information Today, 2013, pp. 1-9.
- BILAL E BOEHM 2013 = Dania Bilal, R. Boehm, *Towards New Methodologies for Assessing Relevance of Information Retrieval from Web Search Engines on Children's Queries*. Qualitative and Quantitative Research Methods in Libraries: An International Journal [Internet]. 2013, pp.93-100.
- BILAL E JACEK 2016 = Dania Bilal, Gwizdka Jacek, *Children's eye-fixations on google search results*, Proceedings of the Association for Information Science and Technology, 53, 1, 2003, pp. 1-6.
- BISHOP 2006 = C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- BJÖRNSSON 1968a = C.H. Björnsson, *Bokforlaget Liber*, Läsbarhet, Stockholm, Sweden, 1968.
- BJÖRNSSON 1968b = C.H. Björnsson, *Lesbarkeit durch Lix*, in Technical Report, Pedagogical Center, Stockholm, Sweden, n. 61, 1968.
- BJÖRNSSON 1974 = C.H. Björnsson, *Vas' barhetsprbvning av engelsk skoltext*, in Technical Report, n. 53, Pedagogical Center, Stockholm, Sweden, 1974.
- BJÖRNSSON 1983 = C.H. Björnsson, *Readability of newspapers in 11 languages*, in Reading Research Quarterly, vol. 18, n. 4, 1983, pp. 480-497.
- BOLASCO 2013 = S. Bolasco, *L'analisi automatica dei testi. Fare ricerca con il text mining*, Carocci, Milano, 2013.
- BORGHETTI ET AL. 2011 = C. Borghetti, S. Castagnoli, M. Brunello. *I testi del web: una proposta di classificazione sulla base del corpus PAISÀ*. In M. Cerruti, E. Corino, and C. Onesti, editors, *Formale e informale. La variazione di registro nella comunicazione elettronica*, Carocci, Roma, 2011, pp. 147–170.
- BORMUTH 1965 = John R. Bormuth, *Optimum Sample Size and Cloze Test Length in Readability Measurement*, in Journal of Educational Measurement, vol. 2, n. 1, gennaio 1965, pp. 111-116.
- BORMUTH 1966 = John R. Bormuth, *Readability: A New Approach*, in Reading Research Quarterly, vol. 1, n. 3, 1966, pp. 79-132.
- BORMUTH 1967 = John R. Bormuth, *Comparable Cloze and Multiple-Choice Comprehension Test Scores*, in Journal of Reading, vol. 10, n. 5, 1967, pp. 291-299.
- BORMUTH 1968 = John R. Bormuth, *Cloze Test Readability: Criterion Reference Scores*, in Journal of Educational Measurement, vol. 5, n. 3, 1968, pp. 189-196.

- BORMUTH 1969a = John R. Bormuth, *Development of readability analysis*, in Final Report, Project n. 7-0052, Contract n. OEC-3-7-070052-0326, U.S. Office of Education, Bureau of Research, U.S. Department of Health, Education, and Welfare, Washington D. C., 1969.
- BORMUTH 1969b = John R. Bormuth, *Factor Validity of Cloze Tests as Measures of Reading Comprehension Ability*, in Reading Research Quarterly, vol. 4, n. 3, 1969, pp. 358-365.
- BORMUTH 1971 = J. Bormuth, De*velopment of standards of readability: Towards a rational criterion of passage performance*, Washington, D. C.: U.S. Office of Education, Bureau of Research, U.S. Department of Health, Education, and Welfare, 1971.
- BORMUTH 1975 = John R. Bormuth, *The cloze procedure: Literacy in the classroom,* in W. D. Page (Ed.), Help for the reading teacher: New directions in research, Urbana, III: ERIC Clearinghouse on Reading, 1975, pp. 60-90.
- BORTOLINI E ZAMPOLLI 1971 = Umberta Bortolini, Antonio Zampolli, *Lessico di frequenza della lingua italiana contemporanea: prospettive metodologiche*, in Società di linguistica italiana, L'insegnamento dell'italiano in Italia e all'estero, in Atti del quarto convegno internazionale di studi, Bulzoni, Roma, 1-2 giugno 1970, Vol. II, 1971, pp.639-648.
- BORTOLINI ET AL. 1972 = U. Bortolini, C. Tagliavini, A. Zampolli*, Lessico di frequenza della lingua italiana contemporanea,* Garzanti, Milano,1972.
- BOYER 1992 = Jean-Yves Boyer, *La lisibilité*, in Revue française de pédagogie, vol. 99, 1992, pp. 5-14.
- BRIGHI ET AL. 2015 = Carla Faralli, Raffaella Brighi, Michele Martoni (a cura di), *Strumenti, diritti, regole e nuove relazioni di cura. Il paziente europeo protagonista nell'eHealth*, Giappichelli, Torino, 2015.
- BRIGO ET AL. 2015 = F. Brigo, W. M. Otte, S. C. Igwe, F. Tezzon, R. Nardone, *Clearly written, easily comprehended? The readability of websites providing information on epilepsy*, in Epilepsy & Behavior, vol. 44, pp. 35 – 39.
- BRIGO E ERRO 2015 = F. Brigo, R. Erro, The readability of the English Wikipedia article on Parkinson's disease, in Neurol Sci., 2015, vol. 36, 6, pp. 1045-1046.
- BRIEST 1974 = W. Briest, *Kann man Verständlichkeit messen?*, in Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, 1974, pp. 543-563.
- BRILL 1995 = E. brill, *Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging*, in Computational Linguistics, 21, 1995, pp. 543-566.
- BROUWER 1963 = R. H. Brouwer, *Onderzock naar de Lees-moeilijkheid van Nederlands Proza*, in Paedagogische Studlen, 40, 1963, pp. 454-46.
- BROWN 1984 = G. D. A. Brown, *A frequency count of 190,000 words in the London-Lund Corpus of English Conversation*. Behavior Research Methods, Instruments, & Computers, 16, 1984, pp. 502-532.
- BROWN 1998 = J. D. Brown, *An EFL readability index*, in JALT Journal, vol. 20, n. 2, 1998, pp. 7-36.

- BRUCE ET AL. 1981 = Bertram Bruce, Andee Rubin, Kathleen S. Starr, *Why Readability Formulas Fail*, in IEEE Transactions on Professional Communication, PC-24, 1981, pp. 50-52.
- BRUGNOLLI ET AL. 2014 = Anna Brugnolli, Giancarla Carraro, Luisa Saiani, *Leggibilità e comprensione delle linee guida sull'igiene delle mani: confronto tra le linee guida OMS (2009) e dei Centres for Disease Control (2002)*, in Assistenza Infermieristica e Ricerca, 2014, 33(4), pp. 183-188.
- BRUNATO 2014 = D. Brunato, *Complessità necessaria o stereotipi del "burocratese"? Un'indagine sulla leggibilità del linguaggio amministrativo da una prospettiva linguistico-computazionale*, in La Lingua Variabile nei testi letterari, artistici e funzionali contemporanei: analisi, interpretazione, traduzione: atti del XIII Congresso della SILFI (Palermo 22-24 settembre 2014), Centro di studi filologici e linguistici siciliani (edit.), Università degli studi di Palermo, 2014.
- BRUNATO E VENTURI 2014 = D. Brunato, G. Venturi, *Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici*, in Diritto, Linguaggio e tecnologie dell'informazione, fascicolo monografico di Informatica e Diritto, D. Tiscornia, F. Romano, M.T. Sagri (a cura di), n. 2014/1, pp. 111-142.
- BRUNATO E VENTURI 2016 = Dominique Brunato, Giulia Venturi, *Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici*, in Informatica e diritto, 23, 1 2014, pp. 111-142.
- BRUNATO ET AL. 2015 = D. Brunato, F. Dell'Orletta, G. Venturi, S. Montemagni, *Design and Annotation of the First Italian Corpus for Text Simplification*, in Proceedings of the 9th Linguistic Annotation Workshopthe (LAW'15), Denver, Colorado, USA, giugno 2015.
- BUCHANAN 1927 = Milton A. Buchanan, *A Graded Spanish Word Book*, University of Toronto Press, Toronto, Ontario, 1927 (nuova edizione 1941).
- CALÒ E FERRERI 1997 = Rosa Calò, Silvana Ferreri, *Il testo fa scuola. Libri di testo, linguaggi ed educazione linguistica*, Quaderni del Giscel n. 18, La Nuova Italia, Firenze, 1997.
- CARELL 1987 = Patricia L. Carrell, *Readability in ESL*, in Reading in a Foreign Language, vol.4, n.1, 1987, pp. 21-40.
- CARLONI 2005 = F. Carloni, *La legge di Zipf sul numero dei significati in italiano e inglese*, in De Mauro e Chiari 2005.
- CARROLL ET AL. 1971 = J. B. Carroll, P. Davies, B. Richman, *Word frequency book*, Boston: Houghton Mifflin, 1971.
- CARVER 1975-1976 = Ronald P. Carver, *Measuring Prose Difficulty Using the Rauding Scale*, in Reading Research Quarterly, vol. 11, n. 4, 1975 - 1976, pp. 660-685.
- CAVALLO ET AL. 2001 = V. Cavallo, M. R. D'Aprile, S. Lanciotti, L. Serianni, *Il referto radiologico e la sua leggibilità,* in La radiologia medica, 2001,101(5), pp. 321-5.
- CAYLOR ET AL. 1973 = J. S. Caylor, T. G. Sticht, L. C. Fox, J. P. Ford, *Methodologies for determining reading requirements of military occupational specialties: Technical report*, n. 73-5, Alexandria, VA: Human Resources Research Organization, 1973.

- CECI E MALERBA 2003 = M. Ceci, D. Malerba, *Web-pages Classification into a Hierarchy of Categories*, in F. Sebastiani (Ed.), Advances in Information Retrieval. Proceedings, Lecture Notes in Computer Science, 2633, Springer, Berlin, Germany, 2003, pp-. 57-72.
- CECI ET AL. 2002 = M. Ceci, D. Malerba, F. Esposito, *Mining HTML pages to support document sharing in a cooperative system*, in R. Unland, A. Chaudri, D. Chabane & W. Lindner (Eds.), XML-Based Data Management and Multimedia Engineering - EDBT 2002 Workshops, Lecture Notes in Computer Science, 2490, Springer, Berlin, Germany, 2002, pp. 190-201.
- CECI ET AL. 2003 = M. Ceci, F. Esposito, M. Lapi, D. Malerba, *Automated Classification of Web Documents into a Hierarchy of Categorie*, in Kłopotek M.A., Wierzchoń S.T., Trojanowski K. (eds) Intelligent Information Processing and Web Mining. Advances in Soft Computing, vol. 22, Springer, Berlin, Heidelberg, 2003.
- CHALL 1947 = Jeanne S. Chall, *This Business of Readability*, in Educational Research Bulletin, vol. 26, no. 1, gennaio 1947, pp. 1-13.
- CHALL 1958 = Jeanne S. Chall, *Readability: an Appraisal of Research and Application*, Ohio State University, Columbus, Ohio, 1958.
- CHALL 1988 = Jeanne S. Chall, The beginning years, in Zakaluk – Samuels, 1988.
- CHANG E LIN 2001 = Chih-Chung Chang, Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm
- CHELBA ET AL. 2012 = Ciprian Chelba, Dan Bikel, Maria Shugrina, Patrick Nguyen, Shankar Kumar, *Large Scale Language Modeling in Automatic Speech Recognition*, Google, 2012.
- CHEN ET AL. 2013 = Yu-Ta Chen, Yaw-Huei Chen, and Yu-Chih Cheng, *Assessing Chinese Readability using Term Frequency and Lexical Chain*, in Computational Linguistics and Chinese Language Processing, vol. 18, n. 2, giugno 2013, pp. 1-18.
- CHIARI 2002 = Isabella Chiari, *La procedura cloze, la ridondanza e la valutazione della competenza della lingua italiana*, in Italica, vol. 79, n. 4, Linguistics and Pedagogy, 2002, pp. 525-540.
- CHIARI 2007 = Isabella Chiari, *Introduzione alla linguistica computazionale*, Laterza, Bari, 2007.
- CICCARELLI E DE VINCENZI 1996 = Laura Ciccarelli, Marica De Vincenzi, *Lo studio della complessità linguistica in rapporto alle strategie cognitive di analisi del linguaggio*, in Colombo e Romani, 1996, pp. 219-230.
- CIMATTI 1986 = F. Cimatti, *L'esperienza del CUD nella redazione e somministrazione di testi didattico-scientifici*, in Linguaggi, III, n. 3, 1986, pp. 123-127.
- CLARKE 1980 = Mark A. Clarke, *The Short Circuit Hypothesis of ESL Reading - Or When Language Competence Interfereswith Reading Performance*, in The Modern Language Journal, vol. 64, n. 2, 1980, pp. 203-209.
- CLARKE ET AL. 2007 = C. L. A. Clarke, E. Agichtein, S. Dumais, R. W. White, *The influence of caption features on clickthrough patterns in web search*. In Proc. 30th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2007, pp. 135–142.

- COIRO E DOBLER 2007 = J. Coiro, E. Dobler, *Exploring the online reading comprehension strategies used by sixth-9 grade skilled readers to search for and locate information on the Internet,* in Reading Research Quarterly, 42(2), 2007, pp. 214-257.
- COLEMAN 1965 = E. B. Coleman, *On understanding prose: some determiners of its complexity*, NSF Final Report GB-2604, D.C.: National Science Foundation, Washington, 1965.
- COLEMAN E LIAU 1975 = M. Coleman, T. L. Liau, *A computer readability formula designed for machine scoring*, in Journal of Applied Psychology, 60, 1975, pp. 283-284.
- COLLINS-THOMPSON 2014 = Kevyn Collins-Thompson, *Computational Assessment of Text Readability: A Survey of Current and Future Research*, in ITL - International Journal of Applied Linguistics, vol. 165, Issue 2, 2014, pp. 97 –135.
- COLLINS-THOMPSON E CALLAN 2004 = Kevyn Collins-Thompson, Jamie Callan, *A Language Modeling Approach to Predicting Reading Difficulty*, in Proceedings of HLT-NAACL, 2004, pp. 193–200.
- COLLINS-THOMPSON E CALLAN 2005 = Kevyn Collins-Thompson, Jamie Callan, *Predicting reading difficulty with statistical language models*, in Journal of the American Society for Information Science and Technology, 56, n. 13, pp. 1448-1462.
- COLLINS-THOMPSON ET AL. 2011 = K. Collins-Thompson, P.N. Bennett, R.W. White, S. de la Chica, D. Sontag, *Personalizing web search results by reading level*, in Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11), ACM, New York, NY, USA, 2011, pp. 403–412.
- COLOMBO 1996 = Adriano Colombo, *Due e tre modi di non capire. Difficoltà di comprensione di testi argomentativi*, in Colombo, Romani, 1996, pp. 189-208.
- COLOMBO E ROMANI 1996 = Adriano Colombo, Werther Romani, *È la lingua che ci fa uguali,* in Lo svantaggio linguistico: problemi di definizione e di intervento, Quaderni del Giscel n. 16, La nuova Italia, Firenze, 1996.
- CONVERTINO ET AL. 1998 = G. Convertino, L. Di Pace, P. Leo, A. Maffione, D. Malerba, G. Vespucci, *Tecniche di web mining per supportare l'attività di navigazione in rete*, in Proceedings *of AICA '98*, Napoli, Italia, 1998, pp. 53-74.
- CORDA COSTA 1984 = Maria Corda Costa, *Comprensione dei testi scritti e capacità espressiva di riformulazione*, in Lettura e scrittura: proposte didattiche, 1984, pp. 207-216.
- CORNOLDI 1996 = Cesare Cornoldi, *Metacognizione e lettura,* in Colombo, Romani, 1996, pp. 85-105.
- CORNOLDI E COLPO 1981 = C. Cornoldi, G. Colpo, *La verifica della lettura*, Organizzazioni Speciali, Firenze, 1981.
- CORNOLDI E COLPO 1998 = C. Cornoldi, G. Colpo, *Prove di lettura MT per la scuola Elementare-2*, Organizzazioni Speciali, Firenze, 1998.
- CORNOLDI ET AL. 1981 = C. Cornoldi, G. Colpo, Gruppo MT, *La verifica dell'apprendimento della lettura*, Organizzazioni Speciali, Firenze, 1981.
- CORNOLDI ET AL. 1986 = C. Cornoldi, G. Colpo, Gruppo MT, *Prove di rapidità e correttezza nella lettura del gruppo MT*, Organizzazioni Speciali, Firenze, 1986.

- CORNOLDI ET AL. 2010 = C. Cornoldi, P.E. Tressoldi, N. Perini, *Valutare la rapidità e la correttezza della lettura di brani. Nuove norme e alcune chiarificazioni per l'uso delle prove MT*, in Dislessia, vol.1, 2010, pp. 89-100.
- COUPLAND 1978 = N. Coupland, *Is readability real?*, *in* Communication of scientific and technical information, aprile 1978, pp.15-17.
- CRAWFORD 1984 = Alan N. Crawdorf, *A Spanish Language Fry-type Readability Procedure: Elementary Level*, in Bilingual Education Paper Series, vol. 7, n. 8, 1984.
- CRAWFORD 1989 = Alan N. Crawoford, *Fórmula y gráfico para determinar la comprensibilidad de textos de nivel primario en castellano*, 1989.
- CRESTI E PANUNZI 2013 = Emanuela Cresti, Alessandro Panunzi, *Introduzione ai corpora dell'italiano*, Bologna, Il Mulino, 2013.
- CULHANE E RANKIN 1969 = Joseph W. Culhane, Earl F. Rankin, *Comparable Cloze and Multiple-Choice Comprehension Test Scores*, in Journal of Reading, vol. 13, n. 3, dicembre 1969, pp. 193-198.
- D'ACHILLE 2016= Paolo D'Achille (a cura di), *Grammatica e testualità : metodologia ed esperienze didattiche a confronto*, Atti del I Convegno-seminario dell'ASLI Scuola, Roma, Università Roma Tre, 25-26 febbraio 2015, Firenze, Franco Cesati, 2016.
- D'AGOSTINO 1998 = Emilio D'Agostino, *Il Lessico di Frequenza dell'Italiano Parlato e la didattica dell'italiano*, in Quaderns d'Italià, n. 3, U.A.B., 1998, pp. 9-28.
- D'ALESSANDRO ET AL. 2001 = D.M. D'Alessandro, P. Kingsley, J. Johnson-West, *The readability of pediatric patient education materials on the world wide web*, in Archives of pediatrics and adolescent medicine, 155(7), 2001, pp.807-12.
- D'AMORE E PINILLA 2016 = B. D'Amore, M. I. Fandiño Pinilla, *Una formula per la misurazione oggettiva della difficoltà di comprensione di un testo di matematica da parte degli studenti. Uso valutativo e uso didattico*, in La matematica e la sua didattica, 24(1-2), 2016, pp. 59-78.
- D'ANTONIS ONOFRI E SALERNI 1987 = D'Antonis Onofri, Salerni, *Padronanza lessicale e comprensione della lettura*, in La ricerca, novembre 1987, pp. 1-14.
- DAHLQVIST 1999 = Bengt Dahlqvist, *The Scarrie Swedish Newspaper Corpus*, in Working Papers in Computational Linguistics & Language Engineering, 1999.
- DALE E CHALL 1948a = Edgar Dale, Jeanne S. Chall, *A Formula for Predicting Readability* in Educational Research Bulletin, vol. 27, n. 1, gennaio 1948, pp. 11-20, 28.
- DALE E CHALL 1948b = Edgar Dale, Jeanne S. Chall, *A Formula for Predicting Readability: Instructions*, in Educational Research Bulletin, vol. 27, n. 2, febbraio 1948, pp. 37-54.
- DALE E CHALL 1995 = Edgar Dale, Jeanne S. Chall, *Manual for The new Dale - Chall Readability Formula*, in Edgar Dale and Jeanne S. Chall, *Readability revised: The new Dale - Chall Readability Formula*, Brookline Books, 1995.
- DALE E TYLER 1934 = Edgar Dale, Ralph W. Tyler, *A Study of the Factors Influencing the Difficulty of Reading Materials for Adults of Limited Reading Ability*, in The Library Quarterly: Information, Community, Policy, vol. 4, n. 3, luglio 1934, pp. 384-412.

- DANIELSON E BRYAN 1963 = A. Wayne Danielson, Sam Dunn Bryan, *Computer automation of two readability formulas*, in Journalism Quarterly, 39, 1963, pp. 201-206.
- DE ANTONI 1997 = Giuseppe De Antoni, *La lettura per lo studio: facilitazione del testo e mediazione dell'insegnante*, in Calò e Ferreri, 1997, pp. 295-315.
- DE GRAFENSTEIN E PIERDONATI 1985 = M. De Grafesnstein, S. Pierdonati, *La tecnica "cloze" come misura della leggibilità e come prova di comprensione della* lettura, in Lettura e scrittura: proposte didattiche, 1985, pp. 77–92.
- DE LANDSHEERE 1963 = G. De Landsheere, *Pour une application des tests de lisibilité de Flesch à Ia langue française*, in Le Travail Humain, 26, 1963, pp. 141-154.
- DE LANDSHEERE 1970 = G De Landsheere, *Einfuirung in die Padagogische Forschung*, (2e edition), Weinheim, Beltz, 1970, pp. 225-234.
- DE LANDSHEERE 1973 = G. De Landsheere, *Le test de closure. Mesure de la lisibilité et de la compréhension*, Nathan, Paris, France; Labor, Brussels, Belgium, 1973.
- DE LUISE ET AL. 2007 = G. De Luise, D. Malerba, G. Convertino, L. Di Pace, P. Leo, A. Maffione, *WebClass: A Web Mining Tool*, 2007.
- DEL VECCHIO E RAPPINI 2009 = Mario Del Vecchio e Valeria Rappini, *La comunicazione aziendale in sanità*, in L'aziendalizzazione della sanità in Italia. Rapporto OASI 2009, 2009, pp. 369-411.
- DE MAURO 1961 = Tullio De Mauro, *Statistica Linguistica*, in Enciclopedia Italiana, Appendice III, vol. 2, Roma, 1961.
- DE MAURO 1979 = Tullio De Mauro, *L'italiano dei non lettori*, in Problemi dell'informazione 3, 1979, pp.419–431.
- DE MAURO 1980 = Tullio De Mauro, *Guida all'uso delle parole*, Editori Riuniti, Roma, 1980.
- DE MAURO 1982 = Tullio De Mauro, *Minisemantica*, Laterza, Bari, 1982.
- DE MAURO 1984 = Tullio De Mauro, *Ai margini del linguaggio*, Editori Riuniti, Roma, 1984.
- DE MAURO 1985 = Tullio De Mauro, *Appunti e spunti in tema di (in)comprensione*, in Linguaggi, n. 3, 1985.
- DE MAURO 1989 = T. De Mauro, *Guida all'uso delle parole*, Editori Riuniti, Roma, 1989.
- DE MAURO 1994 = Tullio De Mauro, *Capire le parole*, Laterza, Bari, 1994.
- DE MAURO 2000 = Tullio De Mauro, *Il dizionario della lingua italiana*. Torino, Paravia, 2000.
- DE MAURO 2003 = Tullio De Mauro, *Guida all'uso delle parole*, 2003.
- DE MAURO 2007 = Tullio De Mauro, *Guida all'uso delle parole: parlare e scrivere semplice e preciso per capire e farsi capire*, Editori Riuniti, Roma, 2007.
- DE MAURO E CHIARI 2005 = Tullio De Mauro, I. Chiari (a cura di), *Parole e numeri*, in Analisi quantitative dei fatti di lingua, Aracne, Roma, 2005.
- DE MAURO ET AL. 1986 = T. De Mauro, M. E. Piemontese, M. Vedovelli, *Leggibilità e comprensione*, Atti dell'incontro di studio, Roma – Istituto di Filosofia, Villa Mirafiori, giugno 1986, in Linguaggi, III, n. 3, 1986.

- DE MAURO ET AL. 1988 = T. De Mauro, S. Gensini, M. E. Piemontese (a cura di), *Dalla parte del ricevente: percezione, comprensione, interpretazione*, in Atti del XIX Congresso Internazionale di Studi, Roma 8-10 novembre 1985, Bulzoni, Roma, 1988.
- DE MAURO ET. AL. 1993 = Tullio De Mauro, Massimo Vedovelli, Miriam Voghera, Federico Mancini, *Lessico di frequenza dell'italiano parlato*, Milano, Etaslibri, 1993.
- DEBOLE E SEBASTIANI 2004 = Franca Debole, Fabrizio Sebastiani, *Supervised Term Weighting for Automated Text Categorization (extended version)*, in Spiros Sirmakessis (ed.), Text Mining and its Applications*,* Physica-Verlag, Heidelberg, DE, 2004, pp. 81-98.
- DELL'ORLETTA 2009 = Felice Dell'Orletta, *Ensemble system for Part-of-Speech tagging*. In Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian, Reggio Emilia, December 2009.
- DELL'ORLETTA ET AL. 2011a = F. Dell'Orletta, S. Montemagni, G. Venturi, *READ-IT: assessing readability of Italian texts with a view to text simplification*, in Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011), Edinburgh, luglio 30, pp. 73-83.
- DELL'ORLETTA ET AL. 2011b= F. Dell'Orletta, S. Montemagni, E.M. Vecchi, G. Venturi, *Tecnologie linguistico–computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria*, in G.C. Bruno, I. Caruso, M. Sanna, I. Vellecco (Eds.), Percorsi migranti: uomini, diritto, lavoro, linguaggi, Milano, McGraw–Hill Editore, 2011, pp. 319-366.
- DELL'ORLETTA ET AL. 2012 = F. Dell'Orletta, S. Montemagni, G. Venturi, *Genre-oriented Readability Assessment: a Case Study*, In R. Mamidi & K. Prahallad (Eds.), Proceedings of the COLING-2012 Workshop on Speech and Language Processing Tools in Education (SLP-TED), December 2012, Mumbai, India, pp. 91-98.
- DELL'ORLETTA ET AL. 2013 = F. Dell'Orletta, S. Montemagni, G. Venturi, *Linguistic profiling of texts across textual genres and readability levels. An exploratory study on Italian fictional prose*. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pp. 189-197.
- DELL'ORLETTA ET AL. 2014a = F. Dell'Orletta, S. Montemagni, G. Ventur, *Assessing document and sentence readability in less resourced languages and across textual genres*, in International Journal of Applied Linguistics (ITL), Special Issue on Readability and Text Simplification. To appear, 2014.
- DELL'ORLETTA ET AL. 2014b = F. Dell'Orletta, M. Wieling, A. Cimino, G. Venturi, S. Montemagni, *Assessing the Readability of Sentences: Which Corpora and Features?*, in Proceedings of 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014), Baltimore, Maryland, USA, giugno 2014.
- DELL'ORLETTA ET AL. 2016 = Felice Dell'Orletta, Franca Orletti, Rossella Iovino, *La leggibilità dei testi di ambito medico rivolti al paziente: il caso dei bugiardini di farmaci senza obbligo di prescrizione medica,* In: Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016: 5-6 December 2016, Napoli.
- DELL'ORLETTA ET AL. 2017 = G. Venturi, F. Dell'Orletta, S. Montemagni, E. Flore, T. Bellandi, *La qualità dei consensi informati. Un'analisi linguistico-computazionale*

della leggibilità dei testi, in Salute e territorio, Anno XXXVIII, Fascicolo 212, – marzo 2017, pp. 35-39.

- DEMPSTER 1977 = P. Dempster, N. M. Laird, D. B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, in Journal of the Royal Statistical Society, serie B, vol. 39, n. 1, 1977, pp. 1-38.
- DEPARTMENT OF EDUCATION AND SCIENCE 1975 = *A language for life*, The Bullock Report. London: Her Majesty's Stationery Office, 1975.
- DIKES E STEIWER 1977 = P. Dikes, L. Steiwer, *Ausarbeitung von Lesbarkeitsformeln fur die deutsche Sprache*, in Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, vol. 9, 1977, pp. 20-28.
- DINI ET AL. (2017) = Guglielmo Dini, Nicola Luigi Bragazzi, Beatrice D'Amico, Alfredo Montecucco, Stanley C. Igwe, Francesco Brigo, Alessandra Toletone, Paolo Durando*, A reliability and readability analysis of silicosis-related italian websites: implications for occupational health* (*Analisi di affidabilità e leggibilità dei contenuti dei siti web italiani sulla silicosi: possibili implicazioni per la salute in ambito occupazionale*), in La medicina del lavoro, 2017, vol. 108, n. 3, pp. 167-173.
- DOUMA 1960 = W. H. Douma, *De Leesbaarheid van Land-bouwbladen: een Onderzoek naar en een Toepassing van Leesbaarheidsformules*, in Bulletin No 17, afd. Sociologie en Socio-graphie van de Landbouwhogeschool te Wageningen, 1960.
- DUBAY 2004 = W. H. Dubay, *The Principles of Readability. Costa Mesa*, CA: Impact Information, 2004.
- DUBAY 2006 = W. H. Dubay, *The Classic Readability studies*, Costa Mesa, CA: Impact Information, 2006.
- DUBAY 2007 = W. H. Dubay, *Smart Language: Readers, Readability, and the Grading of Text*, Mesa, CA: Impact Information, 2007.
- DULLI ET AL. 2004 = S. Dulli, P. Polpettini, M. Trotta, *Text Mining: teoria e applicazioni*, Franco Angeli, 2004.
- DUNN E MARKWARDT 1970 = L. M. Dunn, L. F. C. Markwardt, *Peabody Individual Achievement Test*, Circle Pines, MN: American Guidance Service, 1970.
- EDMUNDS ET AL. 2013 = M. R. Edmunds, R. J. Barry, A. K. Denniston, *Readability Assessment of Online Ophthalmic Patient Information*, in JAMA Ophthalmol. 2013, 131(12), pp. 1610–1616.
- ELTORAI ET AL. 2014 =A. E. M. Eltorai, S. Ghanian, C. A. Adams, C. T Born, A. H. Daniels, Readability of Patient Education Materials on the American Association for Surgery of Trauma Website. Archives of Trauma Research, 2014, 3(2), e18161.
- ERMAKOVA ET AL. 2015 = T. Ermakova, B. Fabian, E. Babina, *Readability of Privacy Policies of Healthcare Websites*, in Wirtschaftsinformatik, 2015, pp. 1085-1099.
- ESPOSITO ET AL. 1999 = F. Esposito, D. Malerba, L. Di Pace, P. Leo, *A Learning Intermediary for Automated Classification of Web Pages*, Proc. of the ICML'99 Workshop on Machine Learning in Text Data Analysis, Bled, Slovenia, 1999, pp. 37-46.
- ESPOSITO ET AL. 2000 = F. Esposito, D. Malerba, L. Di Pace, P. Leo, *A Machine Learning Approach to Web Mining*, in E. Lamma & P. Mello (Eds.), AI IA 99:

Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence,1792, Springer, Berlin, Germany, 2000, pp. 190-201.

- ESTRADA ET AL. 2000 = C. A. Estrada, M. M. Hryniewicz, V. B. Higgs, C. Collins; J. C. Byrd, *Anticoagulant Patient Information Material Is Written at High Readability Levels*, in Stroke, 31, 2000, pp. 2966–70.
- ETZIONI 1996 = O. Etzioni, *The World-Wide Web: Quagmire or Gold Mine?*, in Communications of the ACM 39, 1, gennaio 1996, pp. 65-68.
- FANG 1966-67 = I. E. Fang, *The Easy Listening Formula*, in Journal of Broadcasting, vol. 11, n.1, 1966-67, pp. 63-68.
- FANG 1968 = I. E. Fang, *By computer: Flesch's Reading Ease score and a syllable counter*, in Behavioral Science, vol. 13, n. 3, 1968, pp. 249-251.
- FARR E JENKINS 1949 = James N. Farr, James J. Jenkins, *Tables for Use with the Flesch Readability Formulas*, in Journal of Applied Psychology, XXXIII, giugno 1949, pp. 275-78.
- FARR ET AL. 1951 = James N. Farr, James J. Jenkins, D. G. Paterson, *Simplification of Flesch reading ease formula*, in Journal of Applied Psychology, n.35(5), 1951, pp. 333-37.
- FELLBAUM, 1998 = C. Fellbaum, *WordNet: An electronic lexical database*, Cambridge, MA: MIT Press, 1998.
- FENG ET AL. 2009 = Lijun Feng, Noemie Elhadad, Matt Huenerfauth, *Cognitively motivated features for readability assessment*, in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), 2009, pp. 229–237.
- FENG ET AL. 2010 = Lijun Feng, Martin Jansche, Matt Huenerfauth, No´emie Elhadad, *A comparison of features for automatic readability assessment*, in Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), 2010, pp. 276– 284.
- FERNANDEZ HUERTA 1959 = J. Fernández Huerta, *Medidas sencillas de lecturabilidad*, in Consigna (Revista pedagógica de la sección femenina de Falange ET y de las JONS), 214, 1959, pp. 29-32.
- FERRARI 1991 = Giacomo Ferrari, *Introduzione al natural language Processing*, Bologna, Calderini, 1991.
- FERRERI 1988 = S. Ferreri, *Il problema di matematica: un problema linguistico*, in Ferreri e Guerriero, 1988, pp. 317-329.
- FERRERI 2002 = S. Ferreri (a cura di), *Non uno di meno. Strategie didattiche per leggere e comprendere*, Firenze-Milano, La Nuova Italia, 2002.
- FERRERI E LUCISANO 1996 = S. Ferreri, P. Lucisano, *Indagine IEA sull'alfabetizzazione e svantaggio linguisti*, in Colombo, Romani, 1996, pp. 55-84.
- FERRRI E GUERRIERO 1998 = S. Ferreri, A. R. Guerriero (a cura di), *Educazione linguistica vent'anni dopo e oltre. Che cosa ne pensano De Mauro, Renzi, Simone, Sobrero*, La Nuova Italia, Firenze, 1998.
- FIORINI E PANINI 2014 = F. Fiorini, R. Panini, *La comunicazione delle Aziende Sanitarie*, in Giornale Italiano di Nefrologia, 31, 4, 2014.

- FIORUCCI 1982 = Tiziana Fiorucci, *Si raccomanda un periodare breve: la leggibilità dei libri di base secondo l'indice Flesch*, in E.D.A., A. 3, vol. 2, n. 6, novembre-dicembre 1982, pp. 39-55.
- FLESCH 1946 = Rudolf Flesch, *The art of Plain Talk*, Harper, New York, 1946.
- FLESCH 1948 = Rudolf Flesch, *A new readability yardstick*, in Journal of applied psychology, vol. 32(3), 1948, pp. 221-233.
- FLESCH 1950 = Rudolf Flesch, *Measuring the Level of Abstraction*, in Journal of Applied Psychology, XXXIV, 1950, pp. 384-90.
- FLINTON ET AL. 2018 = D. Flinton, M. Singh, K. Haria, *Readability of internet-based patient information for radiotherapy patients*, in Journal of Radiotherapy in Practice, 2018, 17(2), pp. 142-150.
- FLORANDER 1966 = J. Florander, M. Jansen, *Om letlaeseligheden af nogle dagbladespecielt af Berlingske Tidende* (About the readability of some daily papers especially of Berlingske Tidende), Copenhagen, 1966.
- FOLTZ 1996 = P. W. Foltz, *Latent semantic analysis for text-based research. Behavior Research Methods*, in Instruments, & Computers, 28, 1996, pp. 197-202.
- FRANCHINA E VACCA 1986 = V. Franchina, R. Vacca, *Taratura dell'indice di Flesch su testo di lingue italiano-inglese di unico autore*, in Linguaggi, III, n. 3, 1986, pp. 46-49.
- FRANCIS E KUCERA 1982 = W. N. Francis, H. Kucera, *Frequency analysis of English usage*, Boston: Houghton-Mifflin, 1982.
- FRANÇOIS 2009 = T. L. François, *Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL*, in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, 2009.
- FRANÇOIS E FAIRON 2012 = T. François, C. Fairon, *An "AI readability" formula for French as a foreign language*, In Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012), 2012, pp. 466-477.
- FRANÇOIS E MILTSAKAKI 2012 = T. François, E. Miltsakaki, *Do NLP and machine learning improve traditional readability formulas?,* in Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations, Association for Computational Linguistics, 2012, pp. 49-57.
- FRANÇOIS ET AL. 2014 = T. François, L. Brouwers, H. Naets, C. Fairon, *AMESURE: une plateforme de lisibilité pour les textes administratifs*, In Actes de la 21e Conférence sur le Traitement automatique des Langues Naturelles (TALN 2014), Marseille, 2014, pp. 467-472.
- FRIEDMAN 2001a = J. H. Friedman, *Greedy function approximation: A graidient boosting machine*. Annals of Statistics, 29, 2001, pp. 1189–1232.
- FRIEDMAN 2001b = J. H. Friedman, *Stochastic gradient boosting*. Computational Statistics and Data Analysis, 38, 2001, pp. 367–378.
- FRIEDMAN ET AL. 2004 = D.B. Friedman, L. Hoffman-Goeta, J.F. Arocha, *Readability of cancer information on the Internet*, J Cancer Educ, 2004, 19(2), pp. 117-22.
- FRY 1965 = E. B. Fry, *Teaching faster reading: A manual*, Cambridge University, Cambridge, 1965.

- FRY 1968 = E. B. Fry, *A Readability Formula That Saves Time,* in Journal of Reading, vol. 11, n. 7, 1968, pp. 513-516, 575-578.
- FRY 1969 A = E. B. Fry, *A readability graph validated at primary levels*, in The Reading Teacher, vol. 22, n. 6, 1969, pp. 534-538.
- FRY 1969 B = E. B. Fry, *A readability graph for librarians, part I*, in School Libraries, 19, 1969, pp. 13-16.
- FRY 1975 = E. B. Fry, *A Kernel Distance Theory for Readability,* in Reading in G. McNinch, W. D. Miller, *Reading: Convention and Inquiry*, Eds. 24th Yearbook, National Reading Conference, Clemson University, Clemson, S.C., 1975, pp. 252-254.
- FRY 1977a = E. B. Fry, *Elementary reading instruction*, Mc Graw Hill, New York, 1977.
- FRY 1977b = E. B. Fry, *Fry's readability graph: Clarifications, validity, and extension to level 17*, in Journal of Reading, vol. 21, n. 3, 1977, pp. 242-252.
- FRY 1987 = E. B. Fry, *The Varied Uses of Readability Measurement Today*, in Journal of Reading, vol. 30, n. 4, gennaio 1987, pp. 338-343.
- FRY 1988 = E. B. Fry, *Writeability: The Principles of Writing for Increased Comprehension*, in Zakaluk – Samuels, 1988.
- FRY 1990 = E. B. Fry, *A Readability Formula for Short Passages*, in Journal of Reading, vol. 33, n. 8, 1990, pp. 594-597.
- FRY 2002 = E. B. Fry, *Readability versus Leveling*, in The Reading Teacher, vol. 56, n. 3, 2002, pp. 286-291.
- FUCKS 1955 = W. Fucks, *Unterschiede des Prosastils von Dichtern and anderen Schriftstellern.Sprachforum*, Munster, Min, n. 3/4, 1955.
- GAGATSIS 1995 = Athanasios Gagatsis, *Modi di valutazione della leggibilità dei testi matematici*, in La matematica e la sua didattica, 9 (2), 1995, pp. 136-146
- GAGATSIS 1999 = Athanasios Gagatsis, *Come misurare la leggibilità di testi matematici*, Bologna, Pitagora, 1999
- GARAIS 2011 = E. G. Garais, Web Applications Readability, in Journal of Information Systems & Operations Management, 5, 1, 2011, pp. 114-120.
- GEATZ E ROIGER 2004 = R.J. Roiger, M.W. Geatz, *Introduzione al Data Mining*, McGraw-Hill, 2004.
- GEMOETS ET AL. 2004 = D. Gemoets, G. Rosemblat, T. Tse, R. Logan, *Assessing readability of consumer health information: an exploratory study*, Medinfo 2004, pp. 869-873.
- GILLIAM ET AL. 1980 = Bettye Gilliam, Sylvia C. Peña, Lee Mountain, *The Fry Graph Applied to Spanish Readability*, in The Reading Teacher, vol. 33, n. 4, gennaio 1980, pp. 426-430.
- GISCEL LOMBARDIA 1988 = *Analisi di manuali scientifici ed ipotesi di leggibilità*, in Guerriero 1988, pp. 239-266.
- GISCEL LOMBARDIA 1994 = *Leggibilità e comprensione del manuale di scienze*, in Zambelli 1994, pp. 15-96.
- GISCEL PIEMONTE 1997 = *Il difficile alfabeto del libro di scuola*, in Calò e Ferreri, 1997, pp. 241-260.

- GIULIANI ET AL. 2005 = Giuliani, C. Iacobini, A. M. Thornton, *La nozione di vocabolario di base alla luce della stratificazione diacronica del lessico dell'italiano*, in De Mauro – Chiari, 2005.
- GIULIANO 2013 = Luca Giuliano, *Il valore delle parole. L'analisi automatica dei testi in Web 2.0.*, Roma: Dipartimento di Scienze statistiche, 2013, pp. 116.
- GIULIANO E LA ROCCA 2008 = L. Giuliano, G. La Rocca, *L'analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso*, Milano: LED, 2008.
- GOTTRON 2007 = Thomas Gottron, *Evaluating content extraction on HTML documents*. Theoretical Informatics and Applications – ITA, 2007.
- GOTTRON 2008 = Thomas Gottron, *Content code blurring: A new approach to content extraction*. Database and Expert Systems Applications, International Workshop on, 2008, pp. 29-33.
- GOTTRON 2009 = Thomas Gottron, *Detecting website redesigns via template similarity on streams of documents*. Proceedings of the 3rd International Conference on Internet Technologies and Applications, ITA 09, 2009.
- GOTTRON E MARTIN 2009 = Thomas Gottron, Ludger Martin, *Estimating web site readability using content extraction*. In Proceedings of the 18th international conference on World wide web (WWW '09). ACM, New York, NY, USA, 2009, pp. 1169-1170.
- GOTTRON E MARTIN 2012 = Ludger Martin, Thomas Gottron, *Readability and the Web*, in Future Internet, 4, 2012, p. 238-252.
- GOUGENHEIM ET AL. 1967 = G. Gougenheim, P. Rivenc, R. Michea, A. Sauvageot, *L'élaboration du français fondamental*, in Didier, Paris, 1967, pp. 69-113.
- GRABER ET AL. 2002 = M. A. Graber, D. M. D'Alessandro, J. Johnson-West, *Reading level of privacy policies on Internet health Web sites*, in J Fam Pract. 51(7), 2002, pp. 642-5.
- GRADIŠAR ET AL. 2006 = M. Gradišar, I. Humar, T. Turk, *Factors Affecting the Readability of Colored Text in Computer Displays*, in Proc. 28th international conference on information technology interfaces, 2006, pp 245 – 250.
- GRAESSER E MCNAMARA 2011 = A. C. Graesser, D. S. McNamara, *Computational Analyses of Multilevel Discourse Comprehension*, Topics in Cognitive Science, 3(2), 2011, pp. 371– 398.
- GRAESSER ET AL 2004 = A. C. Graesser, D. S. McNamara, M. M. Louwerse, Z. Cai, *Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods*, Instruments and Computers, 36, 2, 2004, pp. 193-202.
- GRAESSER ET AL. 2011 = A. C. Graesser, D. S. McNamara, J. M. Kulikowich, *Coh-Metrix: Providing Multilevel Analyses of Text Characteristic*', Educational Researcher, 40(5), 2011, pp. 223– 234.
- GRAY 1947 = W. S. Gray, *Progress in the study of readability*, in The Elementary School Journal, University of Chicago Press, Chicago, vol. 47, n. 9, maggio 1947, pp. 491-499.
- GRAY 1958 = William S. Gray, *Summary of Reading Investigations*, 1956 - 1957, in The Journal of Educational Research, vol. 51, n. 6, febbraio 1958, pp. 401-435.

- GRAY E LEARY 1935 = Williams S. Gray, Bernice E. Leary, *What makes a book readable*, University of Chicago Press, Chicago, 1935.
- GRISHMAN 1986 = Ralph Grishman, *Computational linguistics: an introduction*, Cambridge, Cambridge University Press, 1986.
- GROEBEN 1972 = Groeben, *N.Die Verstandlichkeit von Unternchtstexten*, Dimensionen Kriterien rezeptiver Lernstadien.Miinster: Verlag Aschendorff, 1972.
- GROUPE DE RECHERCHE DU CFPJ 1983 = *Lisibilité et écriture télématique*, in Communication et langages, n. 56, 1983, pp. 64-83.
- GUERRIERO 1988 = Guerriero A. R. (a cura di), *L'educazione linguistica e i linguaggi delle scienze*, Firenze, La Nuova Italia, 1988.
- GUERRIERO E SAURO 2000 = A. R. Guerriero, F. R. Sauro, *Leggere ipertesti. Modalità di ricezione delle informazioni elaborate su supporto elettronico e cartaceo*, in Piemontese 2000b, pp. 105-128.
- GUIRAUD 1954a = P. Guiraud, *Bibliographie critique de la statistique linguistique*, Utrecht, 1954.
- GUIRAUD 1954b = P. Guiraud, *Problèmes et méthodes de la statistique linguistique*, Presses universitaires de France, 1960.
- GUNNEWEG ET AL. 1976 = G. Gunneweg, Peter Van Steen, Fritz Zondervan, *De Leesbaarheid van Basisschoolteksten. Objectiev. Ordeningscriteria voor instructieve teksten*, De Nieuwe Toalgids, 62, 1976, pp. 426-445.
- GUNNING 1952 = R. Gunning, *The technique of clear writing*, McGraw-Hill, New York, 1952, 1° edizione e 2° edizione.
- GUNNING 1969 = R. Gunning, *The fog index after twenty years*, in International Journal of Business Communication, vol. 6, n. 2, gennaio 1969, pp. 3-13.
- GUO ET AL. 2011 = S. Guo, G. Zhang, R. Zai, *Integrating readability index into Twitter search engine*, British Journal of Educational Technology, 42(5), 2011, pp. 103-105.
- GUTIÉRREZ DE POLINI 1972 = L. E. Gutiérrez de Polini, *Investigación sobre lectura en Venezuela*, Documento presentado a las Primeras Jornadas de Educación Primaria, Ministerio de Educación, Caracas, 1972.
- HANG 2011 = LI Hang, *A Short Introduction to Learning to Rank*, IEICE Transactions on Information and Systems, vol. E94{D, n.10, ottobre 2011.
- HANSBERRY ET AL. 2014 = D. R. Hansberry, A. John, E. John, N. Agarwal, S. F. Gonzales, S. R. Baker, *A Critical Review of the Readability of Online Patient Education Resources From RadiologyInfo.Org*, in American Journal of Roentgenology, 2014, 202,3, pp. 566-575.
- HANSBERRY ET AL. 2017 = D. R. Hansberry, A. John, E. John, N. Agarwal, P. Agarwal, J. C. Reynolds, S. B. Baker, *Evaluation of internet-based patient education materials from internal medicine subspecialty organizations: will patients understand them?*, in Internal and Emergency Medicine, 2017, Vol. 12, Issue 4, pp 535–543.
- HARMAN E LIBERMAN 1993 = D. Harman, M. Liberman, *TIPSTER Complete,* Linguistic Data Consortium, 1993.
- HARRISON 1980 = C. Harrison, *Readability in the classroom*, Cambridge, Cambridge Univesity Press, 1980.

- HEILMAN ET AL. 2006 = M. Heilman, K. Collins-Thompson, J. Callan, M. Eskenazi, *Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension*, Proceedings of the Ninth International Conference on Spoken Language Processing, 2006.
- HEILMAN ET AL. 2007 = M. Heilman Michael, K. Collins-Thompson, J. Callan, M. Eskenazi, *Combining lexical and grammatical features to improve readability measures for first and second language texts*, in Proceedings of NAACL HLT-2007, 2007, pp.460-467.
- HEILMAN ET AL. 2008 = M. Heilman, K. Collins-Thompson, E. Maxine, *An analysis of statistical models and features for reading difficulty prediction*, in Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (EANL '08), 2008, pp. 71–79.
- HEILMANN 1961 = M. Heilmann, *Statistica linguistica e critica del testo*, in Studi e problemi di critica testuale, Bologna, 1961, pp. 173-182.
- HEND ET AL. 2010 = S. Al-Khalifa Hend, A. Al-Ajlan Amani, *Automatic readability measurements of the arabic text: an exploratory study*, The Arabian Journal for Science and Engineering, vol. 35, n. 2C, 2010.
- HENRY 1979 = G. Henry, *Une méthode de mesure par ordinateur de la lisibilité des textes français*, in Scientia Paedagogica Experimentalis, 16, 1979, pp. 52-58.
- HENRY 1980 = Henry Georges, *Lisibilité et compréhension*, in Communication et langages, n. 45, 1er trimestre 1980, pp. 7-16.
- HUANG ET AL. 2015 = G. Huang, C.H. Fang, N. Agarwal, N. Bhagat, J.A. Eloy, P.D. Langer, *Assessment of Online Patient Education Materials From Major Ophthalmologic Associations*, in JAMA Ophthalmol., 2015, 133, 4, pp. 449–454.
- HUSSAIN ET AL. 2011 = W. Hussain, O. Sohaib, A. Ali, *Improving web page readability by plain language*, in IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011, pp. 315-319
- ILC-CNR 2012 = *READ-IT Documentazione Demo online*, 2012: http://moodle.humnet.unipi.it/pluginfile.php/13172/mod_resource/content/0/Documentazione%20READ-IT%20Demo-1.pdf
- IMPICCIATORE ET AL. 1997 = P. Impicciatore, C. Pandolfini, N. Casella, M. Bonati, *Reliability of health information for the public on the world wide web: sistematic survey of advice on managing fever in children at home*. Bmj vol. 314, pp. 1875-1879, 1997.
- INTRAVERSATO 2013 = Alessandra Intraversato, Pietro Lucisano, *Gli anni di Eco. Riflessioni sull'uso di prove strutturate con risposte chiuse e aperte a margine di una ricerca sulla comprensione della lettura*, ECPS - Educational, Cultural and Psychological Studies, Issue 7, 2013, pp. 23-43.
- INUI E YAMAMOTO 2001 = K. Inui, S. Yamamoto, *Corpus-based acquisition of sentence readability ranking models for deaf people*, in NLPRS, 2001, pp. 159-166.
- IRRSAE EMILIA-ROMAGNA 1995 = *Per capire di non capire*, a cura di L. Lumbelli e P. Senni, Bologna, Synergon.
- JAKOBSEN 1971 = G. Jakobsen, *Dansk Lix 70*, Laesepaedagogen, 1971.
- JAKOBSEN 1976 = G. Jakobsen, *Dansk Lix 75*, Laesepaedagogen, 1976.

- JAKOBSEN 1983 = G. Jakobsen, *Dansk Lix 83*, Laesepaedagogen, 1983.
- JANSEN 1987 = M. Jansen, *Danish Lix: A Danish readability formula*, Unpublished paper, Danish National Association of Reading Teachers, 1987.
- JOACHIMS 1998 = T. Joachims, *Text categorization with support vector machines: learning with many relevant features*, in Proc. of the European Conference on Machine Learning, 1998, pp. 137-142.
- JOHNSON 1930 = G.R. Johnson, *An objective method of determining reading difficulty*, in Journal of Educational Research, 21, 1930, pp. 283-287.
- KANE ET AL. 1974 = R. Kane, M. Byrne, M. Hater, *Helping children read mathematics*, American Book Company, New York, 1974
- KANUNGO E ORR 2009 = Tapas Kanungo, David Orr, *Predicting the readability of short web summaries*, in Proceedings of the Second ACM International Conference on Web Search and Data Mining. ACM, 2009, pp. 202-211.
- KATE ET AL. 2010 = Rohit J. Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos, Chris Welty, *Learning to predict readability using diverse linguistic features*, in Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 546–554.
- KICKMEIER E ALBERT 2003 = M. D. Kickmeier, D. Albert, *The effects of scanability on information search: An online experiment*. In Proc. of HCI, 2003.
- KIDWELL ET AL. 2011= Paul Kidwell, Guy Lebanon, Kevyn Collins-Thompson, *Statistical Estimation of Word Acquisition With Application to Readability Prediction*, in Journal of the American Statistical Association, vol. 106, no. 493, 2011, pp. 21–30.
- KINCAID ET AL. 1975 = J. Peter Kincaid, Lieutenant Robert P. Fishburne, Richard L. Rogers, Brad S. Chissom, *Derivation of new readability formulas for navy enlisted personnel*, in Research Branch Report, TN: Chief of Naval Training, Millington, 1975, pp. 8–75.
- KINTSCH E VIPOND 1979 = Walter Kintsch, Douglas Vipond, *Reading comprehension and readability in educational practice and psychological theory*, in L. Nilsson (ed.), Perspectives on Memory Research, 1979, pp. 329—365.
- KLARE 1952 = G. R. Klare, *A Table for Rapid Determination of Dale-Chall Readability Scores*, in Educational Research Bulletin, vol. 31, n. 2, 1952, pp. 43-47.
- KLARE 1959 = G. R. Klare, *Recensione a Chall 1958*, in Educational Research Bulletin, vol. 38, n. 2, 1959, pp. 49-50.
- KLARE 1963 = G. R. Klare, *The measurement of readability*, Iowa State University Press, Ames, Iowa, 1963.
- KLARE 1966 = G. R. Klare, *Comments on Bormuth's 'Readability: A New Approach'*, in Reading Research Quarterly, vol. 1, n. 4, 1966, pp. 119-125.
- KLARE 1968 = G. R. Klare, *The role of word frequency in readability*, in Elementary English, 45, 1968, pp. 12-22.
- KLARE 1969 = G. R. Klare, *Automation of the Flesch Reading Ease Readability Formula*, *with Various Options*, in Reading Research Quarterly, vol. 4, n. 4, 1969, pp. 550-559.

- KLARE 1974-75 = G. R. Klare, *Assessing Readability*, in Reading Research Quarterly, vol. 10, n.1, 197-1975, pp. 62-102.
- KLARE 1976 = G. R. Klare, *A second look at the validity of readability formulas*, in Journal of Reading Behavior, 8, 1976, pp. 129-152.
- KLARE 1984 = G. R. Klare, *Readability*, in Pearson, 1984.
- KLARE 1988 = G. R. Klare, *The formative years*, in Zakaluk – Samuels, 1988.
- KLARE 2000 = G. R. Klare, *Readable Computer Documentation*, in the *ACM Journal of Computer Documentation*, 2000.
- LAFFERTY E ZHAI 2017 = John Lafferty, Chengxiang Zhai, *Document Language Model*s, Query Models, and Risk Minimization for Information Retrieval, SIGIR Forum 51, 2, agosto 2017, pp. 251-259.
- LANDAUER E DUMAIS 1997 = T. K. Landauer, S. T. Dumais, *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge*. Psychological Review, 104, 1997, pp. 211-240.
- LANDAUER ET AL. 1998 = T. K. Landauer, P. W. Foltz, D. Laham, *An introduction to latent semantic analysis*. Discourse Processes, 25, 1998, pp. 259-284.
- LARSSON 2006 = P. Larsson, *Classification into Readability Levels: Implementation and Evaluation*, 2006.
- LAU E KING 2006 = T.P. Lau, I. King, *Bilingual Web Page and Site Readability Assessment*. In Proceedings of the 15th international conference on World Wide Web (WWW '06), Edinburgh, UK, 22–26 May 2006; ACM: New York, NY, USA, 2006; pp. 993–994.
- LAVRENKO E CROFT 2001 = Victor Lavrenko, W. Bruce Croft, *Relevance based language models*, in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01), ACM, New York, NY, USA, 2001, pp. 120-127.
- LEE 2011 = D. Y. W. Lee, *Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle*, in "Language Learning & Technology", 5(3), 2011, pp. 37-72.
- LENCI ET AL. 2005 = A. Lenci, S. Montemagni, V. Pirrelli, *Testo e computer. Elementi di linguistica computazionale*, Roma, Carocci, 2005.
- LIAU ET AL. 1976 = T. L. Liau, E. B. Bassin, C. J. Martin, E. B. Coleman, *Modification of the Coleman readability formulas*, in Journal of Reading Behavior, 8, 1976, pp. 381-386.
- LIU 2009 = Tie-Yan Liu, *Learning to Rank for Information Retrieval*, in Foundations and Trends in Information Retrieval, vol. 3, n. 3, 2009, pp. 225–331.
- LIU E CROFT 2005 = Xiaoyong Liu, W. Bruce Croft, *Statistical language modeling for information retrieval*, Annual Review of Information Science and Technology, vol. 39, 2005, pp. 1-31.
- LIU ET AL. 2004 = X. Liu, W. B. Croft, P. Oh, D. Hart, *Automatic recognition of reading levels from user queries*, in Proc. of SIGIR 2004, pp. 548-549.
- LIVELY E PRESSEY 1923 = Bertha A. Lively, L. Sidney Pressey, *A method for measuring the "Vocabulary Burden" of textbooks in Educational Administration and Supervision*, in Dubay 2006, n. 9, 1923, pp. 389-98.

- LÓPEZ RODRÍGUEZ 1981 = Natividad López Rodríguez*, Fórmulas de legibilidad para la lengua castellana*, in Tesi di dottorato, Università di Valencia, 1981.
- LÓPEZ RODRÍGUEZ 1983 = Natividad López Rodríguez, *Una técnica para medir la comprensión lectora: el test Cloze*, in Enseñanza: anuario interuniversitario de didáctica, (1), 1983, pp. 299-310.
- LORGE 1939 = Irving Lorge, *Predicting Reading Difficulty of Selections for Children*, in The Elementary English Review, vol. 16, n. 6, 1939, pp. 229-233.
- LORGE 1944 = Irving Lorge, *Predicting Readability*, Teacher's College Record, 40, pp. 404-419.
- LORGE 1948 = Irving Lorge, *The Lorge and Flesch Readability Formulae: A Correction*, in School and Society, n. 67, 1948, pp. 141-142.
- LORGE 1949 = Irving Lorge, *Readability Formulae - An Evaluation*, in Elementary English, vol. 26, n.2, 1949, pp. 86-95.
- LORGE E THORNDIKE 1944 = Irving Lorge, Edward L. Thorndike, *The Teacher's word book of 30.000 words*, Bureau of Publication, Teacher's College, Columbia University, New York, 1944.
- LUCISANO 1985 = Pietro Lucisano, *Leggibilità e lettura in Lettura e scrittura: proposte didattiche*, 1985, pp. 93-103.
- LUCISANO 1989a = Pietro Lucisano, *Scrittura e comprensione*, Torino, Loescher, 1989
- LUCISANO 1989b = Pietro Lucisano, *Il cloze,* In P. Lucisano, A. Salerni, G. Benvenuto, M.T. Siniscalco (a cura di), *Lettura e comprensione*, Torino: Loescher, pp. 152-173.
- LUCISANO 1990 = Pietro Lucisano, *Misurare le parole per farsi capire*, in Rivista IBM, III, 1990, pp. 57-68.
- LUCISANO 1992 = Pietro Lucisano, *Misurare le parole*, Roma, Kepos Edizioni, 1992.
- LUCISANO 1994 = Pietro Lucisano (a cura di), *Alfabetizzazione e lettura in Italia e nel mondo*, Napoli, Tecnodid, 1994.
- LUCISANO E PIEMONTESE 1986 = P. Lucisano, M. E. Piemontese, *Leggibilità dei testi e comprensione della lettura* in Linguaggi, III, n. 3, 1986, pp. 28-38.
- LUCISANO E PIEMONTESE 1988 = P. Lucisano, M. E. Piemontese, *GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana*, in «Scuola e città», n. 3, marzo 1988, pp. 110-124.
- LUMBELLI 1984a = Lucia Lumbelli, *Per la diagnosi della comprensibilità*, in Riforma della scuola, n. 5, 1984.
- LUMBELLI 1984b = Lucia Lumbelli*, Effetti paradossali dell'intenzione di farsi capire*, in Riforma della scuola, n. 9-10, 1984.
- LUMBELLI 1984c= Lucia Lumbelli*, Comprensibilità e qualità dei testi divulgativi*, in Riforma della scuola, n. 12, 1984.
- LUMBELLI 1986 = Lucia Lumbelli, *Il problema della soglia tra comprensione e incomprensione: linguistica e psicologia cognitivista*, in Linguaggi, III, n. 3, 1986, pp.17-27.
- LUMBELLI 1989 = Lucia Lumbelli, *Fenomenologia dello scrivere chiaro*, Roma, Editori Riuniti, 1989.

- LUMBELLI 1996 = Lucia Lumbelli, *Quando la differenza è deprivazione e il recupero è rispetto della differenza*, in Colombo, Romani, 1996, pp. 107-129.
- LUNZER E GARDNER 1979 = E. A. Lunzer, W.K. Gardner, (Eds.) *The effective use of reading*, London: Heinemann, 1979.
- LYDING ET AL. 2014 = V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, V. Pirrelli, *The PAISÀ Corpus of Italian Web Texts*, in Proceedings of 9th workshop on Web as Corpus (WAC-9), Gothenburg, Sweden, April 2014.
- MACGINITIE E TRETIAK 1971 = W. MacGinitie, R. Tretiak, S*entence depth measures as predictors of reading difficulty, in Reading research quarterly*, 6, 1971, pp. 364-376.
- MACQUEEN 1999 = J.B. MacQueen, *Some Methods for Classification Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1967, University of California Press, 1:281-297.
- MANEVITZ E YOUSEF 2001 = Larry M. Manevitz, Malik Yousef, *One-Class SVMs for Document Classification*, Journal of Machine Learning Research 2 (2001), pp. 139-154.
- MANNING E SCHUTZE 1999 = C. Manning, H. Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, 1999.
- MANZO 1986 = Anthony V. Manzo, *Readability: a postscript,* in Elementary English, 1968, pp. 963-965.
- MARCONI ET AL. 1994 = Lucia Marconi, Michela Ott, Elia Pesenti, Daniela Ratti, Mauro Tavella, *Lessico Elementare*. Zanichelli, Bologna, 1994.
- MARELLO 1984 = Carla Marello, *Fare buchi nei testi e poi riempirli. Il cloze nell'insegnamento dell'italiano come lingua madre - I e II parte*, in Lend -Lingua e nuova didattica, vol. 13, n. 2-3, 1984.
- MARELLO 1989a = Carla Marello, *Alla ricerca della parola nascosta*, Quaderni del Giscel, n. 6, La Nuova Italia, Firenze, 1989.
- MARELLO 1989b = Carla Marello, *Le lacune aiutano a capire il testo, in italiano e oltre*, a. 4, n. 5, 1989, pp. 225-230.
- MARELLO E MONDELLI 1991 = C. Marello, G. Mondelli (a cura di), *Riflettere sulla lingua*, Firenze, La Nuova Italia, 1991.
- MARINELLI ET AL. 2003 = R. Marinelli, L. Biagini, R. Bindi, S. Goggi, M. Monachini, P. Orsolini, E. Picchi, S. Rossi, N. Calzolari, A. Zampolli, *The Italian PAROLE corpus: an overview*. In Zampolli A. et al. (eds.), Computational Linguistics in Pisa, Special Issue, XVI–XVII, Pisa-Roma, IEPI. Tomo I, 2003, pp..401–421.
- MARON 1961 = M. E. Maron, *Automatic indexing: an experimental inquiry*, Journal of the Association for Computing Machinery, 8:404{417, 1961.
- MARTINET 1955 = André Martinet, *Economie des changements phonétiques. Traité de phonologie diachronique*, Bern, France, 1955.
- MARTINO E BIANUCCI 1986 = Antonio Martino, Gabriele Bianucci, *Analisi della leggibilità di testi politici* in Linguaggi, III, n. 3, 1986, pp. 56-63.

- MARTINO ET AL. 1986 = Antonio Martino, Gabriele Bianucci, Pietro Mercatali, Daniela Tiscornia, T*rattamento automatico del linguaggio giuridico e politico: analisi di leggibilità. Riflessioni e proposte*, in Linguaggi, III, n. 3, 1986, pp. 50-55.
- MARTINS ET AL. 1996 = Teresa B. F. Martins, Claudete M. Ghiraldelo, Maria das Graças Volpe Nunes e Osvaldo Novais de Oliveira Junior, *Readability formulas applied to textbooks in brazilian portuguese*, Notas do ICMC, N. 28, 1996, 11p.
- MASONI E GUELFI 2017a = M. Masoni, M.R. Guelfi, *La comprensibilità dell'informazione sanitaria in rete*, in La Professione, n. 2, 2017, pp. 113 – 120.
- MASONI E GUELFI 2017b = M. Masoni, M.R. Guelfi, *Going beyond the concept of readability to improve comprehension of patient education materials*, Intern Emerg Med, 2017, 12, 4, pp. 531‑533.
- MASONI ET AL. 2014 = M. Masoni, M.R. Guelfi, A. Conti, G. F. Gensini, *La qualità dell'informazione sanitaria in rete* in L'Infermiere, Vol. I, 2014, pp. 12-21.
- MASONI ET AL. 2017a = M. Masoni, M.R. Guelfi, S. Balzanti, *Il concetto di Readability*, in Toscana medica, Anno XXXV, n.7, 2017, p. 28-29.
- MASONI ET AL. 2017b = M. Masoni, M.R. Guelfi, S. Balzanti, *La Readability delle informazioni sanitarie in rete*, in Toscana medica, Anno XXXV, n.8, 2017, p. 23-24.
- MASTIDORO 1992 = Nicola Mastidoro, *Il sistema Eulogos per la valutazione automatica della leggibilità*, in: LUCISANO P. (a cura di), 1992, pp. 125-141.
- MASTIDORO 1996 = Nicola Mastidoro, *Leggibilità e lessico: il controllo con Eulogos Censor in Cattaneo*, Il cosmonauta. Guida per l'insegnante, Elmedì, Milano, 1996 (nuova ediz. 2003).
- MASTIDORO E AMIZZONI 2005 = Nicola Mastidoro, Maurizio Amizzoni, *Strumenti automatici di analisi e gestione testuale: IntraText, UTM e Censor* in Tullio De Mauro, Isabella Chiari (a cura di), *Parole e numeri. Analisi quantitative dei fatti di lingua*, Aracne editrice, Roma, 2005, pp. 417-438.
- MAXWELL 1979 = M. Maxwell, *Readability: have we gone too far?*, in Journal of Reading, vol. 21, n. 6, 1979, pp. 525-530.
- MCCLUNG ET AL. 1998 = H. J. McClung, R. D. Murray, L. A. Heitlinger, *The Internet as a source for current patient information*, in Pediatrics 101(6) E2, 1998.
- MCCORD 1989 = Michael C. McCord, *Slot grammar: A system for simpler construction of practical natural language grammars*, In Proceedings of the International Symposium on Natural Language and Logic, 1989, pages 118–145.
- MC LAUGHLIN 1969 = G. H. Mc Laughlin, *SMOG Grading-a New Readability Formula*, in Journal of Reading, vol. 12, n. 8, maggio 1969, pp. 639-646.
- MCNAMARA E GRASSER 2012 = Danielle S. McNamara, Arthur C. Graesser, *Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing*, In P. M. McCarthy & C. Boonthum (Eds.), Applied natural language processing: Identification, investigation, and resolution, Hershey, PA: IGI Global, 2012.
- MEHLER ET AL. 2010 = Alexander Mehler, Serge Sharoff, Marina Santini, Genres on the web: computational models and empirical studies, Sordrecht, Springer, 2010.
- MEHRYAR ET AL. 2012 = Mohri Mehryar, Rostamizadeh Afshin, Talwalkar Ameet, *Foundations of Machine Learning*, MIT Press, 2012.

- MELUCCI 1998 = Massimo Melucci, *Passage retrieval: a probabilistic technique*, in Information Processing & Management, vol. 34, n. 1, 1998, pp. 43-68.
- MELUCCI 2012 = Massimo Melucci, *Contextual Search: A Computational Framework*, in Foundations and TrendsR_ in Information Retrieval, vol. 6, n. 4–5, 2012, pp. 257–405.
- MELUCCI 2013 = Massimo Melucci, *Information retrieval*. *Metodi e modelli per i motori di ricerca*, Milano, Franco Angeli, 2013.
- MERCATALI 1986 = Pietro Mercatali, S*trumenti automatici per il controllo della leggibilità di testi giuridici*, in Linguaggi, III, n. 3, 1986., pp. 64-76.
- MERCATALI 1988 = Pietro Mercatali (a cura di), *Computer e linguaggi settoriali. Analisi automatica di testi giuridici e politici*, Milano, Franco Angeli, 1988.
- MERCATALI ET AL. 1979 = Pietro Mercatali, Sandro Ricci, Pierluigi Spinosa, *Un esperimento per il controllo automatico della leggibilità dei documenti di un archivio elettronico di dati giuridici*, in Informatica e diritto, anno V, n. 2, aprile-giugno 1979, pp. 145-155.
- MILLER 1967 = George R. Miller, E. B. Coleman, *A set of thirty-six passages calibrated for complexity*, in Journal of Verbal Learning and Verbal Beahavior, 6, 1967, pp. 851-854.
- MILLER 1972 = George A. Miller, *Linguaggio e comunicazione*, in La Nuova Italia, 1972 [l'edizione originale in inglese è del 1951].
- MILLER 1974 = Lawrence R. Miller, *Predictive powers of the Dale-Chall and Bormuth readability formulas*, in International Journal of Business Communication, vol. 11, n. 2, 1974, pp. 21-30.
- MILLER 1983 = George A. Miller, *Linguaggio e parola*, Il mulino, 1983 [l'edizione originale in inglese è del 1981].
- MILLER E COLEMAN 1967 = G. R. Miller, E. B. Coleman, *A set of thirty-six passages calibrated for complexity*, in Journal of Verbal Learning and Verbal Beahavior, 6, 1967, pp. 851-854.
- MILLER E KINTSCH 1980 = James R. Miller, Walter Kintsch, *Readability and Recall of Short Prose Passages: A Theoretical Analysis*, in Journal of Experimental Psychology: Human Learning and Memory, 6(4), 1980, pp. 335-354.
- MILLER ET AL. 1990 = G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, *Five papers on WordNet* (Tech. Rep. No. 43). Princeton, NJ: Princeton University, Cognitive Science Laboratory, 1990.
- MILONE 2014 = Michael Milone, *Development of the Atos ™ Readability Formula Written for Renaissance Learning* ™, (2014).
- MILTSAKAKI E TROUTT 2007 = Eleni MiltsakakI, Audrey Troutt, *Read-X: Automatic Evaluation of Reading Difficulty of Web Text*, in the Proceedings of E-Learn 2007, Quebec, Canada.
- MILTSAKAKI E TROUTT 2008 = Eleni Miltsakaki, Audrey Troutt, *Real Time Web Text Classification and Analysis of Reading Difficulty*, in the Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications, Columbus, OH, 2008.

- MISRA ET AL. 2012 = P. Misra, K. Kasabwala, N. Agarwal, J.A. Eloy, J. K. Liu, *Readability Analysis of Internet-Based Patient Information Regarding Skull Base Tumors*, in Journal of Neuro-Oncology, vol. 109 (3), 2012, pp. 573-580.
- MISRA ET AL. 2013 = P. Misra, K. Kasabwala, N. Agarwal, J.A. Eloy, D. R. Hansberry, M. Setzen, Readability Analysis of Healthcare-Oriented Education Resources from the American Academy of Facial Plastic and Reconstructive Surgery, in Laryngoscope, 2013, 123(1), pp. 90-96.
- MITCHELL 1997 = T. Mitchell, *Machine Learning*, McGraw Hill, 1997, p. 2.
- MITCHELL 2006 = T. Mitchell, *The discipline of machine learning*, (CMU ML-06 108), Carnegie Mellon University, 2006.
- MONEGLIA E PALADINI 2010 = Moneglia Massimo, Paladini Samuele*, Le risorse di rete dell'italiano. Presentazione del progetto RIDIRE*, in Cresti E. e I. Korzen (a cura di), *Language, Cognition and Identity*, Firenze, Firenze University Press, 2010, pp. 111-128.
- MONTEMAGNI 2013a = Simonetta Montemagni, *Tecnologie linguistico–computazionali e monitoraggio della lingua italiana*, in Studi Italiani di Linguistica Teorica e Applicata (SILTA), 42.1, 2013, pp.145-172.
- MONTEMAGNI 2013b = Simonetta Montemagni, *Estrazione Terminologica Automatica e Indicizzazione: Scenari Applicativi, Problemi e Possibili Soluzioni*, in Guarasci Roberto; Folino Antonietta, *Documenti Digitali*, Milano, Iter, 2013, pp. 241-284.
- MORONY ET AL. (2015) = S. Morony, M. Flynn, K. J. McCafFery, J. Jansen, A. C. Webster, *Readability of Written Materials for CKD Patients: A Systematic Review*, in American Journal of Kidney Diseases, Volume 65, Issue 6, pp. 842-850.
- MORLES 1981 = A. S. Morles, *Medición de la comprensibilidad de materiales escritos mediante pruebas cloze*, in Lectura y Vida, Año 2, n. 4, dicembre 1981, pp. 16-18.
- MÜHLENBOCK E KOKKINAKIS 2009 =Katarina Mühlenbock, Sofie Johansson Kokkinakis, *LIX 68 revisited - An extended readability measure*, 2009.
- NAHSHON 1957 = S. Nahshon, *Readability measurement of Hebrew prose*, Unpublished doctoral dissertation, Teachers College, Columbia University, 1957.
- NESTLER 1977 = Käte Nestler, *Zur Ermittlung von Voraussetzungen für die Formulierung von Normen für die angemessene sprachliche Gestaltung von Lehrtexten*, in Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, 1977.
- NEW ET AL. 2007 = B. New, M. Brysbaert, J. Veronis, C. Pallier, The use of film subtitles to estimate word frequencies. Applied Psycholinguistics, 28(04), 2007, pp.661–677.
- NGUYEN E HENKIN 1982 = Liem T. Nguyen, Alan B. Henkin, *A readability formula for Vietnamese. Journal of Reading*, 1982, 26, pp. 243-251.
- NGUYEN E HENKIN 1985 = Liem T. Nguyen, Alan B. Henkin, *A Second Generation Readability Formula for Vietnamese*, in Journal of Reading, vol. 29, n. 3, dicembre 1985, pp. 219-225.

- NUCCORINI 2001 = Stefania Nuccorini, *Il cloze test in inglese: ricerca, metodologia, didattica*, Roma, Carocci, 2001.
- OBENDORF E WEINREICH 2003 = H. Obendorf, H. Weinreich, Comparing link marker visualization techniques: Changes in reading behavior. In Proc. of 12thIntl. Conference on the World Wide Web, 2003, pp. 736–745.
- OERMAN ET AL. 2003 = M.H. Oermann, N.F. Lowery, J. Thornley, E*valuation of Web sites on management of pain in children*, in Pain Manag Nurs, settembre 2003 Sep, 4 (3), pp. 99-105.
- OLLER 1972 = John W. Oller, Jr, *Assessing Competence in ESL: Reading*, in TESOL Quarterly, vol. 6, n. 4, dicembre 1972, pp. 313-323.
- PAASCHE-ORLOW ET AL. 2003 = Michael Paasche-Orlow, Holly Taylor, Frederick Brancati, *Readability Standards for Informed-Consent Forms as Compared with Actual Readability*, in The New England journal of medicine, 348, 2003, pp. 721-726.
- PAIVIO ET AL. 1968 = A. Paivio, J. C. Yuille, S. A. Madigan, *Concreteness, imagery and meaningfulness values for 925 words*, in Journal of Experimental Psychology Monograph Supplements, 76 (3, Part 2), 1968.
- PALERMO 2013 = Massimo Palermo, *Linguistica testuale dell'italiano*, Bologna, Il Mulino, 2013
- PLAERMO 2016 = Massimo Palermo, *Testi cartacei e digitali: una sfida per il docente di italiano*, in D'Achille 2016, pp. 25-37.
- PALOMBI 1986 = A. Palombi, *L'esperienza di «Spazio linguistico*»: appunti per una riflessione sugli strumenti di calcolo della leggibilità, in Linguaggi, III, n. 3, 1986, pp. 128-135.
- PALOMBI E RAPONI 1984 = Andrea Palombi, Lorenza Raponi, *A proposito di misurazione della leggibilità di un testo*, in Linguaggi, numero unico in attesa di autorizzazione, 1984.
- PANINI E FIORINI 2014 = Roberta Panini, Fulvio Fiorini, *La comunicazione delle Aziende Sanitarie*, in Giornale Italiano di Nefrologia, n. 31, 4, 2014.
- PARK 1974 = Y Park, *An analysis of some structural variables of the Korean language and the development of a readability formula for Korean textbooks*, in Dissertation Abstracts, 1974, 35, 946A.
- PASSAPONTI 1979 = Emilia Passaponti, *Per misurare l'oscurità c'è un metodo: eccolo*, in La Repubblica: dossier, settembre 1979.
- PASSAPONTI 1980 = Emilia Passaponti, *Comprensione del testo nella primaria*, in Riforma della scuola, vol. 2-3, 1980, pp. 51-54 e *Leggibilità del giornale*, in Riforma della scuola, vol. 7-8, 1980, pp. 10-12.
- PASTOR ET AL. 1971 = A. Pastor, R. Guzman Vda de Capo, C. Gomez Tejera, K.B. Hester, *Por el mundo del cuento y la aventura,* in River Forest, IL: Laidlaw Brothers, 1971.
- PATEL ET AL. 2015 = Chirag R. Patel, Saurin Sanghvi, Deepa V. Cherla, Soly Baredes, Jean Anderson Eloy, *Readability Assessment of Internet-Based Patient Education Materials Related to Parathyroid Surgery*, in Annals of Otology, Rhinology & Laryngology, Vol. 124(7), 2015, pp. 523-527.

- PATTERSON 1972 = F. S. Patterson, *Cómo escribir para ser entendido*, in El Paso, Casa Bautista de Publicaciones, Texas, 1972.
- PAUL 2003 = T. Paul, *Guided Independent Reading*, Madison, WI: School Renaissance Institute, 2003.
- PEARSON 1984 = D. P. Pearson, *Handbook of reading research*, Longman inc, New York, 1984
- PENNISI 1986 = A. Pennisi, *Leggibilità e computabilità*, in Linguaggi, III, n. 3, 1986, pp. 88-99.
- PERUGINELLI E RAGONA 2014 = Peruginelli Ginevra, Mario Ragona (a cura di), *L'informatica giuridica in Italia. Cinquant'anni di* studi, ricerche ed esperienze, in Collana: ITTIG, Serie «Studi e documenti», n. 12, Napoli: Edizioni Scientifiche Italiane, 2014.
- PETERSEN E OSTENDORF 2006 = Sarah E. Petersen, Mari Ostendorf, *A machine learning approach to reading level assessment*, University of Washington CSE Technical Report, 2006.
- PETERSEN E OSTENDORF 2009 = Sarah E. Petersen, Mari Ostendorf, *A machine learning approach to reading level assessment*, in Computer Speech and Language (23), 2009, p. 89–106.
- PIANTA ET AL. 2002 = Emanuele Pianta, Luisa Bentivogli, Christian Girardi, *MultiWordNet: developing an aligned multilingual database*, In First International Conference on Global WordNet, Mysore, India, 2002, pp. 292–302.
- PIEMONTESE 1989 = M. E. Piemontese, *Un popolo di "lettori" dimenticati*, in Italiano & Oltre, a. IV, n. 2, marzo-aprile 1989, pp. 69-72 e 81.
- PIEMONTESE 1991 = M. E. Piemontese, *Scrittura e leggibilità: «Due parole»*, in Cortelazzo, Scrivere nella scuola dell'obbligo, Firenze, La Nuova Italia, Quaderni del Giscel/8, 1991, pp. 151-167.
- PIEMONTESE 1996a = M. E. Piemontese, *Capire e farsi capire. Teorie e tecniche della scrittura controllata*, Napoli, Tecnodid, 1996.
- PIEMONTESE 1996b = M. E. Piemontese, *«Due parole»: un approccio allo svantaggio linguistico in termini di semplificazione di strutture*, in Colombo e Romani, 1996, pp. 231-248.
- PIEMONTESE 2000a =M. E. Piemontese, *Leggibilità e comprensibilità dei testi delle pubbliche amministrazioni: problemi risolti e problemi da risolvere*, in Sandra Covino (a cura di), *La scrittura professionale: ricerca, prassi, insegnamento: atti del I Convegno di studi*, Perugia, Università per stranieri, 23-25 ottobre 2000.
- PIEMONTESE 2000b = M. E. Piemontese (a cura di), *Lingue, culture e nuove tecnologie*, in La Nuova Italia, Firenze, Quaderni del Giscel nuova serie n. 3, 2000.
- PIEMONTESE 2005 = M. E. Piemontese, *Misurazioni quantitative degli stili personali e indici di leggibilità*, in T. De Mauro, I. Chiari (a cura di), *Parole e numeri*, Aracne, 2005, Roma, pp. 377-397.
- PIEMONTESE E CAVALIERE 1997 = M. E. Piemontese, L. Cavaliere, *Leggibilità e comprensibilità di sussidiari per le scuole elementari*, in Calò e Ferreri, 1997, pp. 221-240.

- PIEMONTESE E TIRABOSCHI 1986 = M. E. Piemontese, M. T. Tiraboschi, *Problemi di leggibilità e comprensibilità di testi scritti in portatori di deficit intellettivo*, in Linguaggi, III, n. 3, 1986, pp. 107-122.
- PIEMONTESE E TIRABOSCHI 1990 = M. E. Piemontese, M. T. Tiraboschi, *Leggibilità e comprensibilità di testi della Pubblica Amministrazione. Strumenti e metodologie di ricerca al servizio del diritto a capire testi di rilievo pubblico*, in Elisabetta Zuanelli (a cura di), Il diritto all'informazione in Italia, Ricerche promosse dalla Presidenza del Consiglio dei Ministri, Dipartimento per l'informazione e l'editoria, Istituto Poligrafico e Zecca dello Stato, Roma, 1990, pp. 225-246.
- PIEMONTESE E VEDOVELLI 1988 = M. E. Piemontese, M. Vedovelli, *Linguaggio e handicap: problemi della comprensione del linguaggio verbale in situazioni di formazione professionale*, in T. De Mauro, S. Gensini, M. E. Piemontese (a cura di), Dalla parte del ricevente: percezione, comprensione, interpretazione, Atti del XIX Congresso Internazionale di Studi. Roma 8-10 novembre 1985, Roma, Bulzoni, 1988, pp. 303-311.
- PINTO ET AL. 2002 = D. Pinto, M. Branstein, R. Coleman, W. B. Croft, M. King, W. Li, X. Wei, *Quasm: a system for question answering using semi-structured data*. In JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, New York, NY, USA, 2002. ACM, pp. 46–55.
- PITLER E NENKOVA 2008 = Emily Pitler, Ani Nenkova, *Revisiting readability: A unified framework for predicting text quality*, in Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 186–195, 2008.
- PLANTERA 2005 = R. Plantera, *Temperatura informazionale e leggibilità dei testi*, in T. De Mauro – Chiari 2005 (a cura di), pp. 399-415.
- PLATZACK 1974 = C. Platzack*, Spräket och läsbarheten [Language and readability]*, in Lund, Gleerup, Swed, 1974.
- PONTE E CROFT 1998 = Jay M. Ponte, W. Bruce Croft, *A language modeling approach to information retrieval*, in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98), ACM, New York, USA, 1998, pp. 275-281.
- POWERS ET AL. 1958 = R. D. Powers, W. A. Sumner, B. E. Kearl, *A recalculation of four readability formulas*, in Journal of Educational Psychology, 49, 1958, pp. 99-105.
- POZZO 1986 = G. Pozzo, *Comprensibilità dei testi scolastici e apprendimento*, in Insegnare, anno II, n. 9, 1986, pp. 13-19.
- PRADA 2003 = M. Prada, *Lingua e Web*, in I. Bonomi, A. Masini, S. Morgana, La lingua italiana e i mass media, Roma, Carocci, 2003, pp. 249-289.
- PRADA 2015 = M. Prada, L'italiano in rete. Usi e generi della comunicazione mediata tecnicamente, Milano, Franco Angeli, 2015.
- PRASAD ET AL. 2008= R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, B. Webber, *The Penn Discourse Treebank 2.0*, in Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech 2008, pp. 2961-2968.

- QI 2012 = Xiaoguang Qi, Web Page Classification and Hierarchy Adaptation, in Theses and Dissertations, Paper 1386, 2012.
- QI E DAVISON 2009 = X. Qi, B. D. Davison, *Web page classification: Features and algorithms*, in ACM Comput. Surv.41, 2, Article 12, febbraio 2009.
- QV LE ET AL. 2011 = V. Le Quoc, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, Y. Andrew Ng. *Building high-level features using large scale unsupervised learning*, in Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML'12), Omnipress, USA, 2012, pp. 507-514.
- RABIN 1988 = Annette T. Rabin, *Determing Difficulty Levels of Text Written in Languages Other*, in Zakaluk – Samuels, 1988.
- RASTIER 2013 = François Rastier, *La misura e la grana. Semantica del corpus ed analisi del WEB*, Pisa, Edizioni ETS, 2013.
- RADEV E FAN 2000 = D. R. Radev, W. Fan, *Automatic summarization of search engine hit lists*. In Proc. of ACL, 2000.
- RANKIN 1959 = E. F. Rankin, *The cloze procedure: Its validity and utility*, in Farr, Measurement and valuation of reading, Hancourt, Brace and World inc., 1959.
- RANKIN 1965 = E. F. Rankin, *The cloze procedure: a survey of research*, in E. Thurston & L. Hafner (Eds.), Fourteenth Yearbook of the National Reading Conference, 14, 1965, pp. 133-150.
- RANKIN E DALE 1959 = E. F. Rankin, Edgar Dale, *Cloze residual gain – a technique for measuring learning through reading*, in Schick, The psychology of reading, National reading Conference, Yearbook 18, 1959, pp. 17-26.
- REHM ET AL. 2008 = Rehm G., Santini M., Mehler A., Braslavski P., Gleim R., Stubbe A., Symonenko S., Tavosanis M., Vidulin V., *Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems*, in "Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)": http://www.lrecconf.org/proceedings/lrec2008/.
- REVAZ E BRONCKART 1988 = Revaz Françoise, Bronckart Jean-Paul, *Mesurer la lisibilité [Une approche typologique]*, in Revue française de pédagogie, vol. 85, 1988, pp. 37-46.
- REVELLINO 2017 = Rosa Revellino, *La comunicazione medico-paziente. Una storia antica e una prospettiva nuova*, in "La Professione", II, 2017, pp. 149-159.
- RICHAUDEAU 1973 = François Richaudeau, André Conquet, *Cinq méthodes de mesure de la lisibilité*, in Communication et langages, n. 17, 1973, pp. 5-16.
- RICHAUDEAU 1976 = Richaudeau François, *Faut-il brûler les formules de lisibilité?*, in Communication et langages, n. 30, 1976, pp. 6-19.
- RICHAUDEAU 1979 = François Richaudeau. *Une nouvelle formule de lisibilité*, in Communication et langages, n. 44, 4ème trimestre 1979, pp. 5-26.
- RICHAUDEAU E STAATS 1981 = François Richaudeau, Donna M. Staats, *Some French Work on Prose Readability and Syntax*, in Journal of Reading, vol. 24, n. 6, Marzo 1981, pp. 503-508.

- RILOFF E PHILLIPS 2004 = E. Riloff, W. Phillip, *An introduction to the Sundance and Autoslog systems*. Technical Report UUCS-04-015, University of Utah School of Computing, 2004.
- RITTERBAND ET AL. 2009= Lee M. Ritterband, F. P. Thorndike, D. J. Cox, Borsi P. Kovatchev, L. A. Gonder-Frederick, *A behavior change model for Internet interventions*, in Ann Behav Med. 38, 1, 2009, pp. 18‑27.
- ROBERTS ET AL. 2016 = H. Roberts, D. Zhang, D. G. Dyer, *The readability of AAOS patient education materials: evaluating the progress since 2008*, in JBJS, 2016, vol. 98, issue 17, p. e70.
- ROCK 1969 = Ernest Louis Rock, *A readability graph for Russian*, Unpublished doctoral dissertation, Ohio State University, 1969. Si veda anche: Dissertation Abstracts, 1970, 31, 567-A.
- RODRÍGUEZ DIÉGUEZ 1983 = J. L. Rodriguez Diegez, *Evaluación de textos escolares*, in Revista de Investigación Educativa, 1(2), 1983, pp. 259-279.
- RODRIGUEZ TRUJILLO 1980 = Nelson Rodriguez Trujillo, *Determinación de la comprensibilidad de materiales de lectura por medio de variables lingüísticas*, in Lectura y Vida, 1, 1980, pp. 29-32.
- RODRIGUEZ TRUJILLO 1983 = Nelson Rodriguez Trujillo, *El procedimiento "cloze": un procedimiento para evaluar la comprensión de lectura y la complejidad de materiales*, in Lectura y Vida, 4, 1983, pp. 4-13.
- ROGERS 1962 = J. R. Rogers, *A formula for predicting the comprehension level of material to be presented orally*, in The journal of educational research, vol.56, n. 4, 1962, pp.218-220.
- ROHIT ET L. 2010 = Rohit J. Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos, Chris Welty, *Learning to predict readability using diverse linguistic features*, in Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 546–554.
- ROSE ET AL. 2007 = D. E. Rose, D. M. Orr, R. G. P. Kantamneni, *Summary atributes and perceived search quality*. In Proc. of Intl. Conference on the World Wide Web, 2007.
- ROSENFELD 2000 = Ronald Rosenfeld, *Two decades of statistical language modeling: Where do we go from here?*, in proceedings of the IEEE, 88(8), 2000.
- RUSSEL E NORVIG 2003 = Stuart Russell, Peter Norvig, *Intelligenza artificiale: un approccio moderno*, Prentice Hall, 2003.
- SAMUEL 1959 = A. L. Samuel, *Some studies in machine learning using the game of checkers*, in IBM, J. Res. Dev.3, 3, Luglio 1959, pp. 210-229.
- SANTINI 2005 = M. Santini, *Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis*, in "Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK05)", 11 Jan 2005, Manchester, UK.
- SANTINI 2006 = M. Santini, *Common Criteria for Genre Classification: Annotation and Granularity*, in Workshop on Text-based Information Retrieval (TIR-06), Riva del Garda, 2006.

- SANTINI 2007 = M. Santini, *Characterizing Genres of Web Pages: Genre Hybridism and Individualization*, in "Proceedings of the 40th Annual Hawaii International Conference on System Sciences", 2007.
- SANTINI 2008 = M. Santini, *Zero, single, or multi? Genre of web pages through the users' perspective*, in Information Processing & Management. 44, 2, 2008, pp. 702-737.
- SANTINI ET AL. 2009 = M. Santini, G. Rehm, S. Sharoff, A. Mehler, A., *Automatic genre identification: Issues and prospects*, in Journal for Language Technology and Computational Linguistics (JLCL), 24, 1, 2009.
- SANTINI 2011 = M. Santini, *Cross-Testing a Genre Classification Model for the Web*, in A. Mehler, S. Sharoff, M. Santini (eds.), *Genres on the Web: Computational Models and Empirical Studies*, Springer, Dordrecht,2011, pp. 87-128.
- SCARTON ET AL. 2009 = Caroline E. Scarton, Daniel M. Almeida, Sandra M. Aluísio, *Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português*. In Proceedings of STIL-2009, São Carlos, Brazil, 2009.
- SCHOOL RENAISSANCE INSTITUTE 1999 = School Renaissance Institute, *ZPD guidelines: Helping students achieve optimum reading growth*, Madison, WI: School Renaissance Institute, Inc., 1999.
- SCHOOL RENAISSANCE INSTITUTE 2000 = School Renaissance Institute, *The ATOS readability formula for books and how it compares to other formulas*, Madison, WI: School Renaissance Institute, Inc., 2000.
- SCHULTZ 1981 = Renate A. Schulz, *Literature and Readability: Bridging the Gap in Foreign Language Reading*, in The Modern Language Journal, vol. 65, n. 1, 1981, pp. 43-53.
- SCHUTTEN E MCFARLAND 2009 = M. Schutten, A. McFarland, R*eadability levels of health-based websites: from content to comprehension*, in International Electronic Journal of Health Education, 12, 2009, pp. 99-107.
- SCHWARTS 1975 = R.E.W. Schwartz, *An exploratory effort to design a readability graph for German material*, in Unpublished study, State University of New York at Albany, 1975.
- SCHWRM E OSTENDORF 2005 = Sarah E. Schwarm, Mari Ostendorf, *Reading Level Assessment Using Support Vector Machines and Statistical Language Models*, in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05), Ann Arbor, giugno 2005, pp. 523–530.
- SEBASTIANI 1999 = Fabrizio Sebastiani, *A tutorial on automated text categorization*, in Analia Amandi and Alejandro Zunino (eds.), Proceedings of the 1st Argentinian Symposium on Artificial Intelligence (ASAI 1999), Buenos Aires, AR, 1999, pp. 7-35.
- SEBASTIANI 2002 = Fabrizio Sebastiani*, Machine learning in automated text categorization,* ACM Computing Surveys*,* 34(1):1-47, 2002.
- SEBASTIANI 2005a = Fabrizio Sebastiani, *Text categorization, i*n Alessandro Zanasi (ed.), Text Mining and its Applications*,* WIT Press, Southampton, UK, 2005, pp. 109-129.

- SEBASTIANI 2005b =Fabrizio Sebastiani, *Text categorization*, in Laura C. Rivero, Jorge H. Doorn e Viviana E. Ferraggine (eds.), The Encyclopedia of Database Technologies and Applications*,* Idea Group Publishing, Hershey, US, 2005, pp. 683-687.
- SEBASTIANI 2006 = Fabrizio Sebastiani, *Classification of text, automatic*, in Keith Brown (ed.), The Encyclopedia of Language and Linguistics, vol. 14, 2° edizione, Elsevier Science Publishers, Amsterdam, NL, 2006, pp. 457-462.
- SEKINE E GRISHMAN 1995 = S. Sekine, R. Grishman, *A corpus-based probabilistic grammar with only two nonterminals*, In Fourth International Workshop on Parsing Technologies. Prague: Karlovy Vary, 1995, pp. 260-270.
- SHAROFF 2004 = S. Sharoff, *Towards Basic Categories for Describing Properties of Texts in a Corpus*, in "Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004", ELDA, Lisbona.
- SHAROFF 2010 = S. Sharoff, *Analysing similarities and differences between corpora*. In 7th Language Technologies Conference, Ljubljana, 2010.
- SI E CALLAN 2011 = Luo Si, Jamie Callan, *A statistical model for scientific readability*, in Proceedings of the tenth international conference on Information and knowledge management, 2011, pp. 574–576.
- SIERRA ET AL. 1992 = A. E. Sierra, M. A. Bisesi, T. L. Rosenbaum, E. J. Potchen, *Readability of the radiologic report*, in Invest Radiol., 1992, 27(3), pp. 236-9.
- SIMONE 1985 = R. Simone, *Leggere e non leggere*, in Insegnare, anno II, 2, 1985, p. 17-22.
- SMITH 1961 = Edgar A. Smith, *Devereaux readability index*, in The Journal of Educational Research, vol. 54, n. 8, 1961, pp. 289-303.
- SMITH E KINCAID 1970= E. A. Smith, J. P. Kincaid, *Derivation and validation of the automated readability index for use with technical materials*, in Human factors, 12, 1970, pp.457-464.
- SMITH E SENTER 1967 = E. A. Smith, R. J. Senter, *Automated Readability Index,* Wright-Patterson Air Force Base, Aerospace Medical Division, AMRL-TR-66-220, [AMRL = Aerospace Medical Research Laboratories], Ohio, 1967.
- SPACHE 1953 = George Spache, *A New Readability Formula for Primary-Grade Reading Materials*, in The Elementary School Journal, vol. 53, n. 7, marzo 1953, pp. 410-413.
- SPACHE 1966 = G. D. Spache, *Good reading for poor readers*, Garrard, Champaign, Illinois, 1966.
- SPAULDING 1951 = Seth Spaulding, *Two Formulas for Estimating the Reading Difficulty of Spanish*, in Educational Research Bulletin, vol. 30, n. 5, maggio 1951, pp. 117-124.
- SPAULDING 1956 = Seth Spaulding, *A Spanish Readability Formula, in The Modern Language Journal*, vol. 40, n. 8, dicembre 1956, pp. 433-441.
- SPINA 2001 = S. Spina, *Fare i conti con le parole. Introduzione alla linguistica dei corpora*, Perugia, Guerra, 2001.

- SPIRO 2004 = R. J. Spiro, *Principled pluralism for adaptive flexibility in teaching and learning*. In R.B. Ruddell e N. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 654–659). Newark, DE: International Reading Association.
- ŠTAJNER ET AL. 2012 = S. Štajner, R. Evans, C. Orasan, R. Mitkov, *What can readability measures really tell us about text complexity?* In Proceedings of the the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA), Istanbul, Turkey,2012, pp. 14-21.
- STAPHORSIUS E KROM 1985 = G. Staphorsius, R.S.H. Krom, *Cim leesbaarheidsindex voor het basisonderwijs*, Amheim: Central Institute voor Toets Ontwikkeling, 1985.
- STEGER E STEMLE 2009 = J.M. Steger, E. Stemle, *KrdWrd - Architecture for Unified Processing of Web Content*, in I. Alegria, I. Leturia, S. Sharoff (eds), "Proceedings of the Fifth Web as Corpus Workshop", 2009, pp. 63-70. www.sigwac.org.uk/raw-attachment/wiki/WAC5/WAC5_proceedings.pdf
- STENNER E BURDICK 1997 = A. J. Stenner, D. Burdick, *The Objective Measurement of Reading Comprehension*, in Response to Technical Questions Raised by the California Departmentof Education Technical Study Group. Durham, NC: Metametrics, 1997.
- STENNER E WRIGHT 1998 = B. D. Wright, A. J. Stenner, *Readability and Reading Ability*, Paper presented to the Australian Council on Education Research, giugno 1998.
- STENNER ET AL. 1987 = A. J. Stenner, D. R. Smith, I. Horabin, M. Smith, *Fit of the Lexile Theory to Sequenced Units from Eleven Basal Series*, MetaMetrics, Inc., 1987.
- STENNER ET AL. 1988a = A. J. Stenner, D. R. Smith, I. Horabin, M. Smith, *The Lexile Framework*, Durham, NC: Metametrics, Inc., 1988.
- STENNER ET AL. 1988b = A. J. Stenner, D. R. Smith, I. Horabin, M. Smith, *Most comprehension tests do measure reading comprehension: A response to McLean and Goldstein,* Phi Delta Kappan, 1988.
- STENNER ET AL. 1996 = A. J. Stenner, D. R. Smith, I. Horabin, M. Smith, *Measuring Reading Comprehension with the Lexile Framework*, paper presented at the Fourth North American Conference on Adolescent/Adult Literacy, 1996.
- STEVENS 1980 = K. C. Stevens, *Readability formulae and McCall-Crabbs standard test lessons in reading*, in The Reading Teacher, gennaio 1980, pp. 413-415.
- STONE 1956 = Clarence R. Stone, *Measuring Difficulty of Primary Reading Material: A Constructive Criticism of Spache's Measure*, in The Elementary School Journal, vol. 57, n. 1 ottobre 1956, pp. 36-41.
- SZALAY 1965 = T. G. Szalay, *Validation of the Coleman readability formulas*, in Psychological Reports, 17, 1965, pp. 965-966.
- TANAKA-ISHII ET AL. 2010 = K. Tanaka-Ishii, S. Tezuka, H. Terada, *Sorting text by readability*, in Computational Linguistics, 36(2), 2010, pp. 203-227.
- TAVOSANIS 2011 = Mirko Tavosanis, *L'italiano del web*, Carocci, Roma, 2011.
- TAVOSANIS 2018 = Mirko Tavosanis, *Lingue e intelligenza artificiale*, Carocci, Roma, 2018.
- TAYLOR 1953 = W. L. Taylor, *Cloze procedure: a new tool for measuring readability*, in Journalism Quarterly, vol. 30, 1953, pp. 415-433.

- TAYLOR 1956 = W. L. Taylor, *Recent developments in the use of "cloze procedure"*, in Journalism Quarterly, vol. 33, 1956, pp. 42-99.
- TAYLOR 1957 = W. L. Taylor, "*Cloze" readability scores as indices of individual differences in comprehension and aptitude*, in Journal of Applied Psychology, 41, 1957, pp. 19-26.
- TERRANOVA ET AL. 2012 = G. Terranova, M. Ferro, C. Carpeggiani et al., *Low quality and lack of clarity of current informed consent forms in cardiology: how to improve them*, in JACC Cardiovasc Imaging, 2012, 5, pp. 649-655.
- TESTA 2000 = A. Testa, *Farsi capire*, Milano, Rizzoli, 2000.
- THARP 1939 = James B. Tharp, *The Measurement of Vocabulary Difficulty*, in The Modern Language Journal, vol. 24, n. 3, dicembre 1939, pp. 169-178.
- THORNDIKE 1916 = Edward L. Thorndike, *An improved scale for measuring ability in reading*, in Teachers college record, vol. 17, n.1, 1916, pp. 40-67.
- THORNDIKE 1921 = Edward L. Thorndike, *The Teacher's word book*, New York, Bureau of Publication, Teacher's College, Columbia University, 1921.
- THORNTON 1984a = Anna M. Thornton, *Più o meno leggibili* [titolo redazionale], Riforma della Scuola, 1984, 1, p. 44.
- THORNTON 1984b = Anna M. Thornton, *L'indice di leggibilità* [titolo redazionale], Riforma della Scuola, 1984, 2, p. 52.
- THORNTON 1984c = Anna M. Thornton, *Come sono scritti e illustrati i libri* [titolo redazionale], Riforma della Scuola, 1984, 3, pp. 49-51.
- THORNTON 1984d = Anna M. Thornton, *Leggibilità dei manuali* [titolo redazionale], Riforma della Scuola, 1984, 4, pp. 50- 51.
- THORNTON 1984e = Anna M. Thornton, *Cefalocordati divulgati e leggibili* [titolo redazionale], Riforma della Scuola, 1984, 7-8, pp. 64-65.
- THORNTON 1984f = Anna M. Thornton, *L'indagine IEA sulla produzione scritta*, Linguaggi, 1984, 1-2, pp. 67-68.
- THORNTON 1992 = Anna M. Thornton, *Gli studi sulla leggibilità e la riscrittura in Italia*, in Lucisano P. (a cura di), 1992, pp. 45-53.
- TIEMAN E BRADLEY 2013 = J Tieman, S. L. Bradley, *Systematic review of the types of methods and approaches used to assess the effectiveness of healthcare information websites*, in Australian Journal of Primary Health, 2013, 19, 4, pp. 319-324.
- TOBER ET AL. 2015= M. Tober, D. Furch, K. Londenberg, L. Massaron, J. Grundmann, *Search Ranking Factors and Rank Correlations*. Google US 2015.
- TOGEBY 1971 = O. Togeby, *Sproq og laeseproces*, Den Gjellerny, Copenhagen, 1971.
- TOGLIA E BATTIG 1978 = M. P. Toglia, W. R. Battig, *Handbook of semantic word norms*, Hillsdale, NJ: Erlbaum, 1978.
- TONELLI ET AL. 2012 = Sara Tonelli, Ke Tran Manh, Emanuele Pianta, *Making readability indices readable*, in Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations, 2012, pp. 40–48.
- TRESSOLDI 2008 = P.E. Tressoldi, *I brani della batteria MT si possono leggere tutti con la stessa rapidità? Norme trasversali dalla seconda elementare alla terza medi*a, in Dislessia, 5, 3, 2008, pp. 339-345.

- TULBERT ET AL. 2011 = Brittain H. Tulbert, Clint W. Snyder, Robert T. Brodell. *Readability of Patient-Oriented Online Dermatology Resources*, in The Journal of Clinical and Aesthetic Dermatology, 4, 3, 2011, pp. 27–33.
- TURCHI 2014 = Fabrizio Turchi, *Natural Language Processing: modelli e applicazioni in ambito giuridico*, in Ginevra Peruginelli, Mario Ragona (a cura di), *L'informatica giuridica in Italia. Cinquant'anni di studi, ricerche ed esperienze*, Collana: ITTIG, Serie «Studi e documenti», n. 12, Napoli: Edizioni Scientifiche Italiane, 2014.
- UITDENBOGERD 2006 = A. L. Uitdenbogerd, *Web Readability and Computer-Assisted Language Learning*, in Proc. Australasian Language Technology Workshop (ALTW2006), 2006, pp.99–106.
- VACCA 1971 = Roberto Vacca, *Medioevo Prossimo Venturo*, 1971.
- VACCA 1972 = Roberto Vacca, *Per una critica quantitativa: romanzi a chilometri*, in Il Messaggero, dicembre 1972.
- VACCA 1978 = Roberto Vacca, *Smascheriamo gli illeggibili*, in Tuttolibri, luglio 1978.
- VACCA 1981 = Roberto Vacca, *Come imparare più cose e vivere meglio*, in Arnoldo Mondadori, 1981, pp. 172-184.
- VAN HAUWERMEIREN 1972 = Paul Van Hauwermeiren, *De Leesbaarheidsmeting, Toepasselijkheid op het Nederlands*, Tesi di dottorato, Università di Leuven, 1972.
- VAPNIK 1995 = V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
- VAPNIK 1998 = V. N. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998, p. 736.
- VAPNIK 1999 = V. N. Vapnik, *An Overview of Statistical Learning Theory*, in IEEE transactions on neural networks, vol. 10, n. 5, settembre 1999.
- VARI-CARTIER 1981 = Patricia Vari-Cartier, *Development and Validation of a New Instrument to Assess the Readability of Spanish Prose*, in The Modern Language Journal, vol. 65, n. 2 1981, pp. 141-148.
- VEDOVELLI 1980 = Massimo Vedovelli, *Comprensione del testo nella media*, in Riforma della scuola, vol. 2-3, 1980, pp. 48-50.
- VEDOVELLI 1986 = Massimo Vedovelli, *Leggibilità e svantaggio: il caso handicap*, in Linguaggi, III, n. 3, 1986, pp. 100-106.
- VEDOVELLI 1991 = Massimo Vedovelli, I*l progetto CUD per l'insegnamento dell'italiano L2 a distanza: un modello per la formazione linguistica Erasmus*, in Italiano lingua seconda: modelli e strategie per l'insegnamento atti della Giornata di studi del Centro interfacoltà di ricerca sulla didattica delle lingue straniere moderne, Pavia, dicembre 1989, a cura di Marco Mazzoleni e MariaPavesi, 1991, pp. 111-128.
- VEDOVELLI 1995 = M. Vedovelli, *La lingua italiana d'uso: morfosintassi del parlato e dello scritto*, in MILIA. Materiali per gli insegnanti di Lingua Italiana - Aggiornamento, modulo n. 10, Ministero della Pubblica Istruzione, Dir. Gen. Scambi Culturali, d'intesa con Ministero degli Affari Esteri, Genova, IRRSAE Liguria/Sagep, Vol.10, 1995.
- VEDOVELLI ET AL. 1989 = M. Vedovelli, M. Cassandro, M. Pisano, *Una lingua per il made in Italy: banche dati e strumenti didattici*, In Culturiana, n. 2, 1989.

- VENTURI ET AL. 2015= G. Venturi, T. Bellandi, F. Dell'Orletta, S. Montemagni, *NLP-Based Readability Assessment of Health-Related Texts: a Case Study on Italian Informed Consent Forms*, in Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi 2015), EMNLP 2015 Workshop, September, Lisbon, Portugal, 2015, pp. 131-141.
- VIVEKANANTHAM ET AL. 2017 = A. Vivekanantham, J. Protheroe, S. Muller, S. Hider. *Evaluating on-line health information for patients with polymyalgia rheumatica: a descriptive study*, in BMC Musculoskeletal Disorders, 2017, pp. 18-43.
- VOGEL E WASHBURNE 1926 = Mabel Vogel, Carleton Washburne, *Winnetka Graded Book List*, American Library Association, Chicago, 1926.
- VOGEL E WASHBURNE 1928 = Mabel Vogel, Carleton Washburne, *An objective method of determining grade placement of children's reading material*, in The elementary Scholl Journal, vol. 28, n. 5, gennaio 1928, pp. 373-381.
- VOGHERA 2005 = Miriam Voghera, *La misura delle categorie sintattiche*, in Tullio De Mauro, Isabella Chiari (a cura di), *Parole e numeri. Analisi quantitative dei fatti di lingua*, Aracne editrice, Roma, 2005, pp.125-138.
- VOR DER BRUCK ET AL. 2008 = Tim vor der Bruck, Sven Hartrumpf, Hermann Helbig, *A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators*. in Proceedings of the 11th International Multiconference: Information Society - IS 2008 - Language Technologies, Ljubljana, Slovenia, 2008, pp. 92–97.
- WALTERS 1966 = T.W. Walters, *Readability in German. Some levels of difficulty encountered in reading professional theological literature in German*, Dissertation Abstracts, 27*, 2520A*, 1966.
- WANG 2006 = Y. Wang, *Automatic Recognition of Text Difficulty from Consumers Health Information*, in 19th IEEE Symposium on Computer-Based Medical Systems, Los Alamitos, CA, USA, IEEE Computer Society, 2006, pp. 131–136.
- WATAD ET AL. 2017 = A. Watad, N. L. Bragazzi, F. Brigo, K. Sharif, H. Amital, D. McGonagle, Y. Shoenfeld, M. Adawi, *Readability of Wikipedia Pages on Autoimmune Disorders: Systematic Quantitative Assessment*, in J Med Internet Res, 2017, 19(7), p. e260.
- WIIO 1968 = Osmo Antero Wiio, *Readability, comprehension and readership: an experimental study on the readability of Finnish magazine articles, with special reference to readership*, Tampereen Yliopisto, Tampere, Finland, 1968.
- YANG 1970 = Shou-Jung Yang, *A readability formula for Chinese language*, Unpublished doctoral dissertation, University of Wisconsin, 1970.
- YU E MILLER 2010 = C-H. Yu, R. C, Miller, *Enhancing web page readability for non-native readers*, in Proc Users and attention on the web CHI 2010, Atlanta, GA, USA. April 10-15. 2010, pp. 2523-2531.
- ZAKALUK E SAMUELS 1988a = Beverly L. Zakaluk, S. Jay Samuels, *Readability: Its Past, Present, and Future*, International Reading Association, Newark, Delaware, 1988.
- ZAKALUK E SAMUELS 1988b = Beverly L. Zakaluk, S. Jay Samuels, *Toward a New Approach to Predicting Text Comprehensibility*, in Zakaluk – Samuels 1988.

- ZAMANIAN E HEYDARI 2012 = Mostafa Zamanian, Pooneh Heydari, *Readability of Texts: State of the Art*, in Theory and Practice in Language Studies, vol. 2, n. 1, gennaio 2012, pp. 43-53.
- ZAMBELLI 1994 = Maria Luisa Zambelli (a cura di), *La rete e i nodi. Il testo scientifico nella scuola di base*, Quaderni del Giscel n. 14, La Nuova Italia, Firenze, 1994.
- ZAZZARO 2009 = G. Zazzaro, *Data Mining: esplorando* le *miniere alla ricerca della conoscenza nascosta–Clustering con l'algoritmo k-means*, in Matematicamente.it Magazine, n. 9, 2009.
- ZHAI 2008 = Chengxiang Zhai*, Statistical language models for information retrieval: A critical review*, in Foundations and Trends in Information Retrieval, vol. 2, n. 3, 2008, pp. 137–213.
- ZHAI E LAFFERTY 2001 = Chengxiang Zhai, John Lafferty, *A study of smoothing methods for language models applied to Ad Hoc information retrieval*, in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01), ACM, New York, USA, 2001, pp. 334-342.
- ZHAI E LAFFERTY 2002 = Chengxiang Zhai, John Lafferty, *Two-stage language models for information retrieval*, in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '02), ACM, New York, USA, 2002, pp. 49-56.
- ZHENG ET AL. 2002 = W. Zheng, E. Milios, C. Watters, *Filtering for medical news items using a machine learning approach*, Prof. AMIA Symp, 2002, pp. 949-953.
- ZINI 2012 = Andra Zini, *Misurare la competenza lessicale in contesto specifico attraverso prove di cloze*, in Giornale Italiano della Ricerca Educativa - Italian Journal of Educational Research, 9, 2012, pp. 108 – 119.
- ZIPF 1935 = George K. Zipf, *The Psycho-biology of language*, Boston, Houghton, 1935.
- ZIPF 1945 = George K. Zipf, *The meaning-frequency relationship of words*, in The journal of General Psychology, 33, 1945, pp. 251-256.
- ZIPF 1949 = George K. Zipf, *Human Behaviour and the Principle of Least-Effort*, 1949.
- ZU EISSEN E STEIN 2004 = S. M. Zu Eissen, B. Stein, *Genre classification of Web pages*, in Proceedings of the 27th German Conference on Artificial Intelligence. Lecture Notes in Computer Science, vol. 3238, Springer, Berlin, Germany, 2004, pp. 256–269.