UNIVERSITÀ DEGLI STUDI DI TRENTO

Doctoral School in Cognitive and Brain Sciences
XXIX Cycle

DOCTORAL THESIS

VERBS AS NOUNS:

empirical investigations on event-denoting nominalizations

PhD Candidate:
Rossella Varvara

Advisor:
Prof. Roberto Zamparelli

Abstract

In this thesis, I study the differences in form and interpretation presented by event-denoting nominalizations. Frequently, languages have more than one type of event nominalization, such as deverbal nouns derived by means of suffixes (Italian *mutamento / mutazione*, 'change', 'mutation', or English *assignment, explosion*) and their corresponding verbal nouns, e.g. infinitives (*il mutare*, 'the changing') or gerunds (*exploding*). These are usually perceived as alternatives, since their semantic difference is not clearly understood by neither native speakers nor linguists.

The aim of this work is to understand the rationale that leads us to choose one form instead of the other and to define the linguistic features involved. The hypothesis underlying the whole thesis is that different forms are never true synonyms and, thus, present some differences in use, distribution or meaning.

In a first study, I explore the role of the base verb in the nominalization selection. I investigate if the various nominalizations are formed from different types of base verbs and which characteristics define their domain of application. By means of statistical modeling, I highlight how the transitivity of the base verb partially determines which nominalization is preferred. Moreover, I show that NIs are not used to make up for the lack of a corresponding EDN, refuting previous claims.

Then, I move forward analyzing the cases in which both forms are derived from the same base and I try to understand if they differ in meaning. In the second study presented, I use collocation analysis to observe their semantic dissimilarities. With focus on a single syntactic pattern, I find out that nominal infinitives and deverbal nouns inherit only part of the base verb senses. The former usually prefer metaphorical and abstract senses, whereas the latter select more concrete and literal ones.

Lastly, I use distributional semantic models to observe quantitatively the semantic shift of the two processes. I confirm the hypothesis that nominal infinitives are more transparent and more semantically regular than deverbal nouns, given their inflectional nature.

The studies presented have been conducted on Italian and German; however, the findings are relevant for the general treatment of nominalizations and may be replicated for further languages. Overall, my work shows how quantitative analyses of corpus data can help us investigate problems that are hardly addressed by linguists introspection. Moreover, it includes in the study of nominalizations nominal infinitives, non-finite verbal forms which, contrary to English gerunds, have not received the attention they deserve.

Aknowledgments

I would have not been able to come at the end of this journey without the support, the advice, and the help of some incredible people.

First and foremost, I would like to thank my advisor Roberto Zamparelli for the continuous support that he has given me during these years. He has guided me through the difficult steps of becoming a researcher. He has given me the freedom to find my way in the wide field of linguistics, while pointing out constructive criticisms and sharing creative and bright insights.

I would like also to thank the members of my Oversight Committee, Marco Baroni and Raffaella Bernardi, for their fruitful suggestions and for having pushed me toward a more rigorous methodology.

My sincere gratitude goes also to Sebastian Padò, with whom I was given the opportunity to work during my stay at the IMS in Stuttgart. He has been a considerate advisor and a great teacher: he has shown me that it is possible to combine theoretical and computational linguistics together, and thought me how to go deep into my research questions. A particular thank is dedicated to Gabriella Lapesa, without whose encouragement and support I would have been lost. Thank you for your friendship, for your contagious enthusiasm, and for being a living proof that competence and selflessness can live together in the same person. The study in Chapter 5 is the result of my fruitful collaboration with Sebastian and Gabriella.

Many thanks also to the researchers from the B9 project, the TCL group and the whole IMS who have let me feel at home in Stuttgart. A special thank to Max Kisselew for having been a great officemate and friend. Thanks also to all the actual and former members of the Clic lab in Rovereto, particularly to German Kruszewski, Denis Paperno, Marco Marelli, and Sandro Pezzelle for all the help and discussions.

I would like to thank also my former advisor Elisabetta Ježek for having

Contents

Introduction

If you are asked to think of a noun, you will probably think of a word denoting a concrete entity, like 'cat', 'chair' or 'apple'. However, our languages are full of non-prototypical nouns which denote abstract concepts ('freedom', 'intelligence','fear'), and which do not belong to what Lyons (1977: 442) calls *first-order entities*. Among these non-prototypical nouns, event nouns ('party', 'construction', 'running') are interesting for many reasons: they are located in time and space, they are perceivable by senses, but their perceptual properties are not constant and stable over time. They denote what have been described as *second-order entities* (Lyons, 1977: 443). In many cases, such nouns are derived from other syntactic categories, especially verbs, and languages usually present more than one form of event nominalizations.

Two main classes can be identified: nouns built with various suffixes applied to verbal roots (e.g. English *defer-ment*, *activa-tion*), and verbal nouns with nominal functions, such as English *-ing* forms , Italian infinitives and past-participial forms of Romanian supine. In this work I call the former class *event-denoting deverbal nouns* (henceforth EDNs), and the latter *verbal nouns*. These two categories and their differences are the objects of study in this thesis.

Past research has focused mainly on the first class, with particular attention to their argument realization (Grimshaw, 1990), aspectuality (e.g. Brinton, 1995), polysemy (Walinska de Hackbeil, 1984; Grimshaw, 1990; Melloni, 2007, among others) and productivity (Bauer, 2001).

The present work has a more comprehensive scope, since it includes verbal nouns into the analysis. It investigates the differences between these two types of nominalizations by means of empirical studies conducted on Italian and German. The main goal is to understand the rationale that leads

us to choose one form instead of the other and to define the linguistic features involved. The underlying hypothesis is that they are not true synonyms and that they differ in use, distribution or meaning.

The thesis is structured as follows. In chapter 1, I offer an overview of the various types of nominalizations and their main properties. Particularly, I focus on the possible differences in the morphological process used, in the degree of nominal features inherited and in the range of meanings. Since it is common among languages to have more than one form of event nominalization, I explore possible explanations of this coexistence.

In chapter 2, I present a complete description of the two classes of nominalizations in Italian, highlighting their formal differences and similarities.

In chapter 3, I test if they are formed from different types of base verbs and, thus, if they are alternative forms used in complementary distribution. Moreover, I try to identify constraints on their productivity and use. It is shown that, even if there are preference for different classes of verbs which are reflected by token frequency, different nominalizations are mainly derived from the same base verbs.

This result leads us to approaching a different explanation, i.e. the possibility that they differ in meaning. In chapter 4, by means of collocation analysis, I show that they refer to different senses of the base verbs and that they disambiguate the vagueness of the base.

Chapter 5 presents a study conducted on German nominalizations using distributional semantic models. The meaning shift produced by the two processes is observed, supporting the claim that verbal nouns are closer in meaning to the base verbs. This behavior is placed within the broader theoretical perspective of a graded distinction between inflection and derivation. It is shown that verbal nouns behave like other inflectional processes, namely present participles, whereas EDNs are closer to the derivational pole, represented by agent nouns in -er.

Chapter 6 provides a general discussion and conclusions. Moreover, ideas for further directions of investigation are outlined.

Chapter 1

Nominalizations: event deverbal nouns and non-finite verbal forms

1.1 Nominalizations: a classification

The present work focuses on event-denoting deverbal nominalizations, i.e. nouns that may refer to events (in the broadest sense[1]) and that are derived from verbs by means of derivation, inflection or conversion. The same phenomena have been called also *action nominals* (or *nomina actionis*, Comrie, 1976; Comrie and Thompson, 2007; Koptjevskaja-Tamm, 1993, 2006): a shorter name which, however, does not mark the difference with simple (i.e. not derived) action nouns, such as *trip, game*. The term *nominalization*, indeed, meaning "turning something into a noun" (Comrie and Thompson, 2007: 334), points to the transpositional process that takes place when a verb is used as a base for a noun and conveys the idea that we are talking about a complex word, not a simple one.

Not every deverbal nominalization is an event nominalization. Verbs may be base for numerous nominalizations, whose meaning can be linked to different arguments of their base. For example, it is possible to form agentive nouns, like English *singer* (from the verb *sing*), instrumental nouns (*to slice*

---

[1]With the term "event" I refer to every kind of eventuality (Bach, 1986), including states. Thus, event nominalizations may denote activity, achievements, accomplishments and states (following the terminology proposed in Vendler, 1957).

$\rightarrow$ *slicer*), as well as locative, manner, reason or objective ones (Comrie and Thompson, 2007).

Event nominalizations may be derived also by adjectives, e.g. Italian *lontananza* ('separation'), or nouns, like the Italian prefixed noun *sciacallaggio* ('looting'). However, in this work I focus only on deverbal ones, since they are the most common class (Gaeta, 2004: 318). A contrastive analysis of nominalizations derived by the three word-classes deserves undeniably an in-depth study, which, unfortunately, has not found a place in this thesis. Further work is needed to cover this topic and give a more comprehensive view on event nominalizations. In what follows, if not differently stated, I will always refer to event-denoting deverbal nominalizations, even when an abbreviation of this name will be used.

Deverbal event nominalizations can be of different nature among the languages of the world and more than one form can be attested within a specific language. These nominalizations can vary according to: (i) the morphological process involved; (ii) the inheritance of verbal and nominal properties; (iii) the possible meanings expressed.

In this chapter, I introduce the different types of nominalizations attested in the world's languages and the problem linked to their characterization. Specifically, I link their discussion to the distinction between inflection and derivation and to the nature of the word-classes of nouns and verbs. In section 1.5 I introduce the main focus of the rest of this thesis, i.e. the differences and the competition between different event nominalizations in the same language.

1.2 Nominalizations and morphological processes

The most common morphological process involved in nominalization is derivation. It can be defined as the creation of a new lexeme by the addition of an affix (i.e. a bound grammatical morphemes), like in the English event nominalizations *displacement* or *construction*, where the suffixes *-ment* and *-ion* are used. In many languages, more than one eventive affix is available, giving rise to possible competition between words. In English, for example, the suffixes *-al* (*arrival, approval*), *-ance* (*resistance, attendance*), *-ing* (*reading, learning*), *-ation* (*regulation, consultation*), and *-ment* (*recruitment, development*) can all be used to form event nominalizations.

However, this is a narrow definition, which identifies derivation with affixational word-formation, leaving out what has been called *dérivation impropre* (Grevisse, 1969: 162), like replacement and subtraction of morphemes (Lieber, 2000; Dressler, 2000) or identity operations (Bauer 1983:32; Don et al., 2000, also called conversion, zero derivation or transcategorization). Understood in this way, derivation does not incorporate every kind of word-formation processes. In what follows, I will use this narrow definition of derivation, meaning all word-formation processes that use affixational morphology. However, it should be kept in mind that the other word-formation processes listed above can also be involved in nominalizations.

Even with this narrower definition, the boundaries of derivation are not always clear. There are, indeed, cases in which an inflectional affix does form a new word, changing the word-class of the base term. These cases resemble derivation, as we have just defined above, and we may ask if they should count as inflection or not. Few lines above I included *-ing* in the list of possible derivational suffixes in English, but it may be seen as inflectional (Chomsky, 1970), rather than derivational. Haspelmath (1996) discusses various cases from different languages of what he calls *word-class changing inflection*, like German participles and Lezgian masdars:

- German, participle (V → Adj, Haspelmath, 1996: 44):

(1) der im   Wald laut sing$_V$-ende$_{Adj}$ Wanderer
     the in:the forest loud sing-PTCP   hiker
     'the hiker (who is) singing loud in the forest'

- Lezgian (Nakh-Daghestanian), masdar (V → N, Haspelmath, 1993: 153):

(2) wun   fad   q̃arağ$_V$-un$_N$-i   čun   tažub   iji-zwa.
     you:ABS early get.up-MASD-ERG we:ABS surprise do-IMPF
     'That you are getting up early surprises us.'

These morphemes are usually said to be inflectional, since they are totally productive (i.e. they may apply to every base verb) and formally and semantically regular. But, contrary to other inflectional forms and like most derivational ones, they are transpositional, i.e. they change the word-class of the base.

Different explanations have been offered in the past literature for these phenomena. However, every explanation is dependent from a different description of the inflection/derivation dichotomy. Before discussing all these possibilities, we should first focus on defining inflection and derivation.

1.2.1 Defining inflection and derivation

The distinction between inflection and derivation is a classical problem in morphological theory and has been widely debated. For some scholars, it is not possible to draw a clear-cut demarcation between the two, which are better defined as a gradience or continuum (Stephany 1982, Bybee 1985, Corbett 1987, Plank 1994) or as prototypical categories (Dressler 1989, Luraghi 1994). On the contrary, other linguists believe that there is a sharp distinction between them[2] (Perlmutter 1988, Anderson 1982, 1992).

Different properties have been listed as criteria to distinguish among the two categories. A summary can be found in Plank (1994); Booij (2000); Naumann and Vogel (2000); Laca (2001); Bauer et al. (2013).

1. One of the major difference between the two categories is a functional one: derivation creates new lexemes, whereas inflection creates word-forms from known lexemes, marking their role in the sentence. However, it is not always easy to distinguish between a word-form and a lexeme, also because the distinction between the two circularly relies on the distinction between inflection and derivation (Bauer et al., 2013: 8, 533). Take as an example the Italian lexicalized infinitives: forms like *sapere* ('to know, knowdlege'), *dovere* ('must, duty'), *piacere* ('to be liked, pleasure') are inflectional forms of the verb, but they can also be considered as autonomous lexemes (nouns) with their own lexicalized meaning. It could be thought that lexemes and word-forms can be distinguished by the way in which they are processed in the mind: we would expect that lexemes are stored in our mental lexicon and retrieved as unitary element, while word-forms are processed compositionally, splitting the word in the lexeme plus the inflectional information. However, it has been shown that a dual route processing is cognitively plausible, but that it mainly differentiates between high and low frequency items (Stemberger and MacWhinney, 1986).

---

[2]This model is referred to as *split morphology theory*.

2. Inflection is organized in paradigms, i.e. "sets of contrasting forms, none of which is semantically or functionally presupposed by the others" (Laca, 2001: 1215), whereas derivation usually does not. However, there are cases of derivational morphology which seem organized in paradigms, e.g. the case of event suffixes (English *-ment, -ation, -ing, -age*, etc). Moreover, inflectional forms may lack some items in the paradigm or the same word-form may fill more than one cell in the paradigm (syncretism, like Latin *mens-is* which is both the dative and the ablative plural form for *mensa*.

3. Derivation attaches closer to the root than inflection. This feature has been stated as a linguistic universal by Greenberg (1963: 93) (but see Bybee, 1985 for an alternative explanation of the order of affixes).

4. Derivation can be recursive, since a derivative can be basis for further derivation (e.g. It. *storicistico*, 'historicistic', Scalise, 1988).

5. Inflection is the part of morphology that is relevant to syntax, while derivation is not syntactically determined. "A particular syntactic context may necessitate the choice of a particular inflected form, but no syntactic context ever necessitates the choice of a form arising as the effect of a particular word-formation operation" (Stump, 2001: 55). However, derivation may be also relevant for syntax since it often determines the syntactic category and syntactic valency of the derivative. Booij (2000: 365) reports the case of the Dutch prefix *be-* which creates transitive verbs from verbs and nouns: it is derivational since it is transpositional and forms verbs with unpredictable meanings, but the transitivity effect is syntactically relevant.

6. Derivation is transpositional, i.e. it changes the word-class of the derivative, while inflection does not. This statement is problematic in two ways: first, it does not account for cases of transpositional inflection (like the cases outlined above of verbal nouns); second, it neglects cases of non-transpositional derivation.

7. Inflection is said to be more regular than derivation in different respects:

(a) It is formally more regular, since it does not create different allomorphs for the same morpheme.

(b) It is more productive, i.e. it applies without exceptions to all relevant words.

(c) It is more semantically regular, i.e. predictable and compositional; derivation, indeed, tends to acquire some degree of meaning autonomy (or idiosyncracy) from the base and from the general rule it instantiates (Laca, 2001: 1217).

No one of these features is a sufficient property to define a morphological process as inflectional or derivational. As has been seen, counterexamples have been frequently found to the generalizations made. It lead us to believe that inflection and derivation are not distinct categories, but are better represented as a continuum (as argued by Stephany 1982, Bybee 1985, Corbett 1987, Dressler 1989, Plank 1994, Luraghi 1994, among many others). Some instances of event nominalizations, which are the object of this study, are examples of constructs for whom the features listed above configure in a mixed way. They are strange entities which share some properties with derivation, others with inflection, and can be said to be in the middle.

However, from previous studies and from the analysis of our case study, I believe that (ir)regularity (in meaning, form and use) is probably the most suitable feature to distinguish inflection and derivation.

1.2.2 Ambiguous cases between inflection and derivation

Lezgian masdar, German participles, English gerunds, Italian infinitives, are all examples of linguistic forms which are not easily classified as inflectional or derivational processes. We may think of different ways in which we can interpret these phenomena:

- They are instances of derivational morphology;

- They are inflectional, but they do not change the word-class of the base;

- They are cases of word-class-changing inflection;

- They are cases of conversion.

Following the argumentation offered by Haspelmath (1996), the cases considered here cannot be regarded as derivational, since they are regular, general and productive. This feature seems more relevant to the definition of inflection/derivation than the feature of transpositionality. Others have argued that these are cases of inflection, in the traditional sense, i.e. they do not change the word-class of the derivative (e.g. Scalise, 1988: 566). In this sense, the German participle in (1) and the Lezgian masdar in (2) should not be considered, respectively, adjective and noun, but should be said to keep the word-class of the base, i.e. verbs. This view, however, overlooks the fact that they behave like adjectives and nouns in many aspects (e.g. German participles share the inflectional patterns of adjectives, while Lezgian masdars form all sixteen cases that other nouns have), even if some traits are shared with less prototypical members of these classes (e.g. masdars lack plurals, like some other Lezgian nouns). This idea goes together with a syntactic treatment of these phenomena: the word, with its original word-class, would be converted into another syntactic category forcing its use in a syntactic position that it normally does not occupy. In some theories (Farrell, 2001), it is said that a single lexical item (like *hammer* in English) is unspecified for word-class (or is multifunctional) and its category is determined by the syntactic context in which it is used. These theories, however, are strictly dependent with a precise model of syntax and grammar. Plag (2003: 115 and ff.) proposes a theory-independent argument to distinguish syntactic and morphological phenomena. He argues that morphological formations are subjects to idiosyncracies, whereas syntactic patterns are exceptionless and totally compositional. Considering nominal infinitives in Italian, I believe that they should be considered as morphological (specifically, inflectional), rather than syntactic processes since they show some lexicalized cases (e.g. *il sapere* 'knowledge', *i doveri* 'duties'). The same is true for other languages and other phenomena, such as German infinitives and English gerunds. They present some items which have exceptional properties.

The interpretation supported by Haspelmath (1996) regards these phenomena as word-class-changing inflection (or transpositional inflection). They are inflectional, since they are regular and productive, but they do change the word-class of the base.

I believe that this interpretation is equivalent to considering these phe-

nomena as conversion. If we assume that a conversion is in act, we take it for granted that it is the inflected form that is converted, thus inheriting all the regularities of inflectional processes. On the other hand, if we define it as *word-class changing inflection*, we assume that a change of category is taking place, i.e. a conversion or transposition. For this reason, I treat the two terminologies as theoretically equivalent.

1.2.3   The inflection-derivation continuum

Transpositional inflectional processes are, of course, closer to derivation than non-traspositional ones. If we represent the inflection-derivation distinction as a continuum, they will be placed in the middle. The different properties that we have listed above (sec. 1.2.1) will constitute as well a continuum, in which: inflection will be more general and productive, more semantically transparent, more syntactically relevant; derivation will be less productive, semantically more arbitrary and opaque, less relevant syntactically.

Inflection	Transp.Infl.	**Derivation**
+ productive		- productive
+ semantically transparent		+ semantically opaque
+ syntactically relevant		- relevant syntactically
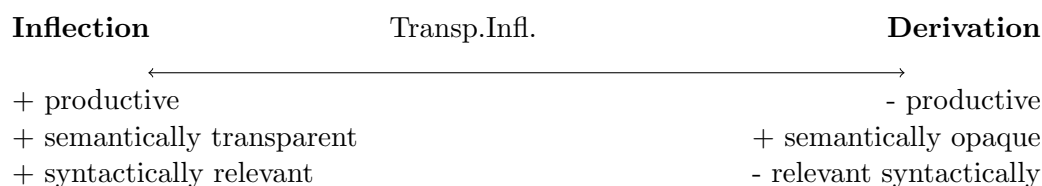
Moreover, Haspelmath (1996) notices an additional characteristics of the inflection-derivation continuum. With words derived by inflectional morphology, the internal syntax of the base is more preserved in the derived word; words derived by derivational morphology, instead, tend to alter the internal syntax of the base and to inherit the internal syntax of the new word-class.

This point is particularly relevant for nominalizations, since the internal syntax corresponds to the way in which the arguments of the base verb are expressed. When in a language two different morphological processes are involved in event nominalizations, different degrees of preservation of the internal syntax will be shown: Haspelmath predicts that the one formed by inflectional means will preserve more the internal syntax, while the derivational one will lose more features.

As will be seen in the next section, this prediction is only partially true. Derivational nominalizations do lose the internal syntax of the base verbs; on

the other hand, nominalizations obtained by transpositional inflection frequently show mixed behavior, as will be seen in the case of English gerunds and Italian nominal infinitives.

It can be noted that, in the case of event nominalizations, another parallel distinction is expressed by this continuum: the noun-verb distinction. Nominalizations will become closer to proper nouns when derivational and will remain closer to base verbs when inflectional. Argument realization, indeed, is one of the features that pushes a nominalization toward the nominal or the verbal pole, and it is the topic of the next section.

1.3 Nominalizations and the inheritance of verbal and nominal properties

In a transposition process, event nominalizations acquire some properties of their new word-class, i.e. nouns, and preserve others of the verbal class. Different languages and different forms of nominalizations show different degrees of inheritance of these properties[3]. As we said at the beginning of the chapter, categorization of different kinds of nominalizations can be based on the presence or absence of nominal and verbal features (categorization that, however, usually reveals parallel to the one based on the morphological process involved).

Like verbs, event nominalizations refer to events and correspond to *second-order entities*, in the terms of Lyons (1977: 443). In some cases, they admit adverbial modification (like verbs), while in others they only accept adjectives as modifiers (like proper nouns). This point is illustrated by English *-ing* forms, which show both adjective and adverb modification.

(3)   Brown's **deft** painting of his daughter is a delight to watch.

(4)   Brown's **deftly** painting his daughter is a delight to watch[4].

However, usually, derivational nominalizations only admit adjectives.

(5)   The complete destruction of the city

(6)   *The completely destruction of the city

---

[3]It is possible to observe a *gradience in verbal traits*, as called by Alexiadou (2013).
[4]Examples from Quirk et al. (1985: 1290-91).

As example (5) illustrates, nominalizations can be headed by determiners. Moreover, they can inherit the arguments of their base verb, but the way in which they realize them can differ significantly and may end up looking more like nominal or sentential realization.

1.3.1 Argument realization

Since the work by Lees (1960) and Chomsky (1970), the relation between pairs like ex. (7) and (8) has been widely investigated.

(7)   The enemy destroyed the city.

(8)   The enemy's destruction of the city

The nominalization in (8) shares the arguments of the corresponding sentence in (7), even if it expresses them in different form. The way in which these arguments are realized has been one of the most investigated aspects of nominalization, since the seminal work by Grimshaw (1990).

In the sentence

(9)   Her revision of the draft was accurate.

'revision' has inherited two arguments from its base verb 'to revise': the object 'draft' and the subject, expressed by the pronoun 'her'. These arguments are the same expressed in a corresponding finite clause, even if they are realized in a different way:

(10)   She accurately revised the draft.

In this case, the nominalization in (9) realizes the internal argument as a prepositional phrase and the subject as a possessive pronoun. Moreover, it cannot be modified by an adverb, like the verbal phrase in (10), but only by an adjective:

(11)   Her accurate/*(accurately) revision of the draft.

However, alternative nominalizations may realize differently the same arguments. In English, the deverbal noun 'revision' can be compared to the gerund 'revising', which presents a different argument realization:

(12)   Her accurate revising the draft.

The gerund itself usually show multiple configurations too. Consider the following examples from Quirk et al. (1985: 1290-1291):

(13)   Brown's deft painting of his daughter is a delight to watch.

(14)   Brown's deftly painting his daughter is a delight to watch.

(15)   I dislike Brown painting his daughter.

In (15) the arguments of 'painting' are realized in the same way as in a sentence (like *Brown paints his daughter*); in (14) the subject is expressed by a Saxon genitive, a typical characteristic of nouns, even if the object 'his daughter' is not expressed by a prepositional phrase like in (13) and is modified by an adverb, instead of an adjective. It seems, indeed, that this construction presents in some domains a clear nominal structure, while in others it preserves some properties of its original class, i.e. verbs.

This fact has led some scholars to posit the distinction of two (not considering cases of *-ing* forms as participles) different gerunds: verbal gerund and nominal gerund (or mixed nominalization, Alexiadou, 2013). However, two types are probably not enough to explain the large morphosyntactic variance displayed by English -ing forms (Pullum and Zwicky, 1999, list twenty-five different construction in which -ing forms are used) and in many cases it is not easy to understand where to posit the boundaries. Reconsidering some of the examples presented above, taken from Quirk et al. (1985: 1290-91) and reported in Table 1.1, we can observe how English *-ing* constructions form a gradience from lexicalized cases (in which they act like common nouns) to verbal configurations.

In the examples in the first and second row, the *-ing* form is a true deverbal noun, with plural endings and a concrete, countable denotation. In this case, *painting* is considered a deverbal noun, like nouns derived with other suffixes, such as *-ation* or *-ment*. In examples 3 and 4, *painting* shows nominal features, i.e. the determiner, adjective and the realization of subject by means of a genitive construction. In this case, however, it is not countable and does not refer to a concrete object; rather, it is an abstract uncountable noun and it refers to something like the act or the manner of painting[5]. While similar in meaning, examples 5 and 6 present more verbal traits than

---

[5]It is also called *nominal gerund*, in contrast with *verbal gerund* of sentences 5 and 6 (Alexiadou, 2013).

1	some paintings of Brown's	Count
2	Brown's paintings of his daughter	noun
3	The painting of Brown is as skilful as that of Gainsborough	Nominal
4	Brown's deft painting of his daughter is a delight to watch	gerund
5	Brown's deftly painting his daughter is a delight to watch	Verbal
6	I dislike Brown's painting his daughter	gerund
7	I dislike Brown painting his daughter	
8	I watched Brown painting his daughter	
9	Brown deftly painting his daughter is a delight to watch.	
10	Painting his daughter, Brown noticed that his hand was shaking.	Present participle
11	Brown painting his daughter that day, I decided to go for a walk.	
12	The man painting the girl is Brown	
13	The silently painting man is Brown	
14	Brown is painting his daughter	

Table 1.1: Gradience of use of English -*ing* forms.

the sentences in 3 and 4. The modifier is not an adjective anymore, but an adverbs, and the internal argument is expressed as a direct object, and not by a PP. The genitive construction is still used, and thus the -*ing* form (called in this case *gerund* or *verbal gerund*) has mixed nominal and verbal features. The meaning of ex.6 can also be interpreted as 'the fact that Brown is painting'. Quirk et al. (1985) refer to the -*ing* form in sentences from 7 to 15 as *present participles* in all cases, without highlighting further distinctions. Here, the -*ing* forms does not present nominal features anymore; it acts like a complement clause of the matrix predicate.

In the case of English gerunds, the different configurations of nominal and verbal features are visible for the same grammatical form, which shows mixed behavior. It is common among non-finite verbal forms used in nominal function, as will be illustrated for Italian nominal infinitives.

However, cross-linguistically, the same gradience of properties has been observed among different forms of nominalization, ranging from a total preservation of the internal syntax of the base verb to the acquisition of the realization typical of nouns. Koptjevskaja-Tamm (1993) offers a complete typological study on the different ways in which the languages around the world encode the arguments of their nominalizations. Her survey, based

on a sample of 70 languages[6], identifies four major types, which were encountered more frequently:

1. Sentential type (SENT): the arguments marking is signalled in the same way as in the corresponding finite clause;

2. Possessive-Accusative type (POSS-ACC): the subject (both of transitive and intransitive verbs) genitivize, while the direct object retains the case assigned in finite clause ("the relation between the subject and the nominalization is expressed in the same way as the relation between the possessor and the possesse in a non-derived NP", Koptjevskaja-Tamm, 2003: 728);

3. Ergative-Possessive type (ERG-POSS): the subject of intransitive verbs and the object of transitive ones are encoded in the same way (as in ergative language, e.g. Dixon), i.e. by genitivization, while the subject of transitive verbs is realized in the instrumental case;

4. Nominal type (NOMN): in a first sub-type, called Double-Possessive, all the subjects and objects are realized in the genitive case; in a second sub-type, called Possessive-Adnominal, the subjects are genitivized, while the direct object gets the same marking as oblique NPs.

Lastly, it should be noted that the realization of arguments is not obligatory. In many sentences a nominalization (either a deverbal noun or a verbal noun) can lack the overt expression of its arguments. Their reference can be grasped from the context or can be generic. Contrary to what argued by Grimshaw (1990), the presence or absence of arguments is not directly linked to the event interpretation of the nominal. In her work, she distinguishes between *simple event nominals* and *complex event nominals*: the former lack full realization of arguments, the latter appear with all of them expressed. She claimed that only complex event nominals can have an event structure, i.e. an event reading, whereas simple event nominals present only result reading. Despite the numerous subsequent works which have gone in the same direction (e.g. Alexiadou, 2001; Roy and Soare, 2011), most of the claims made by Grimshaw have found counterexamples in recent

---

[6] Koptjevskaja-Tamm (2003) enlarges the sample with 60 European languages. The results of her survey are also represented in a map available online: `http://wals.info/feature/62A#2/23.2/148.2` (Koptjevskaja-Tamm, 2013).

studies with corpus data (Grimm and McNally, 2013; Lieber, 2016). It has been shown, indeed, that event nominalizations of different kinds (formed by means of suffixes, by conversion or by an ing form) show event or result readings independently from the arguments they express.

1.4 Meaning-based classification

Since the work by Walinska de Hackbeil (1984); Zubizarreta (1987); Lebeaux (1986); Grimshaw (1990) among others, it has been recognized that event nominalizations show the possibility to express, at least, an event and a result meaning. Usually, they are ambiguous between the two readings, as in the following examples:

(16)   This new construction is ugly.   Result

(17)   The construction of the building took three years.   Event

However, in addition to the result reading, further non-eventive interpretations have been recognized. Melloni (2007) proposes the term *referential* readings to cover the wide range of non-eventive meanings associated with nominalizations. These include *result* (ex. 16, frequently divided into *product* and *result state*), *instrument* (18), *location* (19), *path* (20), *manner* (21) and *fact* (22).

(18)   For **decoration**, three turquoise seahorses descended the wall at a forty-five-degree angle. (Happinees Key 2009, COCA corpus, Bauer et al., 2013: 210)

(19)   It is in the same building as the **dwelling** of Irving Kristol and Gertrude Himmelfarb. (American Spectator 2009, COCA corpus, Bauer et al., 2013: 211)

(20)   In 1924, the United States claimed the North Pole was an underwater **continuation** of Alaska. (Journal of International Affairs 2008, COCA corpus, Bauer et al., 2013: 211)

(21)   The professor's **demonstration** of the technique was deft. (Bauer et al., 2013: 207)

(22)   The professor's **demonstration** of the technique was a scandal.

(Bauer et al., 2013: 207)

1.5 Event nominalizations in competition

We have seen how different types of event nominalizations are attested in the languages of the world and how in the same language more than one form may be available. English, for example, presents nominalizations derived by means of different suffixes, by conversion or by the transpositional use of the *-ing* constructions.

Some questions naturally follow: assuming that a language's history has delivered us a variety of ways to turn verbs into nouns, how do we make use of this variety today? How do we choose between one form or the other? Are they competitors in a single paradigmatic cell of semantic derivation (Booij and Lieber, 2004), i.e. the one expressing event meanings? The rest of this thesis is devoted to the search for an answer to these questions.

Before presenting some empirical investigations, let's hypothesize some possible explanation of this phenomenon.

When two or more suffixes can be used to convey (more or less) the same meaning (as in the case of EDNs), two scenarios can be usually observed. In the first one, the suffixes are in complementary distribution and create new forms from bases belonging to different domains. In the second one, they will not be distinguished by the kind of input they take and will yield multiple EDNs from the same verbal base; in this case, further meaning or distributional distinctions will explain their co-existence.

Let's consider some examples of the first situation. We know, for instance, that in English affixation is preferred when forming event nouns from Latin roots, while conversion is mostly used for native ones (Bauer et al., 2013: 196). Moreover, verbs formed with the suffixes *-ize* and *-ify* prefer nominalizations with *-ation*, instead of *-al*, *-ance*, *-ment*, and *-ure* (Bauer et al., 2013: 196-203). In these cases, for each verbal base we obtain one form of nominalization, which depends on some property of the base.

On the other hand, there are cases in which two or more suffixes can be applied to the same base and understanding which one is preferred to the other may require the analysis of various syntactic, semantic or pragmatic features, whose behavior is not always easy to interpret.

Thai offers a clear context-based distinction among two event nominal-

17

izers (*kaan* and *khwam*), based on a semantic aspect of the output: *kaan* derives process nouns, *khwam* non-process ones (Comrie and Thompson, 2007: p. 336):

(23)   chyâ ('believe'):

     a.   kaan chyâ
          'believing' (process)

     b.   khwam chyâ
          'belief'   (non-process)

In the next chapters, I present some empirical investigations on the differences between various forms of event nominalization in Italian and German. In both languages, I focus on two classes of nominalization: event deverbal nouns (EDNs) derived by means of a suffix from the base verb, and nominal infinitives (henceforth NIs), i.e. nominal forms of the verbs which are used as nouns and which acquire morphosyntactic nominal features. Differently from the example of Thai, if you ask native speakers of these two languages the difference between the two forms of nominalization they will not be able to answer clearly. In many cases, these forms are perceived, indeed, as similar in meaning and use. In what follows, I will test this equivalence of nominalizations by means of empirical studies. Various differences between the two forms will emerge, suggesting possible solutions to this intricate puzzle.

Chapter 2

Italian event nominalizations

In this chapter, I introduce two forms of Italian nominalizations: event deverbal nouns, formed with various suffixes, and nominal infinitives, non-finite verbal forms which can be used as nouns. Here I discuss their morpho-syntactic properties, before investigating the rationale behind their use in chapter 3 and 4. Previous accounts of their differences are outlined and criticized in section 2.3. I show that ontological distinctions among their referents, found by means of introspective judgments, are not suited to explain this puzzle; corpus analyses, instead, can give us a deeper insights.

2.1   Italian event deverbal nouns

As previously specified, with the term *event deverbal noun* (EDN), I refer to every noun derived from a verb which has an event meaning, where event includes every kind of *eventuality* (Vendler, 1957; Bach, 1986). In Italian, the EDNs suffixes available are:

- **-zione**: *venerazione* ('veneration'), *rotazione* ('rotation');

- **-mento**: *annegamento* ('drowning'), *procedimento* ('procedure');

- **-tura**: *spuntatura* ('trim'), *crepatura* ('crack');

- **-aggio**: *smontaggio* ('dismanteling'), *atterraggio* ('landing');

- **-ata**: *sbirciata* ('peek'), *suonata* ('one act of playing');

- **-nza**: *aderenza* ('adherence'), *permanenza* ('permanence, stay');

- **conversion nouns in -o**: *aumento* ('increase'), *viaggio* ('trip').

Thornton (1990, 1991) analyzes the morpho-phonetic properties of the nouns in *-mento* and *-zione*, explaining how the suffixes combine with the verbal root (or, as she suggests, the imperative form of the verb). Gaeta (2002, 2004, 2009) offers a complete description of their morphosyntactic and semantic properties. As English EDNs, they acquire the full set of nominal properties, from argument realization to the possibility of being pluralized. A detailed exemplification of these properties is given in the next section in parallel with the description of the NI. Gaeta also takes into consideration the inheritance of the *aktionsart* of the base verb and describes cases in which it is modified. Melloni (2007: 87-91), however, suggests that deverbal suffixes never modify the actional features of the base verbs. In the same work (as well as in Melloni, 2006, 2008; Ježek, 2007), she further investigates the polysemy of these suffixes, which, as introduced in the last chapter, can acquire a variety of non-eventive readings.

As this list of studies shows, Italian EDNs have been frequently studied. However, their relation and difference with nominal infinitives has received less attention. Only few lines can be found in Gaeta (2002) and Melloni (2007). In works devoted to the analysis of nominal infinitives (such as Skytte, 1983 and Vanvolsem, 1983) the issue is also not analyzed in depth: their use is explained (with few words) as a matter of variation or as a difference in register. The only work devoted to the topic is probably Zucchi (1993): his proposal will be outlined and criticized in section 2.3, after having introduced the properties of NIs.

2.2 Non-finite verbal forms: Italian nominal infinitives

The Italian infinitive is usually considered a mood of the verbal category (Ramat, 2002: 413), which, however, differently from other moods, does not specify any additional semantic information regarding the speaker's attitude or opinion about what is said[1]. This was the reason why the grammarians from ancient Greece called it *aparémphaton*, 'that does not determine clearly' (Ramat, 2002: 410), "neque numeris neque generibus praeserviens[2]"

---

[1]Rephrasing a definition of modality given by Palmer (2001).

[2] "Affected neither by number or gender".

(Aulus Gellius, *Noctes Atticae*, 1,7).

With participles, gerunds and converbs, infinitives are defined as non-finite verbal forms, since they are not marked for categories like tense, mood, aspect, person or number, and because they cannot function as matrix predicates of independent sentences (Koptjevskaja-Tamm, 1999: 146, Ylikoski, 2003: 186, Cristofaro, 2007, Quirk et al., 1985: 149-151). They can be used as complements of the main predicate, even if probably in the more nominal configurations, they are better described as arguments (or as *nominalized complements*, Noonan, 2007: 70).

Italian infinitive, indeed, is not marked for number and person, differently from finite verbal forms. It presents two temporal forms, the present[3] (ex. 1) and the past infinitive (ex. 2), called also respectively *infinito semplice* and *infinito composto* ('simple' and 'complex infinitive'), and can be used in the passive form (ex. 3).

(1) **Andare** in vacanza ti   gioverà.
     Go.INF on vacation to you help.FUT
     'Going on vacation will help you'.

(2) Giulio lo   ringrazió   per **aver pulito** tutto.
     Giulio him thank.PAST for have.INF   cleaned everything
     'Giulio thanked him for having cleaned everything'.

(3) Vuole   **essere richiamato** sul   cellulare.
     Want.3sg be.INF   recalled on the   mobile phone
     'He wants to be recalled on the mobile phone'.

As in other languages, in reason of its being a non-finite form, it cannot occur as the main predicate of a sentence, but it always functions as the predicate of a dependent clause. Skytte (1983) and Skytte et al. (2001) describe the variety of propositions and syntactic patterns in which it may occur. Summarizing the main functions, it may occur as:

- Complement of the matrix predicate (subject, as in example (1), object or as other argument);

- Adjunct of the matrix predicate (e.g. interrogative or purposive propo-

---

[3]The present infinitive, however, does not have always a temporal value of simultaneity with the main sentence. As has been described by Bertinetto (2001b), its temporal and aspectual valency is far more multifaceted. For this reason, I will prefer the terms simple and complex infinitive.

sition, ex. 2);

- Relative proposition:

  (4) Ha   trovato qualcuno con   cui   **andare** al
       Has.3sg found   somebody with whom go.INF to the
       cinema.
       cinema
       'He has found somebody with whom he can go to the cinema'.

- Complement of a noun:

  (5) Ha   voglia di **viaggiare**.
       Has.3sg desire of travel.INF
       'She has the desire of travelling'.

- Complement of an adjective:

  (6) Oreste era stanco di **ascoltare** fandonie.
       Oreste was tired   of listen.INF nonsense
       Oreste was tired of listening nonsense.

- Head of a noun phrase:

  (7) Il   **passeggiare** per   Parigi rese   quel pomeriggio
       The stroll.INF   along Paris make.PAST that afternoon
       indimenticabile.
       unforgettable.
       The strolling along Paris made that afternoon unforgettable.

The last point is the main interest of this thesis. In these cases, the infinitive acquires more nominal properties, such as the presence of a determiner or the possibility of being modified by an adjective, but it retains some verbal ones, like the realization of the direct object as NP.

I will describe these properties in a moment, but first a terminological clarification. To cases in which the infinitive has clear nominal features, first of all a determiner (like ex. 7), I will refer with the label *nominal infinitive* (NI). This corresponds to what is called in Italian "infinito sostantivato". When the infinitive is not preceded by a determiner, but is in a clear nominal function (like as a subject, ex. 1), I will call it *bare nominal infinitive*

(henceforth BNI). As will be shown, it lacks most of the nominal features and, despite its syntactic role, it is equal to the infinitive in verbal uses (which I will call *verbal infinitive*). In the end, I will show how there is a gradation from the more nominal use of the infinitive (i.e. the lexicalized infinitives) to the verbal ones, with in between the NI and the BNI.

2.2.1   Syntactic properties of nominal infinitives

Argument realization

Like EDNs (ex. 9), NIs can express their subject within a prepositional phrase introduced by the preposition *di*, 'of' (or the complex forms *del/della/dello*, 'of the', ex. 8):

(8)   il   rimbombare **della sua voce**
        the resound.INF of the his   voice
        The resounding of his voice.

(9)   il   rimbombo **della sua voce**
        the echo   of the his   voice
        The echo of his voice.

In both nominalizations, the subject can also be expressed by a personal pronoun:

(10)   Il **suo** tacere   valeva   più di   mille   parole.
         The his keep.quiet.INF was worth more than thousand words
         His keeping quiet was worth more than a thousand words.

(11)   La **sua** caduta mi   fece   spaventare.
         The his fall   to me made scared
         His fall scared me.

With a complex nominal infinitive, it is also possible to express the subject in the nominative case, even if it is perceived as more literal:

(12)   L'aver   **egli** compiuto i   primi studi   in Francia
         The have.INF he   finished   the first   studies in France
         Having he finished the first studies in France[4]

BNIs, instead, do not allow any realization of the subject:

---

[4]Examples from Contini, *La letteratura italiana*, p. 289, cited in Skytte et al. (2001).

(13) *Rimbombare della sua voce.
       resound.INF of the his voice

(14) *Suo tacere   valeva   più di   mille   parole.
       His keep.quiet.INF was worth more than thousand words

(15) *Aver   egli compiuto i   primi studi   in Francia
       Have.INF he   finished   the first   studies in France

The interpretation of the subject of BNI is, thus, left to contextual and
pragmatic factors. First, when the BNI has a generic interpretation, the
subject is arbitrary:

(16) Correre fa   bene alla   salute.
       Run.INF make.3sg good for the health.
       Running is good for health.

In the last example, the running event is generic, it is not referred to any
episodic and specific instance of the act of running. The agent of the action
is thus not specified. Note, however, that the same interpretation is possible
with NIs and EDNs:

(17) Il   correre fa   bene alla   salute.
       The run.INF make.3sg good for the health.
       (The) running is good for health.

(18) La corsa   fa   bene alla   salute.
       The run.NOUN make.3sg good for the health.
       Run is good for health.

If the matrix predicate has as subject an experiencer, the subject of the
dependent BNI is the experiencer itself, which controls the BNI:

(19) Danzare   mi   piace.
       Dance.INF to me like
       I like dancing.

In example 19, the only possible interpretation is that 'I' is the subject of
the act of dancing. It is not possible to interpret it as 'I like that the others
dance'. A matrix predicate like *sorprendere*, 'to surprise', sounds odd in
example 20, since one cannot be surprised by its own action:

(20) ??Danzare   mi   sorprese.
        Dance.INF to me surprise
        ??Dancing surprises me.

NIs and EDNs subjects, instead, are not controlled by the matrix predicate. The next sentences can be interpreted both as (i) I like dancing/dance, (ii) I like seing people dancing:

(21)   La danza   mi   piace.
        The dance.NOUN to me like
        I like dance.

(22)   Il danzare   mi   piace.
        The dance.INF to me like
        I like the dancing.

As already noted, with NIs and EDNs the subject can be overtly expressed , and thus any ambiguity is avoided.

Summing up, the need to overtly express the subject of the nominalizations can be a factor in deciding whether to use BNIs or a NIs/EDNs. With matrix predicates which have control on the subjects of BNI, NIs or EDNs can let the speaker express a different subject, like in the example below:

(23)   Il danzare   di Luca mi   piace.
        The dance.INF of Luca to me like
        I like Luca's dancing.

For the realization of the internal argument, NIs and BNIs pattern together and differ from EDNs. NIs and BNIs realize the internal argument as a direct object (ex. 26 and 27), whereas EDNs encode it in a prepositional phrase with *di*, 'of', in the same way as they encode subjects (ex. 24, not grammatical with an NI or BNI, ex. 25):

(24)   La degustazione   di un buon bicchiere di vino era un piacere
        The tasting.NOUN of a   good glass   of wine was a   pleasure
        al quale egli non poteva rinunciare.
        to which he   not could   renounce
        The tasting of a good glass of wine was a pleasure he could not renounce.

(25)   *Il   degustare di un buon bicchiere di vino ...
        The taste.INF of a   good glass   of wine ...

(26) Il   degustare un buon bicchiere di vino era un piacere   al
      The taste.INF a good glass   of wine was a   pleasure to
      quale egli non poteva rinunciare.
      which he   not could   renounce
      (The) tasting a good glass of wine was a pleasure he could not
      renounce.

(27) Degustare un buon bicchiere di vino era un piacere   al quale
      Taste.INF a   good glass   of wine was a   pleasure to which
      egli non poteva rinunciare.
      he   not could   renounce
      Tasting a good glass of wine was a pleasure he could not renounce.

In the case of transitive verbs, thus, EDNs can realize with the same formal
means its internal and external argument, whereas NIs and BNIs differenti-
ate them by the use of a NP or a PP.

Modifiers

With respect to modifiers, NIs have a mixed behavior: they both accept ad-
jectives (typical nominal modifiers) and adverbs (typical verbal modifiers).
However, further distinctions are necessary, based on the position of modi-
fiers. First, the NI can occur with a preceding adjective (ex. 28), like EDNs
(ex. 29) and contrary to BNIs (ex. 33):

(28) L'irreparabile,   continuo   invecchiare del   nostro corpo
      the unrepairable continuous age.INF   of the our   body
      The unrepairable, continuous ageing of our body[5]

(29) L'irreparabile   invecchiamento del   nostro corpo
      the unrepairable ageing.NOUN   of the our   body
      The unrepairable ageing of our body

(30) *Irreparabile   invecchiare del   nostro corpo
      Unrepairable age.INF   of the our   body

Moreover, NIs can be modified by an adjective after it, like in example (31):

(31) il   coincidere   rapidissimo   degli sguardi
      the coincide.INF extremely quick of the gazes

---

[5]Example from Pirandello, *Sei personaggi in cerca d'autore*, p.16, cited in Skytte et al.
(2001).

The extremely quick coinciding of gazes[6]

As for preceding adjectives, EDNs allow a post-N adjective, whereas BNIs do not:

(32)   L'invecchiamento irreparabile del   nostro corpo
           the ageing.NOUN unrepairable of the our   body
           The unrepairable ageing of our body

(33)   *Invecchiare irreparabile del   nostro corpo
           Age.INF   unrepairable of the our   body

BNIs seem to accept a following adjective only when it refers to an implicit subject:

(34)   Disegnare chini sul   foglio
           Draw.INF bent on the sheet of paper
           Drawing bent on the sheet of paper

However, when NIs express the internal argument, they do not allow a following adjective (ex. 35); if the adjective precedes the NI, it sounds completely acceptable (ex. 36).

(35)   *Il   loro riprodurre   accurato l'immagine
           The their reproduce.INF accurate the picture

(36)   Il   loro accurato riprodurre   l'immagine
           The their accurate reproduce.INF the picture
           Their accurate reproducing the picture

NIs accept also adverbs as modifiers (both when the NI overtly express the subject or the object, ex. 37 and 38), in the same way as BNIs (ex. 39) and differently from EDNs (ex. 40):

(37)   Il   lavorare   continuamente di Luigi
           The work.INF continuously   of Luigi
           Luigi's working continuously

(38)   Il   costruire   continuamente nuove case
           The construct.INF continuously   new   houses
           The constructing continuously new houses

---

[6]Italian example from the novel *Un amore*, by D.Buzzati, p.26, cited by Skytte et al. (2001).

(39)   Costruire   continuamente nuove case
           Construct.INF continuously   new   houses
           Constructing continuously new houses

(40)   *La   costruzione   continuamente nuove case
           The construction continuously   new   houses

Negation is, indeed, acceptable with all the three forms; in Italian, the same negation form is used both for verbs and nominals, even if its use with EDNs is often marginal:

(41)   Il   suo non parlare   preoccupava i   suoi genitori.
           The his not talk.INF worried   the his   parents.
           His not talking worried his parents.

(42)   Non parlare   più   con lui era strano.
           Not talk.INF anymore with him was odd
           Not talking with him anymore was odd.

(43)   Era il   punto di non ritorno.
           Was the point of no   return
           It was the point of no return.

2.2.2   Lexicalized infinitives

A restricted number of nominal infinitives have been lexicalized, i.e. they have acquired diachronically the status of independent lexemes. The main difference with non lexicalized NIs is that their meaning differs more from the simple event interpretation, and in some cases lexicalized NIs also denote the result of the action. Moreover, in most cases, it is possible to pluralize the infinitive. The majority of them are modal verbs, which preserve the non lexicalized meaning when used as NIs in conjuction with another verb, i.e. like a modal NI. Let's see a list of lexicalized NIs:

- *sapere*, 'to know': *il sapere* ('the act of knowing' or 'knowledge'), also PL. *i saperi*;

- *dovere*, 'to must': *il dovere* ('duty'), also PL. *i doveri*;

- *volere*, 'to want': *il volere* ('the act of want something' or 'wish/ will'), PL. *i voleri*;

- *potere*, 'to can': *il potere* ('power'), PL. *i poteri*;

- *piacere*, 'to like': *il piacere* ('pleasure'), PL *i piaceri*;

- *dispiacere*, 'to be sorry': *il dispiacere* ('displeasure, sorrow'), PL *i dispiaceri*;

- *parere*, 'to seem, to beleive': *il parere* ('opinion'), PL *i pareri*;

- *avere*, 'to have': as lexicalized NI, used mainly in the plural form *gli averi*, 'belongings';

- *essere*, 'to be': as lexicalized NI, used mainly in the expression *essere umano*, 'human being', PL. *gli esseri umani*, 'the human beings';

- *mangiare*, 'to eat': *il mangiare*, as normal NI 'the act of eating' or in the lexicalized sense of 'food', 'what is eaten';

- *bere*, 'to drink': *il bere*, as normal NI 'the act of drinking' or in the lexicalized sense 'drinks, beverage' or 'alcohol abuse';

- *avvenire*, 'to happen': *l'avvenire*, 'future', no plural form;

- *ammontare*, 'to amount to': *l'ammontare*, 'the amount, sum'

2.2.3   Nominal and verbal features of NIs: summary

Italian infinitives frequently show a mixed behavior between nouns and verbs. In the more extreme case, they acquire plural marks and an idiosyncratic meaning, which make them acquire the status of independent lexemes. In this case, we talk about lexicalized infinitives.

In more common cases, Italian infinitives can be used in nominal functions and acquire only a part of the properties proper of nouns. When they are preceded by a determiner (definite or indefinite articles, or demonstrative pronouns) they are called nominal infinitives (also known in Italian as *infinito sostantivato*). From the category of nouns, they acquire the subject's realization (by means of the prepositional phrase introduced by *di*, 'of') and the possibility of being modified by a preceding adjective; adjectives which follow the NI are acceptable only if the internal argument of the NI is not expressed (whereas Italian nouns accept adjectives in both position). Contrary to nouns, and similarly to verbs, NIs express the internal argument as a direct object (and not as a prepositional phrase) and can be modified by adverbs.

These features and their occurrence with the different kinds of nominalization are summarized in table 2.1.

		EDN	Lex. NI	NI	BNI
Subject realization	Prep. phrase with *di*, 'of'	✓	✓	✓	✗
	Poss.pronoun (e.g. *suo*, 'his')	✓	✓	✓	✗
Modifier	Preceding Adj	✓	✓	✓	✗
	Following Adj	✓	✓	(✓)	✗
	Adverb	✗	✗	✓	✓
Object realization	Direct object	✗	✗	✓	✓
	Prep. phrase with *di*	✓	NA	✗	✗

Table 2.1: Summary of nominal and verbal properties of Italian nominalizations.

2.3 Previous accounts of Italian nominalizations

In the past literature, it has been argued that the difference between various nominalizations can be explained in terms of reference to different ontological entities. In particular, Zucchi (1993) formulates this proposal in a unified approach to both Italian infinitives and English gerunds. He argues that nominalizations can denote two different kinds of entities: events or propositional entities.

With regards to Italian, he divides NIs in four sub-types (1993: 220):

1. Past infinitive:

    (44)   l'   aver   egli   scritto quella lettera
            the have-INF he-NOM written that   letter
            'the having he written that letter'

2. Infinitive + adverb:

    (45)   il   suo   mormorare   sommessamente
            the his/her whisper-INF quietly
            '(the) his whispering quietly'

3. Infinitive + direct object:

(46) il   suo   mormorare parole dolci
      the his/her whisper-INF words sweet
      '(the) his whispering sweet words

4. Infinitive + adjective + prepositional subject:

(47) il   mormorare sommesso del   mare
      the whisper-INF quiet   of-the see
      'The quiet whispering of the see

Zucchi claims that the first two cases refer to propositional entities, in the same way as nominals introduced by 'il fatto che' ('the fact that') do. This would be shown by the possibility to occur with certain kinds of predicates that, in Zucchi's opinion, accept only proposition as their arguments (like 'to explain', 'to imply', examples 48, 49, 50). Moreover, they could not refer to events because they cannot occur with event-predicates (e.g. 'to see', 'to hear', 'to last x times', examples 51, 52, 53, 54) :

(48) l'   aver   egli compiuto i   primi studi   in Francia spiega
      the have-INF he   done   the first   studies in France explains
      come la   sua attività letteraria si   irradi   da
      how the his activity literary   Impersonal-particle spreads from
      Parigi
      Paris
      Having him done his first studies in France explains how his literary
      activity spreads from Paris

(49) Il   suo mormorare sommessamente non implica che egli
      The his whisper   softly   not implies that he
      sia   nei guai.
      is-SUBJ in   trouble
      'His whispering softly does not imply that he is in trouble.'

(50) Il   fatto che egli abbia   compiuto i   primi studi
      The fact that he   have-PAST-SUBJ done   the first   studies
      in Francia spiega...
      in France implies
      'The fact that he has done his first studies in France implies ...'

(51) ??Gianni   ha visto il   fatto che i   nostri giovani partono per
      ??Gianni has seen the fact that the our   kids   leave   for

31

la guerra.
the war
'?Gianni has seen the fact that our kids leave to the war.'

(52) ??Giovanni udì il suo mormorare sommessamente.
??Giovanni listened the his whisper-INF softly-ADV
'??Giovanni listened his whispering softly.'

(53) ??Giovanni vide l'incedere minacciosamente della
??Giovanni saw the-advance-INF threateningly-ADV of-the
folla.
crowd
'??Giovanni saw the advancing threateningly of the crowd.'

(54) ??Il fatto che io sono a corto di denaro è durato alcuni
??The fact that I am at without of money is lasted some
giorni.
days
'??The fact that I am without money is lasted some days.'

On the other hand, he believes that the third and fourth types of infinitive constructions (ex. 46 and 47) refer to events, since they can occur with event-predicates, as in the following examples:

(55) Gianni ha visto l'incedere minaccioso della folla.
Gianni has seen the-advance-INF threatening of-the crowd
'Gianni has seen the threatening advancing of the crowd.'

(56) Gianni ha udito il mormorare sommesso del mare.
Gianni has listened the whisper-INF quiet of-the see
'Gianni has listened the whispering quiet of the see.'

In the very same way, Zucchi argues that English nominal gerunds (i.e. the form with prepositional objects and adjectives) refer to events, whereas verbal gerunds (i.e. the form with direct objects and adverbs) refer to propositional entities, like a *that-clause* does[7].

This analysis presents numerous problems in my opinion. First, a clear definition of what is for Zucchi a 'propositional entity' (which is frequently identified with 'fact') is not given. In the same way, it lacks a more com-

---

[7]Note that Zucchi's proposal is based on Vendler (1967), who claimed that with "loose containers" (what Zucchi indicates as propositional predicates) both forms are accepted, whereas with "narrow containers" (event-predicates, in Zucchi's account) only nominal gerunds (called perfect nominalizations) are acceptable.

prehensive list of propositional or event predicates, larger than the fews reported, and an explanation of why they should be considered as such.

Secondly, some judgments on which he bases his theory seem to me, as an Italian native speaker, quite excessive: sentences like (52) or (53) do not sound to me so unacceptable as the author states, thus not supporting the idea that type 2 cannot refer to an event. Contrary to his opinion, even types 3 and 4 of NIs may occur with propositional predicates:

(57) L'avanzare   della folla implica che la mediazione non è
      The-advance-INF of-the crowd implies that the mediation   not is
      andata a buon fine.
      gone   at good end
      'The advancing of the crowd implies that the mediation has not gone well'.

Third, other diagnostic have been proposed in the literature as tests for event readings (e.g. Grimshaw, 1990: 50-59), and they seem perfectly acceptable with both types of NIs and EDNs:

(58) Il **costante** parlare di Carla mi   esasperò.
      The constant   talk-INF of Carla to me exasperated
      Carla's constant talking exasperated me.

(59) Il   nuotare   50 metri **in 50 secondi** lo   resero fiero di
      The swim.INF 50 meters in 50 seconds   him made proud of
      sé.
      himself
      'Swimming 50 meters in 50 seconds made him proud of himself'.

Note, moreover, that some predicates may be suited for some types of events (e.g. telic events), rather than for events in general.

Fourth, if we assume that propositions or facts may be declared true or false, no types of NIs seem to denote them:

(60) (il   fatto) che lui venga è vero/falso.
      (the fact)   that he comes is true/false

(61) *Il suo venire   è vero/falso.
      The his come-INF is true/false.

Lastly, in some cases a fact interpretation is available for EDNs too. Think of an example like the following, already reported in section 1.4. What

is a scandal is actually the fact that the professor did a demonstration of the technique, rather than the way in which he did it.

(62)   The professor's **demonstration** of the technique was a scandal.

In conclusion, I do not believe that a distinction between events and propositional entities can be relevant in the selection of nominalizations. Both readings seem available for all nominalizations and contextual elements (like matrix predicates or modifiers) can highlight one reading instead or the other. It is implicit in the description of factive predicates offered by Kiparsky and Kiparsky (1970), where the authors talk about factive predicates (i.e. predicates that imply the truth of their complements), rather than factive arguments. As Melloni (2008) notes, the different readings emerge when the nominal is put in the appropriate context, disambiguating its intrinsic vagueness of meaning.

Thus, I assume that both NIs and EDNs can refer to events and that, for this reason, the difference among them should be defined in other terms. Given the difficulties in judging sentences like the ones reported by Zucchi, I adopt a corpus-based approach, relying on naturalistic and quantitative data to inspect the topic under investigation.

Chapter 3

Constraints on productivity

In the search for an explanation of use for Italian NIs and EDNs, a first hypothesis could consist in a difference in their domain of application. We could suspect that the two processes are applied to different base verbs, thus resulting in alternatives used in complementary distribution, rather than in competitors. In this chapter, I investigate this hypothesis, exploring their degree of productivity and their domain of application.

It has been frequently said that verbal nouns such as infinitives and gerunds can be formed from every base verb (e.g. Gaeta, 2004: 321). These are, indeed, inflectional forms of the verbal paradigm and, as such, they are totally productive and their formation is always regular. Event deverbal suffixes, on the other hand, cannot be applied to every base verb. In section 3.1, after having discussed the notion of productivity, I test these statements by means of two measures proposed in the past literature. These values show that NIs have higher potential productivity, even if EDNs have been formed more frequently.

In section 3.2, I try to understand when EDNs are not derived, i.e. which constraints affect their productivity. In a sample of one thousand base verbs, I explore the role of frequency and transitivity in the derivation of EDNs by means of statistical analyses. The model reveals tendencies, rather than sharp restrictions, thus calling into question our notion of constraints.

Given the lower productivity of EDNs, we may suspect that NIs are used to make up for the lack of a corresponding EDN. In section 3.3, I explore the correlation between the presence or absence of the two nominalizations. In the sample considered, there is a correlation between them, but it goes

in the opposite direction from what we would have expected. When EDNs are not attested, NIs are also less frequently formed.

Given their coexistence for the majority of the verbs in our sample, we cannot assume that they are in complementary distribution and derived for different domains. For this reason, probably, it is not correct to talk about constraints on their productivity, but rather about preferences on their use. Consequently, we may ask: when is one form more frequently preferred? Which class of verbs uses it more? In section 3.4, I investigate the correlation between the frequency of the two nominalizations and some features of the base verbs. I show that transitivity has a role in the selection of the nominalization form: transitive verbs prefer EDNs, whereas NIs are more frequently used when derived from unaccusative verbs.

3.1 Productivity and nominalizations

3.1.1 Definition and measures of productivity

The productivity of a morphological process is usually defined as "the potential to be used to create new words and as the degree to which this potential is exploited by the speakers"[1] (Plag, 2006).

However, starting from Corbin (1987), this definition has been split into two distinct notion, i.e. *availability* and *profitability*. The first one refers to the possibility for a given morphological process to be used in contemporary language to form neologism, while the latter describes the extent to which the process is used. Availability has been defined as a yes/no property, whereas profitability is clearly as a gradual phenomenon (Bauer, 2001; Plag, 2006) and quantitatively measured. Different measures of profitability have been proposed (see, e.g., Baayen, 1993, 2009, Plag, 2006), and many of them rely on corpus data. However, as will be shown, each of them gauges a different aspect of profitability.

A first measure is based on the counts of lemma types in the morphological category under investigation. It has been referred to as *extent of use* (Baayen, 1993), *type-frequency* (indicated as V) (Plag, 2006), or *realized*

---

[1]The notion of productivity has been largely debated in the history of linguistics and its definition is still controversial. Although very interesting, I will not dwell on the topic and I refer the readers to the complete treatise offered by Bauer (2001, 2005), in addition to the other references given in this chapter.

productivity (Baayen, 2009). The more a process has produced new complex lemmas, the more it is productive, no matter of the frequency of the new coined words (Bybee, 2001). It has been shown, indeed, that productive processes present a large numbers of low-frequency forms, even if the total number of types attested is high. Baayen and Moscoso del Prado Martìn (2005) show how English verbs with irregular past forms have a total token frequency higher than regular ones; however, the number of types is extremely lower (146) than regular verbs (1454). This example shows that the total token frequency is not informative about the productivity of a morphological process, while type frequency provides a better approximation. It should be noted, however, as pointed out by Plag (2006) and Baayen (2009), that this measure is related to the past productivity of a morphological process, rather than to its use in contemporary language. If a rule has been frequently employed in the past, we cannot infer that is still productive in the mind of contemporary speakers.

A second measure of productivity is the *hapax-conditioned degree of productivity* (Baayen, 1993), also called *expanding productivity* (Baayen, 2009). This measure corresponds to the number of hapax legomena attested in a corpus for a given affix. It is intended as a measure of expanding productivity: the higher the number of hapax, the higher the rate at which the rule is used, lively and productive. The count of hapax legomena is intended as a surrogate of the count of neologisms. The idea is that a productive process is frequently used to coin new words and, thus, shows an higher number of neologisms. However, detecting neologisms is not an easy task. There are dictionaries which report newly coined words, but they rely on the choice of the lexicographer, who decides when a word can be inserted. The counts of hapax in a corpus, instead, is a practical indirect way to count them. Even if not every hapax is a neologism, it has been shown that neologisms are found primarily among hapax legomena (or among very unfrequent words, Baayen and Renouf, 1996; Plag, 2003).

Similarly to the last one, the third measure makes use of the number of hapax legomena, but it computes the ratio of the number of hapax to the total number of tokens. Baayen (2009) calls it *potential productivity* (or *category-conditioned degree of productivity*, Baayen, 1993). It is identified by the capital letter P and its formula is reported below: for a given affix, $n_1{}^{aff}$ indicates the number of hapax and N^{aff} the total number of tokens:

$$P = \frac{n_1^{\text{aff}}}{N^{\text{aff}}} \tag{3.1}$$

Ideally, P shows the degree at which speakers create new well-formed words with that affix, considered its total amount of use. Also in this case, a higher index is associated with higher productivity.

3.1.2 Productivity of Italian event nominalizations

If we consider availability as a discrete property, NIs and EDNs can be said to be both available: both of them can be used to coin new words. However, the general intuition among speakers is that NIs, contrary to EDNs, can be formed from every base verb and are thus more productive. Considering the distinction given above, we could say that NIs are more profitable than EDNs, even if both are available.

To test this claim, I made use of the productivity measures discussed above to understand to what extent these nominalization patterns are used and perceived differently.

Productivity values have been computed for three nominalizations: nominal infinitives[2], deverbal nouns in *-mento* and deverbal nouns in *-zione*[3], relying on data from the corpus Repubblica (Baroni et al., 2004), a corpus of newspaper articles. The values for each measures are shown in table 3.1.

As the measures show, NIs occur with a lower number of lemma types, while the deverbal nouns formed with *-zione* have the highest realized productivity. This means that even if NIs can in principle be formed with every base verb, they have been used in fewer cases than other nominalization strategies. On the other hand, the P measure reveals that NIs do have the higher proportion of hapax legomena. As suggested by Aronoff and Lindsay (2014: pg.74) "Differences in V reflect the extent to which relevant base

---

[2]The frequency counts used to compute the productivity measures were extracted using the CQP corpus query processor (Evert 2009). The query used to extract nominal infinitives is:

[pos="ART|ARTPRE"]   [pos="ADJ|ADV|ADV:mente|PRO:poss|DET:poss|NEG"]? @[pos="VER:infi"];

which takes into consideration the possibility for a nominal infinitives to be modified by adjectives, adverbs, negation and personal pronouns.

[3]For the productivity measures of deverbal nouns in *-mento* and *-zione*, I searched for all the nouns that ended with these suffixes. After this automatic extraction, a manual cleaning was done to do not count the most frequent words that end with the same string but that are not complex words, e.g. *elemento* ('element'), *momento* ('moment').

LaRepubblica

	V	P	$n_1{}^{\text{aff}}$	N^{aff}
NIs	3876	.011	1221	109720
-mento EDNs	4772	.001	2495	1891019
-zione EDNs	10170	.001	5661	4391704

Table 3.1: Productivity values of Italian event nominalizations from the corpus LaRepubblica.

words have been used", i.e. with reference to diachrony, "while differences in P relate to differences in extent that remaining base words can be used to create neologism".

It can be objected that hapax for NIs are not perceivable as distinct lemmas from their base verbs and thus cannot be considered as neologisms. Moreover, in this view, NIs could probably not be compared to EDNs, since the former are cases of syntactic conversion (or word-class changing inflection). I believe that, despite their different nature, the two processes of nominalization can be compared in reason of their semantic and syntactic similarities[4]. In some cases Italian nominal infinitives have evolved into words perceived as different lemmas (i.e. the case of lexicalized infinitives like *il potere*, 'the power', *il sapere*, 'knowledge'), and in other languages this path is even more common (think of English gerunds and German nominal infinitives). I do not believe that NIs (except for the lexicalized cases) can be conceived as autonomous lemmas, and thus their hapax are not neologisms. However, they can be used anyway to measure productivity due to what we know about the relation between frequency and productivity: productive word-formations are less frequent than unproductive ones (Bybee 1985, Hay 2001). Moreover, P measure gives us an idea of their use and profitability in the speakers mind, whose value agrees with the general intuition that linguists have described (e.g. Gaeta, 2004: 321) and with the fact that they are derived from an inflectional/ syntactic process: NIs are more productive than EDNs.

On the other hand, EDNs patterns have given rise to more frequent

---

[4]As described in chapter 1 and 2.

types, i.e. to words that have acquired a more autonomous status. These patterns are still productive, but are less immediately available to the speaker's mind.

Lastly, with respect to the lower type frequency of NIs, recall that when no external argument is expressed a speaker has a third alternative apart from NI and EDN, namely bare nominal infinitive (BNI). While it is difficult to quantify the *nominal* uses of bare infinitives and to tell it apart from more 'verbal' uses, a rough estimation suggests that BNI are actually quite more frequent than NI. Our productivity measures can be used to boost this point. In order to consider only nominal uses of BNIs, I decided to consider infinitives only when they occured as subject in a sentence, looking for cases in which they were sentence initial. The productivity of BNIs occurring as subjects was then compared to EDNs and NIs occurring as subjects as well. The values are reported in table 3.2.

<div align="center">LaRepubblica</div>

	V	P	$n_1{}^{aff}$	N^{aff}
BNIs as subjects	3044	.027	1159	42823
NIs as subjects	458	.184	273	1480
-mento EDNs as subjects	942	.003	339	87109
-zione EDNs as subjects	1988	.003	769	201435

Table 3.2: Productivity values of Italian event nominalizations occurring as subjects (in LaRepubblica corpus).

BNI's types are more numerous than NIs (and EDNs too), but still NIs seem more productive, as the P measure shows.

3.2 Constraints on EDNs productivity

We have just seen that NIs, according to the P measure, are more productive than EDNs, thus meaning that EDNs do have more restrictions on their formation. A question naturally follows: when are EDNs not available? Can we define the constraints on EDNs productivity? Is there a specific class of verbs that do not allow EDNs or is their derivation random? In this

section, I explore possible answers to these questions.

The subset of verbs to which a morphological process can be applied is usually defined as domain of productivity. As an alternative, it is also common to talk about constraints on productivity (or also selectional restrictions), i.e. features that a word should possess to be a base for the morphological process considered. I will use the two terminologies as complementary, since the features which act as constraints define the domain of verbs to which the process can be applied.

Rainer (2005) offers a typology of constraints valid for natural languages in general. In some cases, the phonology of the base word can condition the applicability of a morphological process. An example is the preference of the English suffix -eer for bases ending in [t] (e.g. *musketeer, profiteer, racketeer*, Adams, 1973: 175-178). The morphological structure of the base can also have a role, like for the preference of English verbs in -ize for the nominalizing suffix -ation (Plag, 2003: 63; Rainer, 2005: 345). An example of syntactic constraints is the preference of the suffix -able for transitive base verbs: *visitable* vs *goable*, *observable* vs *lookable* (Rainer, 2005: 348). At the semantic level, both the semantics of the input and the output of the process may be relevant. As example of the first type, the Spanish relational suffix -uno is mostly attached to base nouns referring to animals (e.g. *vaca* 'cow', *vacuno* 'relating to cow').

In the past literature, some constraints on Italian EDNs have been already investigated (e.g. Gaeta, 2004). However, the interest was to find features that explain the use of one suffix (e.g. -mento) instead of another (e.g. -ione). My point, instead, is related to understanding when the whole class of event suffixes cannot be used. In other words, I would like to identify a class of verbs for which neither one of the EDNs is formed. I believe that it would be interesting to investigate more deeply the constraints that guide the selection of each EDN suffix, since previous works do not offer a clear picture; however, it is not the subject of this work and I leave it to further studies.

In order to investigate the restrictions on EDNs productivity, I consider a sample of one thousand verbs. For these verbs, I test if there is a correlation between some of their characteristics and the absence of EDNs. The sample selected is described in the next paragraph, as well as the features considered. Given this dataset, I apply statistical tests and statistical modeling to detect

possible correlations (§ 3.2.2). The analyses are performed with the software R (R Core Team Team, 2015).

3.2.1 Features and sample of verbs considered

The verbs were extracted from a list presented by Ježek (2003: 190-204). The choice was driven by the fact that, for her study, the author annotated these verbs for some characteristics, specifically for their syntactic alternations. Observing their behavior on corpora and their description in dictionaries, she indicated if each verb was attested as transitive (henceforth TR), inergative (henceforth INTR AV, i.e. as an intransitive verb which select the auxiliary *avere*, 'to have'), inaccusative (intransitive verb with the auxiliary *essere*, 'to be', INTR ES) and in the pronominal form (INTR PR, e.g. *lavarsi*, 'to wash oneself'). The combination of these syntactic frames made her describe 15 verb classes:

1. TR
2. INTR AV
3. INTR ES
4. INTR PRON
5. INTR ES / INTR AV
6. INTR AV / INTR PRON
7. INTR ES / INTR PRON
8. INTR ES / INTR AV / INTR PRON
9. TR / INTR AV
10. TR / INTR ES
11. TR / INTR PRON
12. TR / INTR ES / INTR AV
13. TR / INTR AV / INTR PRON
14. TR / INTR ES / INTR PRON
15. TR / INTR ES / INTR AV / INTR PRON

A single verb may, indeed, present multiple syntactic configurations: the verb *bruciare* ('burn', 'be on fire'), for example, can be used transitively (ex.1), intransitively with both auxiliaries (ex. 2, 3), and as an intransitive reflexive verb (ex. 4). It is an exemplar of the class number 15.

(1)   Ho   bruciato tutte le   sue lettere.
           have.1SG burned   all   the his letters.
           I have burned all his letters.

(2)   La   sua casa   è bruciata.
           The his house is burned
           His house has burned.

(3)   La   casa   ha bruciato a   lungo.
           The house has burned   for long.
           The house has burned for long.

(4)   L'arrosto si   è bruciato.
           The roast itself is burned.
           The roast got burned.

Transitivity (TR), inaccusativity (INTR ES) and unergativity (INTR AV) were considered as possible constraints on EDNs productivity[5]. In addition, I consider the frequency of the base verb lemma. We can suspect, indeed, that less frequent verbs, since are more rarely used, will have lower probability to have a nominalization (both an EDN or an NI). The frequency values were extracted from ITWAC (Baroni et al., 2009), a large web-crawled corpus. Some verbs, however, were not attested at all in the corpus, being rare or slang forms. The dataset, thus, reduced to 966 verbs. Their average frequency as lemmas is 46320 (median 6950, s.d. 106980). As for more general frequencies in corpora, we have a Zipfian distribution (Zipf, 1949): few very frequent verbs and numerous unfrequent ones.

For each verb, I searched for a corresponding EDN. First, I extracted from the corpus all the nouns that started with the verbal root (i.e. the infinitive form minus the flexive suffixes *-are,-ere,-ire*) and then manually checked if they denoted an event. Here I report the event suffixes attested and the number of times they produced an EDN among our 1000 verbs:

- *-ione* (e.g. rotazione, diminuizione): 181

- *-mento* (e.g. slittamento, peggioramento): 173

---

[5]The possibility to occur as a pronominal intransitive verb (INTR PR) was not considered as a feature for practical reasons. In these cases we would have considered two corresponding NIs (e.g. *il lavare*,'the wash.INF', and *il lavarsi*, 'the wash.INF oneself'), which would have had the same corresponding base verb and the same EDN. Thus, some verbs would have been repeated twice in the dataset, others only once. This configuration cannot be analyzed neither by linear/logistic models nor by mixed-effect models.

- *-o/a* (zero suffix, e.g. *appoggio, arrivo, recupero)*: 89

- *-tura* (e.g. copiatura, fioritura): 59

- *-nza* (e.g. resistenza, militanza, insistenza): 56

- *-ata* (e.g. scarpinata, ritirata, caduta): 46

- *-ggio* (e.g. arbitraggio, smontaggio): 13

- *-zza* (e.g. arretratezza, brillantezza): 3

346 verbs do not have a corresponding EDN. In the following analyses, I look if there are some features that distinguish these 346 verbs from all the other verbs that have an EDN attested (no matter the suffix used). The features considered as possible constraints are:

1. Frequency of the base verb (extracted from the corpus Itwac);
2. Transitivity (TR);
3. Intransitivity with auxiliary verb *essere* ('to be') (INTR ES);
4. Intransitivity with auxiliary verb *avere* ('to have') (INTR AV);
5. Presence or absence of the NI;

3.2.2   Analyses and results

A first hypothesis I want to test is whether the lack of an EDN is simply due to a lower frequency of the base verb. It would mean that, since the verb is rare and used in few and specific contexts, there is no need to use its corresponding nominalization. We find, indeed, that it is the case: the frequency of the base verb lemma is significantly higher when an EDN is attested (average frequency= 54790), with respect to when the EDN is not attested (average frequency= 32260, U-test: W= 68970, p-value<0.001). Is this enough? Is the syntactic class of the base verb related to the occurrence of an EDN? Given the role of verb frequency, I apply a logistic regression model in order to consider more than one factor at a time.

Our base model is one in which the frequency of the verb lemma predicts the presence or absence of the EDN, which is our dependent variable. Additional predictors are added by forward selection of variables. Starting from our base model, a more complex model is compared by means of likelihood ratio tests[6] and the best one among them is kept. The same procedure is

---

[6]Performed by the built-in function *anova*.

repeated adding at each time one variable. The final best model for our task is reported in table 3.3 and represented in figure 3.1.

	Estimate	Std. Error	z value	Pr($> |z|$)	
(Intercept)	3.582e-01	1.421e-01	2.521	0.011715	*
Verb frequency	2.609e-06	8.029e-07	3.250	0.001155	**
TRyes	4.187e-01	1.430e-01	2.927	0.003420	**
INTR.ESyes	-4.890e-01	1.386e-01	-3.529	0.000417	***

Null deviance: 1278.9 on 965 df; Residual deviance: 1240.7 on 962 df; AIC: 1248.7

Table 3.3: Effects of final model for presence of EDNs

We find, in addition to the base verb frequency, an effect of transitivity. When the base verb is transitive, there are more chances that a corresponding EDN is attested. On the contrary, intransitive verbs have lower probability to be base of an EDN. However, the effect of intransitive verbs seems restricted to unaccusative verbs. The feature of unergativity did not have any effect (and for this reason was not included in the final model). On the other hand, intransitive verbs which select the auxiliary *essere* have lower probability to have a corresponding EDN attested. Some possible explanation of this correlation are presented at the end of this chapter, because they have a role also on what is described in the next sections.

At the moment, we just argue that it is possible to recognize some constraints on EDNs productivity. However, it is probably more suited to talk about preferences, rather than constraints. The analysis has revealed lower probability to derive an EDN for unaccusative verbs, but it does not mean that unaccusative EDN are not attested at all. There are, indeed, 240 unaccusative verbs with an attested EDN. This data can let us reflect on the nature of constraints on productivity. To my knowledge, among previous studies no one discussed how strong usually constraints are and how many exceptions they present. With regards to Italian EDNs, for example, Gaeta (2004) argues that suffixed verbs with *-izz* (e.g. *laicizzare*, 'to secularize') prefer EDNs in *-zione* (*laicizzazione*, 'secularization'), but he also reports exceptions to this constraint (e.g. *volgarizzamento*, 'vulgarization'), as well as to others. I believe that further studies on various constraints on productivity should address this topic, trying to quantify the strength of restrictions and the differences among them.
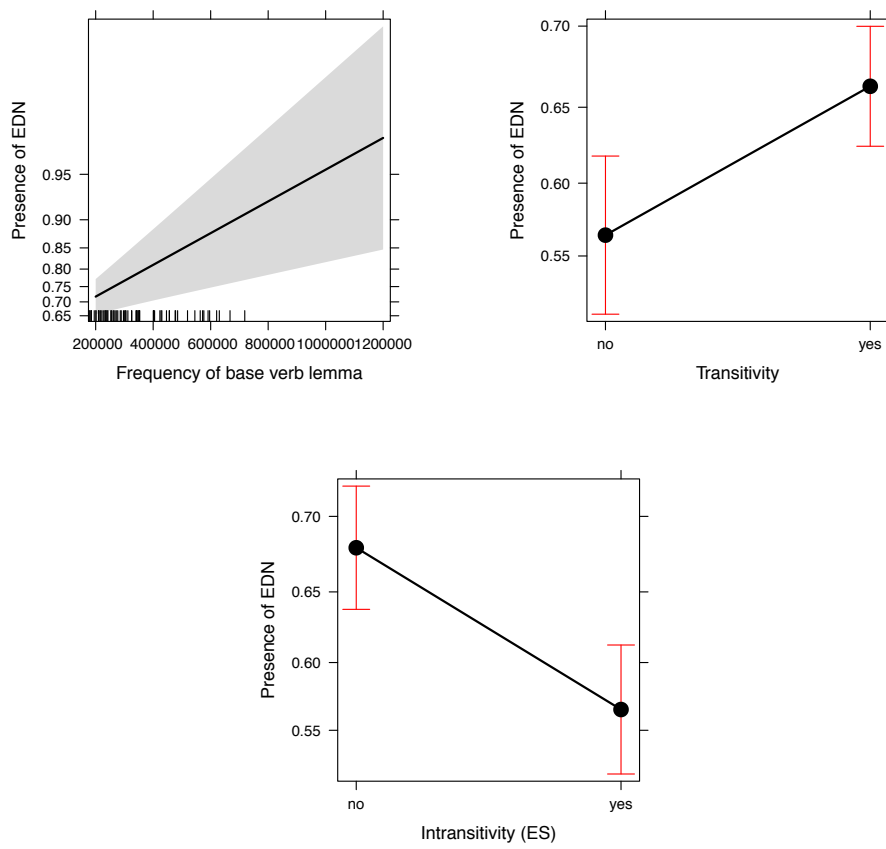
Figure 3.1: Effects in final model for EDNs attestation

3.3 Are NIs used to make up for the EDN's lack?

As frequently stated in the past literature, we have seen that Italian EDNs have lower productivity than NIs and do present some restrictions on their application. We could consequently hypothesize that NIs are used when no corresponding EDN is present and that they make up for this lack, as has been suggested by Simone (2004). This hypothesis, however, turns out to be mostly false.

In the sample of verbs considered in the last section, 346 over 1000 verbs do not have a corresponding EDN. If the hypothesis on the supplementary function of NIs is right, we would expect that NIs are used only when the EDN is missing, thus for around 346 verbs (or otherwise that when NIs are attested no EDN is found). But this is not the case. NIs are attested for 829 verbs, a number larger than 346; moreover, when NIs are attested, EDNs are absent only 276 times. Thus, the hypothesis seems to be not confirmed.

Statistically, we may test if there is a correlation between the absence of EDNs and the presence of NIs. A chi-squared test for independence reveals that there is a significant correlation between the two ($\chi^2 = 44.466$, df $= 1$, $p-value_{two-tailed} < 0.001$), but the correlation goes in the opposite direction from what we predicted. In table 3.4 I report how the presence and absence of NI and EDN combined in our sample: for 87 verbs neither an NI nor an EDN was attested; 50 verbs showed an EDN but not an NI; 276 verbs had an NI but not an EDN; 553 verbs had both nominalizations attested. When we compare these values with their expected values (table 3.5), we see that our hypothesis is not confirmed: when an EDN is missing, NIs are less attested than what we expected, i.e. 276 times versus 311 expected.

Observed frequencies		
	No EDN	EDN attested
No NI	87	50
NI attested	276	553

Table 3.4: Observed frequency of combination of NI/EDN presence and absence.

Expected frequencies		
	No EDN	EDN attested
No NI	51.48	85.51
NI attested	311.51	517.48

Table 3.5: Expected frequency of combination of NI/EDN presence and absence.

In sum, we cannot argue that NIs make up for the lack of a corresponding EDN. They are attested far more frequently and not only when the EDN is missing. On the contrary, when the EDN is not attested, they are used less frequently than what we would expect.

3.4 Frequency and selectional preferences

In the last section, we have seen that frequently both an NI and EDN is formed, more than what we thought. Thus, we cannot assume that they are in complementary distribution and that clear constraints on productivity explain their use.

However, we may suspect that some factors influence their frequency, rather than their productivity. We cannot argue that some features define clearly their domain, but we may guess that some features influence how often they are used.

In this section, I look for an answer to questions like: when is one form more frequently used? Are there verbs which use one form more frequently than the other? The idea is that, even if they are both formed, they present a difference in their frequency. Thus, I test if different verbs use more often one form instead or the other.

For this task, I make use of the same dataset and the same features described above (section 3.2). On a sample of 1000 verbs, I investigate the role of frequency and transitivity on the proportion of occurrences of the two forms. I consider as dependent variable the ratio of the frequency of the NI to the sum of the frequency of both nominalizations, i.e. to the sum of NI and EDN frequency. The frequency of NI alone, indeed, does not tell us the whole story. It may be the case that an NI is quite rare (lets say 30 occurrences), but the corresponding EDN could be even more rare, showing us a preference of that specific verb for the NI. In another case, the same

frequency of NI can be related to a very high frequency of the EDN, thus showing a preference for the EDN.

The proportion considered, instead, will indicate with values closer to 1 the case in which the base verb considered uses mainly NI as form of nominalization and does not have many attestation of the EDN. 1 indicates that only the NI is attested, 0.5 that the NI and the EDN are equally frequent, 0 that only the EDN is present.

From the original dataset, I removed 87 datapoints which did not have any occurrences both for NI and EDN (and for which, thus, was not possible to compute the ratio described above).

A linear regression model was fit with the ratio as dependent variable, and the predictors selected in the final model (after forward selection) are listed in table 3.6, whose effects are also represented in figure 3.2.

	Estimate	Std. Error	t value	$\Pr(> |t|)$	
(Intercept)	0.39574	0.03226	12.265	< 2e-16	***
TRyes	-0.08579	0.03311	-2.591	0.00973	**
INTR.ESyes	0.06953	0.03172	2.192	0.02862	*

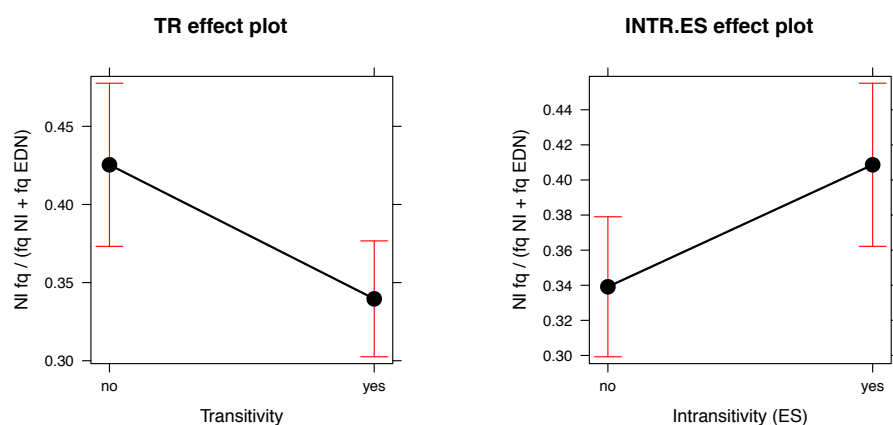Table 3.6: Effects of final model for ratio of NI's frequency to sum of NI's and EDN's frequency

Figure 3.2: Effects in final model for ratio of NI's frequency to sum of NI's and EDN's frequency.

The model is highly signficant (F=7.706, df_1= 2, df_2= 876, p-value $<$ 0.001) and it tells us that transitive verbs have a lower proportion of NI

and thus higher preference for EDN[7]; on the contrary, unaccusative verbs (i.e. intransitive verbs with auxiliary *essere*, 'be') have higher proportion of NIs over EDNs [8]. No effect was found neither for intransitive verbs with auxiliary *avere* ('have') nor for the frequency of base verbs.

These results are parallel to what we have seen in section 3.2 about possible constraints on EDNs productivity. In that section, the model predicted the absence or presence of EDN, rather than their frequency or their proportion. Thus, we may think that the two tasks are similar, even if different in spirit.

Then, how do we explain these results? Why EDNs prefer transitive verbs and NIs unaccusative ones? At this point, we can sketch some possible explanations.

We may suspect that EDNs are more commonly used with transitive verbs because their complement can be omitted, whereas the same is not true for NIs.

(5) La   distruzione (delle città) rimane nella memoria.
     The destruction of the cities persists in the memory
     The destruction (of the cities) persists in memory.

(6) Il   distruggere   ??(le città) rimane nella memoria.
     The destruct.INF the   cities persists in the memory
     The destructing of the cities persists in memory.

Thus, if a nominalization needs to focus on the process itself and to abstract from the referent of the object, the EDN will be preferred, since NIs are less acceptable without the realization of the internal argument. In the case of intransitive verbs, this difference disappears, since no internal argument is present.

Regarding the correlation between NIs and unaccusative verbs, I suspect

---

[7]The mean of the ratio of nominalizations for transitive verbs is 0.33, whereas it is higher, 0.44, for intransitive verbs. If we consider separately the frequencies of NIs and EDNs in the two groups, we see that the frequency's mean for EDNs derived from transitive verbs is higher (20413) than the frequency's mean of intransitive EDNs (14767); NIs, instead, have higher frequency when derived from intransitive verbs (mean=344), with respect to NIs derived from transitive ones (mean= 267).

[8]The mean of the ratio of nominalizations for unaccusative verbs is 0.42; for not-unaccusative verbs it is 0.33. The frequency of NIs with unaccusative verbs (mean=423) is higher than the frequency of NIs from not-unaccusative ones (mean=188). The inverse is true for EDNs: they are less frequent with unaccusative verbs (mean=12322) than with not-unaccusative ones (mean=23459).

that a more semantic difference may be present. It is probably linked to the fact that unaccusative verbs usually denote a telic and dynamic change of state or location, like *affiorare*,'to emerge', or *appassire*, 'to wither'. However, further studies are needed to explain why this class is linked to NIs and to check if EDNs express less frequently than NIs their internal argument.

Chapter 4

Competition among nominalizations: a corpus study

In chapter 3, we have seen that the presence of an NI is not related to the absence of the corresponding EDN. We can, thus, infer that the function of NIs is not to make up for the lack of an EDN.

In this chapter, I present a corpus study focused on cases in which both the NI and the EDN are attested. The aim is to understand when a form is preferred to the other, in which contexts they are used and which semantic differences emerge.

The research for a rationale for their coexistence is motivated by the claim, frequently made in literature (e.g. Bolinger, 1968 or Vennemann, 1972 for the principle of one-to-one relation between form and meaning, also known as Humboltd's universal[1]) that true synonyms do not exist: when two forms are in competition, they always differ in some way, be it their distribution or their meaning.

I will show that, when both nominalizations exist, they have different semantic import, inheriting only part of the senses of the original base verb: NIs usually prefer metaphoric and abstract senses, EDNs literal and concrete ones. I will check the results speakers judgments.

---

[1] Also called *Principle of No Synonymy of Grammatical Forms* by Goldberg (1995: 3).

4.1 Italian event nominalizations: which rationale?

As has been pointed out, it is frequent for a verb to derive both forms of nominalization, i.e. an EDN (even more than one) and an NI. Despite the correlation shown in the last chapter, in some cases the two pattern may be present and potentially in competition. However, are they really competitors? Or do they bring different meanings?

In the existing literature, not enough space has been devoted to NIs and, in particular, to their relation with EDNs. Previous accounts of competing event nominals in Italian have been restricted mainly to alternative EDNs.

Gaeta (2002: 219-221) shows that, when both *-mento* and *-zione* form a deverbal noun for the same base, the former is reserved to the event meaning, while the latter is more polysemous and conveys different readings. For example in the doublets *divaricamento - divaricazione* (from the verb *divaricare*, 'to stretch apart') only the second form displays a resulting state reading (cf. Gaeta, 2002: 318), whereas the former can refer only to the act of stretching apart.

In other cases, like for the EDNs *trattazione / trattamento* ('treatise / treatment'), each member has been lexicalized with distinct senses, each derived from a specific meaning of the polysemous verbal base, i.e. *trattare* ('to treat', 'to debate', 'to concern'). However, both EDNs can refer to the event and to the result of the action; it is the kind of action that is slightly different. It is possible also to have a wider range of readings, e.g. agentive, locative or resultative (see, for an overview of possible senses of EDNs, Melloni, 2006, 2007, 2008; Ježek, 2007), but no difference is seen among the different suffixes.

With reference to nominal infinitives, the previous literature on this issue is scarce and inconclusive. Skytte (1983) takes into consideration the whole range of patterns in which the Italian infinitive occurs, devoting a chapter to the nominal infinitive (also called *infinito sostantivato*). She describes the syntactic characteristics of the construction but, to explain its use and the competition with other deverbal nominalizations, she refers to a stylistic difference: they have exactly the same meaning, but the nominal infinitives has a more formal connotation. She suggests also an aspectual difference: nominal infinitives focus on the duration, the development of the action, as is shown by their occurrence with adjective of duration (1983:533). A

similar idea was expressed in few lines by Vanvolsem (1983), whose book was devoted to NIs but did not address the problem of competition with EDNs.

Gaeta (2002, 2009) and Melloni (2007) (also, in different terms, Simone, 2004) argue that NIs, contrary do EDNs ending in *-(z)ione* and *-mento*, do not allow a telic interpretation, as can be seen by their unacceptability with telic predicates:

(1) *Lo sfruttare i giacimenti auriferi è stato terminato.
     The exploit-INF the gold-deposits has been finished.

(2) Lo sfruttamento dei giacimenti auriferi è stato terminato.
     The exploitation of-the gold-deposits has been finished.

This statement is problematic for various reasons. First, it does not explain cases in which neither EDNs and NIs allow telic interpretations, i.e. cases in which they are derived from non-telic bases, as in *il passeggiare* 'the stroll-INF'. Second, sometimes even EDNs seem marginal with telic predicates, removing the purported difference between them and the corresponding NIs. Gaeta (2009: 116) presents the following examples to show how NIs are not acceptable with telic predicates.

(3) *L'affondare della nave si compì in mezz'ora.
     The sink-INF of-the ship accomplished in half-hour.

(4) *Il guarire è stato raggiunto.
     The heal-INF has been reached.

However, the corresponding EDNs don't seem entirely acceptable (ex. 5 and 6):

(5) ?L'affondamento della nave si compì in mezz'ora.
     The foundering of-the ship accomplished in half-hour.

(6) ?La guarigione è stata raggiunta.
     The healing has been reached.

What is more important, NIs seem deviant with *any* variant of verbs like *avvenire* 'happen', *essere in corso* 'be in progress', *aver luogo* 'take place' (collectively termed the 'happen-class'), even in the absence of any modifier that suggests the existence of a natural endpoint (7), and even when the sentence is generic (8). Thus, the deviance of Gaeta's cases is probably

independent of telicity.

(7)   *L'affondare   della   nave è in corso.
    The sink-INF of-the ship is in progress.

(8)   a. *L'affondare   della   nave avviene spesso in un blocco navale.
     The sink-INF of-the ship happens often   in a   block   naval.

    b. ??In un western, l'arrivare   dei   soldati ha   luogo
     In a   western, the arrive-INF of_the soldiers takes place
     quando la   situazione è   disperata.
     when   the situation   is hopeless

On the other hand, NIs seem sometimes compatible with verbs that explicity invoke a 'thelos', like *terminare* ('ending') or *richiedere settimane* ('take weeks'):

(9)   Il   lento affondare delle   trivella nel   terreno gelato terminò
    the slow   sink-INF of_the drill   in the ground frozen ended
    solo quando la   punta raggiunse il   giacimento petrolifero.
    only when   the tip   reached   the oil field

(10)   Il   guarire   completamente richiese due intere settimane.
    The heal-INF completely   took   two whole weeks.

Last but not least, NIs can embed temporal phrases that mark precise endpoints, as long as the main predicate is not in the *happen*-class. EDNs do not seem equally felicitous in these cases (12), though once again the judgments are delicate.

(11)   [L'affondare   della   nave in mezz'ora] ci sorprese.
    The sink-INF of-the ship in half-hour us surprised.

(12)   ?[L'affondamento della   nave in mezz'ora] ci sorprese.
    The foundering   of-the ship in half-hour us surprised.

These examples show how elusive the telic-atelic distinction is and how tests with container predicates[2] are not always easily interpretable.

   Zucchi (1993) deals with some aspects of nominal infinitives, dividing them in different classes, according to their verbal or nominal properties (e.g., acceptability with adjectives or adverbs, argument structure). However, for the most nominal class (corresponding to cases taken into consider-

---

[2]As they are called in Vendler (1957).

ation in this study), he does not assume the existence of a difference between EDNs and NIs. In other cases, he assumes that NI should be mapped to propositions, not events (*il suo venire* 'his coming' would then be '(the fact) that he comes/is coming'). While this seems to offer a promising way to exclude the 'happen'-class predicates above, it raises an even more serious problem: propositions may be declated true or false, but NIs cannot.

(13)   a.   (il   fatto) che lui venga è vero/falso.
                (the fact) that he comes is true/false.

        b.   *Il suo venire   è vero/falso.
              the his come-INF is true/false

Therefore, the solution to the puzzle of NIs/EDNs use remains to be found.

A list of features that could be possibly involved in explaining the distribution of our nominalizations is given by Martin (2010) in a study on French EDNs (derived with the suffixes *-ment/ -age/ -ion*). Even if not previously applied to nominal verbs, some of these features can play a role in our case and can be easily verified with corpus analysis. Following Kelling (2001), she argues that *-age* EDNs are more agentive than *-ment* ones[3], since they do not tolerate non-agentive subjects (Martin, 2010: 122):

(14)   Le décollement   des   tuiles par le vent/ par
        The unsticking/removal of-the tiles by the wind/ by
        l'ouvrier.
        the worker.

(15)   Le décollage   des   tuiles #[par le vent]/ par
        The unsticking/removing of-the tiles #[by the wind]/ by
        l'ouvrier.
        the worker.

Moreover, she suggests that *-age* suffix is associated with a physical domain, while *-ment* is ontologically unmarked. Then, psych-verbs like *penser* ('think'), *préoccuper* ('preoccupy') and so on, do not select the *-age* suffix. When both a physical and an abstract reading are possible for a base verb, *-age* will select the physical one, *-ment* the other:

---

[3]She suggests also a different version of the agentivity parameter, but we will not consider it in the present study since it does not sound totally convincing to me. For further details, we refer the readers to her article (Martin, 2010: 122-124)

(16)   Le   gonflement des   prix.
        The inflating   of the prices.

[abstract Theme]

(17)   Le   gonflage du   ballon.
        The inflating of the baloon.

[physical Theme]

In what follows, I will keep all of these features into account as we look at the corpus distribution of Italian NIs and EDNs. However I will not consider other potential distinctions, such as actionality or aspect. Even if interesting and entirely reasonable, these properties, in my opinion, would be better investigated with formal acceptability judgments experiments, not with corpus studies or informal judgments. Intuitions are in these cases quite weak and cannot be based on just a few speakers, whereas corpora may not offer attestation of all the linguistic contexts needed to assess these intricate hypotheses.

4.2   Semantic corpus analysis

4.2.1   Hypotheses and dataset

Summarizing the main points of section 4.1, a number of hypotheses have been put forth by researchers and can be considered in the case of NIs/EDNs competition: (i) the former refers only to the event reading, while the latter denotes also other concrete and resultative readings; (ii) the former refers to the event as an atelic event, while the latter preserves the base verb actionality; (iii) they realize different senses of the corresponding predicate; (iv) they express a difference in the agentivity of their subject; (v) they are associated to different degrees of concreteness (different 'ontological domains' in Martin's characterization). As previously stated, point (ii) will not be considered since I believe it would be better addressed with an acceptability judgments experiment. I plan to address it in further studies.

To investigate these possibilities, a sample of 36 pairs of NIs and their corresponding EDNs was individuated. Half of them was selected among the most frequent NIs, in order to be sure to have enough tokens, since their frequency is usually very low. The other pairs were extracted from the most frequent deverbal nouns (with the suffixes *-mento*, *-ione*, *-ggio*),

58

since I did not want to introduce a bias by considering only the most frequent NIs. Table 4.1 shows the items selected with their frequency. Where present, more than one EDN for the same base verb was taken into consideration (e.g. *apparizione*, 'appearance', *apparenza*, 'semblance'), listing the corresponding NI twice and increasing our sample to 39 pairs.

4.2.2 Methodology: the distributional hypothesis

In this work, collocation analysis is used to detect semantic differences between the two types of nominalization. This methodology relies on the hypothesis that examining the syntagmatic environments in which a word occurs let us know more about the properties of the word we deal with. We can observe that two words are similar if they occur with the same words, or that multiple senses of a word can be disambiguated by their different contexts of occurrence. Consider as example the English word *bank*: it is ambiguous between the meaning of 'financial institution' and 'side of the river'. Stubbs (2002: 15) notes that the words that occur with the word *bank* disambiguate its meaning, as can be seen by the following list:

- bank account, bank balance, bank robbery

- canal bank, sand bank, river bank

In the same way, we may think that collocates of NIs and EDNs can inform us about their semantics and that differences in the syntagmatic environments of the two nominalizations correspond to differences among their meanings.

These facts were first noted by Harris (1954), who stated that "difference of meaning correlates with difference of distribution". The idea that words that occur in similar contexts tend to have similar meanings is usually referred to as the *distributional hypothesis*. In the highly-cited words by Firth (1957: 11), "You shall know a word by the company it keeps!". However, the relation between the meaning of a word and its context of use was already present in the second Wittgenstein (1953: 43), who asserted that "in most cases, the meaning of a word is its use". This hypothesis has been highly influential in linguistics and cognitive science. It has led linguists to study lexical-semantic aspects of words by means of collocations from corpora, and it has pushed the computational linguistics community

NI		Freq. NI		EDN		Freq. EDN	
		Total fq	Pattern fq			Total fq	Pattern fq
affermare	'assert'	2276	3	affermazione	'utterance, achievement'	45276	14170
apparire	'appear'	2732	592	apparenza	'appearance, semblance'	9139	1248
apparire	'appear'	2732	592	apparizione	'appearance'	7797	1834
aumentare	'increase'	2902	2013	aumento	'growth'	110116	70849
avanzare	'advance'	1960	1057	avanzamento	'advancement'	7042	3472
battere	'beat'	150	44	battitura	'beating, typing'	280	127
cambiare	'changing'	1008	278	cambiamento	'change'	57511	15224
conoscere	'know'	2208	11	conoscenza	'knowing, knowledge'	130311	47171
costruire	'construct'	2109	21	costruzione	'construction'	114676	64940
creare	'create'	2756	57	creazione	'creation'	96859	62141
accrescere	'grow'	3448	2082	accrescimento	'increse'	3506	1914
crescere	'grow'	3448	2082	crescita	'growth'	111387	35495
discutere	'discuss'	600	110	discussione	'discussion'	124473	20560
emergere	'emerge'	4316	1506	emersione	'emergence'	4776	2041
esplodere	'esplode'	455	276	esplosione	'esplosion'	12519	3132
esprimere	'express'	2814	7	espressione	'expression'	68204	20099
fiorire	'bloom'	591	285	fioritura	'flowering'	2051	427
insorgere	'arise, rise up'	3790	2065	insorgenza	'insurgence, onset'	5109	2841
irrompere	'burst into'	626	362	irruzione	'irruption'	2963	1003
mutare	'change'	1829	1519	mutamento	'changement'	12824	5857
mutare	'change'	1829	1519	mutazione	'mutation'	5545	873
nascere	'be born, arise'	3549	489	nascita	'birth'	73222	28960
operare	'operate, work'	3491	497	operazione	'operation, procedure'	102375	18937
pagare	'pay'	439	9	pagamento	'payment'	88425	52982
passare	'pass'	5980	5600	passaggio	'passage'	49469	11054
permanere	'remain'	1932	871	permanenza	'permanence'	16262	4331
persistere	'persist'	1356	998	persistenza	'persistence'	4257	2287
precipitare	'fall'	608	424	precipitazione	'haste, precipitation'	877	158
prevalere	'prevail'	1124	545	prevalenza	'prevalence'	5930	2891
procedere	'proceed'	3293	960	procedimento	'procedure'	108554	15227
produrre	'produce'	1295	11	produzione	'production'	196572	63911
progredire	'advance'	1691	1168	progresso	'progress'	29946	4324
proliferare	'proliferate'	1392	990	proliferazione	'proliferation'	4357	2480
riconoscere	'recognise'	2341	4	riconoscimento	'recognition'	94710	45743
risorgere	'resuscitate'	340	199	risorgimento	'resurgence'	582	41
scadere	'expire'	5024	3780	scadenza	'expiration date, duty'	51449	27011
scorrere	'flow'	3110	2203	scorrimento	'flow'	2242	1203
scrivere	'write'	4117	61	scrittura	'writing'	37355	3533
vedere	'see'	7811	74	visione	'vision'	90555	22763

Table 4.1: Sample of NIs and EDNs considered with total frequencies and frequencies in the pattern considered

towards a reliable representation of words meanings, by means of so-called distributional semantics models[4]. It has gone beyond the linguistic definition of meaning, and it has proven to affect also the way in which we learn meanings. McDonald and Ramscar show that "learning the meaning of a word is thought to be dependent, at least in part, on exposure to the word in its linguistic contexts of use" (2001: 611).

4.2.3 Methodology: cooccurences and statistical association measures

In this study, I focused the analyses on a single syntactic pattern, i.e. the prepositional phrase which encode most frequently the nominalization's arguments. In Italian, these are expressed after the nominalization and are introduced by the preposition *di* ('of', also in the compounded version with the articles, i.e. *del, della, dello*, 'of the').

From previous analyses, this pattern proved to be the most frequent one for both nominalizations. Moreover, it encodes their arguments, whose semantics allows us to specify the meaning of the derivatives. As previously noted, however, this pattern can encode both the internal or the external argument of the base verb with EDNs (ex. 18, 19), while with NIs it usually expresses only the external one (ex. 20, 21). With NI, direct objects must be realized as DPs, not PP, while other internal complements (e.g. *prevalere su qualcuno* 'prevail over someone') are realized as PPs, just as in the tensed verbal form.

(18)   La   produzione di gelati.
        The production of ice-creams.
        'The production of ice-creams'.

(19)   La   produzione di Giovanni.
        The production of Giovanni.
        'Giovanni's production'.

(20)   Il   produrre   di Giovanni.
        The produce.INF of Giovanni.
        'Giovanni's producing'.

---

[4]In Chapter 5 a study with distributional semantics models is presented and more information are given with regards to this computational method of meaning representation.

(21) *Il   produrre   di gelati.
     The produce.INF of ice-creams

(22) Il   produrre   i   gelati.
     The produce.INF the ice-creams
     'The producing ice-creams'.

Unfortunately, NIs of transitive verbs do not have many tokens in the pattern considered. As a consequence, we do not have enough data to include in our analysis for the following verbs: *affermare, riconoscere, esprimere, conoscere, produrre, vedere, costruire, scrivere, pagare, creare, unire.*

Using Python[5], I extracted all the nouns occurring in this pattern (e.g. "*la costruzione della **casa**", 'the construction of the **house**) from the corpus Itwac and computed their co-occurrence frequency with the two nominalizations.

However, we cannot directly compare the co-occurrence of a certain argument with the two forms, since they usually have very different overall frequencies. Typically, the EDN has higher frequency than the NI. A first strategy to understand which nominalization an argument occurs with more often is to consider its relative frequency in each construction, i.e. the ratio of its frequency with NI (or EDN) to the overall frequency of arguments of NI (or EDN). However, this measure does not tell us when a difference in relative frequency is significant.

This problem is well known in corpus linguistics (see e.g. Evert, 2005, 2009; Gries et al., 2005).A large amount of different measures have been proposed to compute the association of two words. These measures are usually called *association measures*. Evert (2005) lists around 40 different association measures. To this day, there is no agreement on which measures is the most accurate.

For our purpose, we need, first, a measure that weights our co-occurrence frequencies in a way that keeps into account the overall frequency of the two nominalizations; next, we want to know which argument nouns present a difference in their co-occurrence with the two nominalizations which is statistically significant.

One of the most used and accepted association measures for purposes similar to ours is the log-likelihood ratio. It has been proposed as an asso-

---

[5]Python Software Foundation, https://www.python.org/

ciation measure by Dunning (1993: 67), and is used for different purposes, e.g. by Zinsmeister and Heid (2004) for noun-verb collocation, to assess the degree of regularity of compounds by their collocational preferences. For a mathematical discussion see Evert (2005: 83-84).

In order to understand how log-likelihood (also referred to with the symbol G^2) is computed, consider the example of the pair *il fiorire* (NI) / *la fioritura* (DN). The noun *pianta* ('plant') occurs twice as argument of the NI (O_{11})[6] and 21 times with the EDN (O_{12}). The overall frequency of NI in the pattern considered (NI + *di/del*+ NOUN) is 285, while EDN occurs 427 times. This information can be summarized in a contingency table:

	NI	EDN	
noun "pianta" as argument	2 (O_{11})	21 (O_{12})	=23 (R_1)
other nouns as argument	283 (O_{21})	406 (O_{22})	=689 (R_2)
	=285 (C_1)	= 427 (C_2)	= 712 (N)

Table 4.2: Observed frequencies of the noun "pianta" as argument of NI "fiorire" and EDN "fioritura"

The log-likelihood values for each collocation is then computed as follows:

$$log\text{-}likelihood = 2 \sum_{ij} O_{ij} log \frac{O_{ij}}{E_{ij}} \tag{4.1}$$

where E_{ij} refers to the expected frequency of the two components if they were not associated. This formula reports higher values both for positive or negative association, that is it signals if the two components are occurring together more or less than chance. In our case, we are interested in keeping separated positive and negative associations, since a positive association will indicate a relation between the word considered and the NI, while a negative value will indicate an association with EDN. The standard formula will give high values for association with both NI and EDN. An easy solution[7] is to

---

[6]This notation stands for "Observed frequency of element 1 with construction 1."
[7]Suggested by Evert (2005).

multiply by -1 the cases in which the observed frequency of noun and NI (O_{11}) is lower than expected (lower than E_{11}).

For a collocation to be significant the log-likelihood value should be higher than 3.84 (see Oakes 1998: 266, who discusses the equivalence with chi squared test with one degree of freedom). In our case, a value lower than -3.84 will be significant too, but for the EDN form.

Given the values in table 3, the log-likelihood value for the noun *pianta* is computed as follows:

$$E_{pianta \ with \ NI} = \frac{23 * 285}{712} = 9.206461 \tag{4.2}$$

$$E_{pianta \ with \ EDN} = \frac{23 * 427}{712} = 13.79354 \tag{4.3}$$

$$log\text{-}likelihood_{pianta} = 2[(2 * log(\frac{2}{9.206461}) + (21 * log(\frac{21}{13.79354}))] = 11.5465 \tag{4.4}$$

Since the expected frequency of this noun with the NI is higher than the observed one, the log-likelihood value should be multiplied by -1. In this case, the result shows a stronger association with the EDN[8].

Note that, in this study, I am computing the association strength between a form of nominalization (i.e. a more grammatical unit) and a lemma, not the association between two lemmas. A pair of words which are positively associated is usually called a collocation[9]. The analysis of the strength of a collocation is called collocation analysis. Stefanowitsch and Gries (2003) and Gries and Stefanowitsch (2004) talk about "collostruction analysis" or "collexeme analysis" to refer to the analysis of the association between a word and a construction. However, since the statistical basis is the same, I will only use the term collocation analysis, when needed, or just refer to statistical tests for association strength.

---

[8]Other association measures were tested too, e.g. Fisher exact test (Pedersen, 1996). However, the results were similar to log-likelihood ones, which has the advantages of a easier computation and interpretation.

[9]As Evert (2009) points out, in this case the term collocation means "empirical collocation", not a multiword expression, which is a idiosyncratic or semi-compositional word combinations, e.g. *heavy smoker*. Thus, a collocation is just a recurrent and predictable words pair, as described by Firth (1957).

4.2.4 Results and qualitative analysis

Are the two nominalizations semantically different?

As stated above, log-likelihood values that are higher than 3.84 or lower than -3.84 indicate a statistically significant association between the argument and one nominalization, which correspond to a p-value lower than 0.05. For each pair of nominalizations, I calculated in how many cases their arguments were statistically significantly associated with one of the nominalization. The higher this number, the more the two nominalizations have different meanings.

Then, I computed the ratio of the number of tokens of NI and EDN with significantly different arguments to the total number of occurrences (significantly different or not)[10]. A probability score was obtained (reported in table 4.3[11]), which indicates in how many contexts/occurrences the two forms have different arguments (and thus, based on distributional hypothesis, different meanings).

The mean among all pairs is 0.81 (sd 0.2), showing that our pairs of nominalizations do have different distributional contexts and then, as consequence, different meanings. However, it could be argued that the difference of arguments is due to the type of argument EDNs and NIs can take. We have seen as EDNs can take both external and internal arguments in this pattern, while NIs usually take the external one. To overcome this possible objection, I calculated the mean of the probability scores in table 4.3 excluding pairs derived from transitive verbs (marked in the table with an asterisk): the mean and the standard deviation are exactly the same as for the full set (mean=0.81, sd=0.2). Note that, in some cases (*passare*, 'to pass', *aumentare*, 'to grow'), the verb can show a transitive/intransitive alternation. In this case, we excluded them from the calculation of the means.

Moreover, the meanings of some pairs seem more distinct, while other do have some arguments that are not clearly associated with one nominalization or the other. This suggests that the various pairs show different degrees of

---

[10]As an alternative, I calculated also the ratio of the number of arguments types that were significantly different to the total number of arguments types attested. I believe, however, that this measure is less appropriate, since it gives the same importance to arguments that occur only once and to arguments that occur for the half of the total number of occurrences.

[11]The infinitive-noun pairs in this table are the same reported in table 4.1. For this reason, I do not provide here their glosses again.

NI-DN	Total freq. of significant arguments
Apparire-apparenza	0.7334
Accrescere-accrescimento	0.9482
Apparire-apparizione	0.8973
Aumentare-aumento *	1
Avanzare-avanzamento	0.9734
Battere-battitura *	0.5343
Cambiare-cambiamento *	0.5244
Crescere-crescita	1
Discutere-discussione	0.5798
Emergere-emersione	0.9518
Esplodere-esplosione	0.4618
Fiorire-fioritura	0.5056
Insorgere-insorgenza	0.8887
Irrompere-irruzione	0.62
Mutare-mutamento *	0.9616
Mutare-mutazione *	0.8901
Nascere-nascita	0.8163
Operare-operazione *	0.9881
Passare-passaggio *	1
Permanere-permanenza	0.97
Persistere-persistenza	0.6824
Precipitare-precipitazione	0.8715
Prevalere-prevalenza	0.8375
Procedere-procedimento	1
Progredire-progresso	0.9674
Proliferare-proliferazione	0.8463
Risorgere-risorgimento	0.3046
Scadere-scadenza	0.9911
Scorrere-scorrimento *	0.9496

Table 4.3: Probability scores of significantly different arguments for pairs of nominalizations.

competition: in some cases, the competition has been solved over time, resulting in the two nominalizations that have well-separated application; in others, the competition is still in action, with the two forms that are not yet separated for meanings in the speaker mind, but do show some preferences of occurrence. In the next paragraph, we will address this point in more details.

How do the meanings of NI and EDN differ?

Assuming that a difference in selectional preferences correspond to a difference in meaning, I looked at how the arguments occurring with NIs and EDNs differ from each other. We present here some examples.

Four pairs had a probability of divergence of 1 (i.e. 100%): *aumento-aumentare, crescita-crescere, passaggio-passare, procedimento-procedere*. It means that their arguments are totally distinct and show clear preferences for one form or the other. See as example (table 4.4) the different arguments associated with the NI *il passare* ('the pass.INF') and the EDN *il passaggio* ('the passage').

Nouns arguments of NI			Nouns arguments of EDN		
Noun	Fq	LL	Noun	Fq	LL
Tempo ('time')	2750	5331	ruolo ('role')	808	-662
anno ('year')	1489	3124	potere ('power')	419	-343
giorno ('day')	323	691	proprietà ('property')	361	-295
ora ('hour')	268	545	competenza ('expertise')	568	-235

Table 4.4: Most significant arguments for *il passare - il passaggio* ('the pass.INF' - 'the passage')

The analysis of the arguments of the two nominalizations shows us that the two nominalizations frequently denote different events, in much the same way as the base verb does. Verbs may refer to various types of events, which are usually listed in dictionaries as different senses.

Consider the dictionary entry of the verb *passare* ('to pass')[12]:

1. To pass through, to go across a place, a location;

2. To move, to shuttle from a place to another;

---

[12]Adapted from the Sabatini-Coletti online dictionary.

3. To seep through;

4. To pass (of time).

The arguments of the nominalizations are divided in such a way that those occurring with the NI *il passare* ('the passing') refer only to the fourth sense of the base verb, while the EDN *il passaggio* ('passage','transition') occurs with arguments linked to the first three senses, i.e. 'to pass through, to go across, to move'.

Another example can be seen with the verb *fiorire* ('to bloom'), which has two senses: one referred to flowers or plants and a figurative one, with the sense of 'flourish'. The two are kept distinct by the NI 'il fiorire' and the EDN 'la fioritura' (table 4.5).

Nouns arguments of NI			Nouns arguments of EDN		
Noun	Fq	LL	Noun	Fq	LL
iniziativa ('venture')	15	27	ciliegio ('cherry-tree')	27	-27
attività ('enterprise')	7	12	pianta ('plant')	21	-11
scienza ('science')	4	7	alga ('swaweed')	10	-10
rapporto ('relationship')	3	5	albero ('albero')	10	-10

Table 4.5: Most significant arguments for *il fiorire-la fioritura* ('the bloom.INF'-'the blooming')

These and other cases allow me to argue that the different nominalizations resolve the vagueness of the base verb, focusing each on different senses. This finding would be in line with what has been previously argued for the case of *trattazione / trattamento* ('investigation / treatment') (Melloni, 2007, discussed in section 4.1, p. 54), where the two derivatives inherit different senses of the base verb. Similar observation can be found in Lieber (2004: 39) for the English triplets *committal, commitment, commission* "in which individual members have been lexicalized with distinct and idiosyncratic meanings".

However, another question is still hanging: can we predict which senses will be taken up by the NI and which ones by the EDN? Are there regularities in the cases analyzed?

From our observations, it seems that among the most significant arguments for NIs there are more abstract entities, while EDNs arguments refer to more concrete ones. It seems, indeed, that EDNs refer to literal senses

and the NIs to metaphorical ones. In addition to the previous examples, consider the verb *permanere* ('to persist'): the NI refers to the persistence of a state of affair of some abstract entities, while the EDN to the permanence of humans (table 4.6).

Nouns arguments of NI			Nouns arguments of EDN		
Noun	Fq	LL	Noun	Fq	LL
responsabilità ('responsability')	14	50	persona ('person')	179	-56
situazione ('situation')	26	39	straniero ('foreigner')	129	-47
difficoltà ('difficulty')	10	29	docente ('teacher')	115	-42
tensione ('tension')	8	22	giovane ('young person')	96	-35

Table 4.6: Most significant arguments for *la permanenza- il permanere* ('the permanence' - 'the persist.INF')

As discussed above, Martin (2010) suggested that French EDNs are associated with different ontological domains, i.e. abstract and concrete ones. Again, our analysis goes in the same direction. However, no clear difference in agentivity emerged: only few cases (like *permanenza-permanere*, 'the permanence' - 'the persist.INF') show EDNs that are more associated with agentive subjects and NIs which go with non-agentive ones.

It is not always possible to clearly evaluate the abstractness or 'metaphoricity' of arguments. Consider the arguments of the NI *l'aumentare* ('the increase.INF', table 4.7) and the EDN *l'aumento* ('growth/increase'). Among the most significant ones for the EDN we find the noun 'productivity': is 'productivity' more abstract than 'distance', the most significant argument for the NI? How can we be sure of our abstractness judgments?

Nouns arguments of NI			Nouns arguments of EDN		
Noun	Fq	LL	Noun	Fq	LL
distanza ('distance')	118	643	capitale ('assets')	2400	-124
velocità ('speed')	95	257	prezzo ('price')	4310	-101
profondità ('depth')	33	159	spesa ('expense')	1630	-69
dimensione ('dimension')	46	97	costo ('cost')	2599	-62
numero ('number')	160	92	produttività ('productivity')	870	-49

Table 4.7: Most significant arguments for *aumento-aumentare* (´growth/increase'-'increasing')

69

To address this concern, in the next section I test this hypothesis by measuring the concreteness values using existing datasets. From the corpus analysis, we expect EDN's arguments to have higher values of concreteness, while NI's ones should have lower scores.

Note that the reason why NIs tend to be associated with abstract/metaphorical meanings and EDNs to concrete ones is still unexplained. Even if there is no clear answer, I present an hypothesis in the concluding section.

4.3 Difference in concreteness

We have observed a difference in level of abstraction across NIs and EDNs arguments. These facts, however, rely only on personal observations.

To overcome this problem, I decided to extract values for concreteness/abstractness from datasets compiled to: (i) quantify concreteness with numerical values, (ii) aggregate judgments of hundreds of participants. This addresses the possibility of personal bias. Two datasets for Italian nouns were used: the lexvar dataset (Barca et al., 2002), consisting of 626 simple nouns, and the one presented in Della Rosa et al. (2010), containing ratings for 417 nouns.

For each of the 626 nouns, Barca and colleagues report values for a wide range of variables, such as mean age of acquisition, number of ortographic neighbors and so on. I used the values for word familiarity (FAM), imageability (IMAG), concreteness (CONC) and length in letters (LET). The values were obtained from ratings given by 176 participants, all native speakers of Italian between 20 and 30 years of age, using a Likert scale from 1 to 7. 1 is used for very unfamiliar words, hardly imageable words and highly abstract ones. 7 indicates very well known words, highly imageable words and highly concrete ones. Each noun scored by 44 participants, half male and half female. The value reported for each variable is the average among participants.

Similarly, Della Rosa and colleagues collected ratings with the same scale from 250 native speakers. Each speaker rated each noun for only one of the variables above, resulting in each word being rated for each variable by 35 participants. The authors multiplied by 100 the ratings, transforming 7 to 700 and 1 to 100.

From the list of arguments occurring with each pair of NI and EDN, I

did not considered arguments that occurred only once with both the NI and EDN and whose difference of occurrence between the NI and EDN was not significant (i.e. log-likelihood value higher than -3.84 or lower than 3.84). For each pair, the total number of arguments considered was not, in any case, higher than 400. With these constraints, I obtained a list of 2958 nouns that occurred as arguments with our selected pairs (on average, 74 nouns per pair). 334 of them were present in the Lexvar dataset and 260 in the Della Rosa dataset. Some of them were repeated more than once since they occurred with more than one pair of nominalizations.

For the Lexvar dataset, according to a U-test[13], the median concreteness coefficient for NIs (4.650, IQR=1.9100) and the median of the concreteness coefficients for EDNs (4.975, IQR=2.0425) are very significantly different (W=15056, p-value$_{two-tailed}$ = 0.006). EDN's arguments are rated as more concrete than NI's ones, confirming our qualitative analysis. Values for imageability show a parallel very significant distinction: NI's arguments are less imageable than EDN's ones. The median imageability coefficient of the NI's arguments (4.27, IQR=2.27) is significantly different (W = 15022, p-value = 0.006) from the one for the EDN's arguments (5.01, IQR=2.49). The boxplots in Figure 4.1 report the distribution of concreteness and imageability values for the two groups.

No difference was seen for familiarity values (p>0.1). This is interesting since it has been claimed that the difference between NIs and EDNs is a stylistic one (Skytte, 1983; Vanvolsem, 1983): NIs are seen as more appropriate within a formal register, EDNs within a more informal one. If this had been true, we would have found higher values of familiarity for EDNs arguments.

Similar results (with even more significant effect) were replicated with the Della Rosa dataset. The U-test reveals that the difference among the median coefficients for concreteness (NIs: 325, IQR=215.50; EDNs: 497, IQR= 313.75) is highly significant (W=9860, p-value$_{two-tailed}$<0.001). Imageability values behave accordingly : the two groups (NIs: 373.5, IQR=267; EDNs: 476, IQR=342) differ very significantly (W = 9355, p-value = 0.005). The difference in familiarity values was only marginally significant and will

---

[13]A U-test (or Wilcoxon test) was computed, instead of the common t-test, since the distributions of values of the two groups were not normally distributed, as shown by a Shapiro-Wilk test (W = 0.96859, p-value = 1.238e-06).
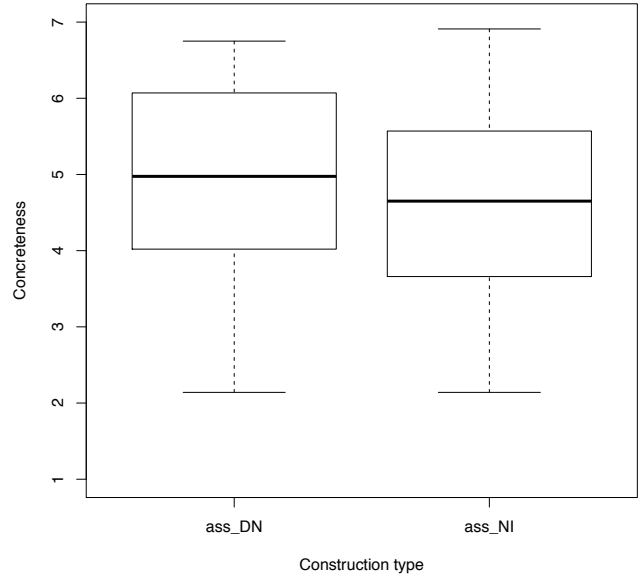
(a) Concreteness values
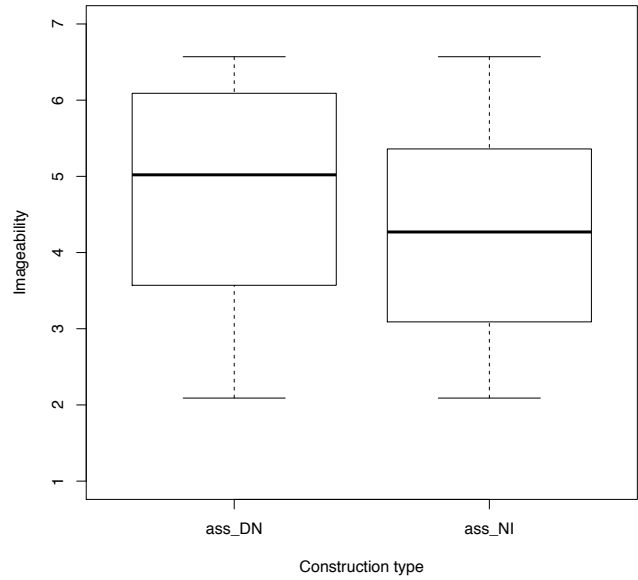
(b) Imageability values

Figure 4.1: Concreteness and imageability values from the lexvar dataset for the arguments of NIs and EDNs.

not be reported. The Della Rosa dataset includes, moreover, an additional variable, called *abstractness*: it can be interpreted as the opposite of the variable *concreteness*, since it assigns higher values (700) to very abstract entities and the lowest (100) to the most concrete ones[14]. The hypothesis is confirmed once more: arguments of NIs received higher values for abstractness (median=424, IQR=214) than EDNs' ones (median=263, IQR=294), and the difference is highly significant, according to a U-test (W = 5513.5, p-value <0.001). Figure 4.2 shows the values for concreteness, imageability and abstractness for the two groups.

4.4 Conclusions

In this chapter I investigated the competition between Italian NIs and EDNs, focusing on cases in which both forms exist for the same base verb. The aim was to observe differences in the way in which the two competitors are used and to understand which semantic features best explain it. I took into consideration various hypotheses from the existing literature regarding the disambiguation of base senses in the derivation and the role of arguments, specifically their concreteness level. I found out that NIs usually take as arguments more abstract nouns and select metaphorical senses of the base verb, whereas EDNs prefer literal senses, taking as arguments more concrete nouns.

We may ask ourselves why this association goes in the direction observed, and not in the other direction. These findings may be linked to the bounded/unbounded distinction discussed in section 4.1 and advocated by Gaeta (2002), Simone (2004) and Melloni (2007). One way to see it is that NIs might introduce the event as unbounded and indefinite, with focus on its duration, whereas EDNs might represent the event as a bounded or perfective process. Indeed, it is easier to imagine an event as bounded, perfective, with clear temporal delimitation, if it is anchored to a concrete object (like in 'the passage of the ship'), while abstract and metaphorical senses do not show a clear end-point ('the passing of time'). Abstract events seem more correlated with process rather than telic interpretations. Indeed, it is probably not a coincidence that NIs cannot pluralize, exactly like mass

---

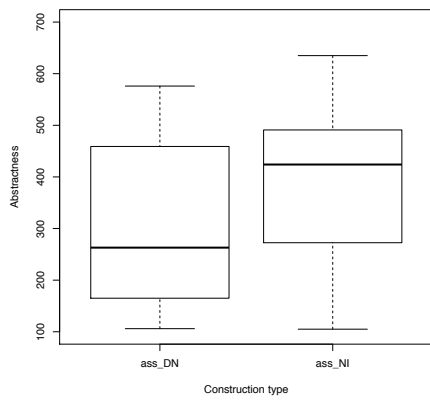[14]Remember that in the Della Rosa dataset participants were asked to rate either "abstractness" or "concreteness". This makes the judgments more robust with respect to the way the question was formulated.
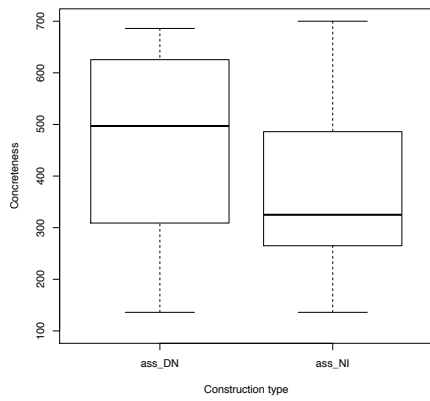
(a) Concreteness values
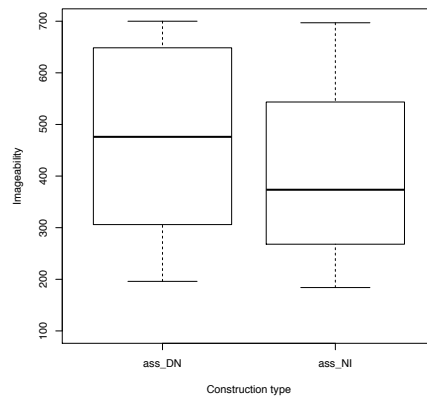
(b) Imageability values

(c) Abstractness values

Figure 4.2: Concreteness, imageability and abstractness values from the Della Rosa dataset for the arguments of NIs and EDNs.

nouns (*water*, *happiness*) which are the prototipical examples of unbounded denotation in the nominal domain (Link, 1983).

However, I suggest that it is preferable to think of the relevant distinction as pertaining to aspect rather than aktionsart. From the examples discussed in section 4.1, I do not believe that NIs (or EDNs) modify the aktionsart of the base verb, but rather convey a different aspectual value. Thus, I suggest that the distinction between NIs and EDNs should be cast in terms of perfectivity / imperfectivity, not telicity / atelicity. As Bertinetto (2001a) points out, these two levels are frequently confused.

Further studies should identify the most appropriate way to bring out this distinction, test the hypothesis above and provide evidence to support it. A multi-factorial analysis could then be used to take into account ontological domains, aspectuality and agentivity at the same time, stressing correlations between them.

I have shown that different complex nominalizations inherit only part of the polysemy of the base, which may be disambiguated by competing morphological processes. Further work could investigate if the degree of polysemy of the base verb has a role in the productivity of deverbal suffixes. We could hypothesize, indeed, that more polysemous verbs would be in a better position to derive doublets or triplets in the same semantic paradigmatic slot.

Chapter 5

Modeling meaning shift in distributional semantics space

5.1 Introduction

In this chapter, I present a quantitative study on the semantic differences between two event nominalizations of the German language, i.e. nominal infinitives (NIs, e.g. *das Evaluieren*, 'the evaluating') and *-ung* EDNs (henceforth UNGs, e.g. *die Evaluierung*, 'the evaluation'). As for Italian, these forms are quite similar to each other and even a native speaker cannot explain why and when to prefer one instead of the other. They are both transpositional and they both refer to the event of the base verb.

The study presented in this chapter sets out to investigate this distinction with methods from distributional semantics, a quantitative methodology that makes use of huge amount of data taken from corpora to represent the meaning of words as numerical vectors. A clearer introduction to the methodology is given in Section 5.4.1.

The aim is to test the hypothesis that the meaning shift associated with UNGs and NIs are markedly different on the distributional level. More specifically, I try to demonstrate that NIs are more predictable in meaning than EDNs, since their semantics remains closer to the one of the original base verb. EDNs, indeed, are more often objects of a semantic drift, which makes them acquire less compositional additional meanings.

Since NIs were originally inflectional forms rather than derivational ones, I place the comparison between nominalizations within the broader theoretical perspective of a graded distinction between inflection and derivation (e.g. Bybee, 1985). With an additional study, I compare the distributional semantic profile of NIs and UNGs with two other verb-derived grammatical forms, present participles (henceforth PPs, e.g. *evaluierend*, 'evaluating') and agentive deverbal nouns in *-er* (henceforth ERs, e.g. *der Evaluierer*, 'the evaluator'). These latter categories are taken as landmarks for, respectively, more clear inflectional and derivational processes. I hypothesize that NIs will pattern with inflectional PPs, whereas UNGs will be closer to the derivation extreme represented by the ERs. As a consequence, the distributional profiles of NIs will be different from ERs, while UNGs will be distant from PPs. Their distributional semantic representations should, thus, arrange on the inflection-derivation continuum as in Figure 5.1.
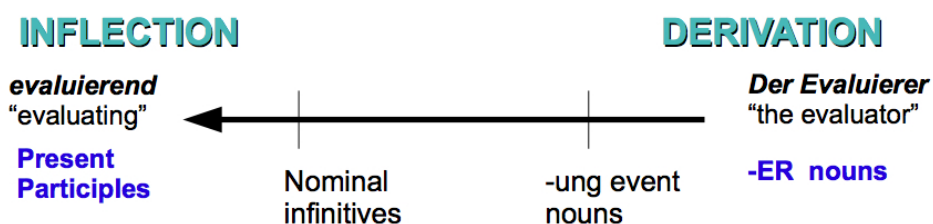
Figure 5.1: The inflection/derivation continuum

In what follows, I will first give an overview of the phenomena under investigation: in Section 5.2 I describe the two morphological processes and their characteristics. In Section 5.3, I introduce the theoretical concepts and predictions, i.e. regularity and meaning predictability (or transparency). In section 5.4, I describe the methodology employed, Distributional Semantics, and the measures used in the study. The first study is illustrated in section 5.5 and 5.6, where the experimental setup and the results are reported. Section 5.7 broadens the perspective to the distinction between inflection and derivation, and presents the second study. Section 5.8 provides discussion and conclusions.

5.2 German event nominalizations

In German, derivation is widely used to form event nominalizations. Different suffixes are available, like the borrowed *-ion* (*Spekulation*, 'speculate/speculation' from the verb *spekulieren*) and *-ur* (*Reparatur*, 'repair' from *reparieren*), or the native *-t* (*Fahrt*, 'drive/ride', from *fahren*), *-e* (*Hilfe*, 'help', from *helfen*) and *-ung* (*Verteidigung*, 'defend/defense', from *vertidigen*[1]). In addition, stem-derived nominals like *Fall* ('fall', from *fallen*) and nominal infinitives (e.g. *das Laufen*, 'walking') can also be used to derive an event noun.

In this study, I focus only on two kinds of German event nominalizations, nominal infinitives and deverbal nouns in *-ung*. The choice was motivated by the fact that they are both productive: NIs, as in Italian, can be formed from every base verb, while *-ung* EDNs are said to be one of the most productive among this class of suffixes (Eisenberg, 1994: 364, Shin, 2001: 297). An higher number of cases in which they are both formed for the same base verb can be, thus, possible. Given the quantitative nature of the present work, I needed a dataset large enough to make inferences, and not limited to few exemplars.

5.2.1 -ung deverbal nouns

The syntactic and morphological behavior of *-ung* nouns is typical of common nouns: they can be pluralized, their arguments can be realized either by a possessive pronoun or by a post-nominal genitive, they can be modified by adjectives and preceded by a definite or indefinite determiner (Demske, 2002; Scheffler, 2005).

From a semantic point of view, they manifest a wide range of meanings[2]:

- Event: *Fertigstellung*, 'completion'
- Result state: *Verärgerung*, 'fury'
- Result object: *Erfindung*, 'invention'

---

[1]Examples from Scheffler (2005: 2).

[2]The basic distinction between event and result state is parallel to what Grimshaw (1990) called complex and simple event nominals, and has been reused, expanded or redefined, for various languages in many different studies, such as Ehrich and Rapp (2000), Alexiadou (2001), Heyvaert (2003), Melloni (2007), Bauer et al. (2013), among many others.

In many cases, it is difficult to discern the different readings in the same token, since they are strictly interconnected: consider as an example (presented in Rossdeutscher and Kamp, 2010) the word *Absperrung* that can denote (i) the event of cordoning off, (ii) the resulting state, and (iii) the barricade that is produced during the event. The description of contextual constraints that allow a disambiguation of the actual reading has been object of numerous studies (Ehrich and Rapp, 2000; Hamm and Kamp, 2009; Kountz et al., 2007; Spranger and Heid, 2007; Eberle et al., 2009).

Nominalizations in *-ung* are usually said to be less productive than nominal infinitives: not all verbs allow suffixation by *-ung*. The constraints that govern the formation of these EDNs are highly debated. Esau (1973); Bartsch (1986); Demske (2002), among others, show that verbs expressing states or verbs referring to the beginning or the repetition of a situation do not allow *-ung* nouns. However, some counterexamples are presented by Knobloch (2003: 338)[3], e.g. *Erblindung*, 'loss of sight', *Erkaltung*, 'becoming cold'. Rossdeutscher and Kamp (2010) notice that most *-ung* nouns are derived from transitive verbs, while from intransitive verbs they are comparatively rare. Moreover, they argue that verbs that do not allow *-ung* nominals can be generally defined as activity verbs, like *arbeiten*, 'to work', or *schreiben*, 'to write'. In addition, applying the distinction between mono-eventive and bi-eventive verbs, they state that bi-eventuality is a discriminating factor for *-ung* nominalizations.

In this study, I do not focus on constraints on *-ung* EDNs productivity, since the main interest is a comparison with NIs when both are formed.

5.2.2 Nominal infinitives

Also called *infinitival nominals* or *nominalized infinitives*, nominal infinitives are inflected forms of verbs (the infinitive form) used in typical nominal syntactic contexts, such as subject or object position[4]. They are usually preceded by a definite or indefinite article and can be modified by adjectives, like nouns do:

(1)   Das Laufen   fiel   ihm   immer   schwerer.
         The walk.INF become.PAST him.DAT evermore hard

---

[3]Cited in Hartmann (2014).

[4]For a summary of the diachronic development of nominal infinitives I refer the reader to Werner (2013) and references therein.

'Walking was getting harder for him[5].'

(2) Es   herrschte ein Laufen   und Springen, ein Rennen   und
     There be.PAST a   run.INF and jump.INF, a   race.INF and
     Hüpfen.
     hop.INF.
     'There was running and jumping, racing and hopping[6].

(3) Das   schnelle Zerstören   der   Stadt war
     The.N rapid   destroy.INF the.GEN.F.SG city   was.3SG
     notwendig.
     necessary
     'The rapid destroying of the city was necessary[7].'

Arguments expressed by a genitive or a possessive pronoun can refer both
to the subject or object of the verb. A subject interpretation is preferred
(Knobloch, 2003, Scheffler, 2005), even though an object reading is possible:

(4) [Dieser Raum enthält   vertrauliches Material.] Sein Betreten   ist
     '[This   room contains confidential   data.]   Its   stepping-in is
     verboten.
     forbidden.'

They can be compounded to produce further nominals:

(5) Wir schreiben Briefe → Briefeschreiben
     We   write   letters → letter-writing

Regarding their semantics, they are truly transpositional, since they keep
only the event reading from the base verb. In a few cases, a result state or re-
sult object reading is also possible (e.g. *Verstehen*, *Ansehen* and *Schreiben*),
but these are usually lexicalized words that are far more frequent than other
nominal infinitives.

5.2.3   The historical development of German event nominalizations

As has been just seen, NIs and EDNs in present day German share the same
morpho-syntactic properties, with the only exception of pluralization. Ger-

---

[5]Example from Scheffler (2005: p. 7).
[6]Example from Scheffler (2005: p. 7).
[7]Example from Koptjevskaja-Tamm (2006: 656).

man NIs, indeed, contrary to Italian ones, do have a more consistent nominal behavior, without multiple alternative patterns. In Italian, as noted in Chapter 2, NIs can have both more verbal patterns (when they are modified by adverbs or express their direct object as a NP) and more nominal ones (where the arguments are expressed by a PP and the NI is modified by an adjective).

However, German NIs did not always had this clear nominal behavior. They underwent a change through history which is parallel to the one experienced by *-ung* EDNs.

In the Early New High German (ENHG) period, *-ung* nominals showed the same argument structure and sortal interpretation of their corresponding base verb (Demske, 2002: 68, but also Göransson, 1911; Behaghel, 1923[8]). This verbal behavior is, indeed, similar to more verbal infinitives. Only in recent times did they evolve a more noun-like character, with increasing restrictions on their productivity. Nominals derived from verbs of states or from inchoative/ingressive verbs, which are not attested in Present Day German (PDG), are attested in ENHG, as Demske (2002: 80) shows with a corpus study on newspapers of the 16th and 17th century. In PDG these missing *-ung* nouns seem to have been replaced by NIs.

In similar way, NIs went through a change from the verbal to the nominal pole (and also, as argued by Gaeta, 1998, from the inflectional to the derivational side). In Old and Middle High German, NIs had more mixed properties: they could be modified by adverbs, prepositional phrases and direct objects (Gaeta, 1998: 6):

(6)   min dort beliben
       my remaining there

(7)   da machet gat ein schaeiden die lieben von den leiden
       there do, friends, a separating the loves from the sorrows

(8)   da wart vil michel grüezen die lieben geste getan
       there was greatly greeting the dear guests

Gaeta (1998) summarizes the evolution of these properties in table 5.1.

German NIs, thus, lost their mixed behavior, preserving only more nominal properties. Thus, NIs in German are closer to other EDNs than Italian

---

[8]Cited in Gaeta (1998).

	OHG, MHG	NHG
nominative subject representation	-	-
accusative object representation	+	-
adverbial modification	+	-
adverbial > adjectival modification	+	+
definiteness operator	+	+
indefiniteness operator	+	+

Table 5.1: Syntactic properties of German infinitives from Old to New High German.

ones.

5.3 Meaning predictability / meaning shift regularity

Unlike Italian NIs, German ones can be said to be truly derivational in present day language (as argued, for example, by Gaeta, 1998). However, their inflectional origin is still present in three characteristics:

1. they do not show constraints on their productivity, i.e. they can be formed from every base verb;

2. they do not pluralize and are non-countable:

   (9) *Die Zerstören der   Stadt waren notwendig.
        'The destroyings the:GEN city   were   necessary.'

3. they present the possibility of being made passive[9]:

   (10)   Das Gesehen-werden ist die Hauptdimension der Kunst.
          'The being seen   is the main dimension of   art'.

In this study, I show that their inflectional nature is still visible also in the regularity and predictability of the meaning shift produced.

It has been frequently claimed, indeed, that the main difference between inflectional and derivational morphology is one of regularity (Haspelmath, 1996; Laca, 2001): inflection is more regular than derivation with respect

---

[9]Examples from Gaeta (1998: 5).

to both productivity and semantic shift. We may refer to the latter aspect as *semantic regularity, semantic transparency, semantic compositionality/idiosyncracy* or *semantic predictability*. I will not assume a difference between these terms and, in what follows, I will use them as synonyms. Booij (2000: 364) describes semantic transparency as a corollary of the general and productive nature of inflection. Whereas inflected forms have highly regular and predictable meanings, derived words, on the other hand, often acquire meanings that are not purely compositional, i.e. are not just a function of the meaning of its morphological constituents. In the words of Bybee (1985: 88), they present more often a *lexical split*. Derivation frequently produces a semantic difference between derivationally related words or between the derivative and its base[10]. For example, "something that is awful does not inspire awe anymore" (Bybee, 1985: 88).

Bybee (1985, but also 1995), moreover, firstly recognized a connection between semantic irregularity and the frequency of the derived word: the more frequent the derivative is, the more likely it is to be subject to lexical split. However, as noted by Hay (2001, 2003), the degree of decompositionality of a given word is better explained by its relative frequency and not by its frequency. By relative frequency she means the ratio of the frequency of the derivative to the frequency of the base term. Words with low relative frequency are more semantically transparent than words with an higher one.

The degree of transparency and the relative frequency also determine the way in which the word is stored and processed. In a dual-route theory of processing (see e.g. McQueen and Cutler, 1998), low frequency words are processed in a decomposition route, in which the meaning of the derivative is composed from the meaning of its constituent; high frequency words, instead, are processed in a whole-word route, since their resting activation is higher than for their bases.

From what we know about NIs and UNG nominals, we can hypothesize that (i) the relative frequency of UNGs is higher than the relative frequency of NIs, (ii) NIs are more transparent than UNGs, i.e. their meaning is closer to the base verb meaning. We know, indeed, that derivational forms can acquire more often idiosyncratic meanings, being subjects of lexical

---

[10]However, some instances of lexical split are found in inflection too: an example from English is the difference between the plural *clothes*, which has the meaning of 'garnments', and the singular form *cloth*, which means 'woven material' (Bybee, 1985: 88, Booij, 2000: 364).

split. Inflectional forms, instead, tend to preserve only their compositional meaning. Moreover, in the nominalization domain, -ung EDNs are more subjects to type shifts (when the derivative denotes an entity, not an event as the base verb) or type ambiguity (when it denotes both the event and an entity linked to the event).

In what follows, I propose some possible measures to quantify the degree of transparency of a nominalization, specifically I compare the derivative to the base verb with distributional semantic models: the closer the distributional representation of the nominalization to the representation of the base, the higher its transparency.

Before describing the measures used and the experimental setup, I will first give an introduction to the methodology, i.e. distributional semantics.

5.4 Methodology

5.4.1 Distributional semantics

Distributional semantics (or distributional semantic models, or vector space model) is a widely used approach in computational linguistics which approximate the meaning of a word as a numerical vector of the occurrences of that word with a series of other words. The values of co-occorrences are extracted from corpora (usually huge ones), thus from collections of naturalistic data (not created directly for the purpose of the research).

As for the corpus-based study presented in Chapter 4, the basic idea underlying this approach is that words similar in meaning tend to occur in similar linguistic contexts. Speakers of English know, for example, that *cat* and *kitten* have really similar meanings and that, indeed, it is possible to use them in the same sentences:

(11)   a.   The **cat** was purring on my lap and I couldn't move for hours.
            b.   The **kitten** was purring on my lap and I couldn't move for hours.

(12)   a.   Mrs Pipchin had a black **cat**, who generally lay coiled upon the centre of the sofa.
            b.   Mrs Pipchin had a black **kitten**, who generally lay coiled upon the centre of the sofa.

As previously stated, this idea is usually referred to as the *distributional hypothesis* and finds his origins in the works by Harris (1954) and Firth (1957). With the advent of statistical methods in computational linguistics in the nineties, it has found numerous applications in different Natural Language Processing tasks and in the cognitive sciences. Among the first and most influential distributional models of meaning, we can cite *Latent Semantic Analysis* (LSA, Landauer and Dumais (1997)) and *Hyperspace Analogue to Language* (HAL, Lund and Kevin (1997)).

But what exactly does the vector of a word tell us? And how do we obtain it? Consider as example the English word *cat*: if we extract some sentences in which it appears from the corpus Ukwac (Baroni et al., 2009), they look like that (printing only 5 words par side):

ignoring each other , like a	**cat**	that puts its head under
of our beloved dogs or	**cats**	having worms.
They all loves Latin , as	**cats**	loves milk.
There are strange frogs ,	**cats** ,	rats , lizards , and even more
a cheeky , affectionate and sociable	**cat**	and we just love him.

Table 5.2: Five sentences corpus for the word *cat*.

From the sentences attested in our corpus, we count how many times the word *cat* (our target word) occurs with all the other words (our context words) in the lexicon (our vocabulary, which corresponds to the length of the vector, indicated by $|V|$). Usually we restrict the number of context words to a feasible amount for processing and storage (usually between 10,000 and 50,000 words). Both lemmas or word forms can be considered as target or context words. Moreover, we can decide to which distance context words should be considered, e.g. in a window of 5 words to the left and to the right from our target word. In general, the window ranges from 1 to 8 words on each side and it has been noted that the shorter the window, the more syntactic the representation of the word, since the relevant information comes from words linked by direct syntactic links to our target word (Jurafsky and Martin, 2016). On the contrary, longer windows give us a more semantic representation. From our small five-sentences corpus in 5.2, we can extract a toy vector for the target word *cat*: we consider as context words (the dimensions of the vector) lemmas actually attested in our examples and others which are not, but which we decided to consider in our lexicon of context

86

words. We, thus, lemmatize them and count co-occurrences[11].

	dog	worm	philosophical	Latin	love	milk	frog	fly
cat	1	1	0	1	3	1	1	0

Table 5.3: A toy vector for the word *cat*.

This procedure is done for all our context words (in the study reported below, I considered ten thousand context words), producing thus very long vectors. Moreover, a single vector is not so useful even if it can tell us that cats are lovable entities and probable linked in some way to other animals (but also linked to Latin!). Thus, we extract vectors for other target words we are interested in, possibly even for all the words that we already used as dimensions. What we obtain this way is a huge matrix (a vector space) where for every target word we know how many times it occurs with all the context words. However, co-occurrences values are misleading in cases in which one word is very frequent or very unfrequent; for this reason, they are usually transformed into PPMI (Positive Point-wise mutual information) scores[12] (Church and Hanks, 1990). PPMI is a measure of association which computes how often two events (in our case, two words) occur, compared with how often they would have occurred it if they were independent, i.e. only based on their independent frequency. As it can be noted, its aim is the same as the log-likelihood measure described in Chapter 4. Considering a target word t and a context word c, we define the PMI value of their association as the logarithmic ratio of the probability of seeing the two words together to the probability of encountering them separately:

$$PMI(t,c) = \log_2 \frac{P(t,c)}{P(t)P(c)}$$

PPMI, differently from PMI, just ignore negative values, transforming them to 0.

Further operations can be applied to improve results, like dimensionality reduction with the SVD (Singular Value Decomposition) technique or NMF

---

[11]Note, however, that there are also other types of distributional models which do not simply count co-occurrences (so called *count models*), but rather predict them with neural networks (*predict models*, see e.g. Bengio et al., 2003; Baroni et al., 2014). However, in this thesis, I use only count models, since they allow a more linguistic inspection of the prominent dimensions, whereas predict models do not.

[12]Other measures can be used for the same purpose, e.g. Positive Local Mutual Information (PLMI, see e.g. Evert, 2005).

(Negative Matrix Factorization). These techniques collapse together similar dimensions in "latent" dimensions to reduce the size of the matrix (Landauer and Dumais, 1997). However, as in the case of the present study, these reductions do not always improve the models and sometimes a simple full matrix is preferred. Moreover, some of the measures applied, cannot be computed for dense spaces, requiring a full space.

Once that we have extracted our co-occurrences matrix, we can compute similarity values in different ways.

5.4.2 Distributional semantic measures and quantitative predictions

Distributional semantics has been mostly used to compute the similarity between words or between documents. However, as pointed out by Padó and Lapata (2003), the notion of similarity that distributional semantics uses has been frequently criticized because of its vagueness. DSMs will report similarity between words that are synonyms, hypernyms and hyponyms, antonyms, and so on, without a clear differentiation. In other words, DSMs can tell us whether two words are similar but they cannot discern why and how. On the other hand, as Miller and Charles (1991) argue, the concept of word similarity, even if vague, has been demonstrated to be intuitively clear to people who were asked to give judgments on that. So, it is true that DSMs do not provide us with further specification on the kind of similarity that is attested between two words, but this vague notion is already informative in many ways. In addition, numerous works are going into the direction of disambiguating the possible types of similarity (see, e.g., Baroni and Lenci, 2011).

The most common measure to compute semantic similarity (in this broad sense) is the cosine of the angles between two vectors. The values of the cosine similarity measure range from 1 for really similar vectors, over 0 for orthogonal vectors, to -1 for opposite vectors. However, if a PPMI transformation has been applied, the values of cosine will range between 0 and 1, since there will not be negative values. Cosine is based on the dot product operator (also called inner product): the dot product of two vectors will be high when they have large values in the same dimensions; on the contrary, if two vectors have zeros in different dimensions, they will have a dot product of 0, which represents their strong dissimilarity. Cosine similarity measure

is the normalized version of the dot product, i.e. it normalizes the value for the vector length. More frequent words have longer vectors, i.e. less zeros and higher values, since they occur more often and then have higher probability to occur with an higher number of different context words. The raw dot product will be, thus, consequently higher for more frequent words. To overcome this problem, the cosine measure divides the dot product by the lengths of each of the two vectors. The vector length is defined as the squared root of the sum of the quadratic values of each dimension:

$$|\vec{t}| = \sqrt{\sum_{i=1}^{N} t_i^2}$$

Consequently, the cosine of vector t and c is computed as follows:

$$cosine(\vec{t}, \vec{c}) = \frac{\vec{t} \cdot \vec{c}}{|\vec{t}||\vec{c}|} = \frac{\sum\limits_{i=1}^{N} t_i c_i}{\sqrt{\sum\limits_{i=1}^{N} t_i^2} \sqrt{\sum\limits_{i=1}^{N} c_i^2}}$$

Despite the wide use of cosine as a measure of semantic similarity, it should be noted that it is a symmetric measure: it reports the same values if we consider how similar word x is to word y and how similar word y is to word x. However, frequently a word x can be similar to y, but word y can have other closer words. It is usually the case of hypernyms - hyponyms (Lenci and Benotto, 2012): an hyponym like *lion* occurs in many of the contexts of its hypernym *animal*; on the other hand, *animal* has a wider meaning and it will have a number of contexts which will not be shared by *lion*. The hyponym will be closer to its hypernym, but the opposite will not be true in the same measure. Lenci and Benotto (2012) propose an asymmetric (or directional) similarity measure to account for this problem, the *InvCL* measure. It counts how included the vector of a word x is in its hypernym y, i.e. in which proportion its contexts are the same of y, but it then takes into consideration how much of the vector of y is not included in x. InvCL is based on the so-called *ClarkeDE* measure[13] (Clarke, 2009), which compute the degree of inclusion of word x into word y:

---

[13]The ClarkeDE measure was at its turn based on the measure proposed by Weeds and Weir (2003); Weeds et al. (2004). For this study, I tested also the original ClarkeDE measure, but it had worse results than InvCL, which for this reason will not be reported.

$$ClarkeDE(u,v) = \frac{\sum_{f \in F_u \cap F_v} min(w_u(f), w_v(f))}{\sum_{f \in F_u} w_u(f)}$$

Given this formula, InvCL is computed as follows:

$$InvCL(u,v) = \sqrt{ClarkeDE(u,v) \cdot (1 - ClarkeDE(v,u))}$$

In the present study, I use this measure as a semantic transparency measure. While cosine tells us if two vectors are close or not (and, consequently, if two words are similar in meaning), InvCL (as well as ClarkeDE measure) calculates how much a vector x is included in vector y: if a nominalization is almost totally included into the base verb vector, it means that its meaning is just a transposition of the base verb semantics and that it is totally transparent; if the inclusion value is lower, instead, it means that the nominalization has acquired some additional content and has been subject to meaning shift.

I compute the InvCL values for NIs with respect to the base verb and then predict that these values will be higher than UNGs ones. NIs should be, indeed, more semantically included into the vector representation of the base verbs than UNGs nouns do. The semantic transparency of the derivative is, in this way, quantified in terms of features inclusion.

In addition to transparency, we also thought that other measures could represent a theoretical counterpart of semantic irregularity, i.e. addition of semantic content. Laca (2001) claims that "a derived lexeme presupposes the lexeme it is derived from", implying that derived lexemes have more (or at least the same) semantic content than their bases. It was explicitly stated by Koontz-Garboden (2007) as the *Monotonicity Hypothesis*: derivational affixes may or may not add semantic content to their bases, but never remove it. Since EDNs tend to acquire additional readings, other than the event ones, I hypothesize that UNGs are not only less transparent, but also more rich in terms of semantic content. Thus, transparency and semantic content should be inversely correlated: the less transparent a derivative is, the more addition of content it should present; more transparent words, instead, will tend to leave the semantic content of the base unaltered.

By means of a metric from information theory, i.e. entropy, Padó et al.

(2015) tested this hypothesis for German derivational patterns, while Kisselew et al. (2016) inspected the directionality of conversion processes in English. Intuitively, if the entropy is low, the vector will have a big difference between its values: some dimensions will have high values, whereas others will have very low ones. On the contrary, when the entropy of the vector is high, the values of the different dimensions are, more or less, close to each other, not presenting big variance. The formula is reported below.

$$H(\vec{w}) = -\sum_i \vec{w}_i \cdot \log \vec{w}_i$$

They show how higher entropy corresponds to less semantic content, that is less restrictions and specification on contexts, whereas lower entropy indicates more specific context of use and thus more semantic content. However, we are not interested in the semantic content of the derived word par se, but rather to its relation to the base verb. For this reason, the delta entropy is computed: it corresponds to the difference between the entropy of the base and the entropy of the derived term. Positive values of Δ entropy mean that the derivative has not higher entropy of the base, that is it has more specific contexts and thus more semantic content. On the contrary, negative values of Δ entropy indicate that the base is more specific and that the derivative does not acquire any additional content.

With this respect, we expect that transparency and semantic content go in the opposite direction: UNGs should be less transparent with respect to the base verb, but should occur in more specific contexts.

Lastly, the neighborhood density was taken as measure of semantic content (Marelli and Baroni, 2015; Sagi et al., 2009). Given a word x we define its nearest neighbor as the word whose vector is closer to the vector of x, where closeness is computed by means of cosine. For each word, the top n neighbors can be defined as the neighborhood of x. However, for each word in our vector space, the distance with the top neighbors can vary. Some words will be more isolated, while others will have a dense neighborhood. The idea is that words with more semantic content (thus, in a sense, more polysemous words) have denser neighborhood, since there will be more words that can be considered closed in meaning to them. The neighborhood density is computed as the average cosine similarity between the target word and the

neighbors considered[14]. We expect that UNGs have denser neighborhood than NIs. However, as for entropy, we considered the relation between the base verb and the derivatives. Thus, we computed the difference between the neighborhood density of the base verb and the neighborhood density of the derivative. We expected NIs to have higher Δ density than UNGs, since they should import less semantic content and their neighborhood should be less dense.

5.5 Experimental setup

5.5.1 Dataset

The experimental items for the dataset are sampled from the SdeWaC corpus (Faaß and Ekart, 2013), which was POS-tagged and lemmatized using Tree-Tagger (Schmid, 1994). Pairs of NIs and UNGs for the same base verb are extracted from the DErivBase (Zeller et al., 2013), a large-coverage database of German derivational morphology. The pairs were selected among the items whose frequency was in the middle range (two standard deviations) both for the NI and the EDN. Unusually frequent items are, as we already noted, uncommon element of the category and usually have already undergone a lexicalization of their meaning; for this reason, we did not want to include them in the dataset. Given this control among the frequency of items in their respective category, it was not possible to balance for the frequency among categories. NIs and EDNs have, thus, very different frequencies, and like Italian EDNs tend to be more frequent than NIs. To overcome this problem, the analyses are done with a regression model, taking frequency as a control variable. Moreover, the sample was restricted in order to have the same dataset for the experiment that will be described in Section 5.7, in which two additional categories are taken into consideration: present participles and agent nouns in *-er*. This further step is described below. The sampling procedure resulted in a dataset containing 115 pairs of NI and EDN.

---

[14]Another variant of the neighborhood density was also taken into consideration. In this case (to which I refer to as the Δ N.density, or neighbors density), for each neighbor, the cosine similarity with every other neighbors of the target word is calculated; thus, the average is computed. However, no difference between density and N.density was seen. They did not have an effect in the prediction task that will be described in the section below. For this reason, I report only the results related to density.

5.5.2 Models and Tasks

For each word in the dataset (derived and base terms), distributional count vectors have been extracted from the SdeWaC corpus (Faaß and Ekart, 2013), adopting a symmetric 5-words context window with PPMI as a scoring function and the 10 thousand most frequent content lemmas as context dimensions.

For each pair, it was further computed the value for each measure described above and summarized as follows:

- Cosine:

  - Cosine similarity between NI and the base verb;
  - Cosine similarity between EDN and the base verb;

- InvCL:

  - InvCL measure of the NI's inclusion in the base verb vector;
  - InvCL measure of the EDN's inclusion in the base verb vector;

- Δ entropy:

  - Difference between base's entropy and NI's entropy;
  - Difference between base's entropy and EDN's entropy;

- Δ density:

  - Difference between the neighborhood density of the base and the neighborhood density of the NI;
  - Difference between the neighborhood density of the base and the neighborhood density of the EDN;

For Δ density, 5 different numbers of top neighbors were considered: 10, 20, 30, 40 and 50. However, no significant difference was found among them and thus I will report the results only for density computed on the base of 20 neighbors.

As additional information, the number of synsets (i.e. different semantic senses) and the semantic class of the base verbs were extracted from GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010), a lexical resource similar to WordNet (Fellbaum, 1998). These two variable were taken as covariates in the analysis.

5.6 Analyses and results

The data collected were analyzed in a logistic mixed-effects model (Baayen et al., 2008) with the R software (R Core Team Team, 2015).

The task for the regression model was to predict for each datapoint the type of the nominalization (NI or EDN), given as predictors the values of transparency and specificity. Moreover, the log-transformed frequency of the base and the derived term were given as control variables (since we know that frequency influences semantic transparency, see Section 5.3), as well as the number of synsets and the semantic class of the base verb. All numeric variables were scaled on their mean. The base verb lemma was taken as random variable, since for every base verb two observations are given, one for the corresponding NI and one for the corresponding EDN.

In order to identify the model which better predicts the type of nominalization, I proceeded by forward selection of the significant variables. Since it is known that frequency has a big role in derivational processes, the frequency of the derivative was the only predictor (except for the random effect variable) in the null model. At each step, one predictor was added and, by means of a likelihood ratio tests performed with the anova function, it was evaluated if the more complex model was better than the simpler one, and thus if the new predictor had to be kept.

A summary of the model selection is summarized in Table 5.4.

Model	Predictors	R^2	AIC	BIC	Anova
M1	Fqbase (***) + Fqderived (***)	0.720	154.68	168.43	
M2	Fqbase (***) + Fqderived (***) + Δ density (n.s.)	0.726	154.91	172.10	n.s.
M2	Fqbase (***) + Fqderived (***) + Δ entropy (n.s.)	0.743	155.66	172.86	n.s.
M2	Fqbase (***) + Fqderived (***) + cosine (n.s.)	0.738	155.60	172.79	n.s.
M2	Fqbase (***) + Fqderived (***) + InvCL (*)	0.757	149.73	**166.92**	**
M3	Fqbase (***) + Fqderived (***) + InvCL (*) + Numsyn (n.s.)	0.757	151.68	172.31	n.s.

Table 5.4: Model selection for the prediction of the type of nominalization (NI or UNG).

94

The best model, based on the BIC values[15], has as predictors only the frequencies of both the bases and the derivatives and the InvCL measure[16]. Neither entropy, density or cosine had a significant effect in the prediction. The number of synsets or the semantic class of the base had neither an effect, as well as the interaction between the predictors.

The final best model is reported in Table 5.5 and represented in Figure 5.2.

	Estimate	Std. Error	z value	Pr(> |z|)	
(Intercept)	0.2960	0.2385	1.241	0.2147	
fqbase	-1.2942	0.2653	-4.879	1.07e-06	***
fqderived	3.6460	0.4793	7.606	2.82e-14	***
InvCL	-0.8149	0.3177	-2.565	0.0103	*

Table 5.5: Best model for the prediction of the type of nominalization (NI or UNG).

As predicted, higher inclusion (InvCL) values are connected to NIs (which correspond, on the y axis, to the 0), whereas UNG nouns present lower values of inclusion. The effect is significant and strong, since the model has also taken into account the strong effect of frequency. The semantic content measures are not significant in predicting the category[17], probably because there is no sufficient difference in semantic content between the two types of nominalization. It is interesting to note, instead, that the more general similarity measure based on cosine distance is not able do predict the category correctly. We can suspect that, at a more general semantic level, NIs and UNGs are similar, since they both refer to the event of the base verb. However, at the more specific level of features inclusion, investigated by the InvCL measure, a difference of semantic transparency emerges, and NIs turn out to be closer than UNGs to the base verb meaning.

---

[15] The BIC value is preferred in place of the AIC one since the former privileges simpler models, i.e. model with a lower number of variables.

[16] The R formula is thus:
m2 ← glmer(Category ∼ fqbase + fqderived + InvCL + (1—Base), family="binomial")

[17] Other settings of the semantic space were used to compute these measures: NMF and SVD dimensionality reduction were applied, as well as for different numbers of neighbors for the density measure. The negative results persisted in every condition. Moreover, I tried also to built the logistic model with the residualized values of the measures. No improvement was found. Thus, we can effectively argue that there is no difference between the two groups for these measures.
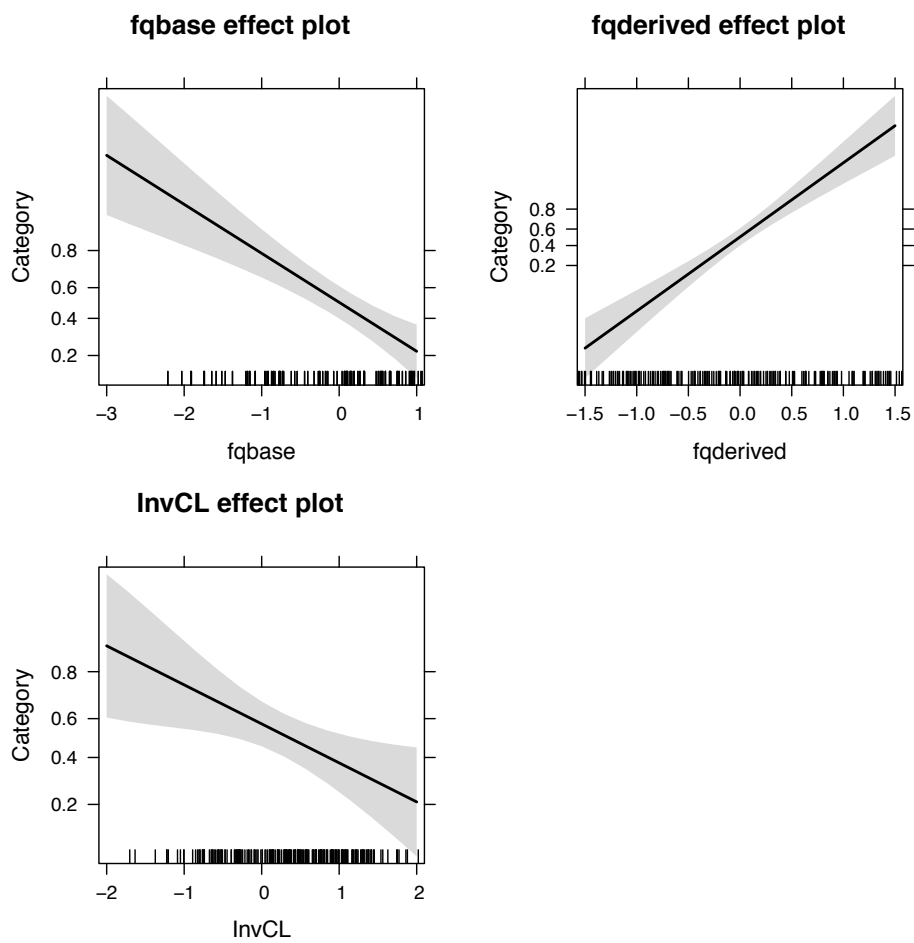
Figure 5.2: Effects of the final model for the prediction of nominalization's type (NI vs UNG). On the y-axis, 0 corresponds to the NI category, 1 to UNG.

5.7 Comparing the transparency of morphological processes

In the last section, the results have shown that NIs are significantly more semantic transparent than their corresponding UNG nouns. This behavior is, in my opinion, inherited by their inflectional nature. It has been frequently argued that inflectional processes are more transparent than derivational ones (see Chapter 1 and Section 5.3) and it seems the most unproblematic statement about the inflection - derivation distinction. As claimed by Laca (2001: 1217) "the semantics of derivatives is usually only partly compositional, since derivatives, as lexemes, tend to acquire some degree of autonomy from their bases and from the process they instantiate".

In this section, I present a study in which I compare the distributional representation of the two nominalizations investigated above to two other morphological processes: present participles (henceforth PPs) and agent nouns in -er (henceforth ERs). The aim is to show that NIs share with PPs the higher degree of semantic transparency, since both are inflectional processes. UNGs, instead, will act as ERs, the derivational landmarks, which will show in turn lower degree of transparency. Inflectional operations tend to be semantically regular, while derivation ones have more frequently unpredictable semantic effects.

The hypothesis is that the four items will arrange on a continuum of semantic transparency in the following way:

Higher semantic transparency			Lower semantic transparency
PPs	NIs	UNGs	ERs

5.7.1 Experimental setup

The semantic space used for this experiment is the same described above (§ 5.5). Quadruples of NIs, UNGs, ERs, and PPs for the same base verb were extracted from the DErivBase (Zeller et al., 2013), considering only items which, for each category, were in the middle range of frequency (two standard deviations). The dataset was, thus, composed of 115 quadruples. For each derivative, the distributional vector was extracted from the SdeWaC

corpus (Faaß and Ekart, 2013), adopting a symmetric 5-words context window with PPMI as a scoring function and the 10 thousand most frequent content lemmas as context dimensions. Transparency and semantic content measures described in § 5.5.2 were computed for each datapoint.

5.7.2 Analyses and results

In addition to the comparison between NIs and UNGs reported in § 5.6 (task 1), three different tasks were given to the statistical model: predicting the distinction between UNG vs ER (task 2), predicting the distinction between NI and PP (task 3), predicting the distinction between ER vs PP (task 4).

Task 2 With regard to task 2 (Table 5.6), Δ entropy and Δ density do not have a signficant effect on predicting the category. InvCL and Cosine, instead, do (Table 5.7 and 5.8). As can be seen from Figure 5.3 and 5.4, UNG nouns have higher cosine similarity with the base verbs than ER nouns. InvCL goes in the same direction: UNG nouns have higher values of inclusion into the base verb vector than ER, as expected.

Model	Predictors	R^2	AIC	BIC	Anova
M1	Fqbase (***) + Fqderived (***)	0.358	257.00	270.00	
M2	Fqbase (**) + Fqderived (***) + Δ density (n.s.)	0.359	258.89	276.08	n.s.
M2	Fqbase (**) + Fqderived (***) + Δ entropy (n.s.)	0.363	258.1	275.29	n.s.
M2	Fqbase (**) + Fqderived (n.s.) + cosine (***)	0.502	219.91	**237.10**	***.
M2	Fqbase (*) + Fqderived (***) + InvCL (***)	0.484	231.33	248.53	***

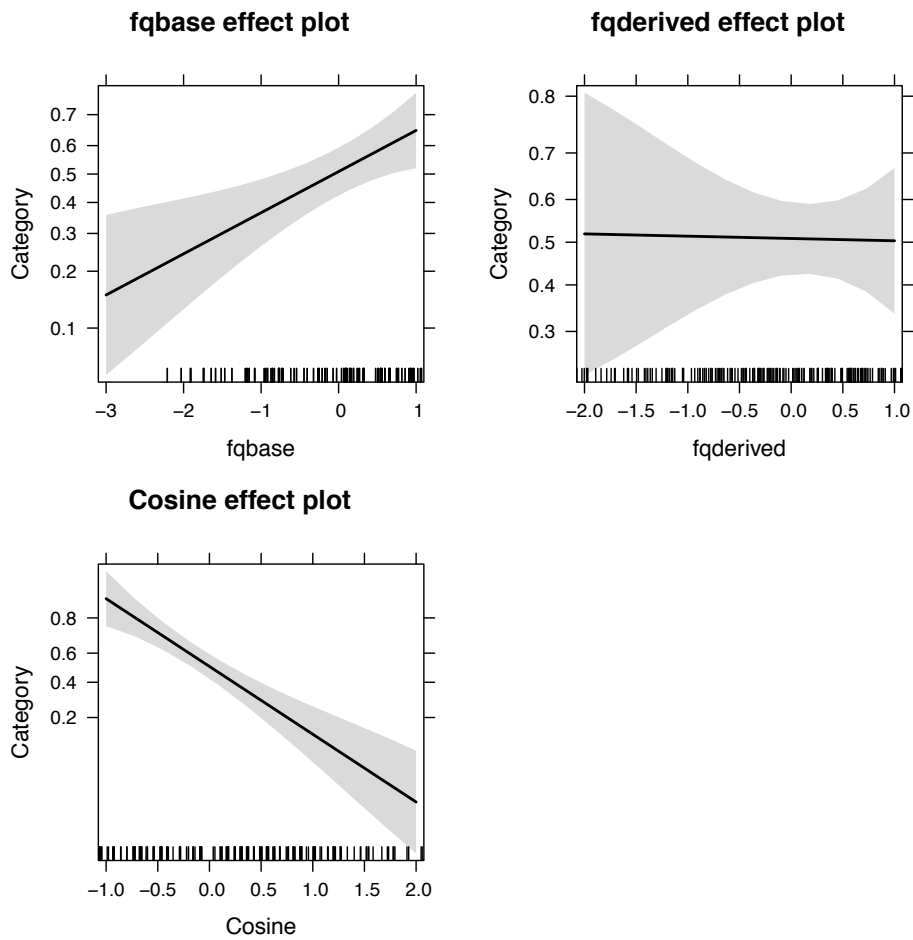Table 5.6: Model selection for the prediction of UNG vs ER.

Figure 5.3: Effects plots for the prediction of category UNG vs ER (task 2), with cosine as predictor. On the y-axis, 0 corresponds to the UNG category, 1 to ER.
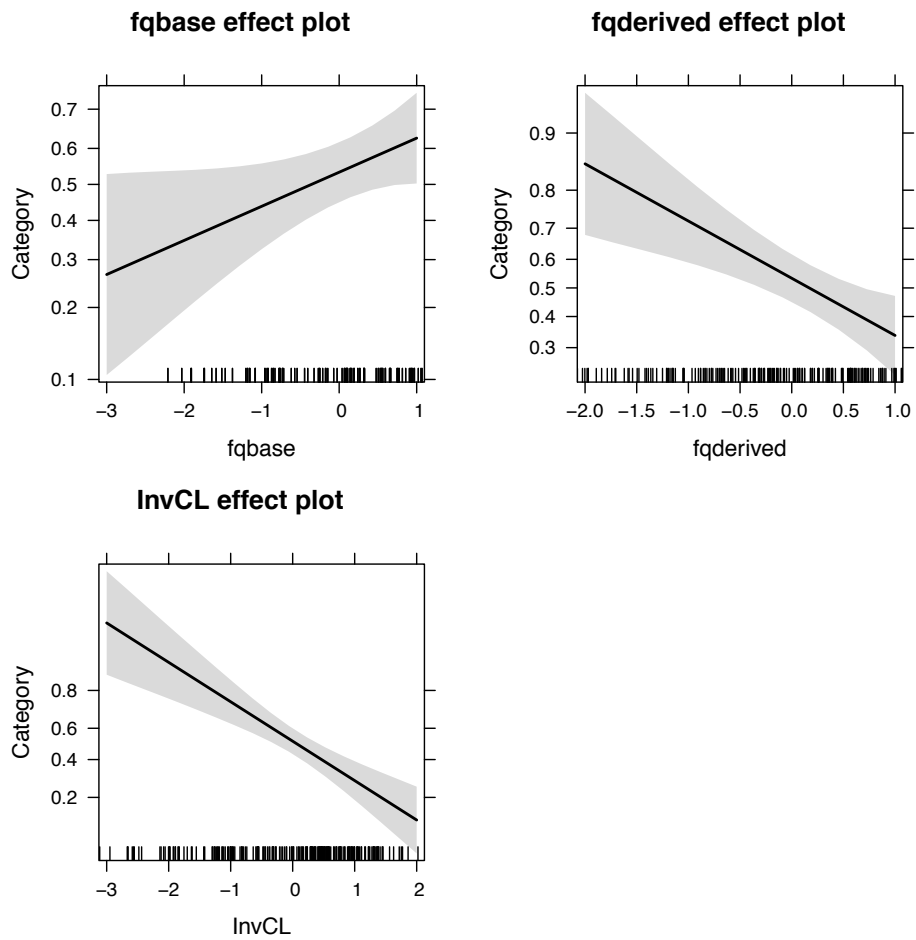
Figure 5.4: Effects plots for the prediction of category UNG vs ER (task 2), with InvCL as predictor. On the y-axis, 0 corresponds to the UNG category, 1 to ER.

	Estimate	Std. Error	z value	Pr($> |z|$)	
(Intercept)	0.03527	0.17379	0.203	0.8392	
fqbase	0.58686	0.19250	3.049	0.0023	**
fqderived	-0.02189	0.32313	-0.068	0.9460	
Cosine	-1.88714	0.35326	-5.342	9.19e-08	***

Table 5.7: Fixed effects of the model for the prediction of morphological class (UNG vs ER), with Cosine as predictor.

	Estimate	Std. Error	z value	Pr($> |z|$)	
(Intercept)	0.07071	0.16926	0.418	0.676144	
fqbase	0.38417	0.18712	2.053	0.040062	*
fqderived	-0.81159	0.23819	-3.407	0.000656	***
InvCL	-1.02179	0.21453	-4.763	1.91e-06	***

Table 5.8: Fixed effects of the model for the prediction of morphological class (UNG vs ER), with InvCL as predictor.

Task 3 In task 3, the model predicts the distinction between NIs and PPs. As shown by Table 5.9, no variable helps in the prediction of the category (the model fit is, indeed, very low), not even frequency. NIs and PPs are thus really similar on the distributional profile.

Model	Predictors	R^2	AIC	BIC	Anova
M1	Fqbase (n.s.) + Fqderived (n.s.)	0.011	324.73	338.48	
M2	Fqbase (n.s.) + Fqderived (n.s.) + density (n.s.)	0.011	326.73	343.92	n.s.
M2	Fqbase (n.s.) + Fqderived (n.s.) + entropy (n.s.)	0.241	324.62	341.81	n.s.
M2	Fqbase (n.s.) + Fqderived (n.s.) + cosine (n.s.)	0.024	324.37	341.56	n.s.
M2	Fqbase (n.s.) + Fqderived (n.s.) + InvCL (n.s.)	0.014	326.12	343.31	n.s.

Table 5.9: Model selection for the prediction of NI vs PP.

Task 4 The results for task 4 are summarized in Table 5.10, where a summary of the model selection is given.

Model	Predictors	R^2	AIC	BIC	Anova
M1	Fqbase (.) + Fqderived (***)	0.109	306.58	320.33	
M2	Fqbase (n.s.) + Fqderived (.) + density (n.s.)	0.114	307.76	324.95	n.s.
M2	Fqbase (.) + Fqderived (*) + entropy (n.s.)	0.111	308.26	325.45	n.s.
M2	Fqbase (n.s.) + Fqderived (***) + cosine (***)	0.328	267.74	284.93	***
M2	Fqbase (*) + Fqderived (***) + InvCL (***)	0.423	251.50	268.69	***

Table 5.10: Model selection for the prediction of ER vs PP.

As for the other tasks, Δ entropy and Δ density do not have any significant effect; InvCL and Cosine, instead, have an highly significant role in the prediction. Their values are higher for PPs and lower for ERs. The fixed effects of the two models are reported in table 5.11 and 5.12 and represented in Figure 5.5 and 5.6.
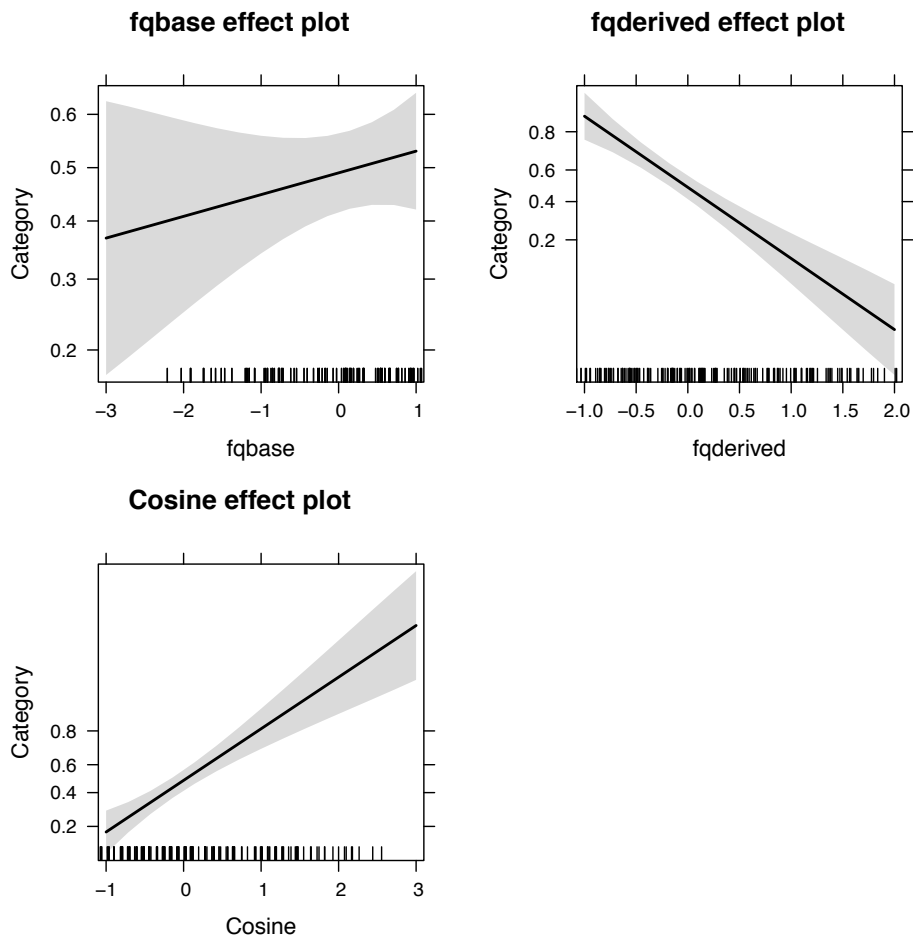
Figure 5.5: Effects plots for the prediction of category ER vs PP (task 4), with cosine as predictor. On the y-axis, 0 corresponds to the ER category, 1 to PP.
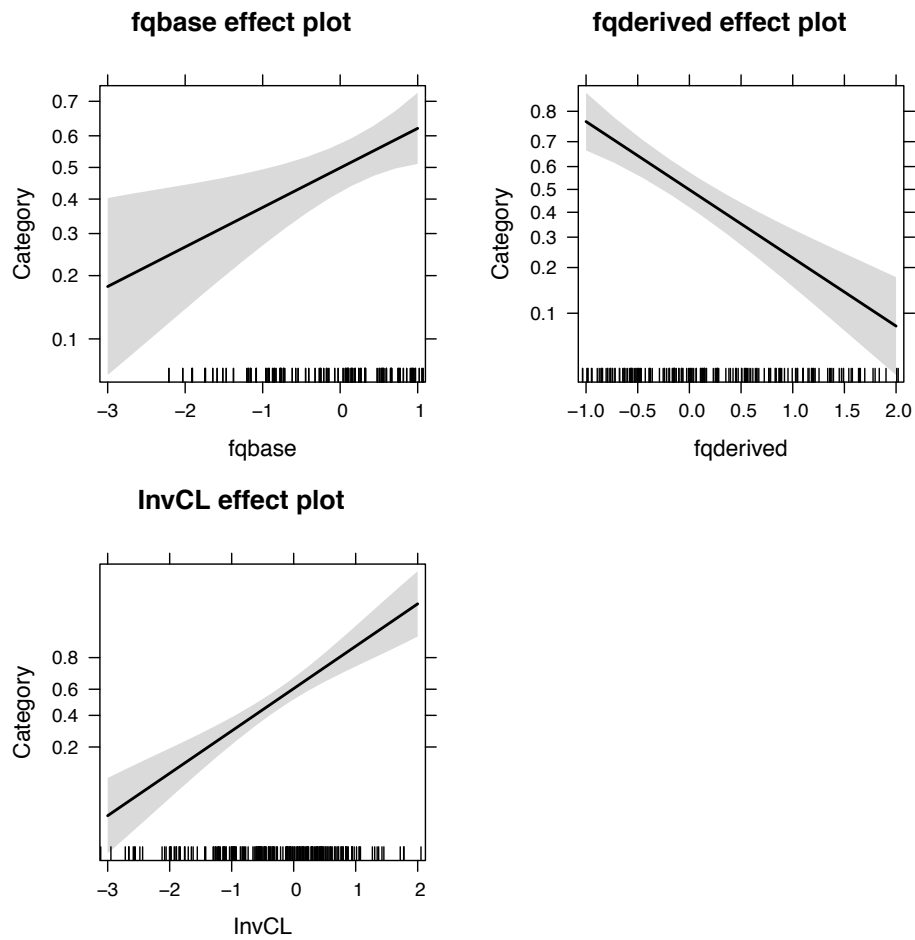
Figure 5.6: Effects plots for the prediction of category ER vs PP (task 4), with InvCL as predictor. On the y-axis, 0 corresponds to the ER category, 1 to PP.

	Estimate	Std. Error	z value	Pr($> |z|$)	
(Intercept)	-0.05017	0.15209	-0.330	0.741	
fqbase	0.16531	0.16931	0.976	0.329	
fqderived	-1.82459	0.28078	-6.498	8.13e-11	***
Cosine	1.49918	0.26750	5.604	2.09e-08	***

Table 5.11: Fixed effects of the model for the prediction of morphological class (ER vs PP), with Cosine as predictor.

	Estimate	Std. Error	z value	Pr($> |z|$)	
(Intercept)	0.4253	0.1707	2.491	0.01273	*
fqbase	0.5067	0.1821	2.782	0.00541	**
fqderived	-1.2097	0.2075	-5.831	5.50e-09	***
InvCL	1.3144	0.2139	6.145	8.02e-10	***

Table 5.12: Fixed effects of the model for the prediction of morphological class (ER vs PP), with InvCL as predictor.

5.8 Conclusions

The studies presented in this chapter confirmed the hypothesis, taken from the existing literature, that NIs are more semantically transparent than -*ung* nouns. Their meaning is closer to the meaning of the corresponding base verb, whereas UNGs drift more apart their bases. To conduct the experiments, distributional semantics was used: it allows us to represent quantitatively the meaning of words, task which would have been difficult otherwise. Thus, the meaning of NIs and UNGs was represented by numerical vectors of their collocations and quantitative measures could be applied. Distributional semantics, which has so far been primarily used to characterize the meaning of words or word compositions, proves to be a strong methodology even when applied to the meaning of constructions or to the comparison between morphological forms like UNGs and complex constructions like NIs.It thus offers a mathematical and statistical way to evaluate linguistic theories with corpus data.

The case of German nominalizations was particularly appropriate for this study because, due to historical changes, their syntactic behavior has become quite similar. German NIs, indeed, shifted toward the derivational pole in PDG.

The two nominalizations were compared to other inflectional and deriva-

tional processes, i.e. present participles and agent -er nouns. The results proved that, in terms of semantic transparency, NIs and PPs (the inflectional processes) act in similar way, whereas ERs are the less transparent derivatives. An interesting distinction was found between the cosine similarity measure and InvCL. As we expected, the asymmetrical InvCL measure is more appropriate to measure inclusion, whereas cosine seems a more generic semantic measure. This is shown by the results of the prediction between NIs and UNGs: the two nominalization patterns are similar in terms of cosine similarity, since they both share the event meaning; on the other hand, they present different degrees of inclusion into the base verb, as highlighted by the InvCL measure. This fine grained distinction of similarity measures is lost when more different morphological processes are compared: in task 2 and 4 the categories differ both in terms of inclusion and similarity, since they are markedly different.

No difference between the four groups was found in terms of addition of semantic content. The degree of context specificity seem the same for all the morphological processes investigated.

Chapter 6

Conclusions

In this work, I have investigated the competition between two classes of nominalization: event deverbal nouns and nominal infinitives. The main question I have been concerned with is how speakers choose one form instead of the other and which differences are perceived. I have presented three novel studies that address the phenomenon by means of quantitative methods (thus, differently from previous works) and provide new insights into this interesting topic, showing facts on nominalizations never disclosed before.

In a first study (chapter 3), we found that a common belief on the function of NIs is not true. NIs, indeed, are not used to make up for the lack of an EDN, contrary to what has been previously claimed (e.g. Simone, 2004). It follows that NIs and EDNs compete in most cases. Moreover, I presented a method to measure quantitatively the constraints on productivity of a morphological process. We found that the choice of nominalizations can be influenced by the transitivity of the base verb, a fact that has not been shown in the past literature. In addition, the results lead us to reconsider the nature of constraints, which are probably better described as tendencies or preferences, rather than as strict rules.

In chapter 4, I analyze a sample of cases in which both forms are attested. The study attempted to detect semantic differences between the two nominalizations analyzing their collocational preferences for arguments. This work assumed an underlying hypothesis, namely that examining the syntagmatic environments in which a word occurs reveals more about the properties of the word we deal with. By means of collocation analysis, we found that NIs and EDNs inherit different senses of the base verb, thus

107

disambiguating its vagueness of meaning. Values obtained by association measures (namely, by the log-likelihood ratio measure) showed that NIs are significantly more associated with abstract arguments, which make them express a metaphorical meaning of the base verb; EDNs, instead, rank among their top collocates arguments with concrete referents and express more literal senses. It could be argued that, despite the statistical association measure applied, these results are based on my personal intuition. For this reason, concreteness values taken from existing datasets have been used to check if this interpretation was correct. The results confirmed my hypothesis and show that it is reliable and strong. Further studies should investigate if vagueness disambiguation is more frequent with more polysemous base verbs. Probably, indeed, more polysemous verbs are more biased for the derivation of multiple nominalizations.

In chapter 5, I made use of distributional semantic models to compare the meaning of NIs and EDNs. In the past twenty years it has been shown that this methodology efficiently approximates the meaning of words into numerical vectors of their colllocations, and has led to big improvement in any task of natural language processing. However, outside the computational linguistics community, its appeal has not been recognized as frequently as it deserves. I have tried to show with this study how distributional semantics can be used to test more theoretical hypotheses and I hope that it will contribute to make this approach more popular among linguists.

Usually, research with DSMs deals with meanings of single lemmas; in this study, instead, I used vector space to inspect properties of the meaning of a whole morphological process. A similar point of view has been taken only recently by, for example, Lazaridou et al. (2013); Marelli and Baroni (2015); Kisselew et al. (2015); Padó et al. (2015). I believe that studies in this direction will improve both our NLP systems and our understanding of linguistic phenomena.

The study sheds light also on differences between similarity measures, specifically between cosine measure (the most common and used one) and the InvCL measure (Lenci and Benotto, 2012), which was tailored to identify relation between hypernyms and hyponyms. Cosine showed to be more suited to detect general semantic similarities, whereas InvCL describes what we have called transparency, i.e. the inclusion of the meaning of a word into another.

Considering these studies as a whole, the picture that emerges is a complex one. We know that both the types of base verb and the semantics of the nominalizations play a role in the selection process. EDNs, indeed, are preferred for transitive verbs and to express more literal senses, and the semantic shift they present from the base is larger. However, we have seen that all these facts are a matter of gradience and tendencies, rather than sharp distinctions. In future works, it could be possible to consider all these findings together and look at the interaction of the factors involved.

Moreover, additional features and hypotheses should be taken into account. In chapter 4, I referred to a possible difference in the aspectuality of nominalizations and suggested that it would be better investigated with formal acceptability experiments. Moreover, in the recent literature (e.g. Grimm and McNally, 2015; Heyvaert, 2008), it has been proposed that a difference in genericity may be observed between types of nominalization. Further efforts, however, are needed to investigate these points.

Bibliography

Adams, V. (1973). *An Introduction to Modern English Word-Formation.* London: Longman.

Alexiadou, A. (2001). *Functional structure in nominals: nominalization and ergativity*, Volume 42. John Benjamins Publishing.

Alexiadou, A. (2013). Nominal vs. Verbal -ing Constructions and the Development of the English Progressive. *English Linguistics Research 2*(2), 126–140.

Aronoff, M. and M. Lindsay (2014). Productivity, blocking and lexicalization. In S. Lieber and P. Štekauer (Eds.), *The Oxford Handbook of Derivational Morphology*, pp. 67–83. Oxford: Oxford University Press.

Baayen, H. (1993). On frequency, transparency and productivity. In G. Booij and J. van Marle (Eds.), *Yearbook of Morphology 1992*, pp. 181–208. Dordrecht: Springer Netherlands.

Baayen, R. H. (2009). Corpus linguistics in morphology: morphological productivity. In A. Lüdeling and M. Kyto (Eds.), *Corpus linguistics. An international handbook*, pp. 900–919. Berlin: De Gruyter Mouton.

Baayen, R. H., D. Davidson, and D. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language 59*(4), 390–412.

Baayen, R. H. and F. Moscoso del Prado Martìn (2005). Semantic density and past-tense formation in three Germanic languages. *Language 81*, 666–98.

Baayen, R. H. and A. Renouf (1996). Chronicling The Times: Productive lexical innovations in an English newspaper. *Language 72*, 69–96.

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy 9*(1), 5–16.

Barca, L., C. Burani, and L. S. Arduino (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, & Computers 34*(3), 424–434.

Baroni, M., S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, and M. Mazzoleni (2004). Introducing the La Repubblica corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian. In *Proceedings of LREC*, Volume 2, Lisbon, pp. 5–163. ELDA.

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation 43*(3), 209–226.

Baroni, M., G. Dinu, and G. Kruszewski (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, Baltimore.

Baroni, M. and A. Lenci (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Edinburgh, Scotland, pp. 1–10. Association for Computational Linguistics.

Bartsch, R. (1986). On aspectual properties of Dutch and German nominalizations. In V. Lo Cascio and C. Vet (Eds.), *Temporal structure in sentence and discourse*, pp. 7–39. Dordrecht: Foris.

Bauer, L. (2001). *Morphological Productivity*. Cambridge: Cambridge University Press.

Bauer, L. (2005). Productivity: theories. In P. Štekauer and R. Lieber (Eds.), *Handbook of word-formation*, pp. 315–334. Dordrecht: Springer.

Bauer, L., R. Lieber, and I. Plag (2013). *The Oxford reference guide to English morphology*. Oxford University Press.

Behaghel, O. (1923). *Deutsche Syntax*, Volume 1. Heidelberg: Winter.

Bengio, Y., R. Ducharme, P. Vincent, and C. Janvin (2003). A neural probabilistic language model. *Journal of Machine Learning Research 3*, 1137–1155.

Bertinetto, P. M. (2001a). On a frequent misunderstanding in the temporal-aspectual domain: the perfective-telic confusion. In C. Cecchetto, G. Chierchia, and M. T. Guasti (Eds.), *Semantic Interfaces: Reference, Anaphora and Aspect*, pp. 177–210. Stanford University: CSLI Publications.

Bertinetto, P. M. (2001b). Sulle proprieta tempo-aspettuali dell'infinito in italiano. In *Atti del 35 Congresso Internazionale della Societa di Linguistica Italiana*.

Bolinger, D. L. (1968). Entailment and the Meaning of Structures. *Glossa 2*, 119–127.

Booij, G. (2000). Inflection and derivation. In G. Booij, C. Lehmann, and J. Mugdan (Eds.), *Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung. An International Handbook on Inflection and Word-Formation*, Volume 1, pp. 360–369. Berlin, Boston: De Gruyter Mouton.

Booij, G. and R. Lieber (2004). On the paradigmatic nature of affixal semantics in English and Dutch. *Linguistics 42*(2), 327–357.

Brinton, L. J. (1995). The aktionsart of deverbal nouns in English. In *Temporal Reference, Aspect and Actionality*, Volume 1, pp. 27–42.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*, Volume 9. John Benjamins Publishing.

Bybee, J. L. (1995). Diachronic and typological properties of morphology and their implications for representation. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing*, pp. 225–246. Hillsdale, NJ: Erlbaum.

Bybee, J. L. (2001). *Phonology and language use.* Cambridge: Cambridge University Press.

Chomsky, N. (1970). Remarks on Nominalization. In R. Jacobs and P. Rosenbaum (Eds.), *Readings in English Transformational Grammar*, pp. 184–272. Waltham, Mass: Ginn and co.

Church, K. and P. Hanks (1990). Word association norms, mutual information, and lexicog- raphy. *Computational Linguistics 16*(1), 22–29.

Clarke, D. (2009). Context-theoretic semantics for natural language: an overview. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, Athens, pp. 112–119. ACL.

Comrie, B. (1976). The syntax of action nominals: A cross-language study. *Lingua 40*(2), 177–201.

Comrie, B. and S. A. Thompson (2007). Lexical nominalization. In *Language Typology and Syntactic Description vol. III: Grammatical Categories and the Lexicon*, pp. 334–381. Cambridge University Press.

Corbin, D. (1987). *Morphologie dérivationnelle et structuration du lexique*. Tübingen: Niemeyer.

Cristofaro, S. (2007). Deconstructing categories: Finiteness in a functional-typological perspective. In *Finiteness: theoretical and empirical foundations*, Volume 1, pp. 91–114.

Della Rosa, P., E. Catricalà, G. Vigliocco, and S. F. Cappa (2010). Beyond the abstract–concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods 42*(4), 1042–1048.

Demske, U. (2002). Nominalization and argument structure in Early New High German. In E. Lang and I. Zimmermann (Eds.), *Nominalisations*, Number 27 in ZAS Papers in Linguistics. Berlin: ZAS.

Don, J., M. Trommelen, and W. Zonneveld (2000). Conversion and category indeterminacy. In G. Booij, C. Lehmann, and J. Mugdan (Eds.), *Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung. An International Handbook on Inflection and Word-Formation*, Volume 1, pp. 943–952. Berlin, Boston: De Gruyter Mouton.

Dressler, W. U. (2000). Subtraction. In G. Booij, C. Lehmann, and J. Mugdan (Eds.), *Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung. An International Handbook on Inflection and Word-Formation*, Volume 1, pp. 581–587. Berlin, Boston: Mouton de Gruyter.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics 19*(1), 61–74.

Eberle, K., G. Faass s, and U. Heid (2009). Corpus-based identification and disambiguation of reading indicators for German nominalizations. In *Proceedings of Corpus Linguistics.*

Ehrich, V. and I. Rapp (2000). Sortale Bedeutung und Argumentstruktur: ung-Nominalisierungen im Deutschen. *Zeitschrift für Sprachwissenschaft 19*(2), 245–303.

Eisenberg, P. (1994). German. In E. König and J. Van Der Auwera (Eds.), *The Germanic Languages*, pp. 349–387. New York: Routledge.

Esau, H. (1973). *Nominalization and Complementation in Modern German.* Amsterdam: North Holland.

Evert, S. (2005). *The statistics of word cooccurrences : word pairs and collocations.* Ph. D. thesis, University of Stuttgart.

Evert, S. (2009). Corpora and collocations. In A. Ludeling and M. Kyto (Eds.), *Corpus linguistics. An international handbook*, pp. 1212–1248. Berlin: De Gruyter Mouton.

Faaß, G. and K. Ekart (2013). SdeWaC - a corpus of parsable sentences from the web. In *Language processing and knowledge in the Web*, pp. 61–68. Berlin - Heidelberg: Springer.

Farrell, P. (2001). Functional shift as category underspecification. *English Language and Linguistics 5*(1), 109–130.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930-55. In *Studies in Linguistic Analysis*, pp. 1–32. Oxford: The Philological Society.

Gaeta, L. (1998). The Inflection vs. Derivation Dichotomy: the Case of German Infinitives. In B. Caron (Ed.), *Proceedings of the XVI International Congress of Linguists*, Oxford, pp. Paper No. 0199. Pergamon.

Gaeta, L. (2002). *Quando i verbi compaiono come nomi. Un saggio di Morfologia Naturale.* Milano, Franco Angeli.

Gaeta, L. (2004). Derivazione nominale deverbale. In M. Grossmann and F. Rainer (Eds.), *La formazione delle parole in italiano.* Tübingen: Max Niemeyer Verlag.

Gaeta, L. (2009). A metà tra nomi e verbi: i nomi d'azione tra morfologia, sintassi e semantica. In E. Lombardi Vallauri and L. Mereu (Eds.), *Spazi linguistici. Studi in onore di Raffaele Simone*, pp. 111–123. Roma: Bulzoni.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Göransson, C. E. (1911). *Die doppelpropositionalen Infinitive im Deutschen.* Göteborg: Wald. Zachrissons Boktrykeri.

Greenberg, J. (1963). Some Universals of Grammar, with Particular Reference to the Order of Meaningful Elements. In J. Greenberg (Ed.), *Universals of Language*, pp. 73–113. Cambridge, MA: MIT Press.

Grevisse, M. (1969). *Le bon usage: grammaire française avez des remarques sur la langue française d'aujourd'hui* (14 ed.). Bruxelles: De Boeck & Larcier.

Gries, S. T., B. Hampe, and D. Schönefeld (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics 16*(4), 635–676.

Gries, S. T. and A. Stefanowitsch (2004). Extending collostructional analysis: A corpus-based perspective on alternations. *International journal of corpus linguistics 9*(1), 97–129.

Grimm, S. and L. McNally (2013). No ordered arguments needed for nouns. In *Proceedings of the 19th Amsterdam colloquium*, pp. 123–130.

Grimm, S. and L. McNally (2015). The-ing dynasty: Rebuilding the semantics of nominalizations. In *Semantics and Linguistic Theory*, Volume 25, pp. 82–102.

Grimshaw, J. (1990). *Argument structure.* Cambridge, MA: MIT Press.

Hamm, F. and H. Kamp (2009). Ontology and inference: The case of German ung–nominals. *Disambiguation and Reambiguation 6*, 1–67.

Hamp, B. and H. Feldweg (1997). GermaNet - a Lexical-Semantic Net for Germa. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP applications*, Madrid.

Harris, Z. (1954). Distributional structure. *Word 10*, 146–62.

Hartmann, S. (2014). The diachronic change of German nominalization patterns: An increase in prototypicality. In *Selected Papers from the 4th UK Cognitive Linguistics Conference*, Lancaster, pp. 52–171. Cognitive Linguistics Association.

Haspelmath, M. (1993). *A Grammar of Lezgian.* Berlin: Mouton de Gruyter.

Haspelmath, M. (1996). Word-class changing inflection and morphological theory. In *Yearbook of Morphology 1995*, pp. 43–66. Springer.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics 39*(4), 1041–1070.

Hay, J. (2003). *Causes and consequences of word structure.* New York & London: Routledge.

Henrich, V. and E. Hinrichs (2010). GernEdiT - The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, pp. 2228–2235.

Heyvaert, L. (2003). *A cognitive-functional approach to nominalization in English*, Volume 26. Walter de Gruyter GmbH & Co KG.

Heyvaert, L. (2008). On the constructional semantics of gerundive nominalizations. *Folia Linguistica 42*(1), 39–82.

Ježek, E. (2003). *Classi di verbi tra semantica e sintassi.* ETS.

Ježek, E. (2007). Polysemy of italian event ominals. *Faits de Langues, Numéro special "Nominalisations" 30*, 251–264.

Jurafsky, D. and J. H. Martin (2016). Speech and Language Processing.

Kelling, C. (2001). Agentivity and suffix selection. In M. Butt and T. Holloway (Eds.), *Proceedings of the LFG01 Conference*, Stanford, pp. 147–162. CSLI Publications.

Kiparsky, O. and C. Kiparsky (1970). Fact. In M. Bierwisch and K. Heidolph (Eds.), *Progress in Linguistics*, pp. 43–73. The Hague: Mouton.

Kisselew, M., S. Padó, A. Palmer, and J. Šnajder (2015). Obtaining a Better Understanding of Distributional Models of German Derivational Morphology. In *Proceedings of IWCS*, London, UK.

Kisselew, M., L. Rimmel, A. Palmer, and S. Padó (2016). Predicting the Direction of Derivation in English Conversion. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Berlin, pp. 93–98. Association for Computational Linguistics.

Knobloch, C. (2003). Zwischen Satz-Nominalisierung und Nennderivation: -ung-Nomina im Deutschen. *Sprachwissenschaft 27*(3), 333–362.

Koontz-Garboden, A. (2007). *States, changes of state, and the Monotonicity Hypothesis.* PhD Thesis, Stanford University.

Koptjevskaja-Tamm, M. (1993). *Nominalizations.* London: Routledge.

Koptjevskaja-Tamm, M. (1999). Finiteness.

Koptjevskaja-Tamm, M. (2003). Action nominal constructions in the languages of Europe. In F. Plank (Ed.), *Noun phrase structure in the languages of europe*, pp. 723–759. Berlin: Mouton de Gruyter.

Koptjevskaja-Tamm, M. (2006). Nominalizations.

Koptjevskaja-Tamm, M. (2013). Action Nominal Constructions. In M. S. Dryer and M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Kountz, M., U. Heid, and K. Spranger (2007). Automatic sortal Interpretation of German Nominalisations with-ung Towards using underspecified Representations in Corpora. In *Proceedings of Corpus Linguistics*.

Laca, B. (2001). Derivation. In M. Haspelmath, E. König, W. Oesterreicher, and W. Raible (Eds.), *Language Typology and Language Universals*, Volume 2, pp. 1214–1227. Berlin: Walter de Gruyter.

Landauer, T. and S. Dumais (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Lazaridou, A., M. Marelli, R. Zamparelli, and M. Baroni (2013). Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics. In *ACL (1)*, pp. 1517–1526. Citeseer.

Lebeaux, D. (1986). The interpretation of derived nominals. *Chicago Linguistic Society 22*, 231–247.

Lees, R. (1960). *The grammar of English nominalizations*. The Hague: Mouton.

Lenci, A. and G. Benotto (2012). Identifying hypernyms in distributional semantic spaces. pp. 75–79. Association for Computational Linguistics.

Lieber, R. (2000). Substitution of segments and features. In G. Booij, C. Lehmann, and J. Mugdan (Eds.), *Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung. An International Handbook on Inflection and Word-Formation*, Volume 1, pp. 567–576. Berlin, Boston: Mouton de Gruyter.

Lieber, R. (2004). *Morphology and Lexical Semantics*. Number 104 in Cambridge Studies in Linguistics. Cambridge University Press.

Lieber, R. (2016). *English Nouns: the Ecology of Nominalization*. Cambridge: Cambridge University Press.

Link, G. (1983). The logical analysis of plurals and mass terms: a lattice-theoretical approach. In R. Bauerle, C. Schwarze, and A. von Stechow (Eds.), *Meaning, Use, and the Interpretation of Lan- guage*, pp. 302–323. Berlin - New York: De Gruyter.

Lund, C. B. and Kevin (1997). Modelling parsing constraints with high-dimensional context space. *Language and cognitive processes 12*(2-3), 177–210.

Lyons, J. (1977). *Semantics.* Cambridge: Cambridge University Press.

Marelli, M. and M. Baroni (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review 122*(3), 485–515.

Martin, F. (2010). The semantics of eventive suffixes in French. In M. Rathert and A. Alexiadou (Eds.), *The semantics of nominalizations across languages and frameworks*, Number 22 in Interface explorations, pp. 109–139. Berlin/New York: Mouton de Gruyter.

McDonald, S. and M. Ramscar (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, University of Edinburgh, pp. 611–616. Erlbaum.

McQueen, J. M. and A. Cutler (1998). Morphology in word recognition. In A. Spencer and A. Zwicky (Eds.), *The handbook of morphology*, pp. 406–427. Oxford: Blackwell.

Melloni, C. (2006). Logical polysemy in word formation: E and R suffixes. *Lingue e linguaggio 2*, 281–308.

Melloni, C. (2007). *Polysemy in word formation: the case of deverbal nominals.* Ph. D. thesis, University of Verona.

Melloni, C. (2008). Per una tassonomia dei nominali "risultato". In E. Cresti (Ed.), *Prospettive nello studio del lessico italiano. Atti del IX congresso SILFI*, pp. 573–581.

Miller, G. and W. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*(1), 1–28.

Naumann, B. and P. M. Vogel (2000). Derivation. In G. Booij, C. Lehmann, and J. Mugdan (Eds.), *Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung. An International Handbook on Inflection and Word-Formation*, Volume 1, pp. 929–943. Berlin, Boston: De Gruyter Mouton.

Noonan, M. (2007). Complementation. In T. Shopen (Ed.), *Language Typology and Syntactic Description*, Volume II: Complex Constructions, pp. 52–150. Cambridge: Cambridge University Press.

Padó, S. and M. Lapata (2003). Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Volume 128-135.

Padó, S., A. Palmer, M. Kisselew, and J. Šnajder (2015). Measuring Semantic Content To Assess Asymmetry in Derivation. In *Proceedings of IWCS Workshop on Advances in Distributional Semantics*, London, UK.

Palmer, F. R. (2001). *Mood and Modality* (2nd ed.). Cambridge: Cambridge University Press.

Pedersen, T. (1996). Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference (SCSUG-96)*.

Plag, I. (2003). *Word-formation in English*. Cambridge University Press.

Plag, I. (2006). Productivity. In B. Aarts and A. McMahon (Eds.), *The handbook of English linguistics*, pp. 537–556. Blackwell.

Plank, F. (1994). Inflection and Derivation.

Pullum, G. K. and A. Zwicky (1999). Gerund participles and head-complement inflection conditions. In P. Collins and D. Lee (Eds.), *The Clause in English: In Honour of Rodney Huddleston*, pp. 251–271. Amsterdam: John Benjamins.

Quirk, R., S. Greenbaum, G. Leech, J. Svartvik, and D. Crystal (1985). *A comprehensive grammar of the English language*. Longman.

Rainer, F. (2005). Constraints on productivity. In P. Štekauer and R. Lieber (Eds.), *Handbook of word-formation*, pp. 335–352. Dordrecht: Springer.

Ramat, P. (2002). La natura dell'infinito. In *L'infinto e oltre. Omaggio a Gunver Skytte*. Odense.

Rossdeutscher and H. Kamp (2010). Syntactic and semantic constraints in the formation and interpretation of ung-nouns. In *The semantics of nominalizations across languages and frameworks*, pp. 169–214.

Roy, I. and E. Soare (2011). Nominalizations: new insights and theoretical implications. *Recherches linguistiques de Vincennes 40*(1), 7–23.

Sagi, E., S. Kaufmann, and B. Clark (2009). Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pp. 104–111. Association for Computational Linguistics.

Scalise, S. (1988). Inflection and derivation. *Linguistics 26*, 561–581.

Scheffler, T. (2005). Nominalization in German.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of NeMLaP*, Manchester, UK.

Shin, S.-S. (2001). On the event structure of -ung-nominals in German. *Linguistics 39*, 297–319.

Simone, R. (2004). L'infinito nominale nel discorso. In P. D'Achille (Ed.), *Generi, architetture e forme testuali: atti*, pp. 73–96. Firenze: Cesati.

Skytte, G. (1983). *La sintassi dell'infinito in italiano moderno*. Museum Tusculanum Press.

Skytte, G., G. Salvi, and M. Manzini (2001). Frasi subordinate all'infinito. In L. Renzi, G. Salvi, and A. Cardinaletti (Eds.), *Grande grammatica italiana di consultazione*, Volume II: I sintagmi verbale, aggettivale, avverbiale. La subordinazione, pp. 483–569. Bologna: Il Mulino.

Spranger, K. and U. Heid (2007). Applying Constraints derived from the Context in the process of Incremental Sortal Specification of German ung-Nominalizations. *Constraints and Language Processing*, 65.

Stefanowitsch, A. and S. T. Gries (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics 8*(2), 209–243.

Stemberger, J. P. and B. MacWhinney (1986). Frequency and the Lexical Storage of Regularly Inflected Forms. *Memory & Cognition 14*(1), 17–26.

Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Stump, G. T. (2001). *Inflectional morphology: A theory of paradigm structure.* Cambridge: Cambridge University Press.

Team, R. C. (2015). R: A language and environment for statistical computing.

Thornton, A. M. (1990). Sui deverbali italiani in-mento e-zione (I). *Archivio glottologico italiano 75*(2), 169–207.

Thornton, A. M. (1991). Sui deverbali italiani in-mento e-zione (II). *Archivio Glottologico Italiano 76*(1), 79–102.

Vanvolsem, S. (1983). *L'infinito sostantivato in italiano.* Firenze: L'Accademia della Crusca.

Vendler, Z. (1957). Verbs and times. *The philosophical review 66*(2), 143–160.

Vendler, Z. (1967). Facts and events. In *Linguistics in philosophy*, pp. 122–146.

Vennemann, T. (1972). Phonetic analogy and conceptual analogy. In *Schuchardt, the neogrammarians, and the transformational theory of phonological change*, pp. 181–204. Frankfurt: Athenaeum.

Walinska de Hackbeil, H. (1984). On two types of derived nominals. In D. Testen, V. Mishra, and J. Drogo (Eds.), *Papers from the parasession on Lexical Semantics*, Chicago, pp. 308–332. Chicago Linguistic Society.

Weeds, J. and D. Weir (2003). A general framework for distributional similarity. In *Proceedings of the EMNLP 2003*, Sapporo, Japan, pp. 81–88.

Weeds, J., D. Weir, and D. McCarthy (2004). Characterising measures of lexical distributional similarity. In *Proceedings of COLING 2004*, Geneva, Switzerland, pp. 1015–1021.

Werner, M. (2013). Nominalizing infinitives in the history of German: state-of-the-art & open questions. *Bavarian Working Papers in Linguistics 3*, 127–134.

Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford: Blackwell.

Ylikoski, J. (2003). Defining non-finites: action nominals, converbs and infinitives. *SKY Journal of linguistics 16*(1), 185–237.

Zeller, B., J. Šnajder, and S. Padó (2013). DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of ACL*, Sofia, Bulgaria. Association for Computational Linguistics.

Zinsmeister, H. and U. Heid (2004). Collocations of complex nouns: Evidence for lexicalisation. In *Proceedings of Konvens*, Vienna, Austria.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Zubizarreta, M. (1987). *Levels of Representations in the Lexicon and in the Syntax*. Dordrecht: Foris.

Zucchi, A. (1993). *The Language of Propositions and Events. Issues in the Syntax and the Semantics of Nominalization*. Number 51 in Studies in Linguistics and Philosophy. Netherlands: Springer.