UNIVERSITÀ
DEGLI STUDI
FIRENZE

Università degli Studi di Firenze
Dipartimento di Lettere e Filosofia (DILEF)
Corso di Dottorato in Filologia,
Letteratura Italiana, Linguistica
Curriculum: Linguistica

Annotation of linguistically derived action concepts in computer vision datasets

Candidate

Andrea Amelio Ravelli

Supervisors

Massimo Moneglia

Lorenzo Seidenari

ciclo XXXII, 2016-2019

Università degli Studi di Firenze, Dipartimento di Lettere e Filosofia (DILEF).

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Filologia, Letteratura Italiana, Linguistica.

To Gisella

Nobody exists on purpose.
Nobody belongs anywhere.
We're all going to die.
Come watch TV.
(Morty Smith)

Acknowledgments

I would like to thank everyone who have supported me, encouraged me and believed in me. There is no need of an explicit list, because all of them know for what I'm grateful to each one. But the bigger *thank you* goes to those who didn't, because they forced me to enhance my self-confidence and skills, and to strengthen my ideas.

Abstract

In the present work, an in-depth exploration of IMAGACT has been traced, with the focus of exploiting the resource in NLP tasks. Starting from the Introduction, the idea of making use of IMAGACT multimodal action conceptualisation has been drawn, with some reflections on evidences of the deep linking between Language and Vision, and on the fact that action plays a key role in this linkage. Thus, the multimodal and multilingual features of IMAGACT Ontology of Action Verbs have been described, with also some details on the framework of the resource building. It followed a concrete case-study on IMAGACT internal data, that led to the proposal of an inter-linguistic manual mapping between the Action Types of verbs which refer to *cutting* eventualities in English and Italian. Then, a series of experiments have been presented, involving the exploitation of IMAGACT in linking with other resources and building deliverable NLP products (such as the Ref-vectors of action verbs). One of the experiments has been described extensively: the visual enrichment of IMAGACT through instance population of its action concepts, making use of Audio Description of movies for visually impaired people. From this last experiment it emerged that dealing with *non-conventional* scenarios, such as the one of assessing action reference similarity between texts from different domains, is particularly challenging, given that fine-grained differences among action concepts are difficult to derive purely from the textual representation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter gives a brief introduction to the Thesis and presents its initial objectives. Then, the Index of the Thesis is explained and a brief description is provided for each Chapter composing this dissertation.

When we need to explain something to someone who does not speak our language, we automatically look around for examples to provide an ostensive definition (Wittgenstein, 1958). In fact, human language is ambiguous by nature and a visual input is more explanatory than any gloss, allowing to overcome communication barriers. This intuition is at the basis of the IMAGACT Multilingual and Multimodal Ontology of Action, which exploits video prototypes as action concepts to whom verbs in different languages are linked, on the basis of their primary pragmatic extension of reference.

One evidence of the deep connection between vision and language in the human brain is the so-called McGurk effect (McGurk and MacDonald, 1976). It is a basic perceptual phenomenon, which might also be considered trivial, but it demonstrates a strong interaction in speech processing of both the language and vision systems. The illusion occurs when the auditory component of one sound is paired with the visual component of another sound, leading to the perception of a third sound.[1] It demonstrates that the visual input is an important contribution in the processing of acoustic information, to the point that it can interfere in case of visual/acoustic mismatch.

Human language and vision systems are deeply linked together due to evolutionary basis. Arbib and Rizzolatti (1997) assert that any plausible account of language evolution must root language in forms of behaviour that are *pragmatic* rather than *communicative* but must also offer a bridge from *doing* to *communicating about doing*. In other words, language is a form of action, and it is probably starting from action *capabilities* that language emerged during the course of the human evolution. In this respect, the *Mirror System Hypothesis* (Arbib and Rizzolatti, 1997) states that the mechanism which supports language in the human brain evolved atop the mirror system for grasping, taking advance of its ability to recognize a set of actions and adapting it to deal with linguistic acts (i.e. utterances) and to discriminate linguistic objects (i.e. audio patterns for words). Since their first discovery in monkeys (Di Pellegrino et al., 1992), and the identification of homologous ones in the Broca's area in the human brain (Fadiga et al., 1995), mirror neurons have been object of study for a vast literature in neuro-linguistics.[2] The most emerging fact is that action is the linkage point between vision and language, i.e. humans developed language by

[1] As a demonstration, the video in the following URL shows a man performing the syllable *ga*, dubbed with the audio of the syllable *ba*, but the perceptual effect is to hear the sound *da*. (Many people are affected differently by the McGurk effect based on many factors, including brain damage and other disorders.) `https://www.youtube.com/watch?v=aFPtc8BVdJk`

[2] In the recent years, plenty of researches on mirror neurons have been published, to the point that mirror neurons are the most hyped concept in neuroscience (Jarrett, 2012).

adapting the pattern recognition system, used by the vision module of the brain to recognise actions, to identify and imitate audio patterns, and to link them with real world entities (i.e. objects and events) and their mental representations.

For these reasons, the hypothesis of investigating action reference into language must take into account a more extended view, that comprises also vision, in order to deliver a complete multimodal representation. The IMAGACT Multimodal and Multilingual Ontology of Action is a good example of this kind of framework, and its categorisation of *eventualities* in language may be of great help in studies on actions that are other than Linguistics, such as Computer Vision.

The objective of this thesis is to explore the possibility of exploiting the fine-grained action categorisation offered by IMAGACT in the annotation of multimodal data, such in dataset for Computer Vision. In this regard, a first attempt in annotating IMAGACT action concepts, which are linguistically derived from spoken corpora, in a Computer Vision datased made of video-caption pairs derived from Audio Description of movies for visually impaired people, namely LSMDC. The possible outcome is twofold: on one side, IMAGACT Ontology would be populated of multimodal instances of actions in the form of video-caption pairs; on the other, LSMDC would be enriched with a fine-grained action discrimination derived from a handcrafted linguistic resource.

The Thesis is organised as follows. Chapter 2 provides a comprehensive description of IMAGACT Ontology of Action verbs. First, the theoretical framework at the basis of the resource is explained; then, the building strategies of its multilingual network are described. Moreover, some experiments implementing the resource are presented.

Chapter 3 illustrates a proposal for the inter-linguistic mapping in IMAGACT between Action Types of the verbs *tagliare* (it), *cut* (en), and their equivalent verbs in the two languages. This example is useful to show the methodology behind IMAGACT in a concrete application: through the analysis of the standardisations of each one, Action Types have been aligned and linked together, and all the local equivalences have been verified.

Chapter 4 reports some experiments, conducted during the three-years PhD program spent in the LABLITA Lab, in exploiting IMAGACT for NLP. In particular, my colleagues and I have conducted two linking experiments with other resources (namely, BabelNet and T-PAS) in order to extend the information conveyed by each action concept; moreover, we built an action concepts vector space model by exploiting IMAGACT multilingual feature, and action verbs embeddings based on co-reference vectors.

Chapter 5 introduces and describes the work and research produced as focus of PhD program, that is the visual enrichment of IMAGACT through Audio Description for visually impaired people and the annotation of IMAGACT action concepts in LSMDC dataset, which is a dataset built for Computer Vision. Starting from the analysis and processing of the target dataset (LSMDC), the experiments to annotate the correct IMAGACT action concepts are described, and an analysis of the results is proposed.

Chapter 6 summarises the contribution of the Thesis and discusses which would be possible future outcomes.

Chapter 2

IMAGACT Ontology and the categorization of action verbs

This chapter provides a comprehensive description of IMAGACT Ontology of Action verbs. First, the theoretical framework at the basis of the resource is explained; then, the building strategies of its multilingual network are described. Moreover, some experiments implementing the resource are presented.

The IMAGACT project (Multilingual Visual Ontology of actions for L2 acquisition) had as primary objective to make explicit the relations between action concepts and their concrete representation in natural language systems. The action concept is represented in the form of a short video or 3D rendering, i.e. prototypical scenes depicting the action, to which verbs in different languages are connected, on the basis of pragmatic and semantic competence of native speakers' judgements. The two-year project, which ended in September 2013, was made possible thanks to funding from the Tuscany Region, under the PAR/FAS 2007-2013 program, and it involved three research units: the Linguistics Laboratory (LABLITA) of the University of Florence, the Department of Communication Sciences of the University of Siena and the Computational Linguistics Institute “Antonio Zampolli” (ILC-CNR) of Pisa.

2.1 The IMAGACT web interface

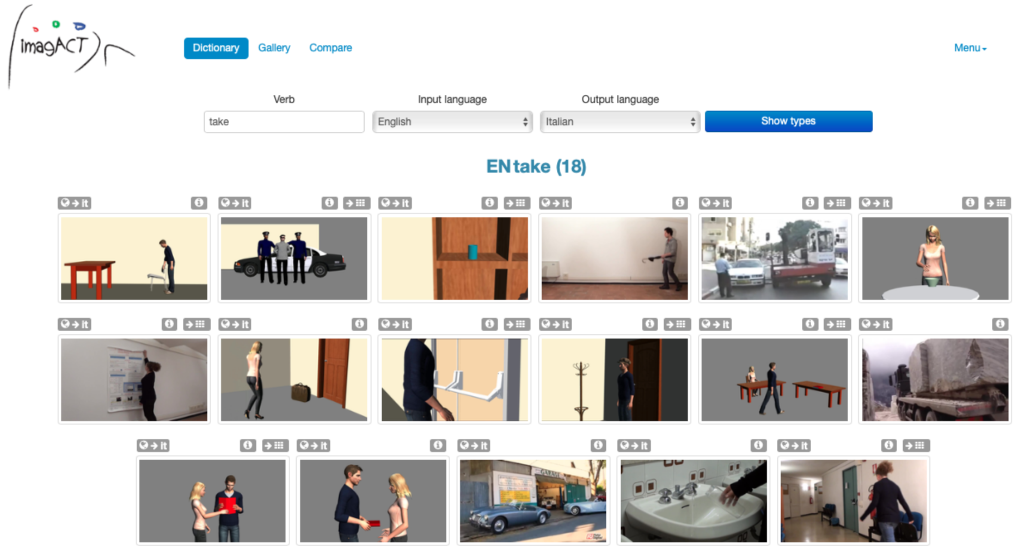
The IMAGACT resource is available on the web,[1] and it can be explored in multiple ways by L2 learners. Indeed, it enables the visitor to conduct analyses at various levels:

- intra-language analyses, by showing all the possible action concepts predicated by one verb;
- inter-language analyses, by offering a translation between lemmas in different languages linked to the same action concept;
- comparisons between two verbs, by showing shared and unshared action concepts both in the same language or different ones;
- heuristic researches, by directly exploring the action concepts and then obtaining verbs in all languages which describe the eventualities depicted in the videos.

Starting from the homepage, it is possible to navigate through the resources by clicking on one of the 3 buttons on the top-left of the page: *Dictionary*, *Gallery*, *Compare.*

[1] www.imagact.it

2.1.1 IMAGACT as Dictionary



Verb: take

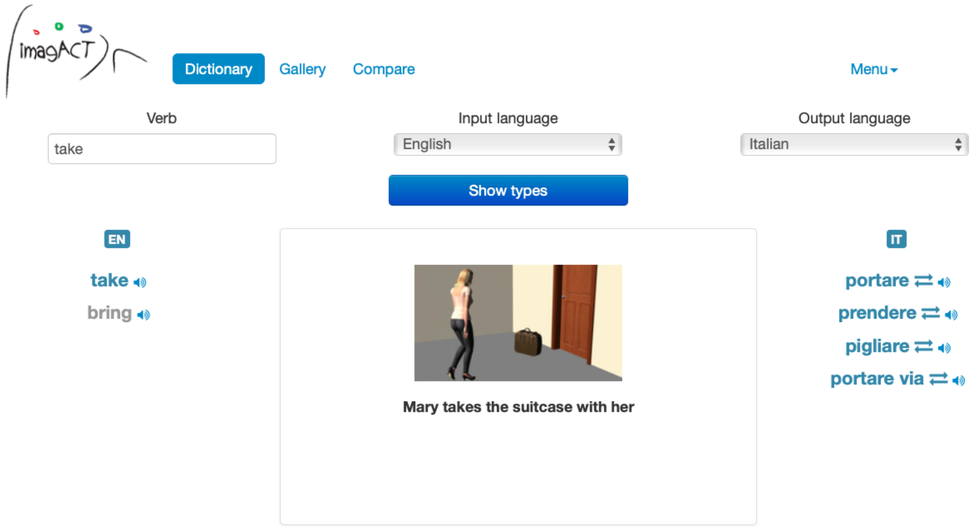
Input language: English

Output language: Italian

Figure 2.1: An example of query in IMAGACT Dictionary mode: Action Types of the verb *take*.

Dictionary is the *default* mode of navigating the IMAGACT resource (Figure 2.1). By selecting the source language, it is possible to perform a lemma lookup, similarly to a common dictionary. The result of the query is the collection of all the Action Types the verb extends to, which are the class of actions that specific verb predicates.

On the top of each video, from left to right, there are three buttons: translation button (the one with the target language abbreviation), information button (the one with the “i”), the *family group* button (the one with a small grid). The first one redirects to the translation page in the desired target language selected from the *Output language* drop-down menu in the top of the page; the second one makes the caption appear in the bottom of the video thumbnail; the third one opens a pop-up window containing all the action concepts gathered in the same Action Type, if it comprises more than one atomic action concept.[2]



Verb: take

Input language: English

Output language: Italian

take
bring

Mary takes the suitcase with her

portare
prendere
pigliare
portare via

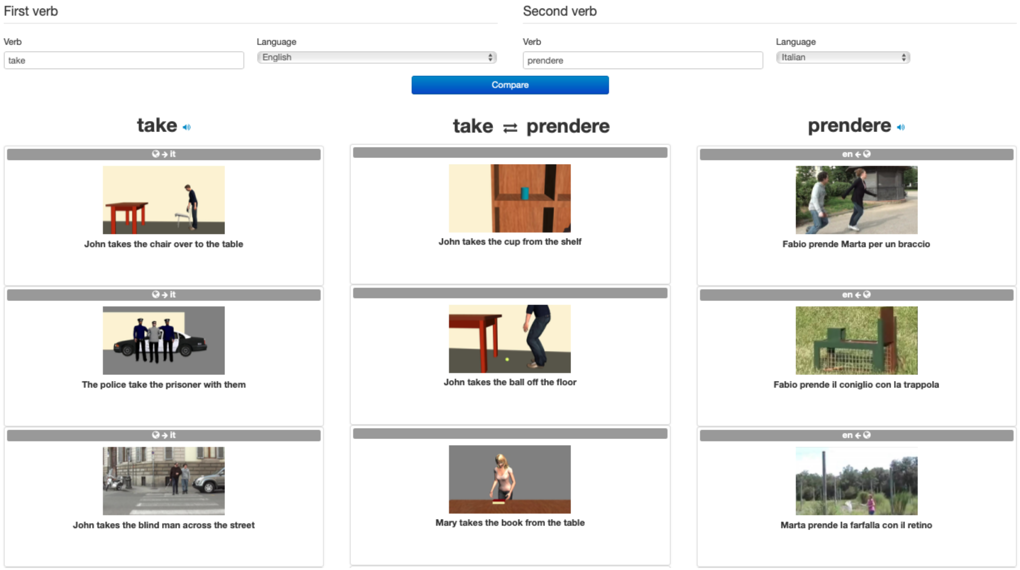
Figure 2.2: An example of translation offered by IMAGACT: one Action Type of *take* translated to Italian.

The translation page is organised in three columns (Figure 2.2). In the centre it is placed the video prototype of the action concept and its caption with the starting verb; on the left, the starting verb followed by synonyms in the source language, i.e. verbs that equally extend their reference to the selected action concept; on the right, all the verbs in the target language that can be used to predicate the action concept. From here, it is possible

[2] This is due to the fact that some general verbs extend to action concepts within the same Action Type without taking into account distinctive differences. As an example, consider the Action Type of the verb *attach* that refers to applying something onto something else: it is possible to attach a sign on the wall by sticking it with some glue, or by using nails or tape. From the point of view of *attach*, all the three actions are the same, because the verb focuses on the result, while the verbs *glue*, *nail* and *tape*, which encode a tool, can predicate only one of the concepts each.

to access directly to the comparison between the starting verb and one of its translation verbs by clicking on the double-arrow button on the right of each verb in the target language.

2.1.2 IMAGACT as Comparison



First verb

Verb: take

Language: English

Second verb

Verb: prendere

Language: Italian

Compare

take

John takes the chair over to the table

The police take the prisoner with them

John takes the blind man across the street

take ⇄ prendere

John takes the cup from the shelf

John takes the ball off the floor

Mary takes the book from the table

prendere

Fabio prende Marta per un braccio

Fabio prende il coniglio con la trappola

Marta prende la farfalla con il retino

Figure 2.3: An example of query in IMAGACT Compare mode: action concepts exclusively predicated by *take* on the left, by *prendere* on the right, and by both verbs in the centre. The figure shows only part of the action concepts connected to the two verbs.

The *Compare* button in the starting page of the IMAGACT website, or the double-arrow button in the translation page, enables the visitor to perform comparison between two verb lemmas, both in the same language and in different languages, on the basis of the action concepts they extend their reference to. The visitor has to type the desired lemmas in the *First verb* and *Second verb* fields in the top of the page, select the respective language of each verb and click on the *Compare* button. As shown in Figure 2.3, two verbs commonly considered as equivalent (*take* (EN) and *prendere* (IT)) can be analysed in depth and in a contrastive way, highlighting commonalities (action concepts in the central column of the compare view) and distinctive references (action concepts in the left and right column refer uniquely to one of the compared verbs). In the specific example, and considering only the action concepts shown in the Figure, it is possible to discover that *take* and *prendere* are equivalent when referring to action concepts pertaining to the semantic space of *gaining possession of an object by grasping*, while *bringing* actions are exclusively predicated by *take* and *catching* ones only by *prendere*.

2.1.3 IMAGACT as Gallery

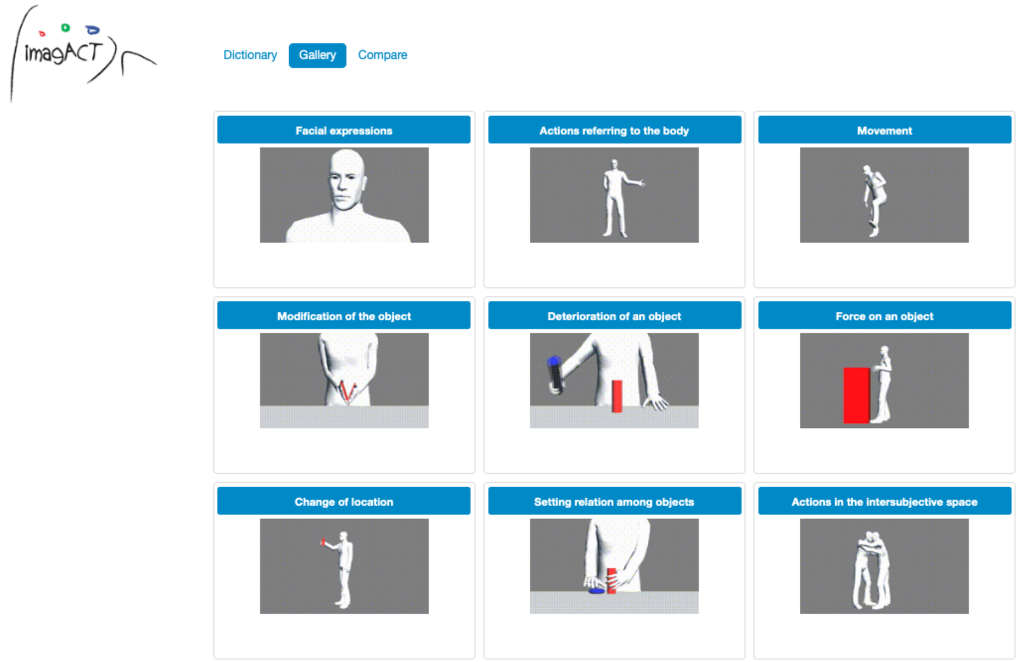

imagACT
Dictionary
Gallery
Compare
Facial expressions
Actions referring to the body
Movement
Modification of the object
Deterioration of an object
Force on an object
Change of location
Setting relation among objects
Actions in the intersubjective space

Figure 2.4: Gallery mode in IMAGACT: meta-categories from where to start exploring the resource in visual-only mode.

In *Gallery mode*, the visitor lands to a page where 9 meta-categories of actions are proposed, which stand for a general classification of all the action concepts comprised in the IMAGACT Ontology.

Figure 2.4 shows the starting page of the Gallery mode, with the 9 meta-categories:

- facial expressions;
- actions referring to the body;
- movement;
- modification of the object;
- deterioration of an object;
- force on an object;
- change of location;
- setting relation among objects;
- actions in the intersubjective space.

Each meta-category is represented by a looped animation which roughly shows various kinds of actions gathered in it. They work similarly to top-ontologies' categories: higher level abstract entities which pull together lower level concrete entities (i.e. action concepts) with common features and peculiarities. Exploring IMAGACT in Gallery mode enables the visitor to identify the desired action concept by exploiting the unambiguousness of the pure visual reference, without any language-specific constrain. Thus, this mode particularly suits visitors whose language is still not implemented in the resources, or those who do not have the possibility to use one of the already implemented languages as a vehicular language.

By entering in one of the meta-categories (Figure 2.5), the collection of the comprised action concepts is shown, and the visitor can find the desired one by watching the videos. By clicking on the translation button on

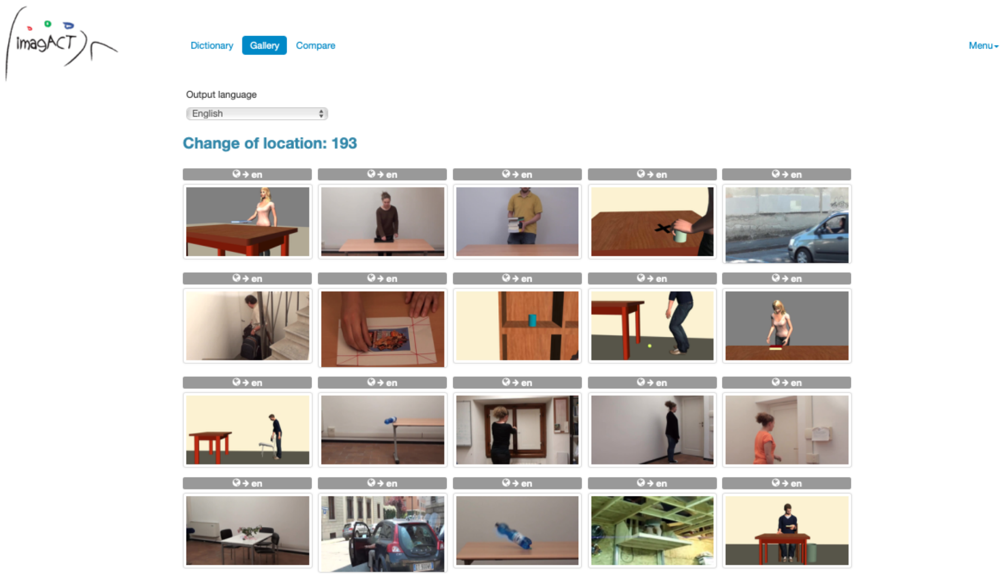


Figure 2.5: Gallery mode in IMAGACT: some action concepts comprised in *Change of location* class.

top of the desired action concept (the button with the target language abbreviation), all the verbs in the selected languages appear, and the visitor can continue the exploration by clicking on one of them and accessing the respective Dictionary page.

2.2 Theoretical Background

In the mind of a speaker, real world entities (i.e. objects and events) are linked to their mental representation through symbols (i.e. words), which are part of his own mother-tongue language. This picture becomes complicated due to cognitive economy constraints (Rosch, 1978), because frequently one lexical item productively applies to a set of objects or events, resulting in a one-to-many relations (Moneglia, 1996). Consequently, the mental representation of the world consists in a complex network of connections between entities, thoughts and lexicon.

To give an idea of this complex network, we can take into account all the possible pragmatic actions that an English speaker can correctly refer to with the verb *put*. In fact, this verb activates in the mind of the speaker a series of possible events, often quite distant in terms of pragmatic execution. For example, *putting a book on the table* and *putting some jam on the bread*, from the lexical point of view of the verb, are similar events, despite their differences: in the first one, no tools are required,[3] and few motoric activations are needed to perform the action; on the contrary, in the second one, a tool is needed (e.g. a knife) even if not explicitly lexicalised, and a sequence of various short actions are performed to complete the task. Moreover, only the second event can be predicated correctly also by the verb *spread* which, in its turn, extends its application to a series of events, some of which are other than those activated by *put*, e.g. *people spreading around the room.*

If we extend the focus from one single speaker to a community of speakers, we can see that the mental representation is roughly shared and still holds. But what happens if we try to compare the mental representation of two or more speakers of different languages? We immediately observe a variation in the linking between items of the conceptual space and words of the lexical dimension.

In fact, if we ask to a Japanese speaker to lexicalise those three actions, he would use 置く(*oku*) (common translation of *put*) to predicate the event of *putting a book on the table*, 付ける(*tsukeru*) for *putting/spreading some jam on the bread*, 散らばる(*chirabaru*) for *people spreading around*, and none of these verbs are overlapping or interchangeable in their primary pragmatic predication.

[3] According to other theoretical frameworks, the arm and the hand could be considered as tools (Pastra and Aloimonos, 2012).

These examples suggest that even a pure lexical discrimination is able to finely segment the conceptual space, and to highlight closeness or distance between concepts, both in monolingual or multilingual setting. In fact, at the basis of the IMAGACT project, there is the observation of different lexicalisation strategies operated by different natural languages, often aimed at making explicit certain traits or characteristics of the denoted objects (whether they are entities or events). An example of these differences can be observed for the lexicalisation of events of *playing an instrument*. In English,[4] the verb *play* applies correctly and productively to a wide range of musical instruments, such as *guitar*, *trumpet*, *drum*. On the contrary, in Chinese there is no verb that can extend its application to all of them (Yi, 2015). In fact, there is one distinct verb to use for each of those musical instruments: the verb 弹(tan) is used when music is produced with a *guitar*, 吹(chui) with a *trumpet*, 打(da) with a *drum*. Within these simple examples it is possible to observe the different segmentations operated in the semantic space from these two languages. English predicates the result of the action (i.e. to produce music out of an instrument) and applies *play* with all the possible music instruments; Chinese, instead, highlights the way an agent creates sound, and it predicates the act of *plucking* a string (on a guitar), *blowing* through a trumpet, *beating* a drum.

Understanding, describing and modelling the relationship between all the phenomena falling under the umbrella of *eventuality* (i.e. physical actions, movements, activities) is crucial if considering that verbs are the most frequent part of speech in spoken language (Halliday, 1989). In fact, it has

[4] The same applies to other Indo-European languages such as Italian, Spanish, Portuguese, German, which have a general verb that refers to the event of *playing an instrument*: *suonare* (IT), *tocar* (ES), *tocar* (PT), *spielen* (DE).

been observed that references to physical actions are very frequent, since spontaneous speech in natural contexts mainly deals with practical domains. Moreover, it has been observed that verbs referring to actions, movements and events are first obtained in early language acquisition (Tomasello, 2003), and for this reason they are crucial for the semantic competence of a speaker (Moneglia, 2014).

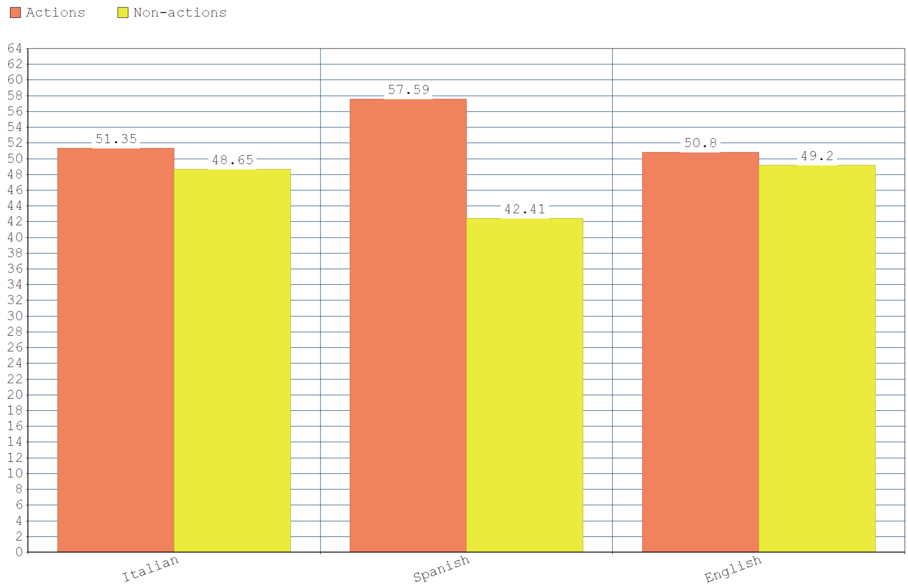


Figure 2.6: Frequency of verbs referring to actions in spoken corpora from 3 languages: Italian, Spanish, English. Adapted from Moneglia and Panunzi (2010).

Foundational studies for the development and set-up of the IMAGACT project (Moneglia and Panunzi, 2007, 2010) observed and compared reference corpora of Italian, English and Spanish spoken languages. It came out that, within the fundamental lexicon of those spoken languages, more than 60% of verb lemmas are action verbs, and that about 50% of the times in which a verb is used,[5] it refers to an action. Figure 2.6 shows the frequency of action reference in the analysis of those studies.

[5] With exclusion of the occurrences of modal, auxiliary and copula verbs, which work as functional elements, and thus they do not have a clear and salient semantic potential.

2.3 Action concepts gathering and resource building

In order to collect examples of use of action verbs in natural contexts, the first step in the building of the IMAGACT resource has been the analysis of natural spoken language corpora. The process has been multilingual from its very beginning: in fact, corpora of Italian and English have been bootstrapped in parallel.[6]

For English, the spoken subsection of the British National Corpus (BNC) (British National Corpus, 2007) has been considered. The BNC-spoken corpus consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins. In total, the BNC-spoken counts ∼2 millions of tokens.

For Italian, prior to IMAGACT building, no spoken corpora of enough dimension were available, and multiple resources have been merged together. In order to obtain a corpus comparable to the English counterpart, data from three different corpora have been used:

- LABLITA Corpus of Spontaneous Spoken Italian (Cresti and Moneglia, 2005): collection of orthographic transcriptions for a total of ∼100 hours of recorded sessions;

[6] This section is partially a translation and a short summary of the information reported by Gloria Gagliardi (2013). She completed her PhD as member of the team from University of Florence that worked on the IMAGACT Project, and she strongly contributed to the design of the Ontology and the annotation of the Italian section of the resource.

- LIP Corpus (Lexicon of Spoken Italian)(De Mauro et al., 1993): ∼60 hours of monologues and conversations from various contexts;
- CLIPS Corpus (Corpora and Lexicon of Spoken and Written Italian) (Albano Leoni, 2003), spoken subsection: ∼100 hours of recordings, sampled in 15 Italian cities (distributed along the whole national territory).

The sum of the three corpora resulted in ∼1.6 millions of tokens. Beside the relative difference in terms of dimension, the two corpora are highly comparable, due to the fact that both result in a balanced sampling on the axis of formal/informal, public/private speech.

After the design of the resources from where to extract action references, common practices of pre-processing have been applied: first, markup symbols (used for the transcription) have been removed through regular expressions; then, corpora have been tokenized and POS-tagged.[7] From cleaned corpora, the selection of data to annotate has been conducted through the identification of high/mid-frequency verbs which, prototypically, encode the reference to concrete and pragmatic actions, events and movements. It is possible to group and summarise in four categories all the verbs included into this analysis: Activity verbs, General verbs, Logical verbs, Movement verbs.

Activity verbs

The Activity verbs class (Dowty, 1979; Vendler, 1957, 1967) refers to verbs which predicate a unique and distinct action concept from a cognitive point of view (Moneglia, 2010). No substantial differences exist between the *eventualities* pointed out by one of these verbs. In other words, changing motor

[7] TreeTagger (Schmid, 1994) has been used for this task.

scheme, object or any other property does not require to change verb in order to predicate the new resulting action. As an example, consider some discrete actions predicated by the verb *cook*:

(2.3.1) John cooks the potatoes

(2.3.2) John cooks the sausages

(2.3.3) John cooks the bread

It is possible to predicate 2.3.1 with *boil*, 2.3.2 with *grill*, 2.3.3 with *bake*. But, if we consider the three events from the point of view of the verb *cook*, no differences emerge from the lexical realisation; therefore, all of the three are equally instances of *cooking*.

General verbs

Within the class of General verbs we can consider those verbs that extend their reference to a wide set of distinct actions, often quite distant, which can be classified independently from each other and do not share common properties, but they still undergo the same lexical categorisation (Moneglia, 2010). The main difference with Activity verbs' class is that the actions referred to by a general verb are totally different and cognitively distant one from each other, thus they are not instances of the same action. Consider the following examples:

(2.3.4) Mary opens the door

(2.3.5) Mary opens the book

(2.3.6) Mary opens the umbrella

It is immediately clear that 2.3.4, 2.3.5 and 2.3.6 tell about three events that, beside the common reference of the verb *open*, are cognitively and pragmatically distant. General verbs extend the same lexical categorization to classes of distinct activities (Choi and Bowerman, 1991; Moneglia, 1998; Wittgenstein, 1958).

Logical verbs

This class collects all the verbs referring to primitive operations, related to logic and arithmetic, such as *add*, *join*, *divide*, *split*, which do not define a specific activity. These verbs do not show a true and distinct typological variation.

Movement verbs

Movement verbs refer to eventualities in which one of the verb's actants modifies its own position in the space, or causes the modification of other's position (e.g. *turn*, *rotate*, *carry*). Two sub-classes of verbs fall under the Movement verbs class: verbs that codify the path of a movement, and those which codify the manner (Talmy, 1985).

2.3.1 Annotation

After the pre-preprocessing steps and the collection of all action verbs occurrences, a team of annotators (Italian and English native speakers) have manually checked, analysed and annotated all the verb occurrences in their context. This phase of the work consisted in two subsequent tasks, for each occurrence:

- production of synthetic sentences;
- assignment of a *variation field*.

First, the annotator selected one of the verbs assigned to him/her from the annotation interface; then, he/she opened every occurrence of the selected verb in the form of concordance (*left context*, *verb*, *right context*) through a KWIC (*Key-Word In Context*; Luhn, 1960) visualization.

Thanks to the simple visualization, the annotator has been able to read and analyse the extended context of each occurrence and, for the processable ones,[8] to write a simple synthetic sentence (henceforth, *standardisation*) that summarises the minimum syntactic/semantic elements necessary to express the action that occurrence is referring to. The writing of standardisations followed some simple rules. A standardisation must be composed:

- in positive form, active voice, present tense, third person (singular or plural);
- with one unique verb (no modal or auxiliary verbs allowed);
- with only the necessary elements, i.e. arguments, to complete the meaning of the verb and fulfil its valency[9] (adjuncts may be added, but written in square brackets);
- without making use of generic expressions (e.g. *someone*, *something*, *a man*);

[8] Given that data collection has been totally based on automatic techniques, some tagging errors might result in false positives.

[9] In linguistics, the notion of valency has been imported from chemistry by Tesnière, who describes verb valency as follows:

> "The verb may therefore be compared to a sort of atom, susceptible to attracting a greater or lesser number of actants, according to the number of bonds the verb has available to keep them as dependents. The number of bonds a verb has constitutes what we call the verb's valency" (Tesnière, 1959, Chapter 97, §3)

- if it is possible, opting for *basic level expressions* (Rosch, 1978), or for proper names if it is not;
- with a linear word order, i.e. avoiding dislocations.

Once the standardisation has been composed, the annotator has been asked to assign a variation field, choosing among: primary, marked, support, sublemma. The most important one, from the point of view of the resource building, is the *primary* variation field. The criteria for tagging a standardisation as *primary* are:

- the standardisation of the occurrence refers to a concrete physical action;
- the occurrence is referring to a proper instance of the action verb, i.e. it is possible to use it as a prototypical example of when to correctly use that verb.[10]

The *marked* variation field has been assigned to those occurrences referring to a metaphoric usage or constituting a phraseology, e.g. *take the bus*. The *support* variation field has been used to tag those occurrences of light verb constructions. The *sublemma* variation field has been used to identify occurrences of polysemic cases and phrasal verbs.

At the end of this annotation task, only standardisations of *primary* and *sublemma* occurrences have been further processed, and used for the action classes induction. The variation field assignment has been evaluated with 2 expert annotators on a sample of 974 occurrences, resulting in an overall inter-annotator agreement greater than 0.8.[11]

[10] This criterion has been fulfilled if the sentence passed a simple pragmatic test, adapted from Wittgenstein (1958): the event described by the sentence s of language L and similar events are instances of what in L the verb X means.

[11] For details on the data evaluation campaign in IMAGACT, see Chapter 4 in Gagliardi (2013).

2.3.2 Action classes induction, ontology building and video production

Once finished the manual check of action verbs occurrences and the production of standardised sentences, all of them have been manually clustered on the basis of the distinct action concept they refer to. From the verb point of view, the reference to an action concept has been defined *Action Type*, that is the action (or set of actions) instantiated by that specific verb as an unitary sense.[12] Action Types have been created with a bottom-up approach, considering cognitive similarities among each eventuality described by standardisations, such as motor scheme, spatial relations, property of the action, verb equivalence. In Wittgensteinian terms, all standardisations referring to an instance of a particular action concept share *family resemblances* (Wittgenstein, 1958, §67). Each Action Type, in order to be validated, must be *productive*, i.e. it could be applied to an open set of objects.

Following the gathering of standardisation in Action Types, one *Best-Example* (BE) have been chosen among all the standardisation as the most representative example of that verb used to predicate that particular action concept. In some cases, more than one BE have been chosen, given that sometimes more than one syntactic scheme could be applied, or various argument alternation are possible, i.e. one argument may be instantiated either as a direct or oblique argument.

[12] In order to clarify what an Action Type is, examples 2.3.4-2.3.6 report on three Action Types of the verb *open*: three distinct action concepts instantiated by the same verb, thus one *open*'s Action Type for each. It is also possible that one Action Type incudes more than one action concept, as it happens with the examples 2.3.1-2.3.3.

As an example, consider the well-known *Spray/Load* argument alternation among object and oblique (Levin, 1993):

(2.3.7) [Someone]$_{\text{AG}}$ [loads]$_{\text{V}}$ [the bricks]$_{\text{TH}}$ [on the truck]$_{\text{DE}}$

(2.3.8) [Someone]$_{\text{AG}}$ [loads]$_{\text{V}}$ [the truck]$_{\text{TH}}$ [with the bricks]$_{\text{TH}}$

In 2.3.7 *the truck* is the destination of the theme argument *the bricks*, while in 2.3.8 both *the truck* and *with bricks* are theme arguments. In these cases, one BE for each realisation is assigned to the verb's Action Type.

For each verb's Action Type, once all the occurrences have been gathered, a short script has been written, with the objective of highlighting peculiar action features and facilitating the video production.

The analysis proceeded in parallel for Italian and English. At the end of the previous steps, Action Types from both the languages have been compared, discussed and mapped, both intra-linguistically and inter-linguistically.[13] As a result, language-independent action concepts have been identified, which constitute the visual taxonomy of the IMAGACT Ontology.

For each concept, a short video has been produced. These videos work both as prototypes for action concepts and as glosses for the verbs' Action Types. This is one of the key feature of IMAGACT: instead of explain the meaning in context of one action verb with words, a short video unambiguously shows the performing of the action, thus everyone could understand the action verb reference without the need of mastering the language of that specific verb, or imagining and guessing the action by interpreting a linguistic explanation. Videos have been produced both with real recordings and 3D

[13] Chapter 3 reports on an example of intra-linguistic and inter-linguistic mapping of language specific (and verb specific) Action Types for the verbs *cut* (EN) and *tagliare* (IT).

animation graphics.[14] Likewise standardisations, video production followed some specifications as well:

- objects involved in the actions must be real-world objects, i.e. books, bottles, boxes (not abstract or too generic objects, like spheres or cubes);
- neutral settings for recording, in order to avoid distracters in the background, and to keep the focus on the foreground and the action performing;
- long take recording technique, thus reducing the necessity of video cutting and editing;
- natural speed action performances, i.e. no slow motion or time-lapse recording.

Figure 2.7 illustrates the result of the building of IMAGACT. It shows how action verbs are represented as Action Types, and how these Action Types encode action concepts in the form of prototype videos. For the sake of simplicity, only few Types for the verbs *prendere* and *take* are presented. Furthermore, it highlights the equivalence relation between two verbs (either intra-linguistically and inter-linguistically): some Action Types of the two verbs encircle the same action concepts, while other Types remain distinct, thus the compared verb is not applicable.

[14] Most of the action concepts have both the real video and the 3D animated one. But for some concepts only the 3D video is available, due to technical reasons: e.g. for the verb *stab* (BE of the Action Type: *John stabs Simon*) it is obvious that harming someone with a knife was not feasible, thus the concept has been represented only with a 3D video.

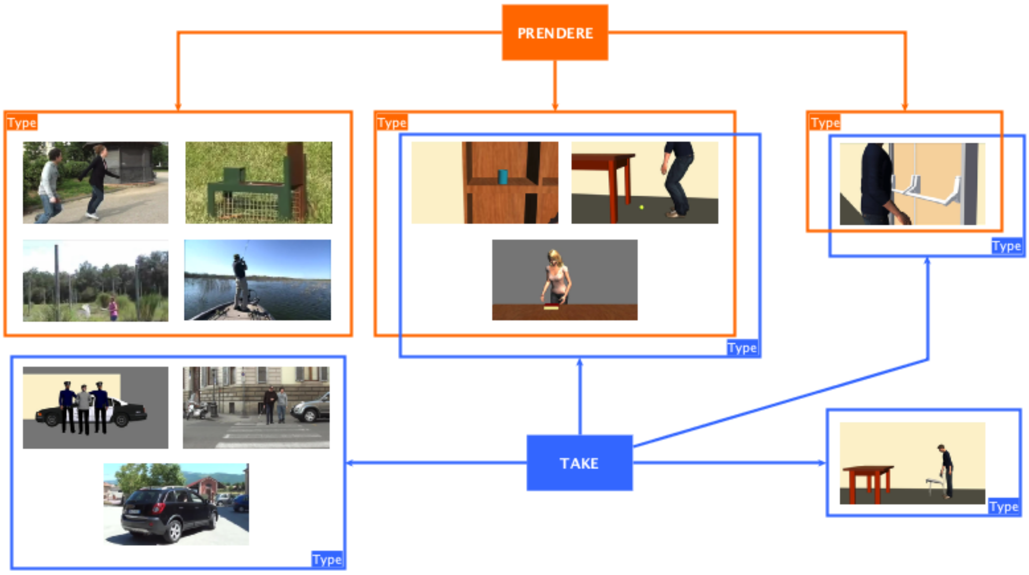


Figure 2.7: An example of relations among verbs and between verbs and action concepts. Partial extensions of the verbs *prendere* and *take*.

2.3.3 Competence-based Extension to many languages

Italian and English represent the core languages on which the IMAGACT system has been built. As stated in the previous section, a language-independent constellation of action concepts has been identified, formalised and linked to prototype videos on the basis of the analysis of occurrences of action verbs in spontaneous speech in those two languages. With this setting, and following the same Wittgensteinian paradigm for the identification of primary uses of action verbs, it has been possible to extend the resource to other languages. This phase has been called IMAGACT4ALL (Brown et al., 2014; Moneglia et al., 2014a,b), in which a competence-based extension (CBE) technique has been successfully applied to map the verbal lexicon of other languages in the multilingual network of IMAGACT.

CBE is a direct implementing method that relies on pragmatic and semantic competence of native speakers. An informant accesses the action concepts as input, provided with the corresponding videos, verbs and captions in Italian and English.[15] The informant assesses the actions depicted in the videos, and provides the verb or verbs in his language that can be used to refer to them. Then, for each verb, the informant is asked to write a simple caption to go with the videos, to be used in IMAGACT as example of use of that verb in the context of the action depicted in the video. Table 2.1 reports all the languages currently mapped in IMAGACT, and their completeness in terms of action concepts covered. With exclusion of Italian and English, all other languages have been mapped through the CBE methodology.

[15] Italian and English data are provided as support, in order to help the informant in identifying and understanding the eventuality of each action a concept.

Language	**Number of verbs**	**Action concepts coverage**
Italian	646	100%
English	662	100%
Arabic (Syria)	571	100%
Chinese (Mandarin)	414	100%
Danish	646	100%
French	656	100%
German	990	100%
Greek	638	100%
Hindi	470	100%
Japanese	736	100%
Polish	1,193	100%
Portuguese	805	100%
Serbian	1,096	100%
Spanish	736	100%
Albanian	5	0.2%
Arabic (Tunisia)	423	99,9%
Assamese	95	14.8%
Bangla	223	25.7%
Finnish	23	1%
Hebrew	160	19.3%
Hungarian	656	42.3%
Indonesian	214	25.2%
Farsi	264	28.7%
Magahi	67	9.9%
Manipuri	62	9.9%
Norwegian	115	16.8%
Oriya	160	10.9%
Russian	86	4%
Tamil	92	9.9%
Urdu	478	99.7%

Table 2.1: Number of verbs and action concepts coverage per language in IMAGACT.

2.4 A comparison with other resources

IMAGACT main feature is its being multimodal and *action-centric*, i.e. it focuses on actions and represents them as visual concepts. Other resources (such as WordNet (Miller et al., 1990), BabelNet (Navigli and Ponzetto, 2012a), FrameNet (Baker et al., 1998)) have more broad scopes, and for this reason senses referred to actions are often vague and overlapping (Panunzi et al., 2018), and often all possible actions can be gathered under one concept. As an example, consider all the possible pragmatic actions that can be predicated by the verb *push* and their representation in some resources.

WordNet is a large lexical database of English, in which the lexicon is organised in *synsets* (i.e., sets of lemmas sharing the same sense), and each synset is a node in a deep web of conceptual-semantic and lexical relations.[16] From a lookup in WordNet (or BabelNet[17]), it emerges that only 4 out of 10 synsets of *push* as verb refer to concrete actions, and some of the glosses are not really exhaustive, thus they apply to a wide set of different actions.

WN Synset	**Lemmas**	**Gloss**
push.v.01	*push*, *force*	move with force
push.v.05	*push*	press against forcefully without moving
push.v.09	*push*	move strenuously and with effort
press.v.11	*press*, *push*	make strenuous pushing movements during birth to expel the baby

Table 2.2: WordNet synsets referring to pragmatic *pushing* actions.

[16] WordNet is freely and publicly available for download: https://wordnet.princeton.edu/download.

[17] BabelNet inherits WordNet taxonomy. For a brief description of BabelNet, see 4.1.1.

Table 2.2 reports all the WordNet synsets referring to pragmatic *pushing* actions, i.e. only those pertaining to the primary variation of the verb *push*. In such framework of categorization, all possible actions referred by *push* can be gathered under the first synset, except from those specifically described by the other three.

FrameNet is focused on *semantic frames*, which are defined as a schematic representation of situations involving various participants, identified by their semantic role (e.g. Agent, Cause, Theme, Place, Manner, Duration, etc.).[18] Words that evoke a semantic frame and fit in it are called Lexical Units (LUs), while those that take part in the event as participants are Frame Elements (FEs).[19] By performing the same lookup in FrameNet, a total of 4 Lexical Units for *push* as verbs are given, belonging to 4 semantic frames: *Manipulation*, *Cause_motion*, *Cause_change_of_position_on_a_scale*, *Subjective_influence*. Out of these, only the first two apply to the description of pragmatic actions execution.

Frame	**Description**	**LUs (verbs)**
Manipulation	The words in this frame describe the manipulation of an *Entity* by an *Agent*. Generally, this implies that the *Entity* is not deeply or permanently physically affected, nor is it overall moved from one place to another.	*caress, clasp, claw, clutch, diddle, finger, fumble, grab_hold, grab, grasp, grip, grope, handle, hold, kiss, knead, lick, massage, nip, nudge, paw, pinch, poke, pull, push, rub, seize, squeeze, stroke, tickle, touch, tug, tweak, wring, yank*
Cause_motion	An *Agent* causes a *Theme* to move from a *Source*, along a *Path*, to a *Goal*. Different members of the frame emphasize the trajectory to different degrees, and a given instance of the frame will usually leave some of the *Source*, *Path* and/or *Goal* implicit. The completion of motion is not required, although individual sentences annotated with this frame may emphasize the *Goal*.	*attract, cast, catapult, chuck, drag, draw, drive, drop, fling, force, haul, hit, hurl, impel, jerk, knock, launch, lift, move, nudge, pitch, press, propel, pull, punt, push, rake, roll, run, scoot, shove, slam, slide, stick, throw, thrust, toss, transfer, tug, wrench, wrest, yank*

Table 2.3: FrameNet semantic frames referring to pragmatic *pushing* actions.

[18] FrameNet follows a theory of meaning called *Frame Semantics*, as it has been defined by Fillmore (1976, 1977, 1982, 1985) and Fillmore and Baker (2010, 2001).

[19] FrameNet data can be browsed online or freely downloaded for research purposes, after filling in a request form. https://framenet.icsi.berkeley.edu/.

Table 2.3 describes the two semantic frames, and lists all the verbs which evoke those frames. It is immediately clear that the main semantic distinction is due to the presence or absence of movement of an object, but it is basically the only distinction expressed. Also, Lexical Units may refers to actions that are totally distant from prototypical actions described by *push* (e.g. *kiss* in the *Manipulation* frame, or *run* in the *Cause_motion* one).

Even if IMAGACT is a smaller resource, its action conceptualization is more fine-grained. In fact, the verb *push* is organised in 18 Action Types, which gather 25 total action concepts. Table 2.4 shows details on the extension of *push* in IMAGACT.

Action concept	Meta-categories	Verbs (EN)	Action Type	Total verbs
9dd0265d	Force on an object, Movement	push, move	push_T1	21
65431186	Movement, Force on an object	push, wheel, move	push_T1	18
cbd1726a	Force on an object, Modification of the object	push, press	push_T2	27
6a8bafd9	Force on an object	push, press	push_T2	21
2decad1a	Actions in the intersubjective space	push, nudge	push_T3	24
52310652	Force on an object, Actions in the intersubjective space	push, press	push_T3	20
f7b9693b	Force on an object, Change of location	push	push_T4	26
7912c776	Force on an object, Change of location	push, ram	push_T5	32
18ad2fa9	Change of location, Force on an object	push, throw, shove	push_T5	19
40374041	Force on an object	push, move, shove	push_T5	33
b4931ed2	Modification of the object, Force on an object	push, squeeze	push_T6	23
28fa1c89	Force on an object, Setting relation among objects	push, insert, put	push_T7	41
950a69d5	Movement, Actions referring to the body	push	push_T8	22
92da1d76	Modification of the object, Force on an object	push	push_T9	23
e017360a	Setting relation among objects, Change of location	push, move	push_T10	39
073129b0	Actions referring to the body	push	push_T11	21
c0c8caf4	Modification of the object, Force on an object	push, squeeze, press, compress, flatten	push_T12	35
4d69cb19	Force on an object, Modification of the object	push, squeeze, compress, press	push_T12	29
d2a1848e	Change of location, Modification of the object	push	push_T13	21
1028b2f9	Force on an object	push	push_T14	33
e2abd551	Force on an object	push, press	push_T15	27
8b2675f8	Force on an object	push, press	push_T15	30
9adb3816	Force on an object	push, press	push_T16	30
f4157bb2	Force on an object	push, insert, put	push_T17	63
838d9dd7	Force on an object, Modification of the object	push, press	push_T18	27

Table 2.4: IMAGACT action concepts linked to the Action Types of the verb *push*, with the detail of meta-categories classification and the total verbs count for all the languages in the resource.

2.5 IMAGACT use cases

This section provides a broad survey on the use cases of IMAGACT in various scenarios, from L2 learning to psychological investigations.

2.5.1 IMAGACT for Chinese Students

This paragraph resumes the translation experiments of Italian general action verbs by Chinese learners of Italian using IMAGACT, as they are described by Yi (2015). Two different experiments have been conducted, and for both the learners have been split in two groups: the first has been trained using the IMAGACT web interface, the second without accessing the information in the resource (i.e. control group). Performances show that the visual information delivered with IMAGACT helped learners in understanding, disambiguating and translating general verbs.

The participants were 16 female students of Italian L2, aged ∼20, with an elementary knowledge of Italian and a good knowledge of English. Both tests have been designed on the disambiguation and translation of 4 Italian general action verbs as target, and they have been administered to participants in 4 different rounds in the course of one month, i.e. one round per verb every week. Target verbs of these experiments are: *mettere* (∼*put*),[20] *prendere* (∼*take*), *prendere* (∼*take*), *girare* (∼*turn*), *alzare* (∼*raise*).

[20] Henceforth I will use the tilde symbol (∼) with the meaning of *approximated* translation in English, i.e. the most likely equivalent in English of a foreign action verb. Given that the whole Thesis relies upon the fact that the correct equivalent in translation of action verbs depends on the context of use and its reference to a specific pragmatic action, it would be restrictive to state that one verb is equivalent to another in general terms.

Test 1

Test 1 is a translation test from Italian to Chinese, and it consists in 60 simple sentences to be translated for each round, half of them containing the target verb, and the other half with random action verbs used as distracters. All the sentences are standardisations taken from the IMAGACT resource. The goal of this test was twofold: evaluate translation performances between general and specific action verbs, and evaluate the difference in performance between the IMAGACT and the control groups.

The test evaluation consists in the scores obtained by students in the translation, with focus on the adequacy of the chosen Italian verb:

- -1 for each test item where the chosen verb is wrong;
- 0 for each test item where the chosen verb leads to an odd but understandable translation;
- 1 for each test item where the source verb is correctly translated.

Table 2.5 shows the performance of the two groups (IMAGACT-trained and control). The IMAGACT-trained group slightly outperforms the control group, especially in the interpretation and translation of general action verbs, which was the main goal of the test.

Test 2

Test 2 is an multiple-choice test, with the possibility of selecting more than one answer. The test consists in 13 Chinese sentences translated to Italian for each target verb, where the Italian sentence is composed with multiple alternative action verbs. Participants have been asked to highlight all the proposed Italian verbs that correctly apply in the translation of the source sentence. In each test item, Italian verbs have been selected on the basis

	mettere	*prendere*	*girare*	*alzare*	Total
IMAGACT (all items)	278	353	391	396	**1418**
general action verbs	120	116	170	161	**567**
specific action verbs	158	237	221	235	**851**
CONTROL (all items)	256	325	353	380	**1314**
general action verbs	84	98	137	149	**468**
specific action verbs	172	227	216	231	**846**

Table 2.5: Results for translation test of Chinese action verbs to Italian (Test 1).

of the local equivalence to the target verb in its Action Types, but not all of the proposed verbs can be used to translate the action described by the Chinese sentence, plus a distracter verb.

The evaluation of the test is based on classical Precision-Recall measure calculated on each sentence for each participant, and Table 2.6 reports on the F-measure obtained from each target verb and the mean of all the results. It is possible to recognize that the IMAGACT-trained group obtains the best results for each target verb.

	mettere	*prendere*	*girare*	*alzare*	Mean
IMAGACT	7.03	6.66	7.31	7.52	7.13
CONTROL	5.51	4.98	6.57	6.36	5.85

Table 2.6: F-measure for multiple-choice test of Chinese action verbs translated to Italian (Test 2).

2.5.2 IMAGACT for Arabic and Italian Students

This paragraph resumes the experiments in translation of general action verbs by Arab learners of Italian and Italian learners of Arabic, as they are described by Mutlak (2019). Similarly to the experiments described in 2.5.1, two translation experiment have been set, and both groups of Italian and Arabic learners have been split into target (trained using the IMAGACT web interface) and control (never used IMAGACT) groups.[21] Performances confirm Yi (2015) results, and thus highlight the potential of IMAGACT and its multimodal approach in early L2 learning.

These experiments focused on 4 Italian general action verbs and their prototypical Arabic equivalents:

- *prendere* / أخذ (~take);
- *portare* / حمَل (~bring);
- *tirare* / سحَب (~pull);
- *spingere* / دفَع (~push).

Test 1

The first test consists in a translation test: 20 sentences in the source language, derived from the IMAGACT resource, to be translated to the target language (AR→IT for Italian participants, IT→AR for Arab participants) for each target verb. Half of the test items contain the target verb, a specific action verb as distracter in the remaining items.

[21] The subjects who took part into these experiments were 10 Arab learners of Italian and 12 Italian learners of Arabic.

Table 2.7 reports on results for the translation test.[22] It is possible to notice that the performances on specific action verbs (e.g. activity verbs, movement verbs), which extend their reference to a small set of action concepts that are normally not too distant in the semantic space, are noticeably higher with respect to performances on general action verbs. This is an obvious result, given the broad extension of reference that general action verbs show, thus equivalent candidates in translation are lower in number, and the possibility to chose a wrong one is lower. However, IMAGACT-trained groups (both Italian and Arab participants) show better performances with respect to control groups. Arab IMAGACT-trained group outperforms the Italian one, which still shows a ∼21% boost with respect to its control counterpart.

	General Ac_V	**Specific Ac_V**
IMAGACT_IT	65.5%	91.44%
CONTROL_IT	44.57%	91.89%
IMAGACT_AR	92.2%	95.38%
CONTROL_AR	51.92%	91.67%

Table 2.7: Results for translation test of Arabic action verbs translated to Italian (X_IT) and Italian action verbs translated to Arabic (X_AR) (Test 1), expressed in percentage of correct answers.

[22] Differently from Yi (2015), Mutlak (2019) calculated the results in terms of percentage, without applying a more fine-grained score system.

Test 2

In line with Yi (2015), Test 2 consists in a multiple-choice test, with the possibility of selecting more than one answer, and it has been designed the same as the Chinese test, but in both translation directions. Mutlak (2019) only reports the total precision-recall scores for Italian and Arab participants, without details on each target verb. Table 2.8 shows the results in terms of F-measure. Also here, the use of IMAGACT in language learning brings to better understanding, disambiguation and translation of action verbs, confirming that its multimodal feature boosts up foreign language learning for early learners of L2.

	F-measure
IMAGACT_IT	0.78
CONTROL_IT	0.6
IMAGACT_AR	0.9
CONTROL_AR	0.48

Table 2.8: F-measure for multiple-choice test of Arabic action verbs translated to Italian (X_IT) and Italian action verbs translated to Arabic (X_AR) (Test 2).

2.5.3 IMAGACT for the assessment of Mild Cognitive Impairment and Aphasia: the SMAAV battery of tests

In this paragraph a short description of the SMAAV (Semantic Memory Assessment on Action Verbs) battery of tests (Gagliardi, 2013) is outlined. This battery has been specifically designed for the assessment of *Mild Cognitive Impairment* (MCI)[23], which is an under-estimated and difficult to diagnose condition, halfway between average cognitive decay linked to ageing and more critical dementia diseases, such as Alzheimer's Disease. Moreover, it has been also used in the assessment of semantic-lexical impairment in aphasic patients. The objective of these tests is to early identify and quantify impairments through the observation and description of the reactions induced by multimedia stimuli. The first one is a *naming* test, aimed at verifying semantic memory access to action verb lexicon and linguistic production in relation to observed actions; the second one is a multiple-choice test, in which multiple sentences are offered, and the patient has to choose

[23] The MCI condition has been labelled with various names by the literature: *Benign Senescent Forgetfulness*, *Age-Associated Memory Impairment* (AAMI), *Age consistent memory impairment*, *Late Life Forgetfulness* (LLF), *Aged-Associated Cognitive Decline* (AACD), *Age-related cognitive decline*, *Cognitive Impairment, No Dementia* (CIND), *Mild (neuro)cognitive Disorder*. Describing and discussing MCI is out of the scope of this Thesis. For a more complete picture of MCI, from the psycholinguistic point of view, please refer to Gagliardi (2013). However, a definition of MCI is as follows:

> "A state in which a single cognitive function, usually memory, is impaired to an extent that is greater than would be anticipated for age, yet the patient does not meet criteria for dementia. These patients are probably at increased risk of developing dementia in subsequent years but are otherwise functioning reasonably well." (Petersen et al., 1997, 66)

the correct one (or ones) to describe the action proposed in the video item; the third one is a *picture task* test, where the patient is asked to select the video depicting the action described by a target sentence. All of the three tests in the SMAAV battery have been designed making use of visual and linguistic information from the IMAGACT Ontology, and for this reason have been reported herein as example of exploiting the resource in scenarios other than early L2 learning. The innovation brought by SMAAV consists in the use of dynamic visual stimuli, i.e. videos, instead of static images, which are generally used in the practice of this kind of tests and experiments. In fact, the use of videos guarantees a better representation of actions, avoiding an high cost in terms of inferencing symbolic meaning, as static images require (Tranel et al., 2008). The items used in this battery of tests are all derived from IMAGACT, and all of them belong to the semantic variation of the verb *girare* (IT) (∼*turn*). All distracter items are obtained from semantically related verbs and general action verbs. The choice of *girare* as target verb derives from some preliminary observations:

- high frequency in spontaneous speech corpora;
- high depictability of the eventualities referred to by the verb;
- extension of the applicability of the verb on a wide range of actions, some of them quite dissimilar from each others (e.g. actions performed with the body, various manipulations and location modification of an object).

The SMAAV battery has been experimentally administered to patients affected by aphasia as consequence of brain damage in the left hemisphere, and to patients affected by dementia (Corsi, 2016; Fanetti, 2018; Lippi, 2018; Shabana, 2018), showing strong effectiveness in highlighting lexical and semantic deficits not emerged from previous screenings.

Test 1

The first test consists in a *naming* test, where the patient is asked to retrieve the correct verb (or verbs) to describe the action showed in each item. This test focuses on the capability of the patient in accessing the mental lexicon, and it checks for the integrity of his/her semantic memory. The test is composed of 25 items in total (19 target plus 5 distracters) and 2 trial items preceding the proper testing session, used to explain the test and get the patient used to it. For each item, the patient is asked to name the action depicted in the video, either by naming a verb or uttering a simple sentence. During the testing, the examiner is not allowed to interfere, but he/she can stimulate the patient by repeating or rephrasing the question.

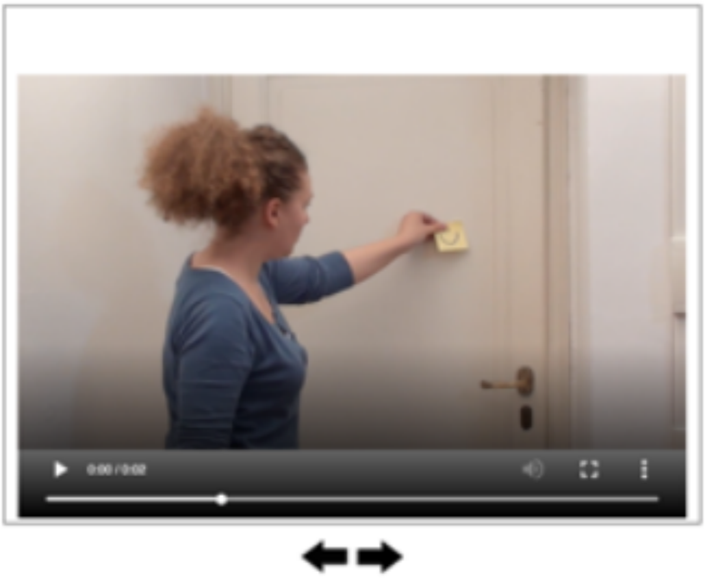
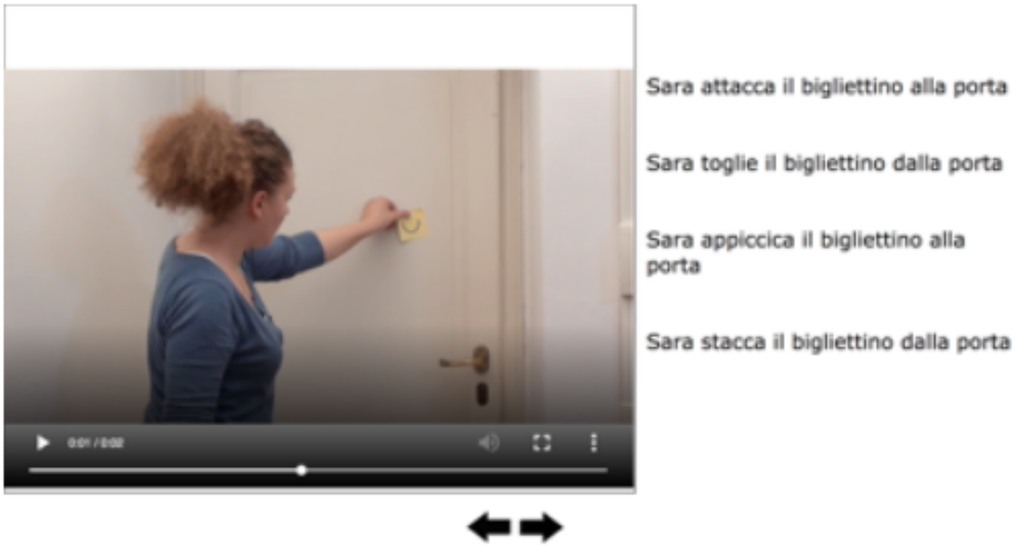


Figure 2.8: One of the items of SMAAV naming test (Test 1).

Test 2

The second test consists in a multiple-choice test (6 items), in which the patient is asked to select the sentences containing the verbs that correctly applies to the action depicted in each video item of the test. For each item, a video depicting one of the action concepts connected to *girare* is shown, and 4 sentences are offered as possible choices. Among them, 2 are correct captions for the video (composed with one specific and one general action verb), one is too generic (it is composed using a verb that is too general, e.g. *muovere* (∼*move*), one contains a totally inapplicable verb.



Sara attacca il bigliettino alla porta

Sara toglie il bigliettino dalla porta

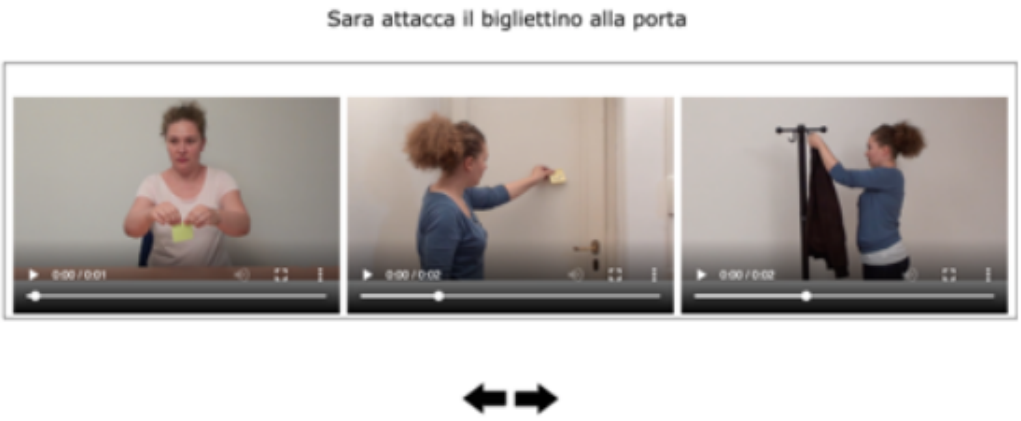
Sara appiccica il bigliettino alla porta

Sara stacca il bigliettino dalla porta

Figure 2.9: One of the items of SMAAV multiple-choice test (Test 2).

Test 3

The third test of the battery consists in a *picture task* test (6 items), in which the patient is presented with a target sentence and 3 videos. The patient is asked to point out the video that depicts the action described by the sentence. One of the three videos is the correct one, while the other two are distracters: one of them shows a completely different action performed on the same object of the sentence; the other one shows another action covered by the primary extension of the target verb, but pragmatically distant and not linkable to the sentence of the test item.



Sara attacca il bigliettino alla porta

Figure 2.10: One of the items of SMAAV picture task test (Test 3).

Chapter 3

Verbs' local equivalence in multi-lingual scenarios: the mapping between *tagliare* (IT) and *cut* (EN) in IMAGACT

This chapter illustrates a proposal for the inter-linguistic mapping in IMAGACT between the Action Types of the verbs tagliare (IT), cut (EN), and their equivalent verbs in the two languages. This example is useful to show the methodology behind IMAGACT in a concrete application: through the analysis of the standardisations of each one, Action Types have been aligned and linked together, and all the local equivalences have been verified.

The conceptual domain of *cutting* and *breaking* events is particularly interesting to analyse,[1] because it is carved up differently in different languages. It might be obvious to state, the most emergent feature of *cutting* events is that the action of modification of an object by interrupting its integrity has to be performed with a sharp instrument, which distinguishes such eventuality from other categories, such as modifications operated by using the hands or heavy tools, or by pulling and tearing. Consider the example in 3.0.1:

(3.0.1) John splits a sandwich

It is possible to split a sandwich using a knife or just with the hands: in the first case, the agent splits by *cutting* the sandwich; in the latter, he splits by *tearing* the object. In English, even if not explicitly expressed, the sharp implement presupposed by *cut* (Bowerman, 2005), is prototypically a knife or a pair of scissors, but it not excludes other objects like a piece of broken glass or pottery, fingernail, wire or thread, even a blade of grass or a piece of paper. The feature shared by all those objects is that they all have a *thin linear edge.* This proposition seems to be generally valid, as it holds for many languages, but it is not always true that the sharp instrument would not impose further restrictions on the verb to be used. As a matter of fact, Dutch (Erkelens, 2003) shows another mandatory distinction between action of cutting with a tool that has *two opposite edges* (scissors, shears, nail clippers, etc.) versus those performed with a *single edge*: *knippen* is used in the first case, *snijden* in the latter.

[1] It is testified by a special issue of *Cognitive Linguistics* (Majid et al., 2007), that collects various contributions which analyse the different lexicalisation strategies operated by languages around the world.

Similar observations to those presented by Bowerman (2005) for *cut* are also valid for the Italian *tagliare*. For this reason, mapping Action Types of the two verbs is not a particularly difficult task, even if some distinctions emerge from the analysis of the data in IMAGACT. The work herein described starts from the analysis of every single Action Type of each verb. All the occurrences and related standardisations have been checked from the corpora used for the building of IMAGACT, and the consistency of each Action Type has been verified. Consequently, locally equivalent verbs have been checked, in order to verify that they actually have one Action Type that shares the reference with the starting verb. Then, other verbs not linked to any of those comprised in the *cutting* semantic network of IMAGACT are also searched and proposed for linking. Once done, those Action Types have been compared inter-linguistically and linked together. In other words, the following work is a practical example of the resource building workflow, applied on the semantic field of *cutting* eventualities.

As a starting point, Tables 3.1 and 3.2 report data available for the verbs *cut* and *tagliare* in IMAGACT. Types are listed with cardinal numbers for the sake of simplicity[2] and for each one, Best Examples, Equivalent Verbs and number of occurrences are shown. The following mapping will be presented with a scheme of the mapping first, and then the explanation. In the schemes, dotted lines and the label *verb-tnew* stand for suggested new Action Types, as shown by the first one in Figure 3.1.

[2] Henceforth, *verb-tn* will be used to refer to one Action Type of the verb, e.g. *cut-t1* stands for Action Type 1 of the verb *cut*.

Cut (EN)			
TYPE	**Best Example**	**Equivalent Verbs**	**Occurrences**
1	**BE1** John cuts the head off of the flower with the knife **BE2** The machine cuts the coal **BE3** The lumberjack cuts the tree down	chop	**68**
2	**BE1** The surgeon cuts along the skin with the knife	slit	**14**
3	**BE1** The sculptor cuts a statue from the rock **BE2** The lathe cuts the wood	carve	**5**
4	**BE1** Mary cuts the bush back in the garden with trimmers **BE2** The lawnmower cuts the grass	trim shear clip	**86**
5	**BE1** John cuts the pizza into quarters **BE2** The machine cuts the sheepskin in half	split	**18**
6	**BE1** John cuts the advertisement out of the photograph	extract	**12**
7	**BE1** John cuts his finger **BE2** The metal cuts john	injure	**4**
8	**BE1** John cuts the string **BE2** John cuts the pinata down	snip	**3**

Table 3.1: Action Types identified by the verb *cut*.

Tagliare (IT)			
TYPE	**Best Example**	**Equivalent Verbs**	**Occurrences**
1	**BE1** Marco taglia la carne con il coltello	suddividere fare in parti	**17**
2	**BE1** Marco taglia le immagini dal libro con il coltello	ritagliare	**15**
3	**BE1** Marco taglia la buccial al formaggio con il coltello	rimuovere togliere	**22**
4	**BE1** Il chirurgo taglia la pancia a Luca con il bisturi **BE2** Marco taglia Luca con le forbici **BE3** Marco si taglia con il coltello	incidere	**34**
5	**BE1** Marco taglia il fieno con la falce **BE2** Cristina taglia i capelli alla bambola con le forbici	trim	**51**
6	**BE1** Il vandalo taglia il cavo telefonico **BE2** Il macchinario taglia il legno	dividere	**14**

Table 3.2: Action Types identified by the verb *tagliare*.

3.1 Cut a shape out of something

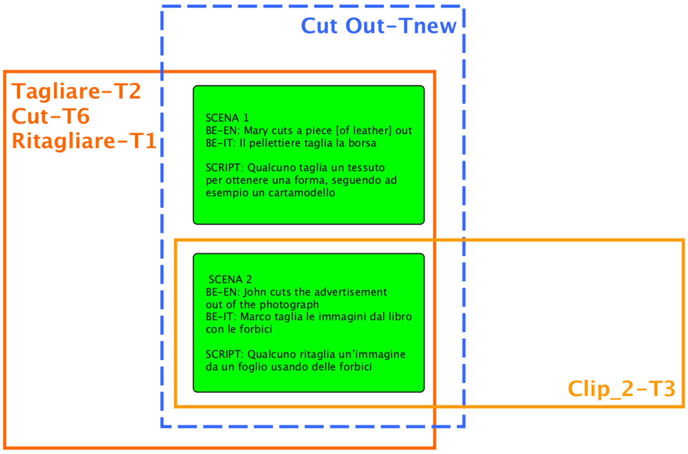

Cut Out–Tnew
Tagliare–T2
Cut–T6
Ritagliare–T1
SCENA 1
BE–EN: Mary cuts a piece [of leather] out
BE–IT: Il pellettiere taglia la borsa
SCRIPT: Qualcuno taglia un tessuto per ottenere una forma, seguendo ad esempio un cartamodello
SCENA 2
BE–EN: John cuts the advertisement out of the photograph
BE–IT: Marco taglia le immagini dal libro con le forbici
SCRIPT: Qualcuno ritaglia un'immagine da un foglio usando delle forbici
Clip_2–T3

Figure 3.1: Mapping of *cut-t6*, *tagliare-t2*, and Action Types of the locally equivalent verbs.

To a speaker of both English and Italian, just by observing Tables 3.1 and 3.2, it is possible to trace similarities between Types that seem to be directly mappable, as in the case of *tagliare-t2* and *cut-t6*. Extending the observation also to locally equivalent verbs, *ritagliare-t2* is perfectly aligned with the prototypical concept expressed by *tagliare-t2* and *cut-t6*. Locally equivalent to *cut* is *extract*, that is part of IMAGACT but its primary variation does not comprise any mappable Action Type. From *ritagliare*, the locally equivalent *cut out* is not present in the resource, but it is highly related and mappable. Moreover, extending the analysis to other IMAGACT lemmas, not retrieved through the actual explicit relations, *clip_2-t3* lists as locally equivalents *cut out* and *ritagliare*, and its BE (*John clips the picture out*) refers effectively to the same eventuality considered for this alignment, even if limited to events that exclusively requires scissors as tool. For this reason, two videos are needed, in order to show both an action performed with a tool with *two opposite edges* (e.g. scissors), and one using a *single edged* tool (e.g. boxcutter).

From the analysis of data collected for each verb in the relative Types, it can be said that *cut*, *tagliare* and *ritagliare* are more general than *clip_2*, while the verb *cut out*, not present in the resource, should refer to both the events. Consequently, the verbs here considered set relations as reported in Figure 3.1.

3.2 Cut to shorten the length of something

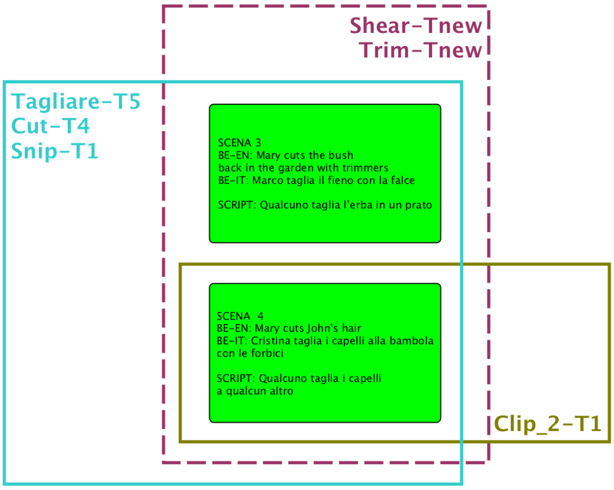

Shear-Tnew
Trim-Tnew
Tagliare-T5
Cut-T4
Snip-T1
SCENA 3
BE-EN: Mary cuts the bush
back in the garden with trimmers
BE-IT: Marco taglia il fieno con la falce
SCRIPT: Qualcuno taglia l'erba in un prato
SCENA 4
BE-EN: Mary cuts John's hair
BE-IT: Cristina taglia i capelli alla bambola
con le forbici
SCRIPT: Qualcuno taglia i capelli
a qualcun altro
Clip_2-T1

Figure 3.2: Mapping of *cut-t4*, *tagliare-t5*, and Action Types of the locally equivalent verbs.

Another case of direct mapping regards *cut-t4* and *tagliare-t5*, where the verbs instantiate the same concept: shorten the length of an object – or a set of objects under a collective name – through a cutting tool. This definition fits the BE of both the types. Moreover, the English BE2 expresses the possibility to use the cutting tool as subject in the construction of the sentence:

in fact, in the IMAGACT annotation infrastructure, the *lawnmower* is indicated as *CAUSER*, i.e. an inanimate agent. Such possibility is productive also in Italian, even if it is not emerged from the corpus. Another interesting aspect, highlighted only by the Italian BE2, but present in the standardisation of both verbs' Types, is the possibility to perform the action on humans (*cutting hair/nails/beard*), as well as on inanimate objects (*grass/hay/field*). It seems that in both languages such distinction does not create semantic restrictions, as testified by the fact that both the annotators, in parallel, did not considered the hypothesis to divide such occurrences in distinct Types. It needs to be considered, in the phase of production of visual contributions for this mapping, the possibility to realise not one but two scenes. In this way, it could be possible to maximise the coverage provided by the action ontology. Such choice is coherent with the IMAGACT meta-categories navigation system[3], which enables the visitor to explore and identify the action eventuality exclusively by means of the visual information and without linguistic constrains. Thus, the scene referred to the action performed on the human body should be categorised as *Action referring to the body*, while the other as *Modification of the object*. In the case of one scene only, it will be excluded the possibility to refer the eventuality depicted by the unique scene to the orphan category, causing the loss of linguistic information. It would be strategic in prevision of extension to other languages through the IMAGACT4ALL interface (Brown et al., 2014; Moneglia et al., 2014a,b). Both the verbs *tagliare* and *cut* show a high degree of generality, and probably other languages should not agree, preferring more specific lemmas for each situation. The verbs *trim* and *shear*, locally equivalent to *cut-t4*, are not covered in the resource, while instead the *snip-t1* results mappable on

[3] See 2.1.3.

the same event, also if it is not specified by the annotators in the equivalences proposed for *cut-t4*. Another locally equivalent verb not considered, but that can be referred to an event performed on a person (i.e. *cutting someone's hair*) is *clip_2-t1*, that picks up circumstances explicitly referred to that eventuality. In Figure 3.2 the resulting mapping.

3.3 Cut to remove

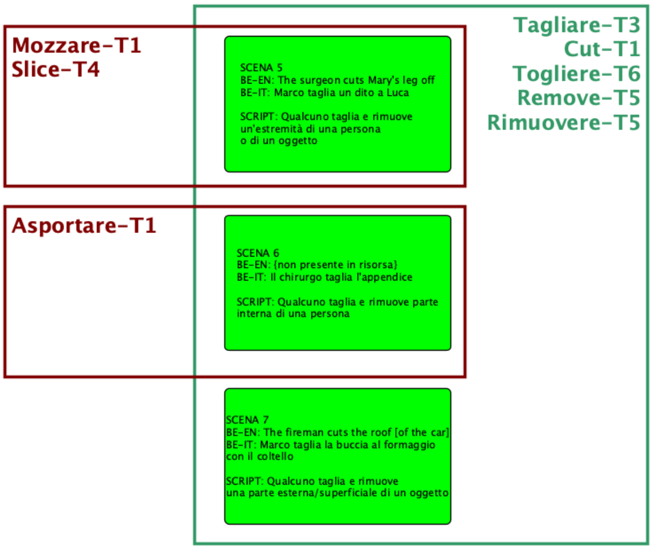

Mozzare–T1
Slice–T4
Asportare–T1
Tagliare–T3
Cut–T1
Togliere–T6
Remove–T5
Rimuovere–T5
SCENA 5
BE-EN: The surgeon cuts Mary's leg off
BE-IT: Marco taglia un dito a Luca
SCRIPT: Qualcuno taglia e rimuove un'estremità di una persona o di un oggetto
SCENA 6
BE-EN: {non presente in risorsa}
BE-IT: Il chirurgo taglia l'appendice
SCRIPT: Qualcuno taglia e rimuove parte interna di una persona
SCENA 7
BE-EN: The fireman cuts the roof [of the car]
BE-IT: Marco taglia la buccia al formaggio con il coltello
SCRIPT: Qualcuno taglia e rimuove una parte esterna/superficiale di un oggetto

Figure 3.3: Mapping of *cut-t1*, *tagliare-t3*, and Action Types of the locally equivalent verbs.

Less intuitive than the previous mapping proposals is the alignment between *cut-t1* and *tagliare-t3*. By analysing BEs and data from the locally equivalent verbs (*rimuovere* and *togliere* for Italian, *chop* for English), a common ground seems to be the fact that BE1 for both the verbs is an instance of the concept of removing a part from the whole: the use of the particle *off* in the English BE (*John cuts the head off of the flower with the knife*) indicates a trait of removal, that is possible to trace also in the Italian BE (*the rind* is *cut/removed* from *the cheese*). The English BEs show that *cut-t1* clusters rather heterogeneous standardisation, some of them well-fitted with the sense of removal, while others evoke a sense of reduction into pieces. Consider the examples in 3.3.1 and 3.3.2:

(3.3.1) removal of a part from the whole:

a. John cuts a portion off of the cheese

b. John cuts the stick from the three

(3.3.2) reduction into pieces:

a. The butcher cuts the meat

b. John cuts the wood

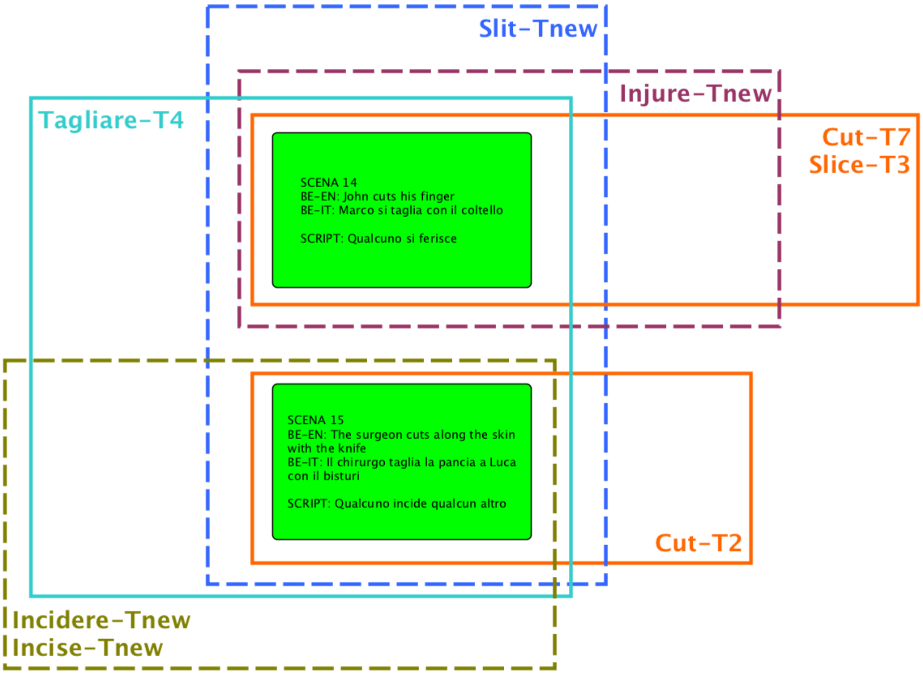
c. The baker cuts the dough

Moreover, they differ also from the aspectual point of view: on one side, actions that are punctual (3.3.1); on the other, actions that are durative or protracted (3.3.2), more close to activities or accomplishments. The set of standardisations in line with the sense of 3.3.2 are clearly out of context if gathered under *cut-t1*. The actions expressed by 3.3.2 are in line with those expressed by *chop-t1*, proposed as locally equivalent to *cut-t1*, but clearly not in line with the sense of removal expressed by the rest of the standardisations of *cut-t1* and *tagliare-t3*. For this reason *chop-t1* has been excluded from the mapping, and the set of standardisation related to 3.3.2 removed from *cut-t1*.[4]

[4] These standardisations, and the locally equivalence with *chop-t1*, have been migrated under *cut-t5*, discussed in 3.6.

By extending the analysis to Italian verbs in relation of local equivalence with *tagliare-t3*, the sense expressed by 3.3.1 is shared also by *rimuovere-t5*, *remove-t5* and *togliere-t6* (more general verbs), and *asportare-t1* (specific verb, referred to surgical removal of organs and similar). Not listed as locally equivalents but highly qualified for this mapping, *mozzare-t1* and *slice-t4* have been added. The result of the mapping is shown in Figure 3.3.

3.4 Cut a surface



SCENA 14
BE-EN: John cuts his finger
BE-IT: Marco si taglia con il coltello

SCRIPT: Qualcuno si ferisce

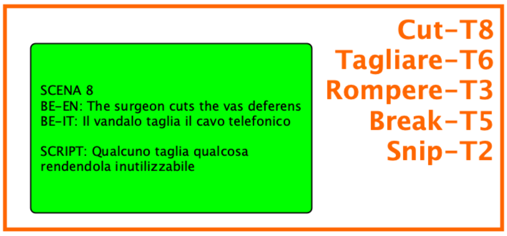
SCENA 15
BE-EN: The surgeon cuts along the skin with the knife
BE-IT: Il chirurgo taglia la pancia a Luca con il bisturi

SCRIPT: Qualcuno incide qualcun altro

Figure 3.4: Mapping of *cut-t2*, *cut-t7*, *tagliare-t4*, and Action Types of the locally equivalent verbs.

The case of *tagliare-t4*, *cut-t7* and *cut-t2* is a clear example of different granularity in action classification operated by two languages. In fact, the Italian *tagliare-t4* refers without distinction to both the eventualities that in English are kept separated by *cut-t7* and *cut-t2*, respectively an *accidental injury* and an *incision*. In other words, the precision in the execution of the action was considered as a prominent feature for English annotators (+[*precision*] in *cut-t2*; −[*precision*] in *cut-t7*); instead, for Italian annotators, this feature was not as strong to require the distinction of the corpus occurrences in two Action Types. It follows that two scenes are needed, both covered by *tagliare-t4*, one each for *cut-t7* and *cut-t2*. For what regards locally equivalent verbs, *injure* is listed for *cut-t7* and *slit* for *cut-t2*, which are highly relevant for the disambiguation of the senses but both are missing in the resource. On the Italian side, there is a very specific verb, *incidere*, that seems to remarks the concept of *cut-t2*. Deepening the analysis, under *incidere* there is only one Action Type, referring of actions related to processing techniques of wood, metal or stone, such as carving, impressing, and sculpting. For this reason, a new Action Type is required, *incidere-tnew*, and the English counterpart *incise* considered for the mapping. The results are shown in Figure 3.4.

3.5 Cut to damage



SCENA 8
BE-EN: The surgeon cuts the vas deferens
BE-IT: Il vandalo taglia il cavo telefonico

SCRIPT: Qualcuno taglia qualcosa rendendola inutilizzabile

Figure 3.5: Mapping of *cut-t8*, *tagliare-t6*, and Action Types of the locally equivalent verbs.

The concept expressed in the mapping between *cut-t8* and *tagliare-t6* is particularly interesting, because it sits in the grey area of the semantic field in between *cutting* and *breaking* events. Moreover, it is particularly complex to treat, because BEs and occurrences gathered under *tagliare-t6* need to be revised. It is enough to look at its BEs (3.5.1-3.5.2):

(3.5.1) **BE1** Il vandalo taglia il cavo telefonico

'The vandal cuts the telephone cable'

(3.5.2) **BE2** Il macchinario taglia il tegno

'The machinery cuts the wood'

To an Italian speaker, it is immediately clear that the two BEs may refer to distinct eventualities. BE1 (3.5.1) and all the similar standardisations describe an event that has a precise effect, the degradation of an object by means of cutting: the object looses its function as a result of the action. This and similar events have parallel interpretation in *cut-t8*, that refers the same. From locally equivalent verbs, *snip-t2* finds a place in the mapping. Given the closeness with *breaking* events, also *break* and *rompere* have been checked, and *break-t5* and *rompere-t3* refers to this kind of eventualities. The eventuality described in BE2 (3.5.2) and related standardisations resulted in not being correctly gathered under *tagliare-t6*, and thus have been moved to *tagliare-t1*;[5] also the locally equivalent *split-t1* refers to that eventuality. Figure 3.5 shows the result of the mapping between *cut-t8*, *tagliare-t6* and locally equivalent verbs.

3.6 Cut to partition

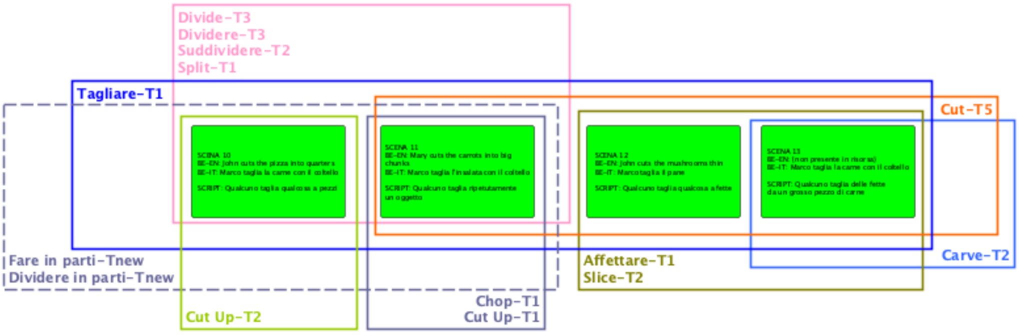

Divide-T3
Dividere-T3
Suddividere-T2
Split-T1
Tagliare-T1
Cut-T5
SCENA 10
BE-EN: John cuts the pizza into quarters
BE-IT: Marco taglia la carne con il coltello
SCRIPT: Qualcuno taglia qualcosa a pezzi
SCENA 11
BE-EN: Mary cuts the carrots into big chunks
BE-IT: Marco taglia l'insalata con il coltello
SCRIPT: Qualcuno taglia ripetutamente un oggetto
SCENA 12
BE-EN: John cuts the mushrooms thin
BE-IT: Marco taglia il pane
SCRIPT: Qualcuno taglia qualcosa a fette
SCENA 13
BE-EN: (non presente in risorsa)
BE-IT: Marco taglia la carne con il coltello
SCRIPT: Qualcuno taglia delle fette da un grosso pezzo di carne
Fare in parti-Tnew
Dividere in parti-Tnew
Affettare-T1
Slice-T2
Carve-T2
Cut Up-T2
Chop-T1
Cut Up-T1

Figure 3.6: Mapping of *cut-t5*, *tagliare-t1*, and Action Types of the locally equivalent verbs.

In the mappings proposed in 3.3 and 3.5, some occurrences have been disjoined from the Action Types due to the fact that they seem to describe eventualities in which the object could be cut in multiple parts, reduced in small pieces, and the action performance does not imply that the object looses its function. This is the case of *cut-t5* and *tagliare-t1*, and the locally equivalent verbs referring this kind of actions (Figure 3.6). Among those, *split-t1* predicates the same eventuality, without an explicit need of a sharp tool, thus it is more general than *cut* in these eventualities. For Italian, locally equivalent to *tagliare-t1* is *suddividere-t2*, that refers to the concept

[5] See 3.6 for details on its mapping.

of partitioning, without specifying a tool. Thus, as for English, also here the local equivalence widens to more general meanings, for which the sharp implement is no more mandatory.

By extending the analysis to verbs not listed as locally equivalents, *divide-t3* and *dividere-t3* are definitely close to the concept behind these Action Types. It is possible to say that *cut-t5* and *tagliare-t1* are in relation of troponymy with *divide-t3* and *dividere-t3*, as they predicate a *manner* of dividing an object by means of a sharp tool. The idea of repeated cutting is expressed also by *chop-t1*, and two Action Types of a sublemma of *cut*: *cut_up-t1* and *cut_up-t2*. The subtle difference between these two is that in *cut_up-t1* the standardisations seem to highlight the repeated cutting with the result of very small pieces, while *cut_up-t2* predicates of reduction into pieces, aimed at partitioning the whole object. This suggests that more than one video prototype is needed for this mapping. Moreover, even if not traceable through the relations expressed by the resource, other verbs can be placed in this mapping, by the fact that they predicate the portioning of an object, and encode the shape of the resulting parts: *slice-t2* and *affettare-t1*. Similar to these is also *carve-t2* but, according to the standardisation it gathers, it seems to refer only to the action performed on big pieces of cooked meat (e.g. roast). Unfortunately, *fare in parti* and *dividere in parti*, linked to these verbs as locally equivalent, are not present in IMAGACT.

3.7 Cut to give shape

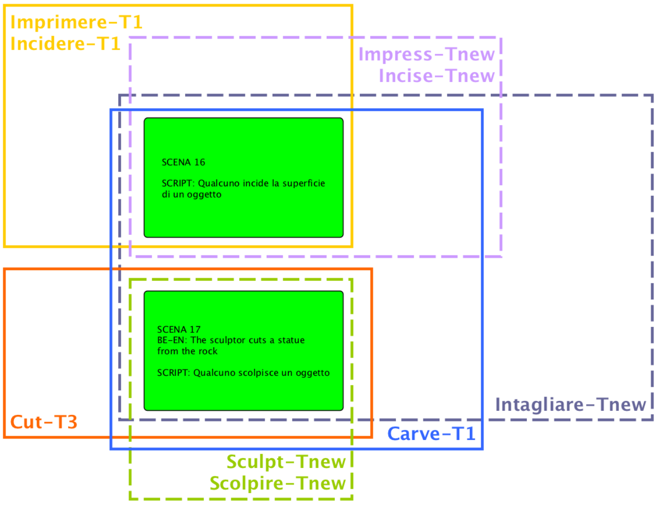

Imprimere–T1
Incidere–T1
Impress–Tnew
Incise–Tnew
SCENA 16
SCRIPT: Qualcuno incide la superficie di un oggetto
SCENA 17
BE–EN: The sculptor cuts a statue from the rock
SCRIPT: Qualcuno scolpisce un oggetto
Cut–T3
Intagliare–Tnew
Carve–T1
Sculpt–Tnew
Scolpire–Tnew

Figure 3.7: Mapping of *cut-t3* and Action Types of the locally equivalent verbs.

One Action Type finds no inter-linguistic parallelism in the comparison between eventualities predicated by *cut* and *tagliare*. In fact, the BEs and the standardisations gathered under *cut-t3* refer to the action of shaping an object by removing superfluous parts with a cutting tool. This meaning is also expressed by *carve-t1*, locally equivalent to *cut-t3*. No Action Type listed under *tagliare* expresses this eventuality.

The concept of *cutting to shape* is already expressed by one of the proposed mappings[6] The subtle distinction resides in the fact that in *cut-t3* the removed part is not the focus (as it is for *cut-t6* and *tagliare-t2*) but, on the contrary, what remains is the goal of the action. A parallel interpretation in Italian concerns actions predicated by *scolpire* and *intagliare*, which are not present in IMAGACT. Present in the resource, but not traceable through the links offered by the actual infrastructure, are *imprimere-t1* and *incidere-*

[6] See 3.1.

t1, while *impress* and *incise*, locally equivalent to those, are not part of the resource. It is proposed to produce 2 videos, one for *cutting a statue* and one for *incision*, as it is possible to refer to both with some of the verbs proposed here. Figure 3.7 shows the resulting mapping.

3.8 Results of the mapping

The mapping proposals traced in the previous sections still need to be reviewed, discussed and revised with English mother-tongue informers, in order to validate and actively implement them in IMAGACT. Missing verbs and linking between them is a consequence of the building process of the resource,[7] that relies on spoken corpora and, for this reason, only verbs that occurred in those corpora are listed.

[7] The builing of IMAGACT is described in 2.3.

Chapter 4

NLP Experiments with IMAGACT dataset

In this chapter experiments in exploiting IMAGACT for NLP are reported. In particular, my colleagues and I have conducted two linking experiments with other resources (namely, BabelNet and T-PAS) in order to extend the information conveyed by each action concept; moreover, an action concepts vector space model has been built by exploiting IMAGACT multilingual feature, and an action verb embedding based on co-reference vectors.

4.1 Linking IMAGACT and BabelNet

This section presents the linking of IMAGACT with BabelNet, with the aim of connecting videos contained in IMAGACT with *actionally*-related concepts in BabelNet. The linking is based on a machine learning algorithm that exploits the lexical information (i.e. verbs in multiple languages) of the two resources. The algorithm has been firstly trained and tested on a manually annotated dataset, and then it was run on the complete set of data, allowing to connect 773 IMAGACT action concepts with 517 BabelNet synsets.[1]

The aim of this experiment is to link IMAGACT and BabelNet interlinguistic concepts through videos prototypes of actions. In fact, the BabelNet objects are already enriched with visual objects, though this information contains static images which are inadequate for representing action concepts. In this way, adding video is very desirable, and it would be a natural extension of BabelNet visual inventory.

BabelNet itself is the product of a mapping between different resources, and other examples of resource mappings are WordFrameNet (Laparra and Rigau, 2009), an integration of both WordNet and FrameNet by means of a knowledge-based WSD algorithm, the Predicate Matrix De Lacalle et al. (2016), which gathers predicate information from multiple lexical resource, and YAGO Fabian et al. (2007), a huge semantic knowledge base derived from WordNet, Wikipedia and GeoNames.

4.1.1 BabelNet

BabelNet[2] (Navigli and Ponzetto, 2012a) is both a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms, and a semantic network which connects concepts and named entities in a very large network of semantic relations, made up of about 16 million entries. It has been created by mapping together the Wikipedia[3] encyclopedia and

[1] The experiment herein described has been published in Gregori et al. (2016). My personal contribution regards the design of the training set, the manual annotation and the evaluation of the results.

[2] http://babelnet.org

[3] https://www.wikipedia.org

the WordNet[4] thesaurus (Miller et al., 1990). At present, BabelNet 4.0 contains 284 languages and it is the widest multilingual resources available for semantic disambiguation (see Table 4.2 for detailed statistics).

Concepts and entities in BabelNet are represented as BabelSynsets (BS), an extensions of WordNet synsets: a BS is an unitary concept identified by several kinds of information (semantic features, glosses, usage examples, etc.) and related to lemmas (in any language) which have a sense matching with that concept. BSs are not isolated, but connected together by semantic relations. Moreover, BabelNet received a large contributions from its mapping with other resources such as ImageNet, GeoNames, OmegaWiki (along with many others), which increased its information beyond the lexicon, and produced a wide-ranging multimedia knowledge base. Table 4.1 reports on the list of linked resources, with a short description for each one; Figure 4.1 shows the first resources linked in the BabelNet network.

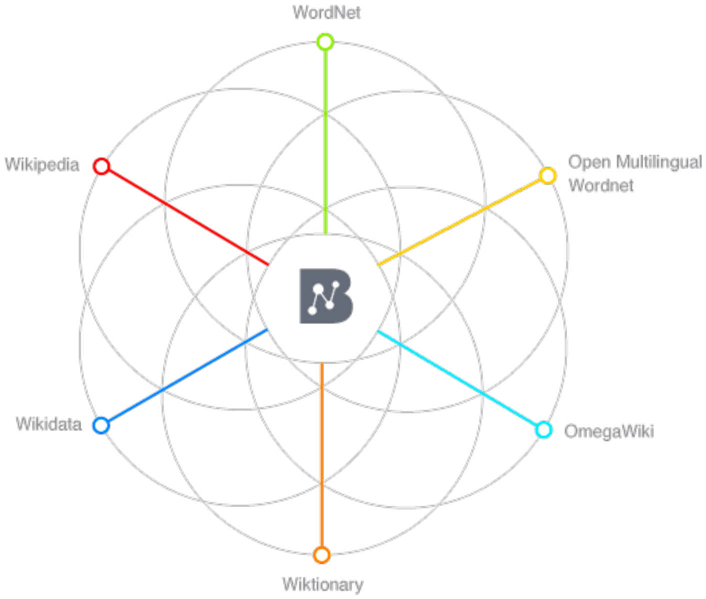


Figure 4.1: BabelNet network of starting resources.

[4] https://wordnet.princeton.edu

Resource	Description
WordNet	a popular computational lexicon of English (version 3.0)
Wikipedia	the largest collaborative multilingual Web encyclopedia (February 2018 dump)
OmegaWiki	a large collaborative multilingual dictionary (January 2017 dump)
Wiktionary	a collaborative project to produce a free-content multilingual dictionary (February 2018 dump)
Wikidata	a free knowledge base that can be read and edited by humans and machines alike (February 2018 dump)
Wikiquote	a free online compendium of sourced quotations from notable people and creative works in every language (March 2015 dump)
VerbNet	a Class-Based Verb Lexicon (version 3.2)
Microsoft Terminology	a collection of terminologies that can be used to develop localized versions of applications (July 2015 dumps)
GeoNames	a free geographical database covering all countries and containing over eight million placenames (April 2015 dump)
ImageNet	an image database organized according to the WordNet hierarchy (2011 release)
FrameNet	a lexical database of English that is both human- and machine-readable (version 1.6)
WN-Map	automatically generated mappings among WordNet versions (2007 release)
Open Multilingual WordNet	a collection of wordnets available in different languages (downloaded in January 2017)

Table 4.1: BabelNet 4.0 collection of linked resources.

Number of languages	284
Total number of Babel synsets	15,780,364
Total number of Babel senses	808,974,108
Total number of concepts	6,113,467
Total number of Named Entities	9,666,897
Total number of lexico-semantic relations	277,036,611
Total number of glosses (textual definitions)	91,218,220
Total number of images	54,229,458
Total number of Babel synsets with at least one domain	2,637,407
Total number of Babel synsets with at least one picture	10,522,922
Total number of sources	47

Table 4.2: BabelNet 4.0 statistics.

4.1.2 Training and test set

A manually annotated dataset of 50 prototype videos of action concepts and 57 BabelSynsets, for a total of 2,850 judgments, was created in order to test the algorithm and evaluate the results. The sampling was carried on in two steps. First of all, a purely actional semantic area has been selected by taking BSs and videos linked to 7 English general action verbs, which are very frequent in the language use.[5] The wide variation of these verbs allowed us to obtain a big set of concepts, with a high variation in terms of frequency and generality. On this set, a second sampling has been performed by preserving the variability in terms of number of connected verbs, that is a measurable parameter in both the resources.

[5] The verbs used to retrive BSs and videos candidates are: *put*, *move*, *take*, *insert*, *press*, *give*, *strike*.

Each ⟨BS,Scene⟩ pair has been evaluated to check if the scene could be appropriate in representing the BS. Three annotators compiled the binary judgment table and values shared by at least 2 have been used. The measured Fleiss' kappa inter-rater agreement for this task was 0.74 [6].

Finally, the dataset has been split in a training set and a test set, with the proportions of 80% and 20% respectively (10 randomly chosen scenes for the test set and the remaining 40 scenes for the training set).

4.1.3 Algorithm

The algorithm developed for this task uses Machine Learning techniques, by exploiting the training set. Similarly to a previous experiment (Gregori et al., 2015), the features are extracted from the lexical items belonging to both the candidate BabelSynset and its neighbours.[7] Table 4.3 reports on the 17 languages common to both BabelNet and IMAGACT, detailing the relative number of verbs in each, and constitutes the quantitative data exploited by the algorithm.

The basic features used for this experiment are:

- ns: the number of verbs connected to the Scene;
- nb: the number of verbs connected to the BS;
- nsb: the number of verbs that are shared between the Scene and the BS;

These 3 features have been calculated for each candidate BS and for the ones which are semantically related to it. To this aim, the 8 BabelNet's verb semantic relations have been considered, and for each BS 8 groups of related

[6] The manually annotated training set is publicly available at http://www.lablita.it/app/bnim/bn-im-dataset2016-en.html

[7] This test is based on BabelNet 3.6; the data was extracted using the Java API (Navigli and Ponzetto, 2012b).

Language	**BN Verbs**	**IM Verbs**
English (EN)	29,738	662
Polish (PL)	9,660	1,193
Chinese (ZH)	9,507	414
Italian (IT)	7,184	646
Spanish (ES)	6,159	736
Russian (RU)	4,975	86
Portuguese (PT)	4,624	805
Arabic (AR)	3,738	804
German (DE)	3,754	990
Norwegian (NO)	1,729	115
Danish (DA)	1,685	646
Hebrew (HE)	1,647	160
Serbian (SR)	858	1,096
Hindi (HI)	831	470
Urdu (UR)	233	478
Sanskrit (SA)	33	276
Oriya (OR)	6	160
Total	86,361	11,130

Table 4.3: The 17 shared languages of BabelNet (BN) and IMAGACT (IM) with verbal lemma counts.

synsets have been extracted, each one containing the set of BS connected to the main one by the same relation. Table 4.4 shows the list of relations between the verbal BSs ranked by their relevance values for this task; this value is measured with Information Gain on the annotated dataset.

Then, ns, nb and nsb are calculated for each group by summing the values of the BSs belonging to it. The feature set is comprised of 27 features: 3 features for the main BS and 3 features for each BabelNet relation. The set of candidates consists of all the possible BSs for each verb connected to the scene. A machine learning algorithm (a Linear Support Vector Machine (SVM) classifier with a RBF kernel) was trained on the annotated dataset.

Beside the algorithm, a baseline is determined by calculating the ratio $\frac{nsb}{nb+ns}$ for each pair and setting a threshold of 0.04, that maximizes the F-measure on this dataset.

BabelNet relations	***IG* value**
Hyponym	0.057
Hypernym	0.026
Also See	0.019
Verb Group	0.019
Gloss Related	0.009
Entailment	0.003
Antonym	0.000
Cause	0.000

Table 4.4: Relations between verbal BSs.

4.1.4 Results

The algorithm was run on the training set and evaluated on the test set; the results are reported in Table 4.5. The results in terms of F-measure are not so satisfying, and the value obtained with the algorithm is barely better than the baseline. Despite this, it is important to consider the difference with the

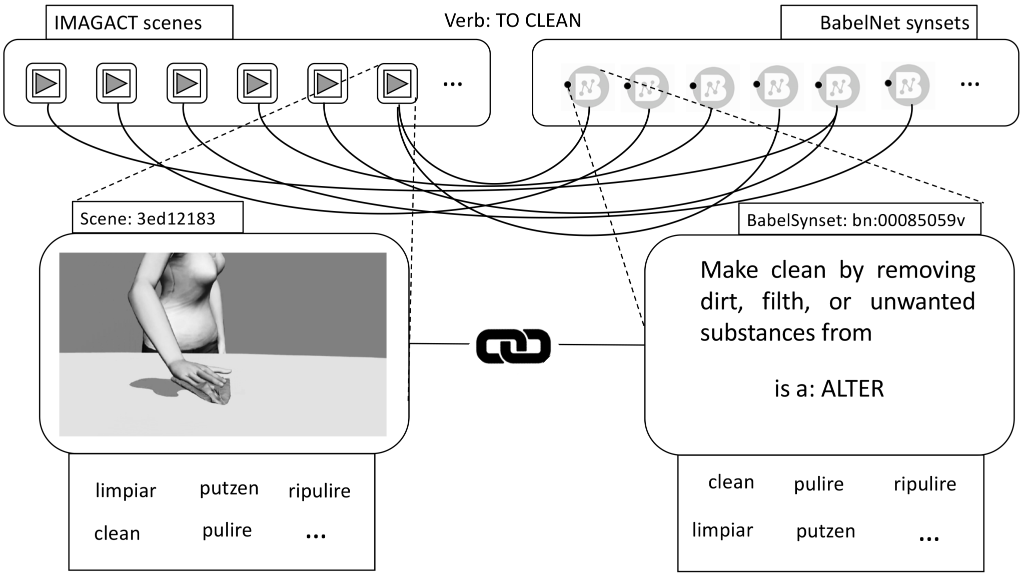


Figure 4.2: An example of the linking between IMAGACT scenes and BabelNet synsets.

baseline in terms of precision and recall, since precision is more important for this task: for this reason, the algorithm provides a much more reliable result compared to the baseline.

It has to be pointed out that a low recall is mainly caused by multiple possibilities in the interpretation of a scene from different points of view. Consider the scene linked to the English verb *to throw* and described by the

	Baseline	**ML Algorithm**
	$th = 0.04$	27 features
Precision	0.580	0.833
Recall	0.529	0.441
F-measure	0.553	0.577

Table 4.5: Precision, Recall and F-measure of BSs-to-scenes linking task calculated on the test set for the algorithm and the baseline.

sentence *John throws the ball to Mark*: it can represent not only a sense of *throw*, but also senses of other verbs, like *to play* or *to catch*, that refer to different semantic concepts. In these cases, the scene in IMAGACT is not linked to the alternative verbs, but it can be described with them (i.e. *John and Mark play with the ball*, *Mark catches the ball*). For this reason, the manual annotation provides more BS-to-scene relations than an algorithm can foresee on the basis of a pure lexical match, causing a low recall value. Table 4.6 reports on statistics about the linking process.[8]

IM Scenes linked to BS	773
BS linked to Scenes	517
IM English Verbs related to Scenes	544
BabelNet English Verbs related to BS	1,100

Table 4.6: IMAGACT-BabelNet linking in numbers.

Switching to Machine Learning had a strong impact on this linking task. The main advantage from the previous linking experiment (Gregori et al., 2015) is that now the number of BSs that can be assigned to each scene is variable, depending on the different reference possibilities that the BSs have. This is coherent with the BabelNet structure where it is possible to find very general concepts, that can be represented by several action prototypes, and specific ones, for which one prototype is enough to provide a clear representation.

As an example, the BS “bn:00090224v” (*Put into a certain place or abstract location*) expresses a general concept and is linked to 72 scenes, com-

[8] The results are browsable at the page `http://lablita.it/app/bnim/`, where it is possible to search for scenes or BS, and retrieve the corresponding linked items.

prising the actions involving one or more objects or a body part, relating to different ways of *putting* (like *inserting*, *throwing*, *attaching*,...) or to different states of the Theme (e.g. solid or liquid). Conversely, the BS "bn:00084326v" (*Fasten with buttons*) is much more specific and is linked to only one scene (ID: c17d7346), which depicts a man fastening his jacket.

4.2 Linking IMAGACT and T-PAS

This section describes the analysis of a mapping between IMAGACT and T-PAS, obtained through a rule-based algorithm which converts argument structures in thematic roles. This experiment is focused on an empirical analysis of argument and thematic structures in Italian verbs to see if, and to which extent, a rule-based system is able to produce thematic structures, and how these results can be exploited for a mapping purpose.[9]

The linking between argument and thematic structures of a predicate is a debated complex task in linguistic theories (Baker, 1997; Bowerman, 1990; Pinker, 2009, among others), which belongs to the *syntax-semantics interface*. The experiment is focused on an empirical analysis of argument and thematic structures in Italian verbs. The objective is to evaluate whether, and to which extent, a rule-based system is able to produce thematic structures, and to verify how these results can be exploited for a mapping purpose. Table 4.7 shows the total and shared quantitative data of the two resources.

[9] The experiment herein described has been published in Ravelli et al. (2017). My personal contribution regards the general design of the experiment, the creation of the Gold Standard and the evaluation of the results.

	IMAGACT	T-PAS
Total Verbs	777	1,000
Total Types - *t-pass*	1,429	4,241
Shared Verbs	248	
Shared Types - *t-pass*	421	1,153

Table 4.7: Comparison of data in IMAGACT and T-PAS.

4.2.1 T-PAS

T-PAS[10], Typed Predicate Argument Structures (Jezek et al., 2014), is a repository of verb patterns acquired from corpora by manual clustering distributional information about Italian verbs. For every typed structure (henceforth, *t-pas*), the specification of the expected Semantic Type (ST) for each argument slot is provided. T-PAS accounts for the following argument positions: *subject, object, indirect object, complement, adverbial* and *clausal.* A description of the sense, in the form of an *implicature*, is also linked to the *t-pas*.

Example 4.2.1 shows the *t-pas*#2 of the verb *abbattere* (∼*knock-down*): the STs [[Human]] and [[Event]] are specified for the subject position (as alternatives) and [[Building]] for the object position.

(4.2.1) [[Human | Event]-subj] *abbattere* [[Building]-obj]
implicature:[[Human | Event]] *distrugge, butta giù* [[Building]]
example: "Il muratore abbatte la parete." ("The bricklayer knocks down the wall.")

The STs aim at generalizing over the set of lexical items observed in a certain position for a particular sense of the verb. For instance, in Example

[10] http://tpas.fbk.eu/

4.2.1, the ST [[Building]] generalizes over the lexical item *parete* (*wall*). STs are drawn from a list of about 230 types[11] and are also organized in a hierarchy, in which the elements are linked by an "IS-A" relation (Jezek et al., 2016). Table 4.8 presents a section of the hierarchy in which it is shown that [[Plane]] IS-A [[Vehicle]], [[Vehicle]] IS-A [[Machine]] and so on.[12] If no generalization is possible, the set of lexical items found in the argument position is listed.

...

- ▷ [[Artifact]]
  - ▷ [[Machine]]
    - ▷ [[Vehicle]]
      - ▷ [[Plane]]
      - ▷ [[Road Vehicle]]
      - ▷ ..

Table 4.8: Section of the STs hierarchy.

Each *t-pas* corresponds to a distinct sense of the verb and it is identified and defined by analysing instances of the verb in a corpus, following the lexicographic procedure called Corpus Pattern Analysis (Hanks, 2004; Hanks and Pustejovsky, 2005).[13] The corpus instances are then associated to the corresponding *t-pas*. T-PAS currently contains 1000 verbs. The reference corpus is a reduced version of ItWAC (Baroni and Kilgarriff, 2006).

[11] For details on the list creation see Jezek et al. (2014).

[12] The same list has been used for the English resource PDEV (Hanks and Pustejovsky, 2005), `http://pdev.org.uk`. The hierarchy can be found in `http://pdev.org.uk/onto`.

[13] According to the CPA procedure, after analysing a random sample of 250 concordances of the verb in the corpus, each *t-pas* is defined by recognizing its relevant structure and identifying the STs for each argument slots.

4.2.2 Mapping through rule-based conversion

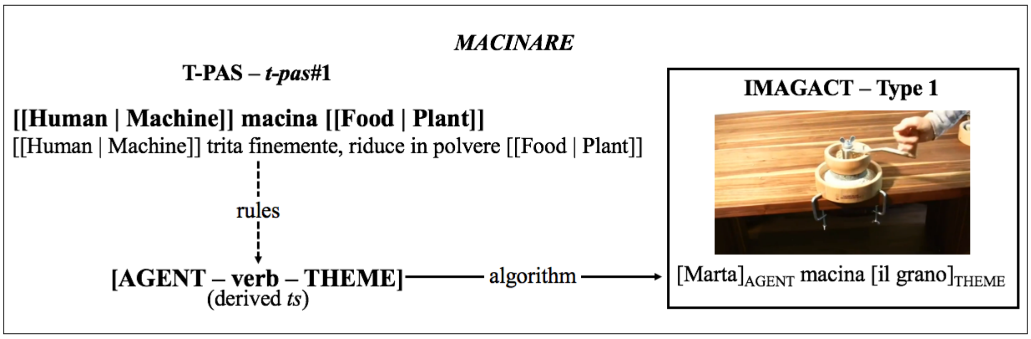

MACINARE
T-PAS – t-pas#1
[[Human | Machine]] macina [[Food | Plant]]
[[Human | Machine]] trita finemente, riduce in polvere [[Food | Plant]]
rules
[AGENT – verb – THEME]
(derived ts)
algorithm
IMAGACT – Type 1
[Marta]$_{AGENT}$ macina [il grano]$_{THEME}$

Figure 4.3: An example of the mapping between IMAGACT and T-PAS for the verb *macinare* (∼*grind*).

The experiment aims at finding the best semantic match between a verb Action Type in IMAGACT and the *t-pas*s of the same verb in T-PAS, the two referring to the same action concept. Notice that it is possible that an Action Type in IMAGACT is mapped to more than one *t-pas* due, for instance, to different possible verb alternations that can occur inside the same Action Type. Figure 4.3 shows an example of this mapping, in which there is a match between Action Type 1 and t-pas#1 of the verb *macinare* (∼*grind*).

The mapping is done as follows. By observing a sample of verbs in the resources, first, a set of simple rules to convert the *t-pas* in a thematic structure has been defined. Considering the ST in the argument positions of the *t-pas* (e.g. [Human]-subj, [Food]-obj]), the rules aim at creating a thematic structure for the *t-pas* of the kind AG-v-TH (dotted arrow in Figure 4.3). Then, an algorithm has been implemented which applies these rules to all the *t-pas*s of a verb, and map the derived thematic structure (*derived-ts*) to the thematic structures (*ts*) of the Action Types in IMAGACT (horizontal arrow in Figure 4.3). Thus, the system compares all the *ts* in IMAGACT with all the *derived-ts* in T-PAS for the same verb, and retrieves the matches.[14] In Figure 4.3, the *t-pas#1* for *macinare* (∼*grind*) has been transformed in the structure AG-v-TH and then mapped to the *ts* of the Action Type.

```
y = ST in argument slot
for y:
if y = IS or IS-A [Abstract | State | ..]
    do not map
if obj:
    y in obj = Theme TH
    if y in subj IS or IS-A [[Animate]]:
        subj = Agent AG
    else:
        subj = Causer CA
else:
    if y in subj IS or IS-A [[Animate]]
    & verb is reflexive:
        subj = Actor AC
    else:
        subj = Theme TH
for y !=subj and obj:
    x = (ImagAct Role != AG, CA, AC, Instrument IN)
    x = y
```

Table 4.9: Rules for mapping.

[14] Notice that the mapping is only considering this information of the resources, and it does not take into account any additional training data, e.g. captions in IMAGACT or examples in T-PAS.

The rules for the conversion of a *t-pas* in a *derived-ts* have been manually created by observing a sample of 15 verbs shared by the two resources (devset). The mapping has been evaluated against a gold standard manually created by pairing the Types of other 14 verbs with the corresponding *t-pas*s. All the 29 verbs have been extracted from the 248 shared by the two resources. The selection was made preserving the variability of the verbs in the two resources, in terms of their number of Types or *t-pas*. For instance, *prendere* (*to take*) is associated with 17 *t-pas*s in T-PAS and 18 Types in IMAGACT; on the contrary *bussare* (*to knock*) has only 2 *t-pas*s and 1 Type.

Table 4.9 summarise the adopted rules. The rules consider both the ST in the argument slot and the argument slot itself, and are meant to associate a ST in an argument slot to a thematic role. For example, line 7 of Table 4.9 has to be interpreted as follows: if for the *subject* position of the *t-pas* the ST [[Animate]] (or a IS-A [[Animate]], according to the hierarchy of ST) is expected, then the AGENT role is selected (line 8). The rules also consider if the verb is in reflexive form (line 13). Moreover, if the *t-pas* registers the ST [[Abstract Entity]] (or a ST that IS-A [[Abstract Entity]]) as unique ST for any argument position (i.e. it is the only ST expected for the position), the *t-pas* was excluded from the mapping, as IMAGACT only accounts for physical actions which do not involve abstract entities.

4.2.3 Results

In order to calculate Precision and Recall of the algorithm, it has been considered that DESTINATION (DE), SOURCE (SO) and LOCATION (LO) roles can not always be discriminated.[15] The same happens for AGENT

[15] As an example, *room* is a DE in *John puts a table in the room*, a SO in *John takes the table from the room*, a LO in *John walks in the room*.

(AG) and ACTOR (AC): a human can be an agent (*John sweeps the room*) or an actor (*John bumps his head*). These limits can not be exceeded by an improvement of the rule definitions, because they are strictly dependent on the verb semantics. When calculating P and R, these derived structures have been grouped together.

Precision	**Recall**	**F-measure**
0.283	0.792	0.418

Table 4.10: Precision, Recall, F-measure of the mapping between IMAGACT and T-PAS.

It is possible to observe good values for Recall, while the Precision is very low (Table 4.10). A deeper analysis shows that in 34.61% of the cases, there is a full match with the gold standard and in 38.46% the results from the mapping include the ones expected by the gold standard. This means that in many cases the system is able to retrieve the correct matches.

Figure 4.4 shows the distribution of the main thematic structures in the Types of the whole IMAGACT ontology (in orange), in the devset (in red), compared with the *derived-ts* from T-PAS (in green). It has been verified *a posteriori* that the distribution of *ts*s in the devset is strictly comparable with the one in the whole ontology, meaning that the devset is also well-balanced in terms of the thematic structures coverage (see orange and red bars in Figure 4.4).

By using the transformational rules it has been possible to recreate all the structures that are used in IMAGACT; however, there are some discrepancies in the production of AG-v-TH, TH-v (too high) and AG-v-TH-[DE|LO|SO] (too low) (see Figure 4.4).

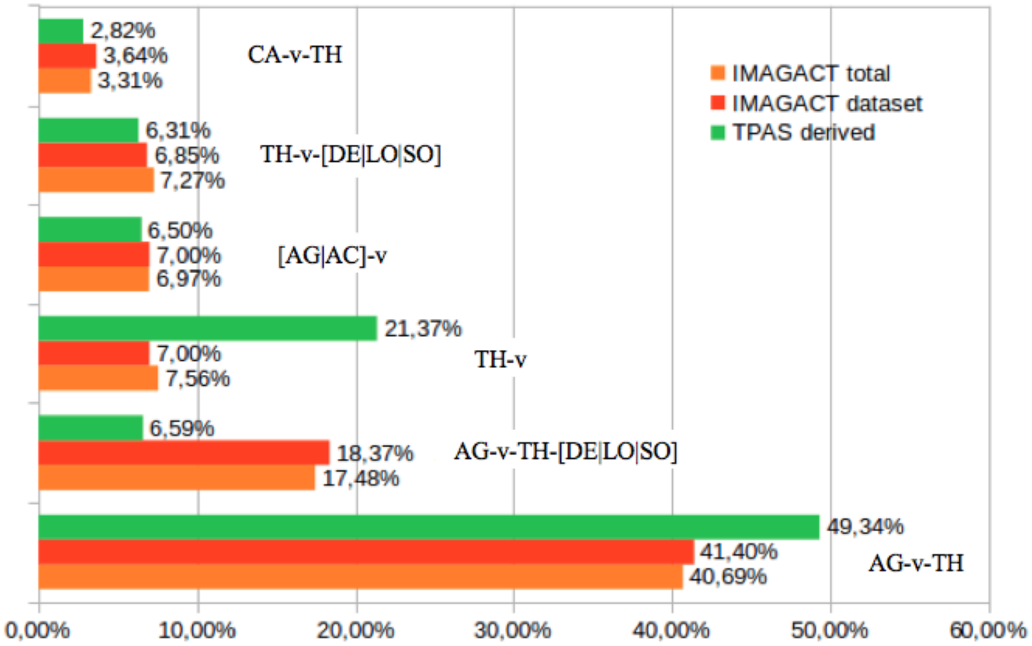


Figure 4.4: Distribution of the thematic structures.

The critical issue is represented by the AG-v-TH structure: this is the most frequent one among the IMAGACT Types and in the test set (112 over 166 Types). As an example, the sentences in 4.2.2-4.2.5 belong to 4 different Action Types of the verb *stringere* (∼clutch|close|tighten), but have the same *ts* AG-v-TH:

(4.2.2) *Marco stringe la mano a Luca* (*Marco shakes Luca's hand*)

(4.2.3) *Marco stringe le gambe* (*Marco closes his legs*)

(4.2.4) *Marco stringe i pugni* (*Marco clenches his fists*)

(4.2.5) *Marco stringe la vite* (*Marco tightens the screw*)

The same happens also for the *t-pass* of *stringere* (∼clutch|close|tighten): 3 over the 5 *derived-ts* are AG-v-TH, therefore the system produces 12 combinations over 3 attested in the gold standard. The high frequency of this structure strongly influences the final Precision and Recall results. Moreover, the *ts* AG-v-TH is not distinctive of intra-verbs Action Types: by taking all the verbs with more than one Action Type, and for which AG-v-TH is a possible *ts*, only in 38,22% of them this *ts* is present exclusively in one Action Type; in the other verbs (61.78%) the AG-v-TH structure appears in more than one Action Type.

4.3 Action Clustering and verbs embeddings

This section presents a vector representation and a clustering of action concepts based on lexical features extracted from IMAGACT.[16] Vectors for 1,010 action concepts have been computed as a matrix $C_{scenes \times verbs}$. In this vector representation, the dimensions correspond to verbs collected for the IMAGACT Ontology through the CBE[17] extension framework in 10 languages. Finally, an unsupervised clustering method has been applied on these data in order to discover action classes based on *typological closeness*.[18] Those clusters are not language-specific or language-biased, and thus constitute an inter-linguistic classification of the action domain. Moreover, action verbs embeddings have been calculated by reading the same Matrix C in the inverse way (i.e. $verbs \times scenes$).

[16] The experiments herein described have been published in Gregori et al. (2019) and Ravelli et al. (2019). My personal contribution regards the analysis and evaluation of action clusters, and the action verbs embedding idea.

[17] See 2.3.3.

[18] A similar approach has been used to represent typological data by Ryzhova et al. (2016): they built a matrix of word references $M_{nouns \times adjectives}$, in which each row corresponds to nouns from a specific semantic field and the dimensions are adjectives from different languages. An intersection of a row and a column is filled with 1 if the adjective can occur in the context and with 0 if it cannot. Given these data, they compute *typological closeness*, that is semantic similarity based on comparison of multilingual data.

4.3.1 Dataset

From the IMAGACT database, a binary matrix $C_{1010\times7881}$ has been derived, with one row per video prototype and one column per verb for the languages considered. Matrix values are the assignments of verbs to videos made by native speakers within the CBE annotation task:

$$C_{i,j} = \begin{cases} 1 & \text{if verb } j \text{ refers to action } i \\ 0 & \text{else} \end{cases} \tag{4.1}$$

In this way, the matrix C encodes the inter-linguistic lexical representation of each video prototype representing an action concept. Table 4.11 shows the number of verbs assigned by the CBE annotators for each language. It is important to notice that the task has been performed on the whole set of 1,010 scenes for each language and the differences between the number of verbs depend on linguistic factors: some examples of verb-rich languages are:

- Polish and Serbian, in which perfective and imperfective forms are lemmatized as different dictionary entries;
- German, that have particle verb compositionality;
- Spanish and Portuguese, for which verbs belong to both American and European varieties.

Judgments of applicability of a verb to a video scene rely on the semantic competence of mother-tongue annotators. An evaluation of CBE assignments has been made for Arabic and Greek in two thesis (Mouyiaris, 2019; Mutlak, 2019); results are summarized in Table 4.12.

Language	**Verbs**
Arabic (Syria)	571
Danish	646
German	990
Greek	638
Hindi	470
Japanese	736
Polish	1,193
Portuguese	805
Serbian	1,096
Spanish	736
TOTAL	**7,881**

Table 4.11: Number of verbs per language.

Language	**Precision**	**Recall**
Arabic (Syria)	0.933	0.927
Greek	0.990	0.927

Table 4.12: Precision and Recall for CBE annotation task measured on 2 languages.

4.3.2 Creating action vectors

In order to provide an exploitable vector representation of action prototypes, an approximated matrix C' has been created from C, by using *Singular Value Decomposition* (SVD) for dimensionality reduction.

SVD is a widely used technique in distributional semantics to reduce the feature space. The application of SVD to the dataset used in this experiment leads to a fixed-size feature space (300 dimensions), that is independent of the number of languages, and to an approximation matrix that smooths language-specific semantic differences. These results are highly desirable, considering that the number of languages in IMAGACT is growing continuously, and that the provided representation should be shared as far as possible, abstracting from lexico-semantic properties of single languages. Moreover, SVD approximation leads to some advantages in terms of computational processing, by removing the matrix sparsity. The output of SVD processing is a dense matrix $C'_{1010\times300}$ that encodes lexical features of action prototypes.

4.3.2.1 Actions clustering

In order to obtain a language-independent classification of action concepts, similarity between action prototype vectors has been computed through Manhattan distance[19], and then an unsupervised clustering algorithm applied to this data. The resulting classification bypasses differences in lexicalisation among languages, in favour of an *average* conceptual representation of actions. In fact, a classification based on data from only one language leads to a representation that is consistent with the semantic space segmentation operated by that language, but it may not be cross-linguistically generalized. Considering more languages together, instead, language-specific differences can be leveraged, highlighting similarities that may remain in shade if comparing monolingual classifications of actions.

[19] Given two vectors p and q, Manhattan distance is computed as:

$$d(p,q) = \sum_{i=1}^{n} |p_i - q_i| \tag{4.2}$$

Data clustering in this scenario is a complex task because there is no initial information on the number of clusters that must be found by the algorithm. Moreover, a proper evaluation of the resulting clusters is not trivial, since comparing one speaker's conceptual representation with the average representation resulting from summing lexical information from multiple languages may be difficult to measure.

4.3.2.2 Clustering algorithm

Action classes induction can be properly considered as a variation of a Word Sense Induction (WSI) task (Agirre and Soroa, 2007): instead of grouping word occurrences with similar meaning based on word contexts, this task aims at grouping similar action occurrences based on lexical features.

In this experiment *Affinity Propagation* (AP) (Frey and Dueck, 2007) has been implemented, a state-of-the-art unsupervised clustering algorithm, that has been successfully applied to accomplish WSI tasks in recent works (Alagić et al., 2018; Arefyev et al., 2018). AP automatically identifies the optimal number of clusters for a given dataset; each cluster consists of one *exemplar* (i.e. one element that is representative of the cluster) plus its neighbouring elements. Results of clustering on C' matrix are summarized in Table 4.13.

Number of clusters	178
Min # of scenes per cluster	2
Max # of scenes per cluster	24
Average # of scenes per cluster	5.67

Table 4.13: Results of the clustering algorithm.

4.3.2.3 The map of action concepts

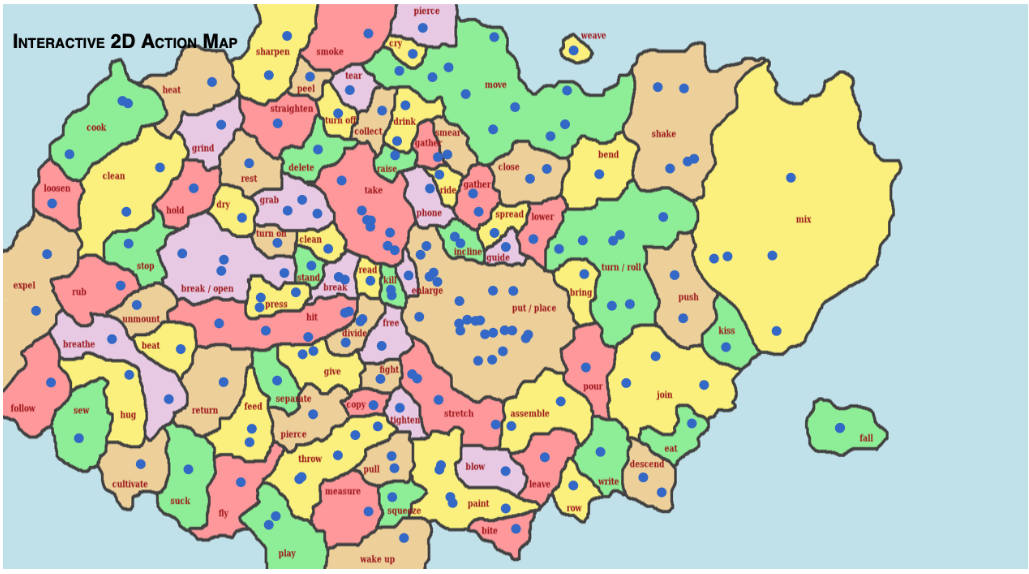


Figure 4.5: Map of the action concepts clustering.

Figure 4.5 shows the visual map that has been created for data exploration purposes. In this map each point is a cluster (i.e. a set of action videos), and the spatial position of points is derived as follows:

- exemplar vectors are chosen as cluster representatives;
- the feature space has been reduced to 2 dimensions with t-SNE[20];
- this 2D representation of exemplars has been projected on x and y axes.

The action map is interactive:[21] by clicking on each point, it is possible to see the set of videos belonging to the cluster. In order to ease data exploration, cluster regions have been manually drawn and named with an English verb that roughly describes the related semantic area.

[20] t-SNE is a dimensionality reduction algorithm, specifically designed for visual representation of high dimensional data (Maaten and Hinton, 2008).

[21] The interactive and explorable version of the map is available at `http://lablita.it/app/imclust/map.php`.

4.3.3 Evaluation

The obtained clusters need to be manually evaluated in order to testify the reliability of the applied methodology. Up to now, no similar previous evaluation work are available and, due to the peculiarity of this kind of work (i.e. language independent clustering obtained by summing of multilingual data), the possibility that a speaker of a language L may find reasonable all the clusters could be far from being positive. Nevertheless, in order to evaluate the obtained Action Type clusters and the applied methodology, an evaluation task specifically tailored for this work has been designed.

The evaluation consists in a two-alternative forced-choice similarity task in which, given as target a scene s from a cluster c, participants are asked to chose the most similar scene to the target among two other scenes. One of the scene used for comparison belongs to the same cluster c of the target, whereas the other has been selected among scenes not belonging to it. Annotators are expected to choose the scene belonging to the same cluster as more similar to the target one. If human judgments mirror the unsupervised algorithm output, the clustering can be interpreted as a reliability result. Figure 4.6 shows one item of the evaluation test: annotators judge which video, between 1 and 2, is more similar to the target.

A preliminary evaluation has been conducted on 11 randomly selected clusters. These were selected considering the number of scenes they contain: the clusters considered were in the middle range for number of scenes (one standard deviation around the mean[22]).

[22] Two reasons lead to prefer clusters with a number of scenes in the middle range: (1) to avoid outlier clusters; (2) to keep the number of items per test small enough, feasible in less then 30 minutes.

ì

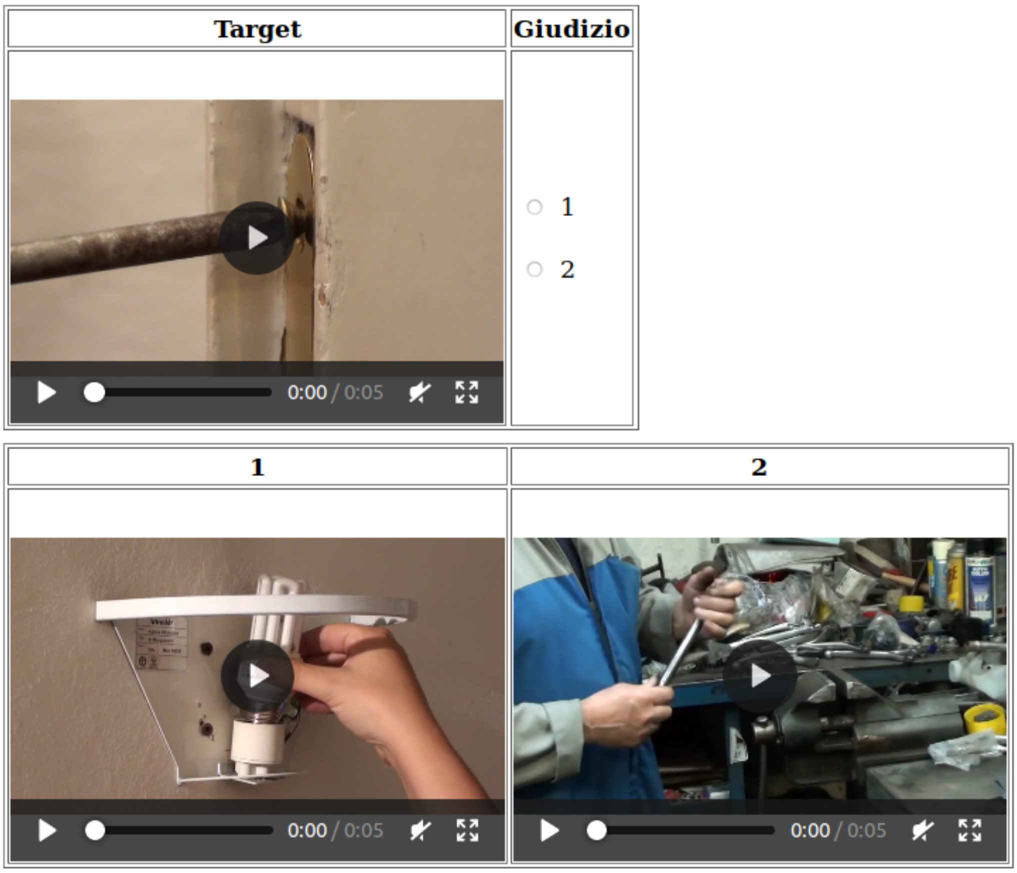


Figure 4.6: One item of the clustering evaluation test.

Participants have been presented every possible couple of scenes belonging to the same cluster, thus resulting in a dataset of 125 data-points. The third scene of each evaluation item was randomly selected among scenes not belonging to the considered cluster. Moreover, the similarity distance of these scenes has been checked from the cluster, and two version of the test have been prepared: the third scene was alternatively selected among the 15th most similar scenes not belonging to the target cluster or among other farther scenes, and this selection was inverted in the second version of the test. In this way, it has been possible to finely evaluate the precision in the categorization of closer concepts. Furthermore, each test has been divided in two, in order to avoid a very time consuming test. As a result, 2 different tests have been obtained, with 2 alternative version each (4 tests in total).

The first test was performed by 10 annotators, 5 for each version, while the second test was performed by 6 annotators, 3 for each version. All the annotators are Italian native speakers. Note that Italian is not among the languages included in the multilingual matrix. This is an additional strength of this evaluation, because results are not biased from the presence of annotators' mother language into the multilingual matrix.

This preliminary evaluation is meant as a first check on the task suitability. An extensive evaluation with a crowd sourcing platform may be planned in future, considering a larger number of clusters and collecting judgments from speakers of different languages. It will be interesting to analyse the influence of participants' mother language in the results, observing language-specific differences.

Tables 4.14 and 4.15 show results of Test 1 for each annotator; tables 4.16 and 4.17 show results of Test 2 for each annotator. Values range from 0 to 1, where 1 indicates that the scenes judged as more similar to the target correspond to the cluster's internal ones. The two versions of the test had a small difference in the percentage of correct pairs (0.88 for Test 1.1, 0.91 for test 1.2 and 0.97 for Test 2.1, 0.92 for Test 2.2), suggesting that the choice of the third scene (the one not belonging to the cluster) is relevant. Moreover, results show a small difference between near and far scenes: if the third scene was farther to the cluster, annotators judged the intra-cluster scene as similar to the target scene more easily.[23] Finally Cohen's k has been

[23] This can be investigated also measuring reaction times: It is reasonable to expect that annotators may take longer to judge triplets of scenes that are closer in such linguistic

measured between each annotator and the automatic clustering assignments; the resulting agreement is high: 0.75 for Test 1.1; 0.81 for Test 1.2; 0.83 for Test 2.1; 0.80 for Test 2.2.

	near	**far**	**total**	***k***
annot 1	0.80	0.92	0.86	0.84
annot 2	0.80	0.88	0.84	0.71
annot 3	0.80	0.92	0.86	0.80
annot 4	0.96	0.88	0.92	0.67
annot 5	0.88	0.92	0.90	0.72
Average	**0.85**	**0.90**	**0.88**	**0.75**

Table 4.14: Evaluation results - Test 1.1.

	near	**far**	**total**	***k***
annot 6	0.88	0.92	0.90	0.84
annot 7	0.92	0.92	0.92	0.84
annot 8	0.88	0.92	0.90	0.80
annot 9	0.88	0.92	0.90	0.79
annot 10	0.92	0.92	0.92	0.79
Average	**0.89**	**0.92**	**0.91**	**0.81**

Table 4.15: Evaluation results - Test 1.2.

Table 4.18 and 4.19 report the distribution of shared judgments among participants, i.e. how many items are evaluated according to the clustering and by how many annotators. For example, the first row of Table 4.18 reports the number of evaluated items where all of the 5 annotators (5/5) identified the scene belonging to the cluster as more similar to the target.

vector space.

	near	**far**	**total**	***k***
annot 11	0.97	0.97	0.97	0.88
annot 12	0.97	1.00	0.98	0.74
annot 13	0.97	0.97	0.97	0.86
Average	**0.97**	**0.98**	**0.97**	**0.83**

Table 4.16: Evaluation results - Test 2.1.

	near	**far**	**total**	***k***
annot 14	0.88	1.00	0.93	0.76
annot 15	0.91	0.95	0.93	0.87
annot 16	0.88	0.97	0.92	0.78
Average	**0.89**	**0.97**	**0.92**	**0.80**

Table 4.17: Evaluation results - Test 2.2.

Pair agr.	**Test 1.1**	**Test 1.2**	**Total**
5/5	39	40	79
4/5	2	4	6
3/5	3	2	5
2/5	2	1	3
1/5	3	3	6
0/5	1	0	1

Table 4.18: Agreement summary on scenes internal to the target cluster in Test 1.

Pair agr.	**Test 2.1**	**Test 2.2**	**Total**
3/3	71	67	138
2/3	1	3	4
1/3	2	1	3
0/3	0	3	3

Table 4.19: Agreement summary on scenes internal to the target cluster in Test 2.

In general, the evaluation confirmed the validity of the obtained clusters, which seem to correspond to classes of similar actions since they mirror human judgments with a high percentage. Thus, the matrix and the obtained clusters can be interpreted as inter-linguistic and cognitively valid representation of action classes. With further evaluations, it would be interesting to investigate to which extent lexical representation mirrors cognitive categorisation, thus providing new insights on how the human brain organises action conceptualisation.

4.3.4 Creating Ref-vectors for action verbs

By inverting the reference matrix used in 4.3.1 from $C_{scenes \times verbs}$ to $C_{verbs \times scenes}$, and adding data for English (662 verbs) and Italian (646 verbs), it has been easy to derive action verbs embeddings calculated on their primary extension to pragmatic action concepts.[24] In this way, the matrix C^{inv} encodes referential properties of verbs, with one row per verb and one column per action concept.

$$C_{j,i} = \begin{cases} 1 & \text{if action } j \text{ is predicable with verb } i \\ 0 & \text{else} \end{cases} \tag{4.3}$$

[24] Various experiments have been set to integrate word embeddings with information other than simple linear contextual co-occurrence. A good example is the work of Levy and Goldberg (2014) on the computing of dependency-based word embeddings.

Similarly to the previous experiment, SVD has been applied to derive C', a dense matrix $9,189 \times 300$.

4.3.4.1 Evaluation and results

The obtained co-referentiality vectors have been compared to state-of-the-art word embeddings in a verb semantic similarity task. In order to do this, the performance of different representation models have been evaluated against the SimVerb-3500 dataset (Gerz et al., 2016) as benchmark, which has been previously used in similar works on verbs similarity (Blundell et al., 2017).

For the evaluation, the full set of 220 English verbs that are shared by the IMAGACT and the SimVerb-3500 dataset has been considered: the comparison dataset (Comp-DS) has been obtained by considering those verb pairs for which similarity scores were present in SimVerb-3500, resulting in 624 verb pairs. Data are reported in Table 4.20.

	SV-3500	**Comp-DS**
Total verbs	827	220
Total pairs	3500	624
Antonyms	111	34
Cohyponyms	190	57
Hyper/Hyponyms	800	185
Synonyms	306	61

Table 4.20: Numbers of the full SimVerb-3500 dataset and of the sampled comparison dataset.

Verb semantic similarity has been automatically estimated for each verb pair in the Comp-DS by computing the cosine similarity between the related Ref-vectors. Then, the correlation between automatic and human judgments about verb pair similarity has been determined through the Pearson correlation coefficient. The result is a positive correlation of 0.37 (Table 4.22). This number highlights the presence of a correlation, but it is not informative without a comparison with other semantic vectors.

To this aim, 6 state-of-the-art word embedding have been considered, created with 3 algorithms (word2vec[25] (Mikolov et al., 2013), fastText[26] (Bojanowski et al., 2017) and GloVe[27] (Pennington et al., 2014)) trained on two big corpora (English Wikipedia[28] (2017 dump) and English GigaWord[29] (fifth edition) (Parker et al., 2011)). In this experiments, lemmatized word embeddings have been used, instead of token-specific representations, in order to obtain vectors that are comparable with SimVerb-3500's verb pairs.

Algorithm	**Corpus**	**Lemmas**	**Window**	**Dimensions**
Word2Vec	EnglishWiki2017	296,630	5	300
Word2Vec	GigaWord5	261,794	5	300
FastText	EnglishWiki2017	273,930	5	300
FastText	GigaWord5	262,269	5	300
GloVe	EnglishWiki2017	273,930	5	300
GloVe	GigaWord5	262,269	5	300

Table 4.21: Numbers of the lemmatized word embeddings used for comparison.

The previous procedure has been repeated by using these embeddings instead than Ref-vectors: cosine similarity has been measured between each pair of the Comp-DS, by using different embeddings. The Pearson correlation

[25] https://code.google.com/archive/p/word2vec/
[26] https://fasttext.cc
[27] https://nlp.stanford.edu/projects/glove/
[28] https://archive.org/details/enwiki-20170920
[29] https://catalog.ldc.upenn.edu/LDC2011T07

with the Comp-DS is reported in Table 4.22. Data show that Ref-vectors are closer to human judgments in estimating verb semantic similarity. All the verb embeddings considered report a lower correlation with Comp-DS.

Ref-vectors	**word2vec**		**fastText**		**GloVe**	
	Wiki	**GigaW**	**Wiki**	**GigaW**	**Wiki**	**GigaW**
0.37	0.20	0.23	0.19	0.21	0.10	0.13

Table 4.22: General correlation results between human judgments from SimVerb3500 and the compared systems.

In the end, the same analysis has been performed per semantic class. SimVerb-3500, and thus Comp-DS, contains the annotation of the semantic relation between the two verbs in each pair. This information has been used to measure the correlation of vector similarity with Comp-DS in verb pairs with specific semantic relations. Table 4.23 shows that Ref-vectors have a stronger correlation (0.32 to 0.36) with human judgments with all kinds of semantic relations. This is not valid for non semantically related pairs, where both Ref-Vectors and word embeddings have a very poor correlation with human judgments.

	Ref-vectors	**word2vec**		**fastText**		**GloVe**	
		Wiki	**GigaW**	**Wiki**	**GigaW**	**Wiki**	**GigaW**
Antonyms	**0.34**	-0.11	0.02	-0.07	0.08	-0.01	0.17
Cohyponyms	**0.32**	0.27	0.18	0.28	0.18	0.07	0.03
Hyper-Hyponyms	**0.36**	0.26	0.28	0.26	0.26	0.02	0.03
Synonyms	**0.34**	0.09	0.16	0.07	0.12	-0.07	0.07
None	0.08	0.11	0.10	**0.12**	0.09	0.09	0.03

Table 4.23: Correlation results between systems and simVerb-3500 dataset based on the semantic relation of verb pairs.

Chapter 5

Visual enrichment of IMAGACT through Audio Description for visually impaired people

> *In this chapter the work and research produced during the PhD program are introduced and described, starting from the initial dataset (LSMDC) analysis and processing, the experiments to annotate the correct IMAGACT action concept, and finishing with the discussion of the results.*

In the experiments described in Chapter 4, the main goal is to make use of the categorisation of action offered by IMAGACT Ontology. Nevertheless, an interesting side result in 4.1 and 4.2 is the enrichment of IMAGACT and the other resources involved by mean of a mutual information exchange. In both experiments, the focus has been the textual information contained in the resources. Until now, no experiments have been settled in order to explore the possibility of enriching the visual information on action concepts. Thus, it is possible to consider the work described in this Chapter as a first attempt in dealing with multimodal instance population of the IMAGACT Ontology.

The intuition of looking for visual instances of action concepts in Movies with Audio Description for visually impaired people came from the fact that such descriptions are by design mainly visual and thus they naturally form an interesting and particularly *action-rich* data source. For this reason, the Large Scale Movie Description Challenge dataset (LSMDC), very popular in Computer Vision, has been acquired, analysed and used as a playground for the experiments.

The work herein described may result, again, in a mutual information exchange between two data sources: on one side, IMAGACT Ontology would be populated of multimodal instances of actions in the form of video-caption pairs; on the other, LSMDC would be enriched with a fine-grained action discrimination derived from a handcrafted linguistic resource.

5.1 Audio description

Audio description[1] (AD) is a media accessibility service that enables millions of blind and visually impaired people to enjoy movies, television programs, live events, and other cultural visual contents along with their peers. In practice, it tells them what they cannot see, such as who is doing what, where and why, by providing an audio narration of the most important aspects of the visual information, such as actions, gestures, scenes, and character

[1] Audio Description is also referred to as Descriptive Video Service (DVS), Audio Narration or Video Description.

appearance (Salway, 2007). The difficulty of the objective of AD resides in the fact that multiple events happen in a same scene or context, and that their descriptions have to fit into tiny temporal bounds between the dialogues of the characters.

As an established practice for more than 20 years (Piety, 2004), AD is a globally spread activity, that is rapidly growing thanks to the work of volunteers, that have played a key role in making many cultural activities accessible for all, and to the development of legislation and regulations, that require broadcast to provide accessibility,[2] which also fostered technical developments to facilitate the delivery of AD (Matamala and Villegas, 2016). Guidelines about the production of AD are available in various countries, namely the UK, Spain, Germany, Austria, France, Poland and the USA, although the quantity and the quality of guidelines might differ from country to country, or from one product to another. However, AD production usually adheres to the following workflow: chose the suitable programme for description, watch the programme, write a draft script, review the script, record the script, review the recording, and finally sync and mix the AD audio track with the original one (ITC, 2000).

[2] The Convention on the Rights of Persons with Disabilities (CRPD) declares:

> "States Parties recognize the right of persons with disabilities to take part on an equal basis with others in cultural life, and shall take all appropriate measures to ensure that persons with disabilities: (a) Enjoy access to cultural materials in accessible formats; (b) Enjoy access to television programmes, films, theatre and other cultural activities, in accessible formats; (c) Enjoy access to places for cultural performances or services, such as theatres, museums, cinemas, libraries and tourism services, and, as far as possible, enjoy access to monuments and sites of national cultural importance." (UNCRPD, United Nations Committee on the Rights of Persons with Disabilities, 2011, Article 30 - Participation in cultural life, recreation, leisure and sport, §1)

Giving the relationships between vision and language, and the translational inspiration of audio description, AD is a vast research issue, which calls for the convergence of multiple disciplines, such as Linguistics, Philosophy, Semiology, Cognitive Studies, Translation and Film Studies, Interpreting Studies and others.[3] It has started to be taught as an Academic discipline in Language and Translation Faculties in a few countries, such as the UK, Spain and Belgium (Arma, 2011).

5.1.1 AD in Linguistics

The language of AD can be classified together with the language of radio and television, which are spoken languages that present traces of textual organisation recalling the written modality, as ADs are scripted before to be vocally recorded by professional speakers. Moreover, it is correct to define AD as an *audiovisual translation* (Orero, 2004), which is a special form of inter-semiotic translation aimed at transfer the information from the visual to the verbal channel, made for an intended specific audience, and with the purpose of communicating the contents of an audiovisual product which would remain otherwise inaccessible (Salway, 2007). It is important to note that source texts such as television programmes and films are complex mixes of codes carried by audio and visual channels, so that audio description, acting as a surrogate for the visuals, must interact appropriately with the existing dialogue and sound (Ballester Casado, 2007).

Turner (1998) conducted an interesting multimodal analysis on a corpus of ADs collected by sampling 27 minutes of AD and video from each of

[3] The first academic record of the concept behind AD appeared in a 1975 master's thesis, where the author drew on some experimental audio productions and theorized that information could be inserted to increase listener comprehension (Frazier, 1975).

a documentary, a drama and a movie. His objective was the potential of implementation of AD in multimodal retrieval and indexing of video data such as television programmes and movies in digital collections. His analysis focussed on two issues that would determine this potential: (i) how well aligned audio description is with the visual content; and, (ii) what aspects of visual content are described. He defines 15 types of information conveyed by audio description: physical description of characters, facial and corporal expressions, clothing, occupation and roles of the characters, attitudes of the characters, spatial relationships between characters, movement of the characters, setting, temporal indicators, indicators of proportions, decor, lighting, action, appearance of titles, text included in images.

Moreover, Piety (2004), adapting spoken discourse analysis techniques to the Audio Description of four movies (23,000 words), highlights a set of relevant aspects:

- 4 structural elements:
  - *Insertion*: a contiguous stretch of description that is uninterrupted by other significant audio content; insertions are usually bounded by dialogue, and they can be either short (few seconds), or much longer (several minutes).
  - *Utterance*: unit of spoken language, which can be arranged by the describer in any way to fill the time available in the insertion. They can be as long as the insertion itself or much shorter. They appear to the consumer as short snapshots of language that describe some visible features, and they are strung together to fill the space between dialogue.
  - *Representation*: a functional grammar description (Halliday, 1985) that includes processes, participants, and circumstances, sub-divided in 7 types of information (see below);

  - *Word*: an extremely restricted set of the words used in spoken or written discourses are used in AD. While most language use deals with information that is not present at the time of speaking, including past and future events and possible conditions (Chafe, 1994), AD relates only to what is actually occurring on the screen at the time (or close to the time) that the words are expressed. Unless they are part of something that is included in a representation of reading, there should be no words indicating conditions, past or future states, or any number of other valid language constructs that do not reflect the immediate reporting of what is showed in the visual modality.

- 7 types of information on representation:

  - *Appearance*: the external appearance of a person, place, or thing.
  - *Action*: something in motion or changing.
  - *Position*: the location of the information that is being described.
  - *Reading*: written information literally read or summarized.
  - *Indexical*: indicates who is speaking or what is making some sound.
  - *Viewpoint*: related to what the viewer would visually perceive in the framing of the scene.
  - *State*: not always visible information, but known to the describer and conveyed in response to visual information.

According to Salway and Tomadaki (2002), on the basis of the 50 most frequent verbs in a corpus of AD from 12 movies (∼70K words), it appears that the majority of events that are referred to by AD are material processes (84%), with some mental processes (10%), a few relational processes (4%) and a few behavioural processes (2%).

Salway (2007), within the Television in Words project, collected the TIWO Corpus of Audio Descriptions, that comprises ADs from 91 movies for a total of ∼619K words. His analysis on the language used in AD is aimed at investigating idiosyncratic linguistic features, by means of common corpus linguistics techniques. He observes that, according to guidelines, utterances in AD are characterised by the use of present tense, avoidance of potentially ambiguous pronouns, adjectives only where they convey relevant information, and use of adverbs to enhance the description of an action.

Moreover, the most manifest result of his analysis is that, on the basis of the deviation of overall frequencies with respect to measures in BNC corpus (British National Corpus, 2007) (*de facto*, reference corpus of British English), it is correct to classify the language of AD as a *Special Language* (SL), with distinct features from *General Language* (GL). Usually in all corpora, the very most frequent words are grammatical words, such as determiners, prepositions, copulas. Indeed, in GL corpora, very few semantically meaningful words tend to appear in the first 100 most frequent words, while many of the top 100 words in the TIWO corpus are noun tokens[4] referred to concrete objects and entities, and the highest frequent verbs are those referred to material processes, i.e. action verbs.

[4] In the work of Salway (2007) the textual analysis has been conducted exclusively on occurrences, thus the numbers he offers referred as *words* have to be considered as token counts. The analysis proposed in 5.2.3.1 and the results in Table 5.6 on the LSMDC dataset are both in terms of tokens and lemmas.

He analyses the top 300 frequent tokens in the TIWO corpus and observes that the most high distribution is registered for:

- characters and their body parts

  e.g. *man*, *head*, *face*, *eyes*, *hand*, *hands*, *men*, *human*, *hair*, *arms*, *arm*, *feet*, *girl*, *mouth*, *boy*, *crowd*, *shoulder*, *officer*, *people*, *lady*, *body*, *police*, *soldiers*, *father*;

- actions

  e.g. *looks*, *turns*, *takes*, *walks*, *goes*, *stands*, *steps*, *smiles*, *stares*, *puts*, *watches*, *opens*, *looking*, *runs*, *sitting*, *comes*, *picks*, *sees*, *holds*, *wearing*, *smile*, *nods*, *standing*, *leans*, *glances*, *gives*, *holding*, *watch*, *beat*, *grabs*, *leaves*, *falls*, *reaches*, *watching*, *drops*, *closes*, *lifts*, *throws*, *shakes*, *passes*, *run*, *follows*, *climbs*, *kiss*, *pushes*, *kisses*, *walk*, *lies*, *staring*, *carrying*;

- objects and scenes

  e.g. *door*, *room*, *car*, *window*, *table*, *water*, *bed*, *house*, *floor*, *gun*, *boat*, *street*, *road*, *ground*, *horse*, *phone*, *desk*, *hat*, *office*, *book*, *bag*, *stairs*, *chair*, *seat*, *sky*, *fire*, *jacket*, *bedroom*, *corridor*.

Raw frequencies, especially in terms of tokens, may not be exhaustive for justify a SL classification for Audio Description, and cannot record idiosyncratic phenomena of the language in the target domain. Thus, Salway (2007) applies the so-called *SL/GL ratio* (Equation 5.1) proposed by Ahmad and Rogers (2001), that is specifically designed for observing relative frequency shifting in the distribution of words in a SL corpus, with respect to a GL reference corpus.

$$SL/GL\,ratio \quad = \quad \frac{rel\_freq(w_{SL})}{rel\_freq(w_{GL})} \tag{5.1}$$

This formula highlights even the shift of words that appear relatively few times in the target SL corpus, but still with a much higher rate than in GL. If $SL/GL \sim 1$, it means that the distribution of the target word shows no substantial differences in both the domains of SL and GL; in contrast, a high score, e.g. $SL/GL = 50$, indicates that the word is being used relatively 50 times more often in the SL corpus with respect to GL reference corpus.

As expected, some of the top ranked words in the TIWO corpus, in terms of *SL/GL* ratio, are action verbs (Table 5.1), confirming that the language of Audio Description is a particularly *action-rich* language, thus a good source for analyse the relationship between *eventualities* and their linguistic translation in the form of a natural language description.

Ratio	**Tokens**
$SL/GL > 100$	saunters, hurries, stares, shoves, clambers, straightens, gazes, kneels, scrambles, leans, glares, nods, periscope, strolls, crouches, tosses, blinks, trots, frowns, hurls, clunk, grabs, pulls, llama, watches, smashes
$50 < SL/GL > 100$	unlocks, hauls, staggers, heaves, minion, stumbles, shakes, wipes, hesitates, pats, haired, lowers, pushes, wanders, crawls, grins, glances, flings, picks, flicks, slaps, hugs, smiles, sniffs, glides, scarecrow, sits, slams, rubs, pours, squeezes, diner, postman, spins, shuts, salutes, drags
$25 < SL/GL > 50$	rips, walks, climbs, closes, sips, strides, slumps, gallops, flashback, leaps, knocks, throws, fades, stirs, rushes, kisses, tugs, creeps, jumps, dives, shrugs, crashes, lifts, turns, licks, opens, silhouetted, elevator, pauses, swings, sighs, bounces, stops, dials, swims, bangs, presses, slips, removes

Table 5.1: Selection of words sampled at $SL/GL > 25$ and $Freq > 30$ in the TIWO Corpus of Audio Descriptions. Adapted from Salway (2007).

5.2 The Large Scale Movie Description Challenge Dataset

Given the interest in collecting many videos through natural language reference to action, The Large Scale Movie Description Challenge Dataset[5] (LSMDC) (Rohrbach et al., 2017) suites perfectly for the purpose. In fact, its current version (LSMDC16)[6] consists in a parallel corpus of more than 128K captions, obtained through the transcription of ADs, aligned to short video clips from 200 movies. LSMDC was first presented in 2015 at the Workshop "Describing and Understanding Video & The Large Scale Movie Description Challenge (LSMDC)".[7] Since then, it has been used as dataset in various challenge tasks in the Computer Vision Community. Table 5.2 reports a short description of the tasks from 2015 to 2019.

5.2.1 LSMDC dataset building

This dataset has been obtained through the combination of two datasets, previously collected independently but on the basis of similar methodologies: Max Plank Institute für Informatik Movie Description Dataset (MPII-MD) (Rohrbach et al., 2015) and Montreal Video Annotation Dataset (M-VAD) (Torabi et al., 2015).

[5] https://sites.google.com/site/describingmovies/home

[6] A new version, LSMDC19, have been announced, but by the time of writing it has not been released yet.

[7] The Workshop was part of the "International Conference on Computer Vision" (ICCV2015).

Task	**Editions**	**Description**
Movie description	2015-2016 2017-2019	Automatically describe videos by generating rich and natural captions. Participants have access to training, validation, public-test and blind-test (only videos, no captions) datasets. Challenge winner is determined on the basis of the result in the blind-test, evaluated both automatically and manually. In 2019 edition, evaluation is based on set of 5 consecutive videos.
Movie Annotation	2016-2017	Annotate and search videos based on natural sentences for variety of human activities. The task is subdivided in two tracks: (1) Multiple-choice test and (2) movie retrieval. For (1), given 5 caption for each video, find the correct one. Evaluation in percentage of correct pairings. For (2), given a caption, rank videos in the dataset (most similar on top). Evaluation through recall@k.
Movie fill-in-the-blank	2016-2017	Given a video clip and a sentence with a blank in it, fill in the blank with the correct word. Detailed evaluation of performance by part-of-speech.
Local characters prediction	2019	Given a set of 5 video-caption pairs with a blank in each caption, fill the blank with the corresponding local character ID (e.g. PERSON1, PERSON2), in order to identify people depicted in videos, and track them across the 5 clips.

Table 5.2: List of challenge tasks based on LSMDC dataset.

5.2.1.1 MPII-MD

To build the MPII-MD dataset, 55 Blu-ray movies from the "Audio Description" section of the British Amazon web-store[8] have been acquired. Movies have been converted in Matroska file format (.mkv), and from there video, original audio and AD audio track has been extracted. Then, by comparing the original audio track with the AD track through Fast Fourier Transform (FFT) analysis, the AD has been segmented on the basis of the differences of its spectrogram with respect to the one of the original audio track. Those audio segments have been transcribed by a crowd-sourced transcription service[9]. Video track has been also segmented following timestamps from the AD audio track segmentation. Then, the alignment between video and AD has been manually checked, in order to minimise misalignment between visual content and its description due to constrains deriving from AD insertion between movie dialogues. Additionally to ADs, MPII-MD is also composed of movie scripts mined from web sources[10], for a total of 50 movies, some of them already acquired with ADs. As for ADs, scripts also have been aligned to movies, but using subtitles as reference, and implementing the method used by Laptev et al. (2008) to infer timestamps for script sentences. These alignments have been also manually checked.

In order to avoid noise in the vocabulary by counting movie specific words, such as proper names of characters, the corpus of captions has been anonymised, by substituting proper names with *SOMEONE* or *PEOPLE*, in case of plurals. As a result, MPII-MD dataset consists in 94 unique movies, with ∼37K video-caption pairs from AD and ∼31K from movie scripts.

[8] https://www.amazon.co.uk/Audio-Description-DVD-Blu-ray/b?ie=UTF8&node=680949011

[9] https://castingwords.com

[10] https://www.weeklyscript.com, https://www.simplyscripts.com, https://dailyscript.com, https://www.imsdb.com

5.2.1.2 M-VAD

Differently from MPII-MD, M-VAD is exclusively composed of ADs, which have been collected on the basis of movie lists provided by "An Initiative of the American Council of the Blind"[11] and "Media Access Group at WGBH"[12] websites. Thus, 92 DVDs have been acquired and processed, with a similar workflow as the one used for MPII-MD, with the only difference that, prior to compare the two audio tracks (original and AD), vocal isolation techniques have been applied, in order to boost vocal signals and clearly separate them from surrounding sounds and soundtrack. Through this process, they obtained more clear audio tracks, and thus the comparison led to more precise automatic alignment of visual and auditory information, and the identification of timestamps did not require human effort. Consequently, following the obtained timestamps, ADs and movies have been segmented, and ADs segments have been manually transcribed using a professional transcription service.[13]

Similarly to MPII-MD, captions have been anonymised by substituting proper names with *SOMEONE* or *PEOPLE*. M-VAD dataset features 92 unique movies with ∼56K video-caption pairs from AD.

5.2.2 LSMDC statistics and analysis

LSMDC is by now the biggest open-domain Computer Vision dataset that features natural captions as linguistic information, paired with realistic videos. In fact, in the field of Computer Vision, where the demand for more and more data leads to constant ongoing campaigns for the collection and building of

[11] http://www.acb.org/adp/movies.html

[12] http://main.wgbh.org/wgbh/pages/mag/dvsondvd.html

[13] https://www.transcribeme.com

new training datasets, it is not a surprise that language gives a great support, adding semantic information about scenes, agents, objects and events that appear or happen in a video.[14]

In most of the datasets, actually, textual information is limited to labels (often referred to as classes), which deliver unambiguous definitions such as *Person-Collecting-Garbage* or *Playing-Trumpet*, where the action could be defined just by a *verb-object* bigram. In the light of this, it is difficult to consider a rough bigram as a natural language expression. Moreover, this approach makes difficult to gather similar actions, given that each bigram is offered as a unique label. As an example, in the Kinetics dataset[15] (Carreira et al., 2019; Kay et al., 2017), one of the biggest open-domain dataset for action recognition (∼650K videos, 700 labels with at least 600 videos each), the class *Applying-Cream* is defined but no other versions of *Applying* are addressed, and *Drinking-shots* is the only reference to the action of *drinking*. On the contrary, two distinct classes refer to the action of *passing* an American football (*Passing-American-Football-(in-game)*, and *Passing-American-Football-(not-in-game)*) and one unique class is defined as *Catching-or-Throwing-Baseball*; it means that it is not possible to address the action of *throwing/passing* a different kind of ball, and *cathcing* and *throwing*, different actions from various points of view (even if some entailment is present), are treated as the same. For this reasons, LSMDC dataset and similar, which feature natural language captions, offer a more natural way of grounding action reference in language, both for linguistics and NLP scopes, and for boosting action recognition performances in Computer Vision.

[14] The *Video Dataset Overview* webpage (https://www.di.ens.fr/~miech/datasetviz/) offers a comprehensive and constantly updated list of existing datasets.

[15] https://deepmind.com/research/open-source/kinetics

Another important feature of LSMDC is the fact that, from the visual point of view, it is an open-domain dataset. Popular datasets in the field, such as YouCook2[16] (Zhou et al., 2018) or EPIC-KITCHENS[17] (Damen et al., 2018), are closed on one single domain (*cooking*, in this case), thus do not offer a variety of scenes, situations, objects, people involved in the eventualities depicted.

Dataset	**Movies**	**Words**	**Captions**	**Clips**	**Length**(h)
MPII-MD (AD)	55	330,086	37,272	37,266	44.0
MPII-MD (script)	50	317,728	31,103	31,071	33.8
MPII-MD (total)	94	647,814	68,375	68,337	77.8
M-VAD (AD)	92	502,926	55,904	46,589	84.6
LSMDC15 train	153	914327	91,941	91,908	124.9
LSMDC15 val	12	63,789	6542	6542	9.6
LSMDC16 train	153	922,918	101,079	101,046	114.9
LSMDC16 val	12	63,321	7408	7408	8.4
LSMDC15-16 public test	17	87,15	10,053	10,053	11.7
LSMDC15-16 blind test	20	83,766	9578	9578	12.0
LSMDC15 (total)	200	1,149,032	118,114	118,081	158.1
LSMDC16 (total)	200	1,157,155	128,118	128,085	147.0

Table 5.3: Statistics for the MPII-MD, the M-VAD, and the two versions (2015 and 2016) of the LSMDC datasets (adapted from Rohrbach et al. (2017)).

[16] http://youcook2.eecs.umich.edu

[17] https://epic-kitchens.github.io/2018

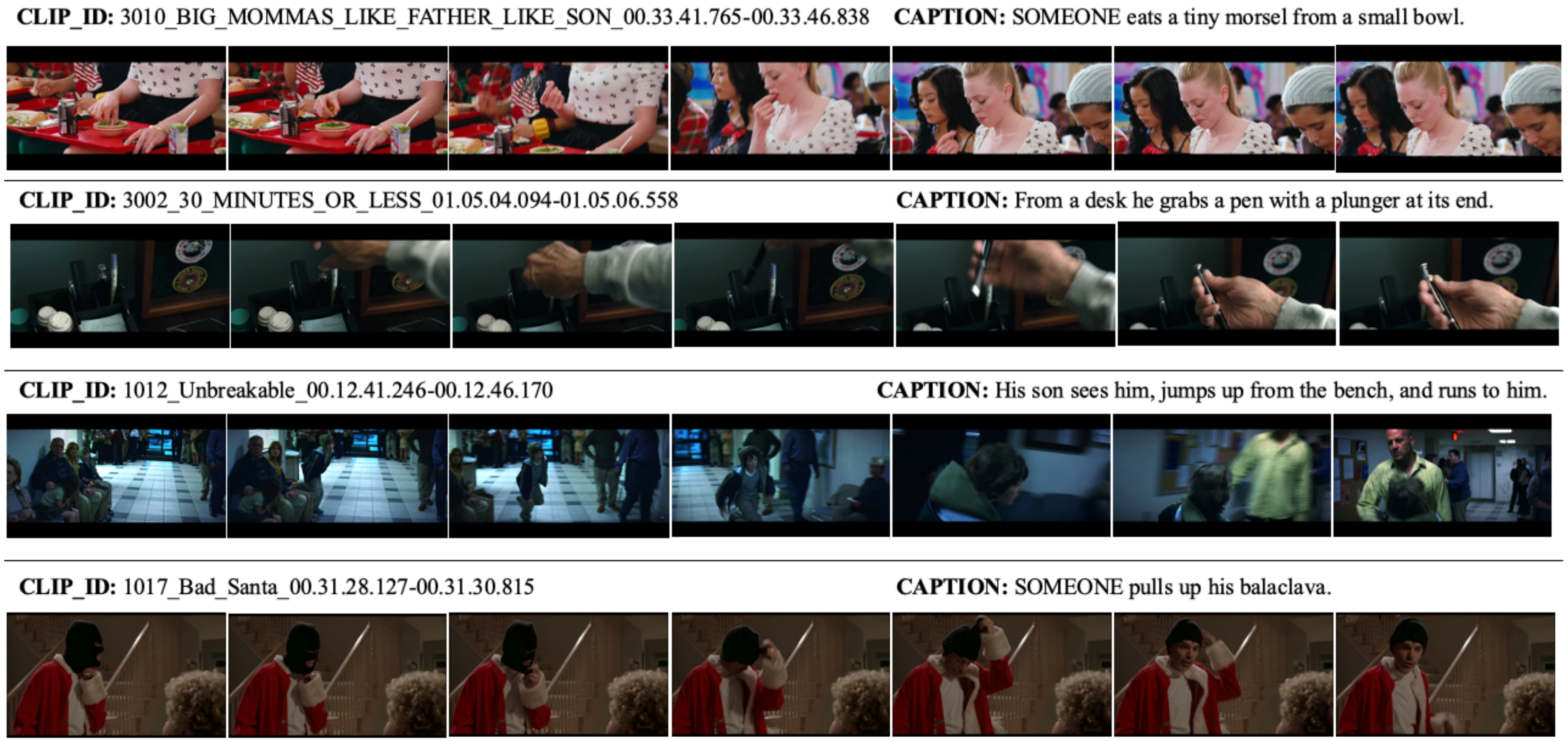
Table 5.3 reports some general statistics for MPII-MD, M-VAD and their merged version LSMDC, in the two variants from 2015 and 2016. Figure 5.1 shows some examples of video-caption pairs in LSMDC. In this regard, it is important to mention that in LSMDC items have different code-names, on the basis of their acquisition from MPII-MD or M-VAD. Depending on their origin, `CLIP_ID`s of the items are named as follows:

- `0XXX_MOVIE_START-END`
  (e.g. `0001_American_Beauty_00.02.29.298-00.02.30.004`)
  Video-caption pairs derived from movie script alignment in MPII-MD;

- `1XXX_MOVIE_START-END`
  (e.g. `1006_Slumdog_Millionaire_01.25.49.077-01.25.54.718`)
  Video-caption pairs derived from AD alignment in MPII-MD;

- `3XXX_MOVIE_START-END`
  (e.g. `3011_BLIND_DATING_01.11.23.496-01.11.25.853`)
  Video-caption pairs derived from AD alignment in M-VAD.

In this way, it is possible to select ADs only (`1XXX` + `3XXX`), scripts only (`0XXX`) or all of them, depending on the research interests and needs, or on the need of using exclusively data from one of the previous datasets.



CLIP_ID: 3010_BIG_MOMMAS_LIKE_FATHER_LIKE_SON_00.33.41.765-00.33.46.838 **CAPTION:** SOMEONE eats a tiny morsel from a small bowl.

CLIP_ID: 3002_30_MINUTES_OR_LESS_01.05.04.094-01.05.06.558 **CAPTION:** From a desk he grabs a pen with a plunger at its end.

CLIP_ID: 1012_Unbreakable_00.12.41.246-00.12.46.170 **CAPTION:** His son sees him, jumps up from the bench, and runs to him.

CLIP_ID: 1017_Bad_Santa_00.31.28.127-00.31.30.815 **CAPTION:** SOMEONE pulls up his balaclava.

Figure 5.1: Some examples taken from LSMDC dataset.

5.2.3 LSMDC data exploration

LSMDC data are freely available from the Challenge website, after requesting access for the use.[18] Data are presented as comma separated values files (.csv) for the textual modality, HD videos as .avi containers for the visual part. Text is divided in multiple files according to splits for the Movie description task of the Challenge:[19] `LSMDC16_annos_training`, `LSMDC16_annos_val`, `LSMDC16_annos_test`, `LSMDC16_annos_blindtest`. Given that the blind test file does not contain captions, it has been ignored, and the concatenation of the other three files resulted in 118,540 video-caption pairs.[20] For the scopes of the experiments herein described, the important information of each item[21] in the files is as follows:

- `CLIP_ID`: the name of the corresponding HD video file (e.g. `1017_Bad_Santa_00.31.28.127-00.31.30.815`);
- `CAPTION`: the caption describing the `CLIP_ID` video (e.g. *SOMEONE pulls up his balaclava*) (Examples from last item in Figure 5.1).

[18] All the informations about the access and the data can be found on the download page: `https://sites.google.com/site/describingmovies/download`.

[19] See Table 5.2.

[20] Numbers are slightly different from the official ones (see Table 5.3). All numbers and data henceforth in this dissertation will refer to this version, downloaded on February 2017.

[21] Not necessary to use, but worth to mention, other information regards timestamps of movie and ADs/scripts: `START_ALIGNED` and `END_ALIGNED` correspond to manually obtained timestamps; `START_EXTRACTED` and `END_EXTRACTED` indicate the automatically obtained ones.

5.2.3.1 Textual data analysis

As preliminary operations, firstly, the textual part of the dataset has been saved to an SQLite[22] DB, where also all the consequent processes, results and analysis have been stored. Secondly, all the LSMDC captions have been processed with StanfordNLP[23] (Qi et al., 2018). This Python Library delivers a full neural network pipeline for robust multilingual text analytics, including processors for tokenisation, multi-word token expansion, lemmatization, part-of-speech (POS) tagging, morphological features tagging, and syntactic structure dependency parsing using the *Universal Dependencies* formalism (McDonald et al., 2013).[24]

The output of the StanfordNLP pipeline is a text file (or `string` variable) where the information is written in CoNLL-U format. Table 5.4 shows the result of the pipeline for the LSMDC caption *Sitting in the passenger seat, the kid stares at him and smiles warmly. SOMEONE looks away from it.* (`1017_Bad_Santa_00.28.36.566-00.28.40.935`). In the output, the information on each word is written on one single line, and blank lines mark sentence boundaries (from full-stop mark to full-stop mark), if any. Word-level annotation fields are separated by tabs, and filled with an underscore if not available (e.g. punctuation has no `FEATS`). Table 5.5 shows in details all the fields of the CoNLL-U word-level annotation with a brief description of each.

[22] https://www.sqlite.org/index.html

[23] https://stanfordnlp.github.io/stanfordnlp/

[24] Details on the Universal Dependencies formalism can be found at: https://universaldependencies.org

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Sitting	sit	VERB	VBG	VerbForm=Ger	9	advcl	_	_
2	in	in	ADP	IN	_	5	case	_	_
3	the	the	DET	DT	Definite=Def—PronType=Art	5	det	_	_
4	passenger	passenger	NOUN	NN	Number=Sing	5	compound	_	_
5	seat	seat	NOUN	NN	Number=Sing	1	obl	_	_
6	,	,	PUNCT	,	_	9	punct	_	_
7	the	the	DET	DT	Definite=Def—PronType=Art	8	det	_	_
8	kid	kid	NOUN	NN	Number=Sing	9	nsubj	_	_
9	stares	stare	VERB	VBZ	Mood=Ind—Number=Sing—Person=3—Tense=Pres—VerbForm=Fin	0	root	_	_
10	at	at	ADP	IN	_	11	case	_	_
11	him	he	PRON	PRP	Case=Acc—Gender=Masc—Number=Sing—Person=3—PronType=Prs	9	obl	_	_
12	and	and	CCONJ	CC	_	13	cc	_	_
13	smiles	smile	VERB	VBZ	Mood=Ind—Number=Sing—Person=3—Tense=Pres—VerbForm=Fin	9	conj	_	_
14	warmly	warmly	ADV	RB	_	13	advmod	_	_
15	.	.	PUNCT	.	_	9	punct	_	_
1	someone	someone	PRON	NN	Number=Sing	2	nsubj	_	_
2	looks	look	VERB	VBZ	Mood=Ind—Number=Sing—Person=3—Tense=Pres—VerbForm=Fin	0	root	_	_
3	away	away	ADV	RB	_	2	advmod	_	_
4	from	from	ADP	IN	_	5	case	_	_
5	it	it	PRON	PRP	Case=Acc—Gender=Neut—Number=Sing—Person=3—PronType=Prs	3	obl	_	_
6	.	.	PUNCT	.	_	2	punct	_	_

Table 5.4: Example of the CoNLL-U output of StanfordNLP pipeline.

ID	Word index, integer starting at 1 for each new sentence (range for tokens with multiple words)
FORM	Word form or punctuation symbol
LEMMA	Lemma or stem of word form
UPOSTAG	Universal part-of-speech tag
XPOSTAG	Language-specific part-of-speech tag
FEATS	List of morphological features from the universal feature inventory or from a defined language-specific extension
HEAD	Head of the current token, which is either a value of ID or zero (0)
DEPREL	Universal Stanford dependency relation to the HEAD (root if HEAD = 0) or a defined language-specific subtype of one
DEPS	List of secondary dependencies (HEAD-DEPREL pairs)
MISC	Any other annotation

Table 5.5: CoNLL word-level annotation tagset.

Table 5.6 shows vocabulary distribution in LSMDC in terms of tokens and lemmas, organised by part-of-speech.[25] Numbers vary a lot from those in Rohrbach et al. (2017). Two considerations must be taken into account: firstly, they performed the processing with Stanford POS Tagger (Toutanova et al., 2003), that is a different tool relying on different models; secondly, they considered stemmed words. It is possible that treating words as stems leads to higher tagging mistakes, which summed with inevitable ones from the automatic tagger, may turn out in problematic results. Lemmas and Tokens are considered standard units of measures in Corpus Linguistics, thus they will be used in this analysis.

[25] Part-of-speech tags refers to the UPOSTAG field of the CoNLL-U output of StanfordNLP pipeline. The complete list of tags can be found at: https://universaldependencies.org/u/pos/.

Tokens distribution						
POS	**NOUN**	**PROPN**	**VERB**	**AUX**	**ADJ**	**ADV**
n	246,625	9,601	189,416	12,797	50,088	51,044
POS	**PRON**	**ADP**	**DET**	**PART**	**INTJ**	**CCONJ**
n	180,857	149,272	133,051	17,254	82	32,113
POS	**SCONJ**	**NUM**	**PUNCT**	**SYM**	**X**	**Total**
n	12,351	6,275	173,854	48	127	1,264,855

Lemmas distribution						
POS	**NOUN**	**PROPN**	**VERB**	**AUX**	**ADJ**	**ADV**
n	10,047	2,970	4,503	23	3,679	1,400
POS	**PRON**	**ADP**	**DET**	**PART**	**INTJ**	**CCONJ**
n	59	118	27	4	42	11
POS	**SCONJ**	**NUM**	**PUNCT**	**SYM**	**X**	**Total**
n	52	286	26	6	33	23,286

Table 5.6: LSMDC tokens and lemmas distribution by Part-of-Speech (`UPOSTAG`).

Given the interest in finding candidates in LSMDC to enrich IMAGACT with new videos, the target in this analysis are LSMDC captions containing at least one verb listed in the IMAGACT resource. Table 5.7 reports more LSMDC statistics in terms of number of sentences (considered as units of text delimited by periods or functionally equivalent marks), number of verbs and the detail of action verbs shared with IMAGACT. It is possible to notice that there are more than one verb and sentence for each caption, thus more than one action may occur in one single item.

	Sentences	Tokens	Verbs freq	IM verbs freq
Count	120,049	1,264,855	189,416	108,779
Mean	1.013	10.670	1.560	0.918
std	0.130	5.615	0.930	0.808

Table 5.7: LSMDC statistics on number of sentences, verbs and IMAGACT action verbs.

Considering the IMAGACT English verb list as reference[26] for action verbs in the English lexicon, the absolute frequency in terms of tokens of IMAGACT action verbs is ∼57% of the total. That is, more than half of the times there is a verb in AD language, it is an action verb, and this datum is in line with what Moneglia and Panunzi (2010) highlighted in speech corpora.[27] Thus, from the point of view of the action reference in the verb class, it is possible to confirm the findings of Salway (2007) on AD, i.e. its language is particularly rich of references to actions. Figures from 5.2 to 5.5 show the distribution of the 50 most frequent verbs and nouns, in terms of tokens (red bars) and lemmas (blue bars). At a glance, it is possible to note that almost all the verbs in the top 50 are action verbs (Figure 5.2 and 5.3).

After this preliminary analysis, LSMDC results in being a good source for investigating action in language, and it is also particularly fruitful to mine this dataset to extract variate multimodal candidates for augmenting IMAGACT ontology, both linguistically, with more captions as example of use in the language, and visually, with more videos depicting instances of action concepts.

[26] Even if IMAGACT vocabulary of action verbs may not be exhaustive, i.e. verbs not found in the corpora have no entry in the resource, one cannot disagree that it is a good repertoire of verbs more frequently used in everyday language interactions.

[27] See Section 2.2 and Figure 2.6.

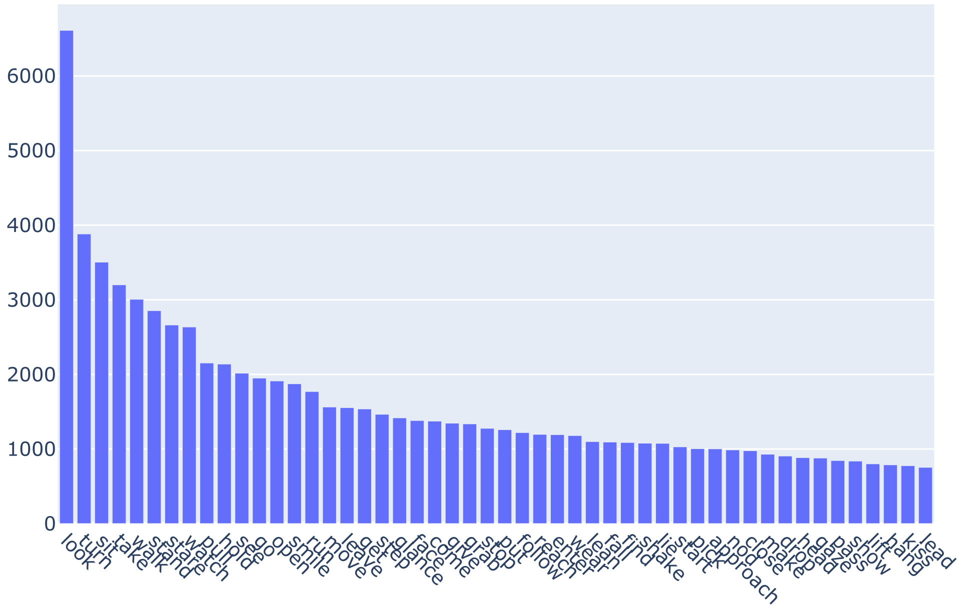


Figure 5.2: LSMDC verbs - lemmas distribution.

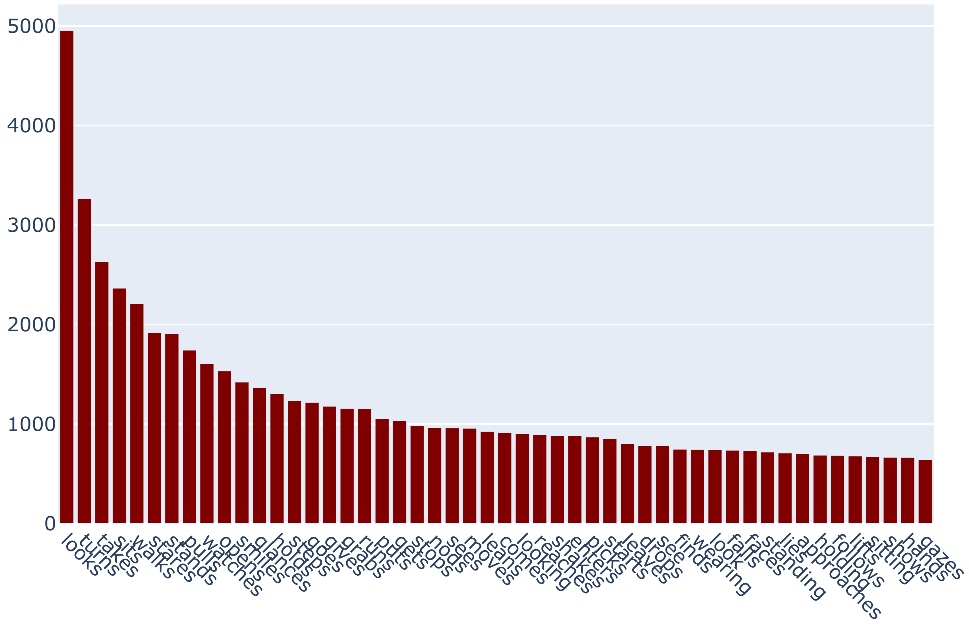


Figure 5.3: LSMDC verbs - tokens distribution.

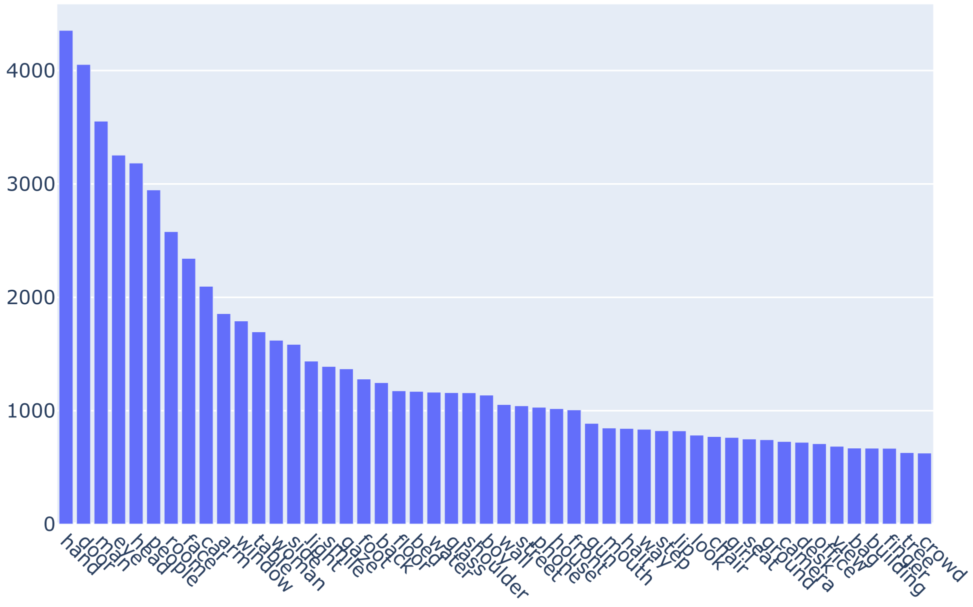


Figure 5.4: LSMDC nouns - lemmas distribution.

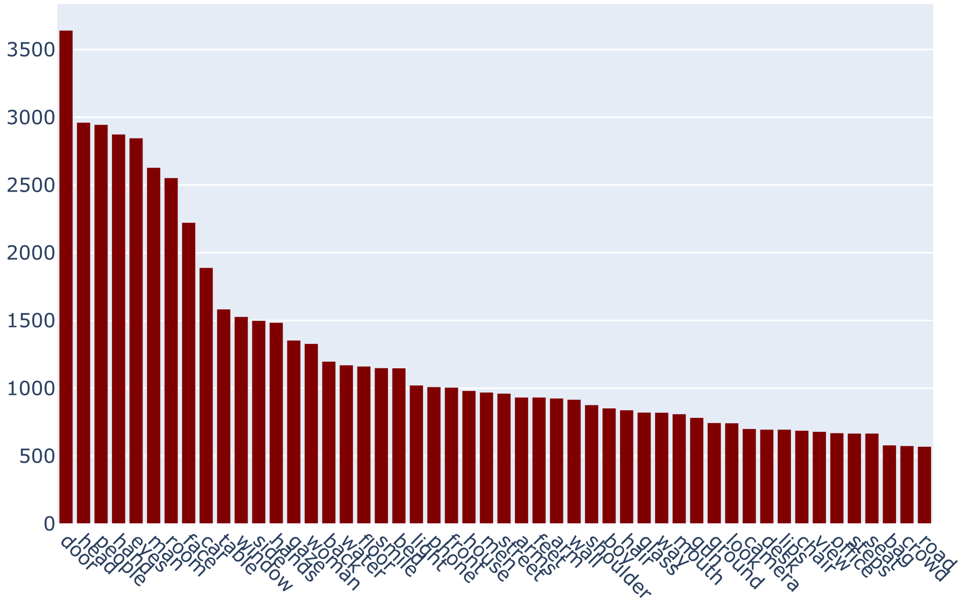


Figure 5.5: LSMDC nouns - tokens distribution.

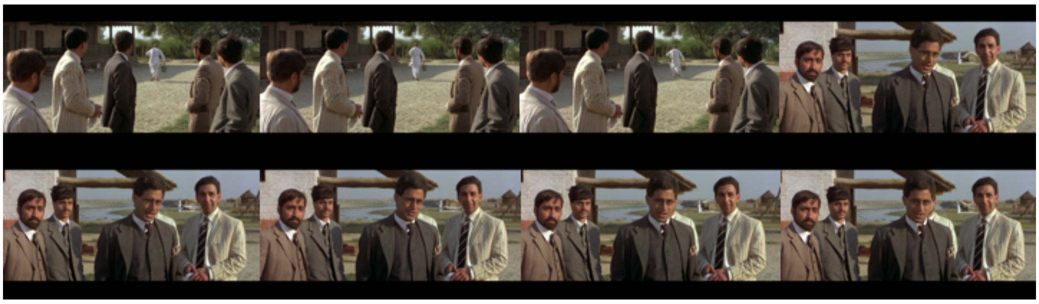
5.2.3.2 Visual data analysis

Before to start working with LSMDC data, its visual part has been also analysed, in order to ensure that the semantic information delivered by the captions is actually grounded in the videos. Video segmentation in LSMDC has been performed on the basis of timestamps obtained from AD audio track. The automatic process has been manually checked, but it is important to remark AD constraints due to its fitting in between movie dialogues. It is clear that it is impossible to avoid misalignments cases between narration in AD and exact happening in the movie.

The idea is to find visual representations of actions referred by natural language, so it has been necessary to verify the visual part of the dataset. Video analysis has been conducted manually, simply by reading the caption and watching the video, and assigning a label to classify the quality of the visual-textual association. Given the presence of various items in which more than one action is involved (e.g. rapid sequences in action movies), a more fine-grained action segmentation has been annotated, in order to assign to the narration of the events the exact happening time in the videos. Due to the dataset dimension (118K+ video-caption pairs), it is impossible to carry on a consistent and complete analysis of videos manually. For this reason, a small sub-portion of the dataset has been selected, and the manual analysis has been carried on this portion exclusively. Even if not exhaustive, this analysis may help to understand to what extent LSMDC effectively conveys in the visual modality the information supplied by language in the captions.

In this regard, the first observation is that script-derived items, i.e. video-caption pairs with `CLIP_ID` starting as `0XXX`, rarely find a concrete match between the actions referred by the caption and the events that actually happen in the video.

An example of these cases is in Figure 5.6: the video shows briefly a backview of one man running away from the camera, then some men watching him and talking, while the caption says *He pushes it away and starts for the kitchen*. It is obvious that from this we can infer that the running man is running into a kitchen from outside the house, but the reality of what is shown cannot be bend to the meaning of the caption: no *pushing* is performed.



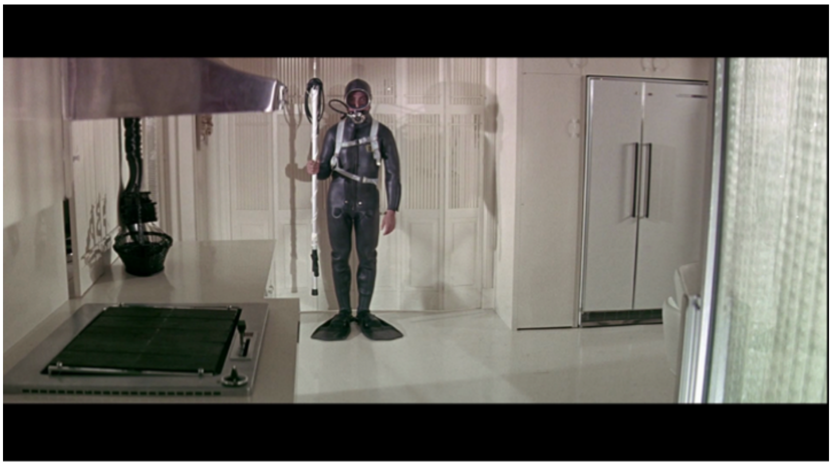
CLIP_ID: 0011_Gandhi_01.01.59.921-01.02.01.558

Caption: He pushes it away and starts for the kitchen.

Figure 5.6: LSMDC example of script-derived item, where actions are not matching in video and caption.

In any case, the target of this work is AD only LSMDC dataset (CLIP_IDs 1XXX and 3XXX), thus the script part of the dataset will not be considered from this point forward. Another reason is that, even if the scripts used for LSMDC have been selected on the basis of the quality,[28] a script is always something written before the movie shooting, used as a guideline by the shooting troupe, but it often happens that changes are made in the making of the shooting.

[28] Only scripts labelled as *Final*, *Shooting*, or *Production Draft* on the web-sources have been included in the dataset.



CLIP_ID: 0029_The_Graduate_00.22.31.807-00.22.32.223

Caption: At the back of the hall, inside the house, stands SOMEONE, dressed in a full length skin diver's wet suit, flippers on his feet, the oxygen tank strapped to his back, the mask pushed up his forehead, the air hose dangling.

Figure 5.7: LSMDC example of script-derived item.

As an example, consider Figure 5.7. The 2-seconds video clip is basically a still shot of a man with diving equipment in the context of a kitchen. The alignment in this item is perfect, but besides the fact that no action is happening and all the action verbs used in the caption are states description and not actions, some discrepancies between textual and visual informations arise. In fact, the caption describes a man with the diving mask pushed up his forehead and the air hose dangling, while in the video the mask is on the eyes of the actor, and the air nose in his mouth. This kind of discrepancies can be found in many script-derived LSMDC items.

The selection of the items to be analysed started with the choice of one verb with comparable distribution, i.e. in the middle of the ranking on the basis of frequency, in both LSMDC and IMAGACT. The verb *push* results in being very frequent in both the resources, 177^{th} on 4,503 verbs with ∼600 occurrences in the ranking in LSMDC, and 60^{th} on 569 with ∼140 occurrences in the English corpus bootstrapped for IMAGACT. Moreover, *push* in IMAGACT shows a high level of generality: this verb is composed of 18 Action Types, for a total of 25 action concepts, resulting in the verb with higher number of Action Types. Given that one action can be referred by more than one verb, from IMAGACT all the verbs linked to at least one of those action concepts have been queried, and then this set of verbs has been used to select the sub-portion of LSMDC items to analyse. In this way, a test-case has been created, focused on the semantic field of *pushing* eventualities.[29]

The workflow of the visual analysis is very simple: given a video-caption pair from LSMDC, evaluate the agreement between the reference to actions in the caption and what is depicted in the video, subdivide the caption in simple sentences (if more than one action is referred), align each sentence to the exact timing of the events showed in the video, when possible.

Given that AD is affected by misalignment problems, deriving from the fact that ADs have to fit into tiny temporal bounds between dialogues,[30] the resulting description of a video clip will inevitably not illustrate every detail of a scene, but it will focus on the most salient entities and events.[31] In this analysis, if there are actions not showed in the video, comment as

[29] Details on the primary variation of *push* and the list of related verbs in IMAGACT are in Table 2.4.

[30] See 5.1.

[31] On the contrary, the stream that flows from the visual medium carries plenty of information (Bruni et al., 2016), and it is difficult to fit everything into the perimeter of a single utterance or sentence. In particular, in the Computer Vision field, the semantics of a video is still hardly subordinated to visual-specific features due to the well known semantic gap (Smeulders et al., 2000).

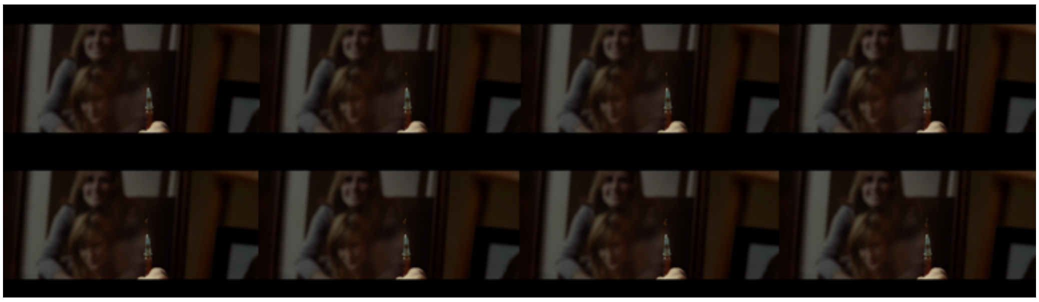
not_showed. Otherwise, if an action referred to by the caption finds its happening in the video, four labels have been defined for assessing the quality of the visual reference:

- good: the video clearly shows the action performance in the foreground, without any issue;
- out_of_framing: the action performance happens out of the boundaries of the shooting frame;
- camera_angle_change: the action performance is not completely showed within one single framing, i.e. the scene is shot with multiple cameras and video editing has been applied;
- background: the action happens in the background of the scene, not in the focus of the foreground.

It is important to underline that this qualitative analysis is necessary to ensure that videos selected from LSMDC satisfy the specifications defined within the IMAGACT project for the production of video prototypes of action concepts.[32] It is necessary that videos to be gathered under IMAGACT action concepts depict the action in an unambiguous and clearly visible way. Finally, if the video meets all the requirements, the corresponding IMAGACT action concept's ID is assigned to the video.

As an example, consider the item in Figure 5.8. Two actions are expressed by the caption, through the verbs *push* and *slide*. While the *small drop* sliding down the needle is at the centre of the frame, the *pushing* event is not happening in the video. The pragmatic and common-sense knowledge of the observer leads to infer that someone is pushing the syringe plunger as the causative event of the sliding drop, but actually it is not showed. In this case, the event predicated by *push* is tagged as out_of_frame, while the event referred to by *slide* is tagged as good.

[32] See 2.3.2 for the video production specifications defined for IMAGACT.

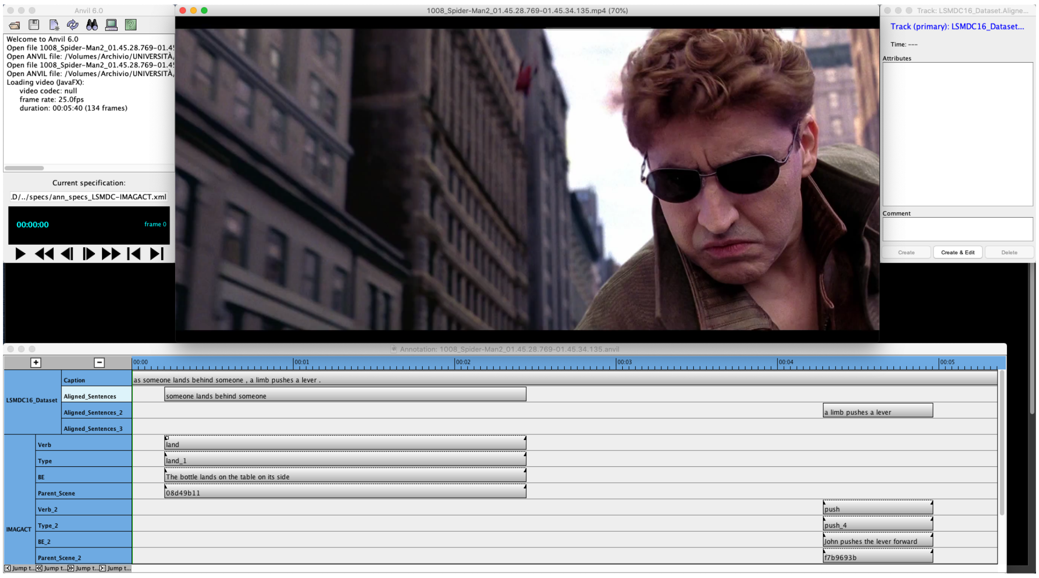


CLIP_ID: 3025_FLIGHT_00.15.53.660-00.15.56.013

Caption: Then pushes out a small drop that slides down the needle.

Figure 5.8: LSMDC example of item tagged as out_of_frame.

The manual analysis has been carried on using Anvil[33] (Kipp, 2001), a simple to use and to configure video annotation tool, that creates annotation files in XML format which are easy to parse. Figure 5.9 shows an example of the Anvil software interface.



Anvil 6.0
1008_Spider-Man2_01.45.28.769-01.45.34.135.mp4 (70%)
Track: LSMDC16_Dataset.Aligne...
Annotation: 1008_Spider-Man2_01.45.28.769-01.45.34.135.anvil
Create
Create & Edit
Delete
Jump t...
Jump t...
Jump t...
Jump t...

```
Welcome to Anvil 6.0
Open file 1008_Spider-Man2_01.45.28.769-01.4
Open ANVIL file: /Volumes/Archivio/UNIVERSITÀ,
Open file 1008_Spider-Man2_01.45.28.769-01.4
Open ANVIL file: /Volumes/Archivio/UNIVERSITÀ,
Loading video (JavaFX):
    video codec: null
    frame rate: 25.0fps
    duration: 00:05:40 (134 frames)
```

Current specification: .D/../specs/ann_specs_LSMDC-IMAGACT.xml

00:00:00 frame 0

Track (primary): LSMDC16_Dataset...

Time: ---

Attributes

Comment

00:00 00:01 00:02 00:03 00:04 00:05

LSMDC16_Dataset

- Caption: as someone lands behind someone , a limb pushes a lever .
- Aligned_Sentences: someone lands behind someone
- Aligned_Sentences_2: a limb pushes a lever
- Aligned_Sentences_3

IMAGACT

- Verb: land
- Type: land_1
- BE: The bottle lands on the table on its side
- Parent_Scene: 08d49b11
- Verb_2: push
- Type_2: push_4
- BE_2: John pushes the lever forward
- Parent_Scene_2: f7b9693b

Figure 5.9: Anvil annotation tool interface.

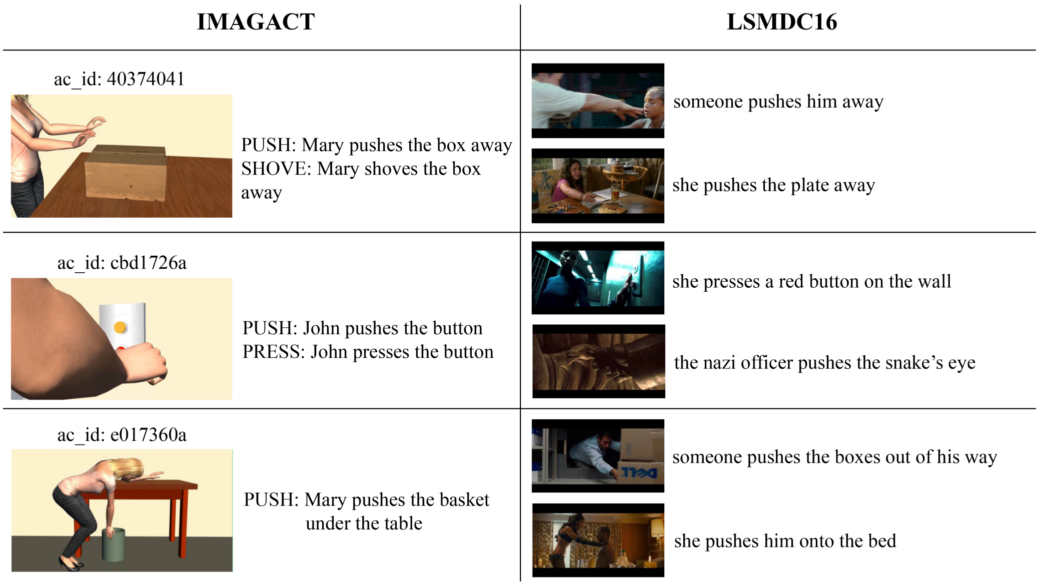
[33] http://www.anvil-software.org

At the end of the manual analysis, only few items resulted in being good candidates for IMAGACT. From an initial selection of 1,090 LSMDC items, with exclusion of 391 script-derived items, only 375 found a correct association with one action concept of the 25 connected to *push*, for a total of 380 instances[34] classified according to the IMAGACT categorisation of action. On the other side, not all IMAGACT action concepts found an example among LSMDC video-caption pairs: only 18 out of 25 action concepts found at least one instance in the domain of Audio Description. Most of the remaining items contain a state description expressed through an action verb,[35] or was not tagged as `good` visual action reference.

Moreover, some of the unclassified items were referring to action concepts not present in IMAGACT, or difficult to assign to an existing one. As an example, IMAGACT do not report resultative constructions, thus expressions like *push the door closed* or *squeeze sb eyes shut* are not present in IMAGACT, and LSMDC items about these eventualities found no association with IMAGACT action concepts. Finally, it is important to remark that action concepts linked to the locally equivalent verbs of *push* but other than the extension of this verb have not been considered.

[34] Some captions present more than one action reference, and also in a restricted set as those in this analysis it is possible to have multiple references within the *pushing* eventualities.

[35] Similarly to the example shown in Figure 5.6, many items are state descriptions, such as in `1005_Signs_01.23.26.304-01.23.28.400`, where the caption is *SOMEONE, sitting between SOMEONE's legs, his head pushed back into his father's chest.* The person is not actually pushing his head, that is simply leant back on the other person's chest.



IMAGACT		LSMDC16
ac_id: 40374041	PUSH: Mary pushes the box away SHOVE: Mary shoves the box away	someone pushes him away she pushes the plate away
ac_id: cbd1726a	PUSH: John pushes the button PRESS: John presses the button	she presses a red button on the wall the nazi officer pushes the snake's eye
ac_id: e017360a	PUSH: Mary pushes the basket under the table	someone pushes the boxes out of his way she pushes him onto the bed

Figure 5.10: An example of alignment between the two resources. On the left, action concepts with prototype videos and captions in IMAGACT; on the right, the video-caption pairs in LSMDC, classified according to the depicted and described action.

This small set of LSMDC items, manually tagged with an IMAGACT action concept, has been used as Gold Standard (GS) for the evaluation of the automatic annotation of action concepts proposed in the following experiments. Figure 5.10 shows an example of LSMDC items aligned with the corresponding action concepts in IMAGACT; Figure 5.11 shows the distribution of the single action instances according to IMAGACT categorisation.

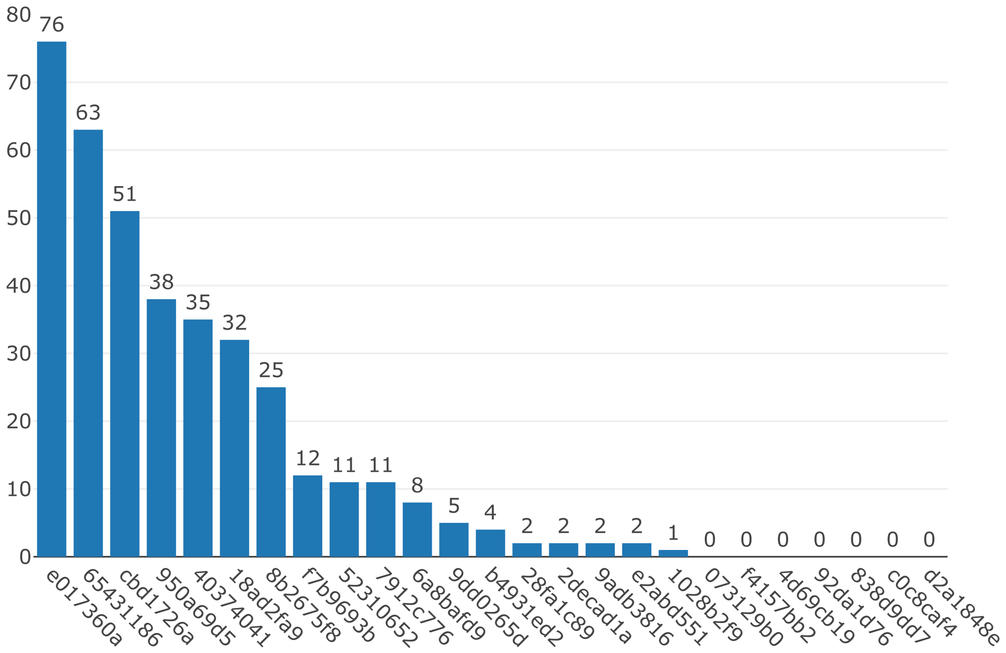


Figure 5.11: Gold Standard distribution of LSMDC action instances in IMAGACT concepts.

5.3 NLP experiments

The objective is to explore the possibility to gather similar visual items from LSMDC under the IMAGACT taxonomy of actions, by assessing the similarity between semantic content of captions that accompany both the items and the target action concepts. The possible outcome of this experiment is twofold: on one side, fine-grained action concept classification for action verbs in LSMDC dataset; on the other, multimodal instance population of IMAGACT Ontology, by importing those video-caption pairs under the corresponding action concepts.

Action verbs are generally ambiguous and complex to treat in NLP tasks, because the relation between verbs and action concepts is not one-to-one (Moneglia, 1996). Consider the following examples of actions referred to by the verb *push* that are cognitively separated:

(5.3.1) John pushes the button

(5.3.2) Mary pushes the table to the corner

The action referred to by 5.3.1 can also be predicated through *press*, while *move* can be used for 5.3.2. These represent two different *pragmatic actions*, despite of the verb used to describe them. Moreover an action concept can be productively applied to a wide set of objects and the same object can be used within several actions: e.g. *pushing a pawn on a chessboard* or *pushing a box into the closet* can be considered as the same action of 5.3.2 performed on different objects; at the same time a *box* can be *pushed/moved to the closet*, *pushed/pressed on the lid to be closed*, *pushed/thrown away*. Nevertheless, it is possible to run into ambiguity problems, such as in a sentence like *John pushes the bottle*: is the agent applying a continuous and controlled force to move the object from position A to position B, or is he carelessly shoving an object away from its location? These are just two of the possible interpretation of this sentence *as is*, without any other lexical information or pragmatic reference.

Given these examples and the observations on action categorisation in 2.2, it is clear that the task of automatically classifying sentences referring to actions in a fine-grained way is not trivial at all, and even humans may need extra information (e.g. images, videos) to precisely identify the exact action and to disambiguate its reference.

5.3.1 Action reference classification as a Semantic Textual Similarity problem

Augmentation of IMAGACT with LSMDC data, or tagging LSMDC data with IMAGACT action concepts (depending on the point of view in observing the problem), can be seen as a classification task, where verbs occurring in LSMDC caption are mapped to the taxonomy of actions defined in IMAGACT. Due to the lack of annotated data in IMAGACT to train a classifier,[36] the action reference classification can be cast as Semantic Textual Similarity (STS) problem, assessing that lexical semantic information encodes, at a certain level, the action referred to by a caption.

STS is a well-known problem in NLP, and various challenges have been set up in recent years (Agirre et al., 2015, 2014, 2016, 2012, 2013; Cer et al., 2017), plus a benchmark dataset for the evaluation of representation models (which also contains a specific sub-set derived from captions of static images), consisting in sentence pairs manually annotated with a similarity judgment raging from 0 (no similarity) to 5 (semantic equivalence).[37]

The simplest way to calculate semantic similarity between captions is to make use of pre-computed word embeddings, which are ready to use for computing similarity between words, sentences and documents. Word embeddings are abstract representations of words in the form of dense vectors, specifically tailored to encode semantic information. They represent an example of the so called *transfer learning*, as the vectors are built to minimise certain objective functions (i.e., guessing the next word in a sentence),

[36] Only 1 caption per each verb connected to every action concept, for a total of 38 captions in 18 concepts for the test-set used.

[37] All the informations about STS tasks and benchmark dataset can be found at `http://ixa2.si.ehu.es/stswiki/`

but they have been successfully applied to different unrelated tasks, such as searching for words that are semantically related. In fact, the comparison of the vectors of two words, sentences or documents is meant to mimic a human score that assesses the grade of semantic similarity between them.

Various models have been presented in the past years that make use of well-known static word embeddings, such as word2vec, GloVe and fastText (Bojanowski et al., 2017; Mikolov et al., 2013; Pennington et al., 2014). In these models a word has always the same representation, regardless of the context where it occurs, and the representation is calculated on the basis of short left-right window of context (word2vec), statistical co-occurrence (GloVe), bag-of-characters (fastText), or other techniques that produce a fixed token representation as a vector.

Recently, the best STS models rely on representations obtained from contextual embeddings, such as ELMO, BERT and XLNet (Devlin et al., 2018; Peters et al., 2018; Yang et al., 2019). Contextualised word embeddings aim at capturing word semantics in different contexts to address the issues of polysemy and context-dependent nature of words. These models produce a different representation for each occurrence of a token, and are trained to predict the next word given the past ones.

5.3.2 Methodology and workflow

The methodology herein described is based on the CAVERAGE algorithm of Lopez de Lacalle et al. (2018), which has been used as a baseline for the evaluation of a multimodal version of STS benchmark dataset (vSTS). Given two sentences, first, the algorithm collects a vocabulary containing all the tokens of both sentences. Then, it reads the pre-computed embeddings for all the words in the vocabulary and builds a tensor of $n \times 300$ for each sentence compiled with its word vectors, where n are the words of the sentence, and 300 the dimension of the corresponding word vector. Subsequently, these tensors are reduced to single vectors, by calculating the centroid of each tensor. The obtained centroid vectors are the representation of the starting sentences. At this point, the semantic similarity is computed, by calculating the cosine similarity of the two centroid vectors, normalised in a 0-5 rage. The following example shows the steps of the process.

Algorithm workflow

Compute the similarity between two sentences:

$$S_1 = A\ man\ pushes\ his\ car$$

$$S_2 = John\ pushes\ a\ button$$

1. Create the tokens dictionary:

$$dict = \begin{bmatrix} a \\ man \\ pushes \\ his \\ car \\ john \\ button \end{bmatrix}$$

2. Read a precomputed word embedding for each token:

$$feats = \begin{bmatrix} a \\ man \\ pushes \\ his \\ car \\ john \\ button \end{bmatrix} \begin{bmatrix} 2.0689 & 0.5411 & \cdots & 6.0002 \\ 3.0114 & 0.0208 & \cdots & 3.2272 \\ 5.6671 & 4.5483 & \cdots & 5.3282 \\ -4.0356 & 1.5464 & \cdots & 7.0774 \\ 0.7437 & -0.9423 & \cdots & -8.0481 \\ 6.8133 & 9.8815 & \cdots & -1.9205 \\ 11.4869 & 6.5412 & \cdots & 0.4032 \end{bmatrix}$$

3. Create the sentence tensors T_1 and T_2:

$$T_1 = \begin{bmatrix} a \\ man \\ pushes \\ his \\ car \end{bmatrix} \begin{bmatrix} 2.0689 & 0.5411 & \cdots & 6.0002 \\ 3.0114 & 0.0208 & \cdots & 3.2272 \\ 5.6671 & 4.5483 & \cdots & 5.3282 \\ -4.0356 & 1.5464 & \cdots & 7.0774 \\ 0.7437 & -0.9423 & \cdots & -8.0481 \end{bmatrix}$$

$$T_2 = \begin{bmatrix} john \\ pushes \\ a \\ button \end{bmatrix} \begin{bmatrix} 6.8133 & 9.8815 & \cdots & -1.9205 \\ 5.6671 & 4.5483 & \cdots & 5.3282 \\ 2.0689 & 0.5411 & \cdots & 6.0002 \\ 11.4869 & 6.5412 & \cdots & 0.4032 \end{bmatrix}$$

4. Calculate tensor centroids C_1 and C_2:

$$C_1 = \begin{bmatrix} 1.4911 & 1.1429 & \cdots & 2.7169 \end{bmatrix}$$

$$C_2 = \begin{bmatrix} 6.5090 & 5.3780 & \cdots & 2.4528 \end{bmatrix}$$

5. Compute the normalised cosine similarity between centroids:

$$sim(S_1, S_2) = norm(cosSim(C_1, C_2))$$

The original algorithm has been adapted, in the input, in order to compute the similarity between each LSMDC item and all the IMAGACT captions linked to the action concepts selected for the test-case. In the output, for each LSMDC caption it ranks all the IMAGACT captions by the obtained similarity score. In this way, the output is a ranking of k most similar IMAGACT captions, labelled with concepts to which the action verb of the LSMDC caption may refer to.

5.3.3 Representation Models

In this section the pre-trained word embeddings used to represent the captions are described. In order to test multiple representations, 4 models have been selected: one-hot encoding, GloVe, BERT and USE. The first two are static representation models, while the latter are two of the so-called *contextual encoders*. Their name derive from the fact that they encode the representation of a piece of text (e.g. word, sentence, document) by analysing each unique token on the basis of the context in which it occurs, i.e. the other tokens that compose that piece of text.

One-hot encoding

This is the most basic textual representation, in which text is represented as binary vectors indicating the words occurring in the context Manning et al. (2008). This way of representing text creates long and sparse vectors, but it has been successfully used in many NLP tasks, thus it has been included in this experiment.

GloVe

The Global Vector model (GloVe)[38] Pennington et al. (2014) is a log-linear model trained to encode semantic relationships between words as vector offsets in the learned vector space, combining global matrix factorization and local context window methods. Since GloVe is a word-level vector model, it has been sufficient to collect in a tensor the embeddings of all words composing the sentence, and then compute the mean of the tensor, in order to obtain the sentence-level representation.

The pre-trained model from GloVe considered in this experiment is the *6B-300d*, counting a vocabulary of 400k words with 300 dimensions vectors and trained on a dataset of 6 billion tokens.

BERT

The Bidirectional Encoder Representations from Transformer (BERT)[39] Devlin et al. (2018) implements a novel methodology based on the so called *masked language model*, which randomly masks some of the tokens from the input, and predicts the original vocabulary id of the masked word based only on its context. Similarly with GloVe, the token embeddings of the last layer of the Neural Net have been collected, and the mean vector to obtain the sentence-level representation has been computed.

The BERT model used in this experiment is the *Large Uncased* (24-layer, 1024-hidden, 16-heads, 340M parameters).

[38] `https://nlp.stanford.edu/projects/glove/`

[39] `https://github.com/google-research/bert`

USE

The Universal Sentence Encoder (USE)[40] Cer et al. (2018) is a model for encoding sentences into embedding vectors, specifically designed for transfer learning in NLP. Based on a deep averaging network encoder, the model is trained for a variety of texts, such as sentences, phrases or short paragraphs, and in a variety of semantic task, including the STS. The encoder returns the corresponding vector of the sentence, and then the similarity has been computed using the usual cosine formula.

5.3.4 Experimental scenario

The experiment has been set up using as data 375 captions from LSMDC, containing 380 instances of *pushing* eventualities, and 38 captions from IMAGACT linked to 18 action concepts, for a total of 14,440 similarity judgements between action references computed by the algorithm. Language in the two resources varies considerably: on one side, captions in IMAGACT are *artificial*, and they only contain minimum syntactic/semantic elements required to describe the action concept; on the other, captions in LSMDC are transcription of more natural spoken language, and usually convey information on more than one action at the same time. The language style in LSMDC is higly descriptive, and lot of details are conveyed through a rich vocabulary of adjectives and adverbs.

It is important to highlight the peculiarities of this experiment, because it is true that representation models have shown very promising results in solving semantic similarity problems, but their performances have always been benchmarked on well-tailored experimental settings.

[40] `https://tfhub.dev/google/universal-sentence-encoder`

The data for which semantic similarity is going to be computed comes from two different domains, i.e. they are formally different and the language and the vocabulary vary between captions from the two sources. STS methods are normally tested on data within the same domain, where often the standard variation resides only in the substitution of elements composing the sentences to be compared, or in the presence/absence of adjuncts arguments.

Examples 5.3.3 and 5.3.4 are caption pairs from STS benchmark dataset:

(5.3.3) *A man is smoking.*
A man is skating.

(5.3.4) *A woman is dancing and singing with other women.*
A woman is dancing and singing in the rain.

Examples in 5.3.5 and 5.3.6, instead, are two captions from the data exploited in this test:

(5.3.5) `LSMDC`: *Finding the prisoner asleep, he shuffles to the cell block door and pushes a button.*

(5.3.6) `IMAGACT`: *John pushes the button*

It is immediately clear that this is a totally different setting from canonical STS tasks. For this reason, this experiment can be considered as a *non-conventional* scenario for semantic textual similarity

5.3.4.1 Pre-processing and test scenarios

As an attempt to leverage some differences between IMAGACT and LSMDC captions, and to improve the performances of the STS algorithm used, some basic pre-processing has been applied.

Length of captions in the two resources vary, and for this reason, it has been necessary to re-wrote one simplified clause for each verb occurrence in LSMDC captions. To do that, it has been sufficient to make use of the CoNLL output of StanforNLP Qi et al. (2018), produced during the textual analysis of the resource (See Table 5.4 for an example). With NLTK[41] Python Library (Loper and Bird, 2002) it is possible to build dependency graphs and parse trees from CoNLL-style annotated data.

Thus, to derive simple clauses from long and complex captions, a simple algorithm has been developed, that acts like a sort of arborist that prunes parse trees. First, it converts the CoNLL `string` to a `DependencyGraph` object, from where it collects all the dependencies of each single word as `dicts`. Then, it selects the dependencies of words tagged as *VERB* in the `UPOSTAG` field of the CoNLL, and from there it re-collect all the words listed in the dependencies of words dependent from each target verb, with exclusion of branches that stem from other verbs. The following example shows how the caption splitter algorithm works.

[41] `https://www.nltk.org`

Splitting Algorithm

Derive simple clauses from the following sentence:

S = Turning her head, she pushes his face off her shoulder and moves an arm from her torso, then quickly retrieves the phone.

1. Read the StanfordNLP output of the sentence in CoNLL format:

```
1   Turning   turn    VERB    VBG    [...]  6    advcl
2   her       she     ADP     PRP$   [...]  3    nmod:poss
3   head      head    NOUN    NN     [...]  1    dobj
4   ,         ,       PUNCT   ,      [...]  _    _
5   she       she     PRON    PRP    [...]  6    nsubj
6   pushes    push    VERB    VBZ    [...]  0    ROOT
7   his       he      PRON    PRP$   [...]  8    nmod:poss
8   face      face    NOUN    NN     [...]  6    dobj
9   off       off     ADP     IN     [...]  11   case
                              [...]
```

2. Parse the CoNLL and create a `DependencyGraph`:

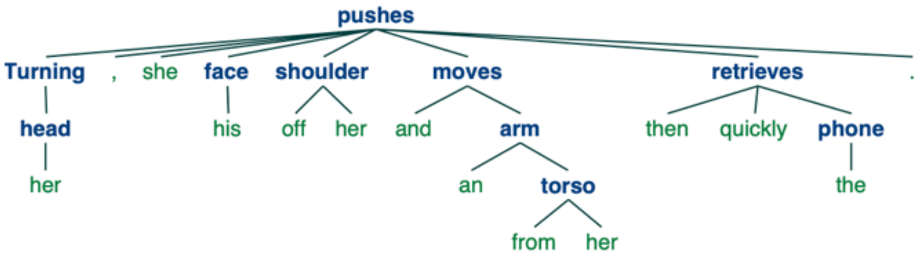


Figure 5.12: Parse tree from `DependencyGraph` of *S*.

3. Locate all the `VERB`s and get their sub-trees:

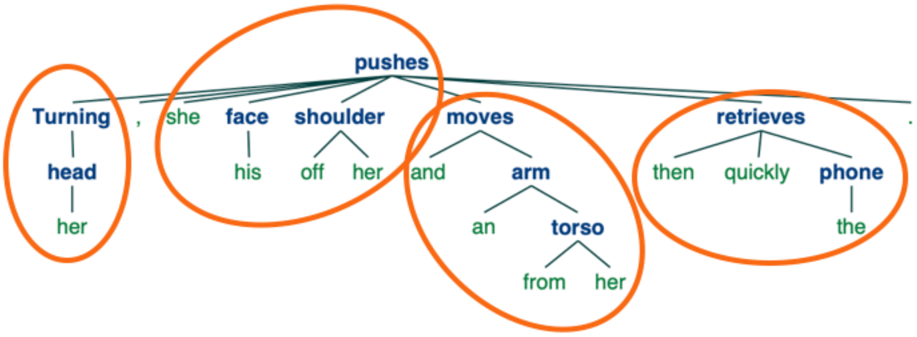


Figure 5.13: `VERB`s sub-trees location.

4. Re-write simple clauses by collecting and ordering all the words corresponding to the sub-trees:

$$S_1 = \textit{Turning her head}$$

$$S_2 = \textit{, she pushes his face off her shoulder .}$$

$$S_3 = \textit{and moves an arm from her torso}$$

$$S_4 = \textit{then quickly retrieves the phone}$$

Table 5.8 shows some examples of the output from the splitting process on complex LSMDC captions.[42]

Video	Original caption	Simplified clause	
3089_XMEN_FIRST_CLASS_01.36.34.076-01.36.40.337	With wide eyes, he *crosses* the bridge, ***pushes*** a fellow officer out of his way, and ***presses*** a button.	With wide eyes , he *crosses* the bridge .	✗
		pushes a fellow officer out of his way	✓
		and ***presses*** a button	✓
1049_Harry_Potter_and_the_chamber_of_secrets_01.07.45.225-01.07.53.561	As he *crashes* onto the platform, SOMEONE *hauls* him to his feet and ***pushes*** him back towards SOMEONE.	As he *crashes* onto the platform	✗
		, someone *hauls* him to his feet .	✗
		and ***pushes*** him back towards someone	✓

Table 5.8: Examples of simple clauses obtained with the splitter algorithm. Only the clauses with the target verb are used (✓), and the rest is ignored (✗).

[42] Henceforth, only simplified clauses from LSMDC will be considered, but for the sake of simplicity they will still referred as LSMDC captions in the description and discussion of experiments and results.

LSMDC dataset is anonymised, i.e. the pronoun *someone* is used in place of all proper names; on the contrary, captions in IMAGACT always have a proper name (e.g. John, Mary). Thus, all proper names in IMAGACT have been automatically replaced with *someone*, without the use of any sophisticated system. IMAGACT captions are written in a controlled language, and they use a restricted set of common proper names (i.e. John, Mary). It has been sufficient to read the corpus of names from NLTK, and then substitute with *someone* each token in IMAGACT captions matching with the list of names.

Stop-words are often the first lexical elements to be pruned out from texts, prior of any computation, because they do not convey semantic information, and they sometimes introduce noise in the process. For these reasons Stop-words have been filtered out. First, a list of stop-words has been read from NLTK and concatenated with the punctuation list from the Python `string` module, for the sake of simplicity, in order to filter any semantically empty token. Then, the removal has been executed in the moment of calculating the similarity between pairs. In this way, contextual models made use of stop-words to calculate features for each LSMDC clause and IMAGACT caption, but those embeddings have been discharged when computing the centroid vector representation of each sentence.

With these pre-processing operations, it has been possible to obtain 4 variants of testing data:

- `plain` (LSMDC splitting only);
- `anonIM` (anonymisation of IMAGACT captions by substitution of proper names with *someone*);
- `noSW` (stop-words removing from both resources);
- `anonIM+noSW` (combination of the two previous ones).

The STS algorithm has been run once for each variant of the testing data, and making use of all the embedding models considered.

5.4 Results

In order to benchmark the performances of the four models, a baseline has been defined that randomly ranks the 18 target action concepts for each LSMDC caption in the GS. To evaluate the results, recall@k has been used, which is a typical evaluation metric in information retrieval. It means that the ratio of relevant/correct action concept labels in the top positions of the similarity ranking, between all IMAGACT captions and the target LSMDC caption, is calculated at different k for each model.

Table 5.9 shows the recall results at different k: 1, 3, 5, and 10. Best overall results (per column of recall@k) are in bold, while the top score for each run on the 4 variants (e.g. `plain`) are underlined, thus it is also possible to observe the effectiveness of the different pre-processing strategies in detail.

Model	Pre-processing	recall@1	recall@3	recall@5	recall@10
one-hot encoding	plain	0.195	0.379	0.484	0.655
	noSW	0.139	0.271	0.411	0.687
	anonIM	0.197	0.4	0.482	0.624
	anonIM+noSW	0.155	0.329	0.453	0.65
GloVe	plain	0.213	0.392	0.553	**0.818**
	noSW	0.182	0.408	0.505	0.755
	anonIM	0.218	0.453	**0.568**	0.774
	anonIM+noSW	**0.279**	0.453	0.553	0.761
BERT	plain	0.245	0.439	0.539	0.632
	noSW	0.247	**0.484**	0.558	0.679
	anonIM	0.239	0.434	0.529	0.645
	anonIM+noSW	0.2	0.384	0.526	0.668
USE	plain	0.213	0.403	0.492	0.616
	noSW	0.171	0.376	0.461	0.563
	anonIM	0.239	0.471	0.561	0.666
	anonIM+noSW	0.179	0.426	0.518	0.637
Random baseline		0.120	0.309	0.447	0.658

Table 5.9: Semantic textual similarity results for the models tested on IMAGACT-LSMDC scenario. In **bold**, the best overall recall@k; underlined, the top score for each variant of the four test sets (e.g. `plain`).

All models show slightly better results if compared to the baseline, but not much higher. Given the difficulty, novelty and *non-conventionality* of the action reference classification task, low results were predictable. In fact, even for a human annotator, the majority of LSMDC-IMAGACT caption pairs are ambiguous, and the correct understanding of the referred action is almost impossible exclusively by mean of the textual information contained in the captions, and without taking into account the visual information conveyed with the videos.

Regarding the pre-processing, any strategy (`noSW`, `anonIM`, `anonIM+noSW`) seems to make no substantial difference in terms of performances. However, anonymisation and stop-words removal (`AnonIM+noSW`) positively affect the results of GloVe with respect to the other models, in the sense that this model works better than the others within this scenario. Another interesting observation is that stop-words removal (`noSW`), with respect to the other pre-processing strategies, slightly improves the individual results of BERT, i.e. the algorithm run with BERT representations on the test set without stop-words obtains the best score if compared to the other runs with the same model on the other test sets. It is important to remark that stop-words have been used with BERT and USE for computing the sentence representation and single tokens embedding, but they have been ignored when the sentence tensor has been built. Thus, the contextual information of stop-words is encoded in the representation of neighbouring words, but the actual representation of the stop-word tokens has not been used. Lastly, anonymisation of IMAGACT captions (`anonIM`), gives few points more to USE, that achieves its best individual performance in this scenario, with respect to the results obtained on `plain` data or applying other pre-processing strategies.

Worthy of attention is the fact that GloVe model outperforms the baseline and other models in recall@10 in all the runs, with a peak of ∼0.2 for the `plain` variant of the test sets. It is not an exciting result, but it shows that STS with pre-trained word embeddings might be effective to speed up manual annotation tasks by filtering all the possible candidates and reduce the pool of possibilities, without any computational cost.[43] Probably, one reason to explain the lower trend in results obtained by contextual models (BERT, USE) could be that these systems have been penalised by the splitting process of LSMDC captions in simple clauses. Examples in Table 5.8 show some good splitting results, while the processing of some other captions leads to less-natural sentence splitting, and this might influence the global result of all the models, especially those which are sensitive to the context and its variations.

Some considerations on the results are necessary. First, the models were not trained *ad-hoc*, i.e. making use of the same (or comparable) data used in the test. Due to the scarcity of examples (i.e. captions) for each IMAGACT action concept, it has been impossible to properly set up a training, thus pre-trained versions have been used for all the models. It means that the representations are obtained from very general and big corpora, thus the target information of action verbs is present, but not prominent enough in the training data for the objective of these experiments.

Moreover, it is important to remark that results obtained in a *non-conventional* scenario, such the one of compare pieces of texts from different domains herein proposed, can be counter-intuitive if compared to results obtained in conventional ones, where contextual encoders usually outperform by far static models. In order to track the behaviour of GloVe, BERT and USE when benchmarked on a more conventional scenario, similar experiments have been run on the STS-benchmark dataset[44] Cer et al. (2017).

[43] More on this regard in Section 6.1.

[44] The STS-benchmark dataset is at the basis of the official evaluation of Semantic Textual Similarity representation models. The dataset is publicly available at `http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark`

The task is similar to the one presented herein: it consists in the assessment of pairs of sentences according to their degree of semantic similarity. On the STS-benchmark dataset, models are evaluated in terms of Pearson correlation between machine scores and human judgments, i.e. the score assigned by the annotators, used as Gold Standard.

Model	**Pre-processing**	**Pearson**
USE	plain	**0.702**
BERT	plain	0.47
GloVe	plain	0.336

Table 5.10: Results on STS-benchmark.

Table 5.10 shows the expected results: the two contextual encoders outperform GloVe model in a consisted way, and USE achieves the best score, with a margin of ∼0.3 with respect to BERT. It confirms that model performances are task-dependent, thus the anomalous behaviour registered in these experiments has to be charged to the difficulty of the task, namely action reference classification of data from different domains.

5.4.1 Limits of IMAGACT for computational implementation

IMAGACT resource is particularly interesting for applications other than its original goal, i.e. support tool for early learners of L2. In fact, IMAGACT has been successfully used in L2 learning scenarios (2.5.1 and 2.5.2), and it proves that its theoretical framework is strong and effective. Moreover, IMAGACT action concepts have been used in a battery of tests for the assessment of semantic memory decay (2.5.3).

The main problem in exploiting IMAGACT in computational implementation, as it also emerged in 4.1 and 4.2, is that the lexical information available with the action concepts is limited to one caption for each verb pointing to each concept. Formally, captions are very similar among them, and this makes more difficult to disambiguate without watching videos to whom they belong.[45]

To illustrate this problem in a practical perspective, consider Figure 5.14. It shows a heat-map of the similarities for captions linked to action concepts of *push* in IMAGACT, computed with the same algorithm used in the previous experiments; numbers stand for the `ID` of each caption in the DB, and their colour is the action concept they belong. The heat-map is a matrix of similarities translated in colours: the darker is the cell corresponding to an `ID`s-pair, the higher is the similarity score.

[45] This problem has been highlighted also in Ravelli (2016), in the scenario of exploiting IMAGACT captions as parallel data for translation memories and machine translation.

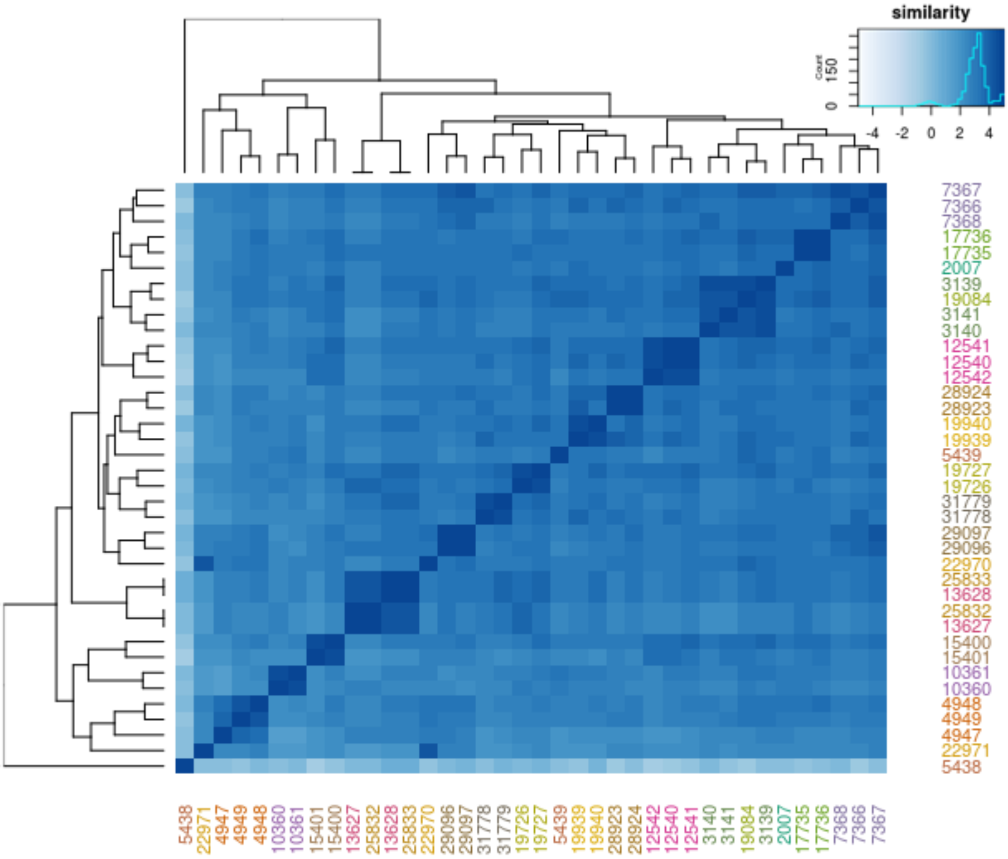


Figure 5.14: Heat-map of the similarities in IMAGACT. Labels on the X and Y axes correspond to caption ids, and their colour denotes the action concept. Above all, overall similarity distribution is shown (range [-5, 5]).

The first observation is that almost the whole matrix is coloured in blue, and it means that the overall similarity of all captions is very high and uniform across all the comparisons, which makes difficult the correct discrimination of action concepts with textual information only. Nevertheless, in most of the cases, the captions linked to the same action concept are the most similar ones, which means that there is some kind of textual coherence when modelling action concepts with making use of captions. It is the case of the three captions coloured in pink (12540, 12541 and 12542) linked to the action concept 65431186, which is part of the extension of *push_T1*. This concept refers to *pushing* an object with the application of a continuous force and with the goal of moving the object.

There are also cases where the captions of the same action concept are not the most similar, meaning that many actions are closely related, and captions on their own do not sufficiently specify the described action. It is the case of the two captions coloured in brown (5438 and 5439) linked to the action concept 2decad1a, which is part of the extension of *push_T3*. The corresponding cell is very light-coloured, meaning that the two caption are not considered as correlated by the algorithm.

Interestingly, looking at the Gold Standard made of LSMDC captions manually annotated, these two action concepts show diametrically opposite distribution. In fact, 65431186 appears in 63 examples, while 2decad1a has only 2 examples.[46] It means that the action reference encoded in concept 65431186 may be easier to model in case of the training of a classifier, because the textual information connected to it is coherent, i.e. IMAGACT captions are highly similar, and there are enough examples. On the contrary, it would be impossible with concept 2decad1a.

[46] For details on the distribution of IMAGACT action concepts in the manually annotated Gold Standard, see 5.11.

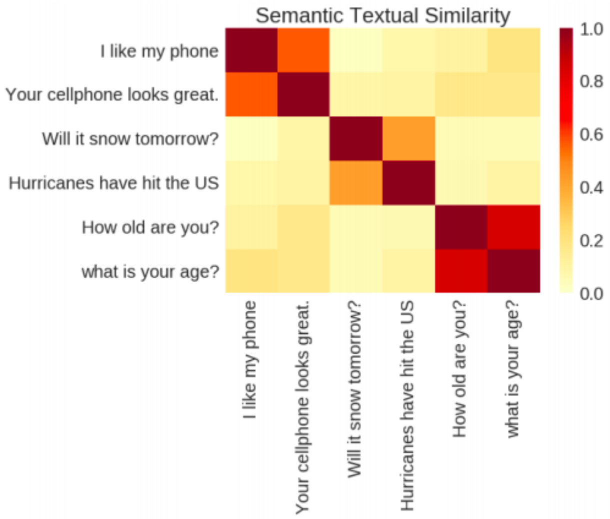


Figure 5.15: Heat-map of the similarities in the STS-Benchmark dataset obtained with USE. Taken from Cer et al. (2018).

Moreover, continuing the comparison with STS-Benchmark dataset from 5.4, the heat-map of the similarities for some captions in that dataset is shown in Figure 5.15, as reported by Cer et al. (2018). The purpose is purely to offer a visual impression of the difficulty in working within a restrict field, limited to the *pushing* eventualities, with textual information such ambiguous from the point of view of representation models, and considering very few examples for each action concept.

Chapter 6

Conclusions

In the present work, a sort of in-depth exploration of IMAGACT has been traced, with the focus of exploiting the resource in NLP tasks. Starting from the Introduction (Chapter 1), the idea of making use of IMAGACT multimodal action conceptualisation has been drawn, with some reflections on evidences of the deep linking between Language and Vision, and on the fact that action plays a key role in this linkage. Thus, the multimodal and multilingual features of IMAGACT Ontology of Action Verbs have been described (Chapter 2), with also some details on the framework of the resource building. It followed a concrete case-study on IMAGACT internal data, that lead to the proposal of an inter-linguistic manual mapping between the Action Types of verbs which refer to *cutting* eventualities in English and Italian (Chapter 3).

Then, a series of experiments have been presented (Chapter 4), involving the exploitation of IMAGACT in linking with other resources and building deliverable NLP products, such as the Ref-vectors of action verbs. One of the experiments has been described extensively (Chapter 5): the visual enrichment of IMAGACT through instance population of its action concepts, making use of Audio Description of movies for visually impaired people.

From this last experiment it emerged that dealing with *non-conventional* scenarios, such as the one of assessing action reference similarity between texts from different domains, is not a trivial task, given that fine-grained differences among action concepts are difficult to derive purely from the textual representation. Moreover, assessing action reference similarity itself has not been addressed yet by the NLP community, and considering the multimodal nature of the data presented in this work, such a challenge could be interesting also for the Computer Vision community.

In fact, a dataset composed of LSMDC video-caption pairs, annotated with IMAGACT action concepts,[1] could be a stimulating playground for modelling action recognition on the basis of a fine-grained categorisation. The annotation of action concepts, instead of canonical verb senses, leads to a more faithful representation of the mental classification of action patterns, gaining more granularity and scalability than a simple taxonomy or a list of *verb_object* tuples. Thus, the gathering of videos and captions under an action concept prototype smooths over the distinction made by verb lemmas. In this way, the focus is on pragmatic action execution instead of linguistic categorisation.

[1] See 5.2.3.2 for the details on the manual annotation of LSMDC data.

6.1 Future Works

A dataset enriched with fine-grained action categorisation may open up to a series of challenging tasks. First, action reference similarity problem could be proposed as a novel STS task. Recently, given the interest in exploring multimodal representations on sentence similarity, Lopez de Lacalle et al. (2018) proposed the Visual Semantic Textual Similarity dataset (vSTS), that is derived from a sub-set of the STS-benchmark dataset composed of caption pairs, linked with the images those captions are referring to. In this view point, a dataset of video-caption pairs, with the plus of a linguistically derived categorisation of the actions depicted, would be highly desirable.

Such a dataset may arouse interest also in the Computer Vision community. In fact, LSMDC is a popular dataset by itself, and various tasks have already been settled making use of it.[2] Given that one of the objectives of the present work was to enrich LSMDC captions with the annotation of IMAGACT categorisation of action, it is sure that these data could be used for at least two challenging tasks. First, a dataset of video-caption pairs semantically enriched with IMAGACT action categorisation could be used in a challenge on multimodal fine-grained action recognition in movies. Second, IMAGACT is also multilingual, i.e. it could be used as a framework for the translation of action verbs.[3] Thus, such a dataset could also be exploited for multilingual automatic captioning of videos.

[2] For more details regarding LSMDC challenges and tasks, see Table 5.2 and the description of the dataset in Section 5.2.

[3] As stated in Chapter 5.3, Action verbs are generally ambiguous and complex to treat in NLP tasks, because the relation between verbs and action concepts is not one-to-one (Moneglia, 1996), and for this reason action verb disambiguation is a critical task for Natural Language Understanding and Machine Translation (Moneglia, 2011).

It is clear that a collection of visual-textual items annotated with IMAGACT categorisation of action verbs would be particularly desirable from various points of view. At the moment, the only limitation is the scarcity of data, i.e. there are no enough exemplars for each action concept. Considering IMAGACT actual data, one synthetic caption for each verb connected to an action concept and one unique prototypical video are not sufficient for any kind of training. That said, IMAGACT *as is* cannot be implemented in any automatic language or visual modelling, and this is the main reason why pre-trained representation models have been used in the experiments described in Section 5.3.

The manual annotation of LSMDC video-caption pairs for the Gold Standard has been particularly time consuming, because it is not an easy task to assign an action concept, and it requires to watch the video prototypes of all the possible action concepts, before to decide to which of them a new video belongs. The Gold Standard used herein is composed of only 375 LSMDC video-caption pairs, for a total of 380 actions correctly annotated. Out of 25 total *push* action concepts, 18 found at least one exemplar in the LSMDC dataset, and just 7 action concepts have been annotated in more than 25 LSMDC examples (Figure 5.11). With these numbers, it would be difficult to train an efficient classifier. Thus, it is necessary to augment the number of annotations in order to build an exploitable dataset.

In this regard, it is mandatory to plan an extensive annotation campaign through one of the numerous crowd-sourcing platforms specialised in the annotation of natural language data. With a small investment, it would be possible to annotate all LSMDC video-caption pairs with the corresponding action concepts in a short time, and thus obtain both a semantically enriched multimodal dataset for action recognition, and the instance population of the

IMAGACT Ontology. As emerged from the experiments in Section 5.3, even a naive Semantic Textual Similarity algorithm with pre-trained word embeddings might be effective to speed up manual annotation tasks. In fact, such an algorithm is able to filter candidates on the basis of the semantic similarity, and thus reduce the pool of possibilities without any computational cost. It means that an annotator, instead of watching and analysing all the videos of action concepts connected to the target verb, may be offered only the most-likely similar ones, thus each item may be processed faster than having to manually explore the complete extension of a verb according to IMAGACT. Considering that many general verbs extend to a wide range of different actions, it may be highly time consuming to examine every concept.

Moreover, in the framework of a real annotation campaign involving multiple annotators, it would be particularly interesting to observe to what extend the visual stimuli offered by IMAGACT video prototypes can be interpreted clearly by more than one annotator and applied uniformly, and thus calculate the inter-annotator agreement among the participants. In this way, the resulting annotated data would be trustable and coherent, and also the IMAGACT conceptualisation would be extensively revised and validated.

Last but not least, it would be fruitful to make use of the information conveyed by the visual modality in IMAGACT, and thus to take advantage of the complementarity of visual and textual representations by combining the two modalities in one single model (Kiros et al., 2014). In this regard, it would be interesting to exploit dense video captioning models, such as videoBERT (Sun et al., 2019), which may obtain better results in assessing action reference similarity among video-caption pairs by accessing also the visual information of each dataset item.

Bibliography

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. (2014). SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A pilot on semantic textual similarity. In **SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics*, pages 385–393. Universidad del Pais Vasco, Leioa, Spain.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Agirre, E. and Soroa, A. (2007). Semeval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 7–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ahmad, K. and Rogers, M. (2001). The Analysis of Text Corpora for the Creation of Advanced Terminology Databases. In Wright, S. E. and Budin, G., editors, *Handbook of Terminology Management*. J. Benjamins, Amsterdam; Philadelphia, PA.

Alagić, D., Šnajder, J., and Padó, S. (2018). Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Albano Leoni, F. (2003). Tre progetti per l'italiano parlato: Avip, api, clips. *Italia linguistica anno Mille. Italia linguistica anno Duemila. Atti del XXXIV Congresso Internazionale di Studi della Società linguistica Italiana (SLI)*, pages 675–683.

Arbib, M. A. and Rizzolatti, G. (1997). Neural expectations: a possible evolutionary path from manual skills to language. *Communication & Cognition*, 29:393–424.

Arefyev, N., Ermolaev, P., and Panchenko, A. (2018). How much does a word weigh? Weighting word embeddings for word sense induction. *CoRR*, abs/1805.09209.

Arma, S. (2011). *The language of filmic audio description: a corpus-based analysis of adjectives*. PhD thesis, Università degli Studi di Napoli Federico II, Napoli.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada.

Baker, M. C. (1997). Thematic roles and syntactic structure. In *Elements of grammar*, pages 73–137. Springer.

Ballester Casado, A. (2007). La audiodescripción: apuntes sobre el estado de la cuestión y las perspectivas de investigación. *Tradterm*, 13:151–169.

Baroni, M. and Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 87–90. Association for Computational Linguistics.

Blundell, B., Sadrzadeh, M., and Jezek, E. (2017). Experimental Results on Exploiting Predicate-Argument Structure for Verb Similarity in Distributional Semantics. In Dobnik, S. and Lappin, S., editors, *Conference on Logic and Machine Learning in Natural Language (LaML 2017)*, Gothenburg.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics (TACL)*, 5(1):135–146.

Bowerman, M. (1990). Mapping thematic roles onto syntactic functions: are children helped by innate linking rules? *Linguistics*, 28(6):1253–1290.

Bowerman, M. (2005). Why can't you "open" a nut or "break" a cooked noodle? Learning covert object categories in action word meanings. In Gershkoff-Stowe, L. and Rakison, D. H., editors, *Building Object Categories in Developmental Time*, pages 209–243. Psychology Press.

British National Corpus (2007). British National Corpus, version 3. BNC XML edition.

Brown, S. W., Gagliardi, G., and Moneglia, M. (2014). IMAGACT4ALL. Mapping Spanish Varieties onto a Corpus-Based Ontology of Action. *CHIMERA Romance Corpora and Linguistic Studies*, 1.

Bruni, M., Uricchio, T., Seidenari, L., and Del Bimbo, A. (2016). Do Textual Descriptions Help Action Recognition? In *the 2016 ACM*, pages 645–649, New York, New York, USA. ACM Press.

Carreira, J., Noland, E., Hillier, C., and Zisserman, A. (2019). A Short Note on the Kinetics-700 Human Action Dataset. *CoRR*, abs/1907.06987.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *CoRR*, abs/1803.11175.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* University of Chicago Press.

Choi, S. and Bowerman, M. (1991). Learning to express motion events in English and Korean: the influence of language-specific lexicalization patterns. *Cognition*, 41(1-3):83–121.

Corsi, G. (2016). *Competenze semantico-lessicali nell'eloquio dei pazienti afasici. Analisi dell'uso dei verbi generali mediante la batteria SMAAV.* Bachelor's thesis, University of Florence.

Cresti, E. and Moneglia, M., editors (2005). *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*. Number 15 in Studies in corpus linguistics. J. Benjamins, Amsterdam; Philadelphia, PA.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., and Wray, M. (2018). Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *The European Conference on Computer Vision (ECCV)*.

De Lacalle, M. L., Laparra, E., Aldabe, I., and Rigau, G. (2016). Predicate Matrix: automatically extending the semantic interoperability between predicate resources. *Language Resources and Evaluation Conference*, 50(2):263–289.

De Mauro, T., Mancini, F., Vedovelli, M., and Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato (LIP)*. Etaslibri, Milano.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental brain research*, 91(1):176–180.

Dowty, D. R. (1979). *Word meaning and Montague grammar : the semantics of verbs and times in generative semantics and in Montague's PTQ*. Dordrecht ; Boston : D. Reidel Pub. Co.

Erkelens, M. (2003). The semantic organization of "cut" and "break" in Dutch. Master's thesis, Free University of Amsterdam.

Fabian, M. S., Gjergji, K., Gerhard, W., and others (2007). Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International World Wide Web Conference, WWW*, pages 697–706.

Fadiga, L., Fogassi, L., Pavesi, G., and Rizzolatti, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *Journal of neurophysiology*, 73(6):2608–2611.

Fanetti, V. (2018). *Le abilità semantiche in pazienti con demenza: Applicazione della batteria SMAAV*. Bachelor's thesis, University of Florence.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.

Fillmore, C. J. (1977). Scenes-and-frames semantics. In Zampolli, A., editor, *Linguistic Structures Processing*, number 59 in Fundamental Studies in Computer Science. North Holland Publishing.

Fillmore, C. J. (1982). Frame semantics. In Hakhoe, H. Ŏ., editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Fillmore, C. J. and Baker, C. (2010). A frames approach to semantic analysis. In Heine, B. and Narrog, H., editors, *The Oxford handbook of linguistic analysis*. Oxford University Press.

Fillmore, C. J. and Baker, C. F. (2001). Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*, Pittsburgh. NAACL, NAACL.

Frazier, G. M. (1975). The autobiography of Miss Jane Pitman: An all-audio adaptation of the teleplay for the blind and visually handicapped. Master's thesis, San Francisco State University.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

Gagliardi, G. (2013). *Validazione dell'ontologia dell'azione IMAGACT per lo studio e la diagnosi del mild cognitive impairment (MCI).* PhD thesis, University of Florence.

Gerz, D., Vulic, I., Hill, F., Reichart, R., and Korhonen, A. (2016). SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. *EMNLP*.

Gregori, L., Panunzi, A., and Ravelli, A. A. (2016). Linking IMAGACT Ontology to BabelNet through Action Videos. In *Proceedings of the Third Italian Conference on Computational Linguistics & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (CLiC-it & EVALITA).*

Gregori, L., Ravelli, A. A., and Panunzi, A. (2015). Linking dei contenuti multimediali tra ontologie multilingui: i verbi di azione tra IMAGACT e BabelNet. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it)*, pages 150–154. Accademia University Press.

Gregori, L., Varvara, R., and Ravelli, A. A. (2019). Action Type induction from multilingual lexical features. *Procesamiento del Lenguaje Natural*, 63:85–92.

Halliday, M. A. K. (1985). *An introduction to functional grammar.* Arnold, New York.

Halliday, M. A. K. (1989). *Spoken and Written Language.* Oxford University Press, Hong Kong.

Hanks, P. (2004). Corpus pattern analysis. In *Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Universite de Bretagne-Sud.*

Hanks, P. and Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue française de linguistique appliquée*, 10(2):63–82.

ITC (2000). Guidance On Standards for Audio Description.

Jarrett, C. (2012). Mirror neurons: the most hyped concept in neuroscience. *Psychology Today*, 10.

Jezek, E., Feltracco, A., Gatti, L., Magnolini, S., and Magnini, B. (2016). Mapping Semantic Types onto WordNet Synset. In *Proceedings of the Twelfth Joint ACL - ISO Workshop on Interoperable Semantic Annotation (Isa '12)*, Portorose, Slovenia.

Jezek, E., Magnini, B., Feltracco, A., Bianchini, A., and Popescu, O. (2014). T-PAS: a resource of corpus-derived Types Predicate-Argument Structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. (2017). The Kinetics Human Action Video Dataset. *CoRR*, abs/1705.06950.

Kipp, M. (2001). ANVIL A Generic Annotation Tool for Multimodal Dialogue. In *Seventh European Conference on Speech Communication and Technology*, pages 1367–1370.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Laparra, E. and Rigau, G. (2009). Integrating WordNet and FrameNet using a Knowledge-based Word Sense Disambiguation Algorithm. *RANLP*.

Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. *CVPR*.

Levin, B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation.* University of Chicago Press.

Levy, O. and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Lippi, A. (2018). *Validazione della batteria SMAAV 2.0 per lo studio e la diagnosi di deficit semantico-lessicali. Analisi delle competenze semantico-lessicali in compiti di comprensione mediante l'uso di verbi generali.* Bachelor's thesis, University of Florence.

Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lopez de Lacalle, O., Soroa, A., and Agirre, E. (2018). Evaluating Multimodal Representations on Sentence Similarity: vSTS, Visual Semantic Textual Similarity Dataset. *CoRR*, abs/1809.03695.

Luhn, H. P. (1960). Keyword-in-context Index for Technical Literature (KWIC Index). *American Documentation*, 11(4):288–295.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Majid, A., Bowerman, M., van Staden, M., and Boster, J. S. (2007). The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics*, 18(2):133–20.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA.

Matamala, A. and Villegas, M. (2016). Building an audio description multilingual multimodal corpus : the VIW project. *MMC2016 (Multimodal Corpora: Computer vision and language processing)*, 11.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Moneglia, M. (1996). Prototypical vs. non-prototypical predicates: ways of understanding and the semantic partition of lexical meaning. *QDLF - Quaderni del Dipartimento di Linguistica*, 7:163–181.

Moneglia, M. (1998). Teoria empirica del senso e partizione semantica del lessico. *Studi di grammatica italiana*, XVII:363–398.

Moneglia, M. (2010). Predicati generali ad alta frequenza nei corpora orali delle lingue romanze. In Iliescu, M., Siller-Runggaldier, H., and P., D., editors, *Actes du XXV Congrès International de Linguistique et de Philologie Romanes*, volume 4, pages 497–506.

Moneglia, M. (2011). Natural Language Ontology of Action: A Gap with Huge Consequences for Natural Language Understanding and Machine Translation. In Vetulani, Z. and Mariani, J., editors, *Human Language Technology Challenges for Computer Science and Linguistics*, pages 379–395. Springer International Publishing.

Moneglia, M. (2014). The variation of Action verbs in multilingual spontaneous speech corpora. *Spoken Corpora and Linguistic Studies*, 61:152.

Moneglia, M., Brown, S., Frontini, F., Gagliardi, G., Khan, F., Monachini, M., and Panunzi, A. (2014a). The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the representation of lexical encoding of Action. *LREC*, pages 3425–3432.

Moneglia, M., Brown, S., Kar, A., Kumar, A., Ojha, A. K., Mello, H., Niharika, Jha, G. N., Ray, B., and Sharma, A. (2014b). Mapping Indian Languages onto the IMAGACT Visual Ontology of Action. In Jha, G. N., Bali, K., Sobha, L., and Banerjee, E., editors, *Proceedings of WILDRE2 - 2nd Workshop on Indian Language Data: Resources and Evaluation at LREC'14*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Moneglia, M. and Panunzi, A. (2007). Action Predicates and the Ontology of Action across Spoken Language Corpora. In Plá, M. A. and Declerk, T., editors, *International Workshop on the Semantic Representation of Spoken Language*, pages 51–58, Salamanca.

Moneglia, M. and Panunzi, A. (2010). I verbi generali nei corpora di parlato. Un progetto di annotazione semantica cross-linguistica. In Cresti, E. and Korzen, I., editors, *Language, Cognition and Identity*, pages 27–46. Florence University Press.

Mouyiaris, A. (2019). I verbi d'azione del greco nell'ontologia IMAGACT. Master's thesis, University of Florence.

Mutlak, M. (2019). *I verbi di azione dell'arabo standard nell'ontologia dell'azione IMAGACT*. PhD thesis, University of Florence, Florence.

Navigli, R. and Ponzetto, S. P. (2012a). BabelNet - The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. and Ponzetto, S. P. (2012b). Multilingual WSD with Just a Few Lines of Code - the BabelNet API. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL).*

Orero, P. (2004). Audiovisual translation: A new dynamic umbrella. In Orero, P., editor, *Topics in Audiovisual Translation*, pages vii–xiii. J. Benjamins, Amsterdam; Philadelphia, PA.

Panunzi, A., Gregori, L., and Ravelli, A. A. (2018). One event, many representations. mapping action concepts through visual features. In Pustejovsky, J. and van der Sluis, I., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English Gigaword Fifth Edition. *Linguistic Data Consortium, LDC2011T07*, 12.

Pastra, K. and Aloimonos, Y. (2012). The minimalist grammar of action. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1585):103–117.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. Stanford University, Palo Alto, United States.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Kokmen, E., and Tangelos, E. G. (1997). Aging, memory, and mild cognitive impairment. *International psychogeriatrics*, 9 Suppl 1:65–69.

Piety, P. J. (2004). The Language System of Audio Description: An Investigation as a Discursive Process - Audio Description - August 2004. *Journal of Visual Impairment Blindness*, 98(8):453–469.

Pinker, S. (2009). *Language learnability and language development, with new commentary by the author*, volume 7. Harvard University Press.

Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2018). Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Ravelli, A. A. (2016). La disambiguazione del riferimento dei verbi azionali attraverso la risorsa multilingue IMAGACT. Prospettive di sfruttamento e integrazione dei dati. Master's thesis, University of Florence.

Ravelli, A. A., Gregori, L., and Feltracco, A. (2017). Evaluating a rule based strategy to map IMAGACT and T-PAS. In *Proceedings of the Forth Italian Conference on Computational Linguistics (CLiC-it)*, pages 281–286. Accademia University Press.

Ravelli, A. A., Gregori, L., and Varvara, R. (2019). Comparing Ref-Vectors and Word Embeddings in a Verb Semantic Similarity Task. *NL4AI@AI*IA*.

Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. (2015). A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., and Schiele, B. (2017). Movie Description. *International Journal of Computer Vision*, 123(1):94–120.

Rosch, E. (1978). Principles of Categorization. In Rosch, E. and Lloyd, B., editors, *Cognition and Categorization*. Lawrence Elbaum Associates.

Ryzhova, D., Kyuseva, M., and Paperno, D. (2016). Typology of adjectives benchmark for compositional distributional models. In *Proceedings of the 10th Language Resources and Evaluation Conference*, pages 1253–1257.

Salway, A. (2007). A corpus-based analysis of audio description. In Díaz Cintas, J., Orero, P., and Remael, A., editors, *Media for All*, pages 151–174. Rodopi, Leiden.

Salway, A. and Tomadaki, E. (2002). Temporal information in collateral texts for indexing video. In *Proceedings of LREC 2002 Workshop on Annotation Standards for Temporal Information in Natural Language*. European Language Resources Association (ELRA).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Shabana, S. (2018). *Validazione della batteria SMAAV 2.0 per lo studio e la diagnosi di deficit semantico-lessicali. Analisi delle competenze semantico-lessicali in compiti di produzione mediante l'uso di verbi generali.* Bachelor's thesis, University of Florence.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. (2019). Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description*, 3(99):36–149.

Tesnière, L. (1959). *Elements de Syntaxe Structurale*. Klincksieck, Paris.

Tomasello, M. (2003). *Constructing a Language*. Harvard University Press.

Torabi, A., Pal, C. J., Larochelle, H., and Courville, A. C. (2015). Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research. *arXiv preprint arXiv:1503.01070*, cs.CV.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.

Tranel, D., Manzel, K., Asp, E., and Kemmerer, D. (2008). Naming dynamic and static actions: Neuropsychological evidence. *Journal of Physiology-Paris*, 102(1-3):80–94.

Turner, J. M. (1998). Some characteristics of audio description and the corresponding moving image. In *Proceedings of the ASIS Annual Meeting*, volume 35, pages 108–17. ERIC.

UNCRPD, United Nations Committee on the Rights of Persons with Disabilities (2011). Implementation of the Convention on the Rights of Persons with Disabilities.

Vendler, Z. (1957). Verbs and times. *The philosophical review*, 66(2):143–160.

Vendler, Z. (1967). *Linguistics in philosophy.* Cornell University Press, Ithaca, N.Y.

Wittgenstein, L. (1958). *Philosophical Investigations.* Basil Blackwell, Oxford, 2nd edition.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*.

Yi, P. (2015). *Verbi di azione in italiano e in cinese mandarino: implementazione e validazione del cinese nell'ontologia interlinguistica dell'azione IMAGACT*. PhD thesis, University of Florence.

Zhou, L., Xu, C., and Corso, J. J. (2018). Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598.