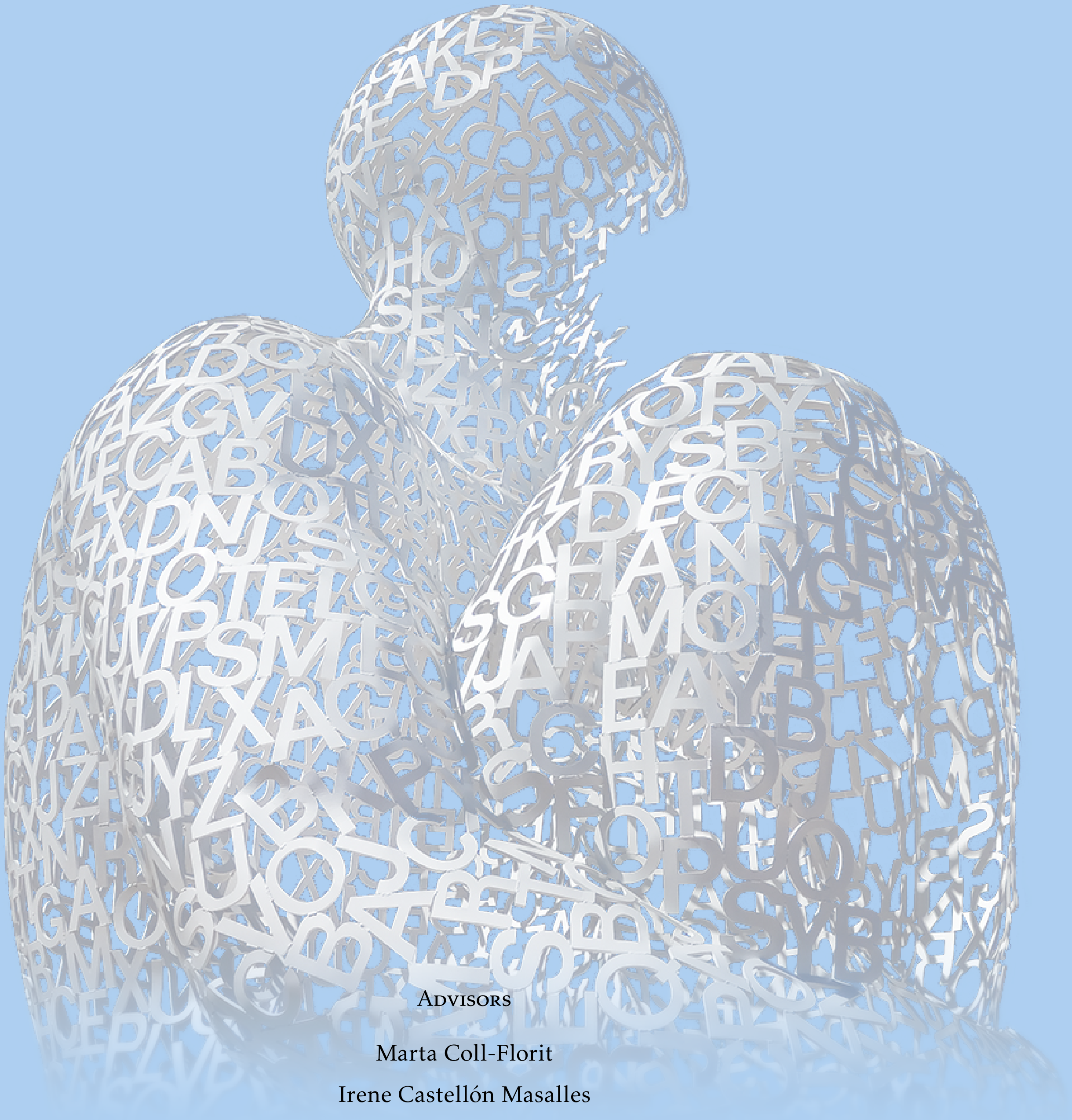
Exploiting verb similarity for event modelling

Doctoral Thesis

Lara Gil Vallejo



Advisors

Marta Coll-Florit

Irene Castellón Masalles

Information and Knowledge Society Doctoral Programme
2019

Universitat
Oberta
de Catalunya

Abstract

The present thesis falls within the scope of Natural Language Processing and aims at exploring the potential of verb similarity, and more specifically of verb classifications, when it comes to capturing and modelling basic information related to events expressed in Spanish.

Verbs are one of the principal means through which events are conveyed. They possess a property that makes them important from the point of view of event conceptualization and expression: they are relational categories, which implies that they occur with entities that are external to them, the event participants, creating structures of relations between them.

This work is concerned with the linguistic materialization of these relations, the predicate-argument structures. These structures are ensembles, composed by a verb and its arguments, that need to be interpreted in order to decode the relevant information of the events expressed in the sentences. This has been frequently summarized as determining "who did what to whom and under what circumstances". Verb arguments are, thus, the carriers of essential information about the participants in the event.

Our research is organised around two studies that examine the ability of verb similarity to model event participant information. We first perform a study of verb similarity with respect to argument structure, looking at its relevant characteristics through the lens of three different perspectives that deal with it (linguistic theory, corpus linguistics and psycholinguistics). Here we examine how each perspective defines verb similarity on the basis of argument structure, paying attention to how much coincidence there is between them and which linguistic features are salient for each perspective. After this analysis is concluded, we find significant correlations between the different perspectives. Besides, we also find that each perspective relies on different linguistic features to structure the similarity relations between verbs: the similarities configured by psycholinguistic data are aligned with the semantic fields of the verbs, the theoretical linguistics approach defines similarities in a way that is congruent with the aspect of those verbs and the corpus linguistic approach is situated in an intermediate position, were both types of information are relevant.

This analysis motivates our choice of features and configurations to be explored in the creation of an automatic classification of verb senses using a clustering algorithm. This automatic classification aims at capturing the argument structure of the verbs and reflecting it in the classes in a way that allows for adequately modelling the participants in the events expressed by those verbs.

This ability to model information related to the participants in the event is evaluated in two steps: we first compare the automatic classification to a gold standard classification of verb senses that is based on semantic roles. Secondly, we test its capacity of generalizing the information gathered in the classes. To do so, we assign verb senses that were not included in the automatic classification to the classes using a similarity metric. Then, we measure the overlap between the semantic roles that are associated to the classes and those that are associated to the new verb senses. These two evaluations confirm the ability of automatic classifications to capture and infer relevant information related to participants in events.

To complete the assessment of the automatic classification taking other facets of argument structure into account, we compare the verb similarities defined by the automatic classes with those obtained when using data coming from the psycholinguistic and the theoretical linguistics approaches, finding that the classification is also able to organize information in a comprehensive way that adapts to different aspects of argument structure.

Finally, we present a study in which we assess the role of the different linguist features used to create the classification in relation to its performance, concluding that both semantic and syntactic features play an important role in order to create a robust and flexible automatic classification from the point of view of argument structure and event information.

Resumen

El presente trabajo se enmarca dentro del ámbito del Procesamiento del Lenguaje Natural. Su objetivo es explorar el potencial de la similitud verbal, y más concretamente de las clasificaciones verbales, a la hora de capturar y modelizar la información básica relacionada con la expresión de eventos en español.

Los verbos son uno de los principales medios para comunicar eventos. Desde el punto de vista de la conceptualización y expresión de los eventos, los verbos poseen una característica esencial: tienen una naturaleza relacional, lo que implica que concurren con entidades externas a ellos, los participantes en los eventos, con los que crean estructuras de relaciones.

Este trabajo se ocupa de la materialización lingüística de estas relaciones, las estructuras argumentales. Estas estructuras son elementos compuestos por un verbo y sus argumentos. Transmiten la información relevante relativa a los eventos expresados en las frases. Decodificar esta información ha sido comúnmente resumido como determinar "quién hizo qué a quién y bajo qué circunstancias". Los argumentos verbales pueden ser vistos, por lo tanto, como los portadores de la información básica sobre los participantes en el evento.

La tesis se articula en torno a dos estudios que examinan la capacidad de la similitud verbal para modelizar la información relativa a los participantes en eventos. En primer lugar elaboramos un análisis de la similitud verbal en relación a la estructura argumental. Para ello tomamos tres perspectivas que tratan este tema (lingüística teórica, lingüística de corpus y psicolingüística) y analizamos cómo cada una de ellas define la similitud entre los verbos. En concreto, nos centramos en la coincidencias y divergencias entre estas tres perspectivas y en las características lingüísticas que resultan importantes para cada perspectiva a la hora de definir el eje similitud-disimilitud. Como resultados del análisis, por un lado encontramos correlaciones estadísticamente significativas entre las tres caracterizaciones de la similitud verbal. Por otro lado vemos que cada una de estas perspectivas privilegia un tipo de información lingüística diferente cuando definen las relaciones de similitud entre los verbos: la similitud definida desde el ámbito de la psicolingüística se alinea con el campo semántico de los verbos, mientras que la similitud definida por la perspectiva de la lingüística teórica viene marcada por la información aspectual. Finalmente, la perspectiva de corpus se sitúa en el medio de ambas, siendo ambos tipos de información relevantes.

Este análisis nos sirve para definir un conjunto de características lingüísticas y configuraciones que se aplican en el segundo estudio. Este estudio consiste en la creación de una clasificación automática de sentidos verbales usando un algoritmo de clustering. El objetivo de esta clasificación es capturar la estructura argumental de los verbos y reflejarla en las clases, de manera tal que permita modelizar los participantes en los eventos expresados por los verbos.

Esta capacidad de modelizar información relacionada con los participantes en eventos se evalúa de dos maneras: primero comparamos la clasificación automática obtenida con una clasificación de referencia que está basada en roles semánticos. El segundo paso consiste en evaluar la capacidad de generalización de la clasificación en base a la información guardada en las clases. Para ello, clasificamos nuevos sentidos verbales (no incluidos en la clasificación) usando una métrica. Después medimos la coincidencia existente entre los roles semánticos asociados a los verbos de la clase y los asociados al nuevo verbo clasificado en ella. Los resultados de estas dos evaluaciones confirman la capacidad de la clasificación automática para capturar y generalizar información relevante de los participantes en los eventos.

Para completar la evaluación de esta clasificación tomando en cuenta otras facetas de la estructura argumental, comparamos las relaciones de similitud verbal definidas por la clasificación verbal y las definidas por la perspectiva psicolingüística y de lingüística teórica. Los hallazgos nos permiten afirmar que la clasificación verbal organiza la información de manera que es capaz de acomodar diferentes aspectos de la estructura argumental.

Finalmente, presentamos un estudio en el que analizamos el rol de las diferentes características lingüísticas usadas para realizar la clasificación con respecto a los resultados obtenidos por la misma, concluyendo que tanto las características semánticas como las sintácticas juegan un rol importante a la hora de crear una clasificación robusta y flexible desde el punto de vista de la estructura argumental y de la información asociada a eventos.

Resum

El present treball s'emmarca dins l'àmbit del Processament del Llenguatge Natural. El seu objectiu és explorar el potencial de la similitud verbal, i més concretament de les classificacions verbals, a l'hora de capturar i modelitzar la informació bàsica relacionada amb l'expressió d'esdeveniments en espanyol.

Els verbs són un dels principals mitjans per comunicar esdeveniments. Des del punt de vista de la conceptualització i expressió dels esdeveniments, els verbs posseeixen una característica essencial: tenen una naturalesa relacional, fet que implica que concorren amb entitats externes a ells, els participants en els esdeveniments, amb els quals creen estructures de relacions.

Aquest treball s'ocupa de la materialització lingüística d'aquestes relacions, les estructures argumentals. Aquestes estructures són elements compostos per un verb i els seus arguments, els quals transmeten la informació rellevant relativa als esdeveniments expressats en les oracions. Descodificar aquesta informació ha estat habitualment resumit com a determinar "qui va fer què a qui i sota quines circumstàncies". Els arguments verbals poden ser vistos, per tant, com els portadors de la informació bàsica sobre els participants en l'esdeveniment.

La tesi s'articula al voltant de dos estudis que examinen la capacitat que té la similitud verbal de modelitzar la informació relativa als participants en esdeveniments. En primer lloc elaborem una anàlisi de la similitud verbal en relació a l'estructura argumental. Amb aquesta finalitat, partim de tres perspectives que tracten aquest tema (lingüística teòrica, lingüística de corpus i psicolingüística) i analitzem com cadascuna d'elles defineix la similitud entre els verbs. En concret, ens centrem en les coincidències i divergències entre aquestes tres perspectives i en les característiques lingüístiques que resulten importants per a cada perspectiva a l'hora de definir l'eix similitud-dissimilitud. Com a resultats de l'anàlisi, d'una banda trobem correlacions estadísticament significatives entre les tres caracteritzacions de la similitud verbal. D'altra banda, veiem que cadascuna d'aquestes perspectives privilegia un tipus d'informació lingüística diferent a'hora de definir les relacions de similitud entre els verbs: la similitud definida des de l'àmbit de la psicolingüística s'alinea amb el camp semàntic dels verbs, mentre que la similitud definida per la perspectiva de la lingüística teòrica ve marcada per la informació aspectual. Finalment, la perspectiva de corpus se situa en una posició intermédia entre totes dues, sent ambdós tipus d'informació rellevants.

Aquesta anàlisi ens serveix per definir un conjunt de característiques lingüístiques i configuracions que s'apliquen en el segon estudi. Aquest estudi consisteix en la creació d'una classificació automàtica de sentits verbals usant un algoritme de *clustering*. L'objectiu d'aquesta classificació és capturar l'estructura argumental dels verbs i reflectir-la en les classes, de tal manera que permeti modelitzar els participants en els esdeveniments expressats pels verbs.

Aquesta capacitat de modelitzar informació relacionada amb els participants en esdeveniments s'avalua de dues maneres: primer comparem la classificació automàtica obtinguda amb una classificació de referència que està basada en rols semàntics. El segon pas consisteix a avaluar la capacitat de generalització de la classificació en base a la informació guardada a les classes. Per a això, classifiquem nous sentits verbals (no inclosos en la classificació) usant una mètrica. Després mesurem la coincidència existent entre els rols semàntics associats als verbs de la classe i els associats al nou verb classificat en aquesta classe. Els resultats d'aquestes dues avaluacions confirmen la capacitat de la classificació automàtica per capturar i generalitzar informació rellevant dels participants en els esdeveniments.

Per completar l'avaluació d'aquesta classificació tenint en compte altres facetes de l'estructura argumental, comparem les relacions de similitud verbal definides per la classificació verbal i les definides per la perspectiva psicolingüística i de lingüística teòrica. Les troballes ens permeten afirmar que la classificació verbal organitza la informació de manera que és capaç d'acomodar diferents aspectes de l'estructura argumental.

Finalment, presentem un estudi en el que analitzem el paper de les diferents característiques lingüístiques usades per realitzar la classificació en relació amb els resultats obtinguts per aquesta classificació, concloent que tant les característiques semàntiques com les sintàctiques juguen un paper important a l'hora de crear una classificació robusta i flexible des del punt de vista de l'estructura argumental i de la informació associada a esdeveniments.

Contents

List of Figures

List of Tables

I

Background

CHAPTER 1

Introduction

Todo está en la palabra... Una idea entera se cambia porque una palabra se trasladó de sitio, o porque otra se sentó como una reinita adentro de una frase que no la esperaba y que le obedeció. Tienen sombra, transparencia, peso, plumas, pelos, tienen de todo lo que se les fue agregando de tanto rodar por el río, de tanto transmigrar de patria, de tanto ser raíces...

Pablo Neruda, Confieso que he vivido

This dissertation revolves around what makes two or more verbs similar and how this information can be used in order to model the events expressed by verbs in a way that adequately captures and generalizes relevant information about event participants.

Exploring this subject implies delving deeper into two basic issues: what it means for two verbs to be similar and what is the contribution of the verb to the interpretation of the event.

In relation to the first issue, we can advance here that the notion of similarity is a slippery territory. This statement might result counterintuitive since it is a concept that we use continuously in our daily lives without any trouble. The problem, however, arises when we try to turn it into a scientifically useful device. For this end, many questions not considered previously need to be clarified: in which respects two elements are similar? how to quantify this similarity? can we look at similarity also from a qualitative point of view? We will see throughout this work that these questions are also central to the study of the similarity between verbs.

Similarity is a relevant device in Natural Language Processing systems (NLP), although many times its role is rather secondary (for example, when used in one of the several steps that might exist in the pipeline of an NLP system). In this work similarity is put in a central position: it is used to learn relations between verbs, to analyze these relations from a qualitative and quantitative point of view and to make generalizations about them. These different goals will be examined in the following chapters and here we will just emphasize that, to reach them, it is necessary to pay special attention to the grounds on which similarity is defined, and their impact in the final outcome.

Particularly, in this work we will study how similarity comes into play in the complex scenario drawn by verbs, which, as linguistic units, have a relational nature. This characteristic implies that they participate in a system together with elements that are external to them, the arguments. Such system can be captured and represented in many ways. Therefore, what follows is an attempt to approach this complexity using similarity as a tool that both enables a comprehensive analysis and, at the same time, sheds light on its specific features.

In relation to the contribution of the verbs to the interpretation of the event, we can start by quoting the classic definition of event extraction (EE) by Grishman (2003), which states that EE consists in two basic tasks: "identifying instances of events of a particular type, and the arguments of each event". This second task, identifying the arguments or event participants is directly aligned with our research. Identifying event participants, usually paraphrased as finding out *who did what to whom, and when, where and how*, implies recognizing key entities and their relationships. As we mentioned earlier, verbs are relational in nature and this has certain implications. Before examining them in more detail, we will define what it is to be *relational*. Following Gentner and Kurtz (2005) and Arthur B. Markman et al. (2013) we can define relational categories as those whose properties are related to elements that are extrinsic to them. Therefore, these categories describe relationships between entities and usually take external arguments (this is the case of verbs or deverbal nouns such as *destruction*) or have meanings centered on relations with other concepts (such as *neighbour*, which involves an entity that is close, or *father*, which involves an entity that is a descendant). This situation contrasts with non-relational categories, which are characterized by having intrinsic stable properties. An example of these kind of categories is *bird*, which can be defined by a set of features such as "have feathers", "have beak" and "have wings" among others. Both types of categories, relational and non-relational, may exhibit some extrinsic and intrinsic properties at the same time, but generally one aspect prevails over the other.

Thus, verbs, as relational categories, are crucial in order to detect the relationships that can be established between the participants in the event and the role that each participant has. This can be seen in examples 1 and 2, where two sentences that only differ in the verb are contrasted:

(1) This man robbed several jewels from my neighbour.

Type of event: rob/steal

Roles of participants
- This man: perpetrator
- Jewels: goods
- Neighbour: victim

Relations between participants:
- *perpetrator* possesses *goods*
- *neighbour* does not possess *goods*
- *goods* are moved from neighbour to *perpetrator*

(2) This man showed several jewels from my neighbour.

Type of event: show/exhibit

Roles of participants
- This man: agent
- Jewels: phenomenon
- Neighbour: source

Relations between participants:
- *agent* shows *goods*
- *goods* are possessed by *neighbour*
- *agent* does not possess *goods*

Therefore the verb is associated to a structure of relations and this enables the basic interpretation of the event expressed in a sentence. This structure of relations is usually known as argument structure in Linguistics or predicate-argument structure in NLP. Arguments may receive a specific label, the *semantic role*, that clarifies the nature of the function it has within the event. The set of elements that participate in the structure of relations, the arguments, also make an important contribution in the interpretation of the event and particularly in its disambiguation.

The idea that argument structure is directly linked with the linguistic codification of the events expressed in the sentence was translated into a practical approach to Information Extraction by Surdeanu et al. (2003), who relied on the good correspondence between verb arguments and event participants to propose a system that was domain-independent, achieving results similar to those obtained by finite state automata (FSAs), a less portable, state-of-the-art approach at that moment.

From the linguistic point of view, the nature of argument structure lies in the interface between syntax and semantics, what has made it a hot spot for syntactic and lexical theories of all flavours. Additionally, since argument structure deals with the overt realization of core elements in event conceptualization, it has also attracted attention from psycholinguistic theories of language acquisition and semantic memory. These different perspectives usually put the focus on particular aspects, although this does not necessarily imply that they are fully incompatible.

Thus, although completely extracting all the information relevant for an event may require going beyond arguments, information about verbs and their arguments is fundamental to process events. This information is complex in nature, it involves relations between several elements built on different linguistic cues. In order to systematize this information in a way that is readily accessible and useful for event modelling and other NLP tasks, lexical classifications are particularly handy. Indeed, classifications offer a natural way to organize and catalogue knowledge. They are able to systematize relationships between the classified elements and they also make explicit non overt relationships. Although handcrafted lexical classifications such as VerbNet (Schuler, 2005) have been successfully used for a number of NLP tasks, they require lots of time and effort in their development. On the other hand, automatically created classifications using corpus data and clustering algorithms are still useful to model and organize lexical information while being less costly in terms of human effort and time.

1.1 RESEARCH MOTIVATION

There is a variety of views on predicate-argument structure that come from different language related fields. Although, at first sight they may seem hard to couple, we think that looking for their connections in the definition of verb similarity and comparing them will enable us to not only gain knowledge of the linguistic properties that are relevant for each perspective, but also to study the possible overlaps between the different perspectives in terms of these properties. Moreover, a multi-perspective study can shed light on the general issue of the configuration of verb similarity.

Besides the particularities of the different approaches to predicate-argument structures when defining verb similarity, it is considered that automatic classifications created using clustering algorithms have a lot of potential for systematizing verb related information. However, this potential has remained largely unexplored for actual applications. A number of verb classifications have been built with the goal of creating semantic classifications based on verb behaviour (more or less enriched subcategorization frames) mainly for English. However many of them fall short when it comes to being applied to other tasks, focusing instead on building an automatic classification that resembles a manually developed classification. Besides, their ability to capture and organize information related to argument structure has been barely scratched. Moreover, classifications are not only a way of organizing knowledge, but they can also be used to infer knowledge. Assigning a new member to an existing class implies also allocating information from the class to this member. In particular, a verb classification automatically created may be useful in order to generalize information associated to verbs using limited resources.

1.2 OBJECTIVES

The main objective of this thesis is to model the events expressed in the sentences in Spanish, specifically the information related to the participants, using the similarity between the verb senses that convey the events as the basic mechanism to create an adequate and generalizable model. In order to do this, we define two basic steps:

The first step is to provide an analysis of verb similarity with respect to argument structure from a multi-perspective approach. In particular, the perspectives that we take into account are the following:

- Linguistic theory: analysis of the constructions in which verbs participate.
- Psycholinguistics: analysis of the words that are typically associated with verbs.
- Corpus linguistics: analysis of the semantic roles that co-occur with verbs in corpora

The relations between these three approaches are considered from a qualitative and quantitative perspective.

The second step is to capture verb similarity from the point of view of argument structure. In order to do this, we create an automatic classification of verb senses based on information related to their argument structure. We evaluate its coherence by comparing it with a gold standard and with the information obtained in the analysis step. Besides, we evaluate its generalization power by looking at its usefulness in identifying participants in events represented by verbs which are outside of the classification.

1.3 HYPOTHESIS

Predicate-argument structures are important in order to determine types of events and to assign roles to the participants in those events. However, they are considered to have a rather complex nature that lives at the interface between syntax and semantics and thus, the interplay between these two dimensions should be taken into account when modelling predicate-argument structures using similarity. Therefore, the main question that we need to address is the following:

1. Can we model effectively the event information present in predicate argument structures using similarity-based techniques?

Our hypothesis is that **predicate-argument structures have regularities that can be captured and generalized with similarity-based techniques**. We can materialize this general hypothesis into two more specific sub-hypotheses:

Firstly, if we focus on the nature of predicate-argument structure and its principal features, we will notice that, as we mentioned, there are many different perspectives that come from different areas of research and put the focus on diverse aspects of these structures. Therefore, the following question needs to be explored first:

1.1. Do different perspectives of predicate-argument structure contribute useful and consistent information when they are applied to the analysis of verb similarity?

Our hypothesis is that **different perspectives of predicate-argument structure exhibit basic consistent behaviour in defining verb similarity**. Besides, we expect that the comparison between perspectives will help discover useful information in order to capture predicate-argument structures using similarity-based techniques.

Secondly, with respect to the ability of similarity-based techniques to generalize predicate-argument structures, it is usually argued that approaches such as automatic verb classifications, which are typically created clustering algorithms and similarity (or divergence) metrics, contain generalizable information. Unfortunately, this has only been tested for manually-developed classifications. This issue is particularly relevant because verb classifications are inherently limited, meaning that they contain a restricted number of verbs and verb-related information and they do not provide information for verbs that are outside the classification. Therefore we need to explore the actual extensibility of these classifications formulating the following question:

1.2. Are automatically built verb classifications useful in order to represent and generalize predicate-argument structures?

Our hypothesis is that **a classification that is automatically created from corpus data can reveal patterns of homogeneous linguistic behaviour which is representative enough to represent and generalize verb behaviour with respect to argument structure**, assigning correct predicate-argument structures to verbs outside the classification.

These hypotheses set up the framework scenario in which this research is carried out. Next, we describe the structure of the thesis in which we explore these research questions stated above.

1.4 THESIS STRUCTURE

This thesis is organized in six chapters. Besides the current chapter, which presents the introduction, the motivation, the objectives and the hypotheses, in chapter 2 we first provide an overview of linguistic approaches to the formalization of event information. Then we put the focus on the notion of similarity and how it has been applied to study and model verbs. We finish the chapter with a review of the basic issues that have arisen in automatic verb classifications. In chapter 3 we detail the methodology that we follow to analyze and research the application of verb similarity to the modelling of event information and we list and describe the resources used to do so. Following this, in chapter 4 we detail the analysis of verb similarity using the different perspectives already mentioned and we present qualitative and quantitative results. In chapter 5 we describe the specific steps followed in order to create several automatic verb classifications. These classifications are evaluated using two main methods and they are also compared to the results of the analysis generated in chapter 4. Finally, in chapter 6 we present the conclusions of the thesis, detailing its contributions and the work that can be developed in the future.

CHAPTER 2

Previous and related work

Verbs play a key role in the expression of events. Besides the contribution of their own meaning, in the previous section we saw that their relational nature implies that they are connected to other elements, which are represented in the argument structure. Since argument structure is basic for event conceptualization and expression, verb similarity becomes an interesting tool in order to analyze, capture and generalize the relevant pieces of information related to events. In this section we first explore the relationship between event conceptualization and argument structure and how this last component has been formalized. Subsequently we introduce the definition of similarity and some caveats about its usage that have been put forward in the literature. We then delve into the issues that have arisen when similarity has been applied to study and model information related to verbs. Finally, literature about automatic verb classifications is critically examined.

2.1 Linking linguistic and conceptual representations of events

Events may be considered at different levels: the actual happenings in the physical world, their conceptualization, which consists on the mental representation of these happenings, and their linguistic expression. The linking between the second and third levels, our mental representation of events and their linguistic expression, has been a recurring endeavour in many of the research fields that deal with language, ranging from theoretical linguistics to cognitive science.

From the point of view of the linguistic approaches, a main interest has been to find properties of our mental representations of events that were relevant for their linguistic expression, including the linguistic realization of the *event arguments*, as Grishman (2003) puts it.

In this line, Levin and Hovav (2005) cite several properties of events that different approaches have considered crucial in order to deal with event conceptualization and expression: spatial motion and location (Jackendoff, 1992), the internal temporal properties of events or *aktionsart* (Vendler, 1957) and causation (Croft, 1998). Linguistic theories differ as to which of these properties is relevant to determine the linguistic realization of the events, being the last two approaches the ones that entail a major contribution towards the study of argument realization, while the first one more involved in issues such as polysemy, according to Levin and Hovav (2005).

Thus, the structure of the event, and more specifically, the expression of the arguments or *argument structure* occupies a central position in the understanding of events. The notion of argument structure refers to how many and what type of arguments are related to a predicate as it can be seen in example 3:

(3) Mary$_{\text{[SUBJECT- EXPERIENCER]}}$ likes$_{\text{[PREDICATE]}}$ apples$_{\text{[OBJECT - THEME]}}$

There are diverse (and sometimes opposing) views of the nature of argument structure. To comment briefly on two influential examples, we can mention Levin (1993) and Goldberg (1995) approaches. Levin's work is based on the hypothesis that "the behaviour of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extend determined by its meaning". Therefore, in this view, argument structure is motivated by the lexical meaning of the verb. A quite different view is that of Goldberg. This author considers that argument structure is a type of construction. Constructions, in this framework, are associated with a specific meaning. Therefore, argument structure, as every construction, has its own independent meaning. Whether the relation between the argument structure and the predicate is driven by the lexical meaning of verbs or by the interplay of lexical meaning and constructional meaning, it is assumed that such a relation exists: argument structure and verbs do not combine at random, but in a specific way.

Research in the field of psycholinguistics has also been concerned with event conceptualization and expression. Particularly relevant for us are a number of studies that gather evidence from several experiments that link event participants and their linguistic expression: McRae, Hare, et al. (2005) found that nouns that refer to entities that represent typical argument roles (agents, patients, instruments and locations) prime verbs that describe events in which they participate. According to this work "event memory is organized so that nouns denoting entities and objects activate the classes of events in which they typically play a role". Ferretti, McRae, et al. (2001) discovered that the converse effect also holds: verbs primed nouns that refer to typical agent, patients and instruments, but not locations. Similar effects are also found when nouns are the categories that refer to events instead of verbs, such as *sale* or *breakfast* (Hare, Elman, et al., 2009). This information was summarized by McRae and Matsuki (2009) in figure 2.1.

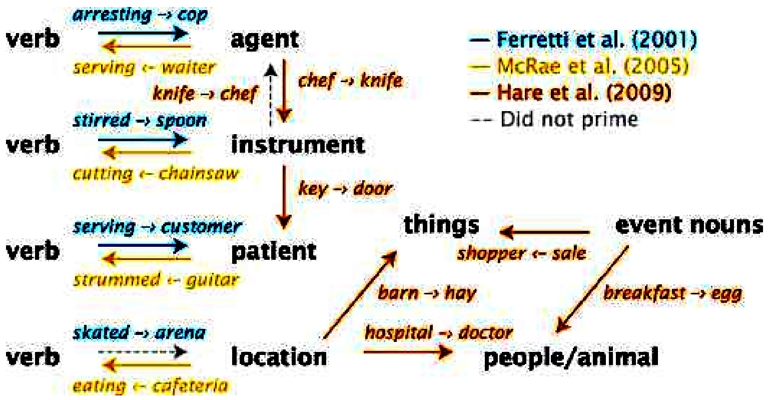


Figure 2.1: Priming effects between event participants and verbs

Also, other complex priming effects can be found in the literature: typical agents also prime patients (Kamide et al., 2003) and combinations of agents and verbs prime patients (Bicknell et al., 2010). Besides, the aspect of the verb also influences the configuration of the event and the priming effects on the specific properties of objects (Altmann et al., 2007; Ferretti, Kutas, et al., 2007).

In addition to priming experiments, Sanz-Torrent et al. (2017) analyze the effects of argument structure on verb recognition using data from eye tracking experiments. Their findings reveal that the number of arguments that combine with a verb augments the processing time needed in order to identify the event referred. One of the possible causes of this effect is the increased representational complexity that a larger number of arguments entails.

Besides the connection between the mental representations of the events and the argument structure present in their expressions, there are authors that explore the notion of *event knowledge* and its importance for sentence processing. McRae and Matsuki (2009) and McRae, Hare, et al. (2005) claim that people possess a generalized knowledge of events that is accumulated through experiences such as acting in those events or observing them. Collectively, these experiences conform *prototypical events* and this knowledge, combined with linguistic cues, plays an important role in online sentence processing. Taken together, these evidences put the focus on the strong connection between argument structure and psycholinguistic phenomena such as event representations and sentence processing.

2.2 CAPTURING ARGUMENT STRUCTURE

In order to capture and model information related to predicate-argument structures, researchers have resorted to annotating corpora with semantic roles. Semantic roles were first introduced in modern linguistics by the works of Gruber (1965), Fillmore (1967) and Jackendoff (1972), who proposed them as mediators between syntax and semantics in the expression of the arguments. This turned out to be a very influential idea, and semantic roles were adopted in generative grammar to interpret the relation that holds between each argument and the predicate, and to account for the similarities between arguments with different syntactic realizations. However, the use of semantic roles is not limited to formal accounts: they have been widely used for semantic annotation in corpora and for NLP tasks such as question answering (Shen et al., 2007, information extraction (Christensen et al., 2010) or machine translation (Liu et al., 2010).

Semantic roles have proven to be a useful level of representation when characterizing languages from a syntactic-semantic perspective. However, a crucial matter for using these categories for corpus annotation is the disagreement on the inventory of roles that should be used. There is a variety of roleset proposals that differ in the number of roles and their degree of detail when identifying participants in the event (e.g. *agent* vs. *writer*, *cook* or *painter*). The pioneering work carried out by Fillmore (1967) introduced a set of semantic roles based on interpretative categories: *agent*, *patient*, *theme* and *beneficiary* among others, which has been the approach of lexical resources such as VerbNet (Schuler, 2005). In the same spirit, but aiming for more neutrality, we find the approach of PropBank (Palmer et al., 2005), which identifies arguments using numbered labels (ARG0, ARG1, etc.). This represents a set of ranked roles that takes into account the order of the arguments and remains constant for a specific verb, but changes across different verbs. Therefore, argument labels, such as A0, do not always refer to the same semantic role across all the verbs. A more fine-grained account of semantic roles can be found in Frame Semantics theory and FrameNet (Fillmore et al., 2003), in which the semantic roles are determined by the specific event: roles such as *cook*, *food* or *recipient* are used to address the participants in a cooking event (frame, in their terminology). A different approach is the one found in D. Dowty (1991), which regards semantic roles not as clear-cut categories, but rather as sets of properties or entailments that are associated to two proto-roles, the Proto-Agent and the Proto-Patient. Finally, we find the framework proposed by LIRICS (Bonial et al., 2011), whose aim is to develop a standard roleset suitable for the different needs of various natural language processing tasks. They propose a hierarchical organization of roles based on different levels of granularity in which coarse-grained roles, such as *actor*, subsume related fine-grained roles such as *cause* and *agent*.

A final note on the relationship between events and argument structure should be added. The relevant information in events may be defined in a different way in NLP and in linguistic theories since there might be elements in an event that are not part of the argument structure associated to the verb. In example 4 *yesterday* does not belong to the argument structure of *give*, although it is relevant from the point of view of the information that we might like to extract from the event. However, although argument structure and event structure are not identical, they show a high degree of correspondence.

(4) Joan gave me the book yesterday.

Besides, argument structure is key to interpret the basic information of the event. Participants in events can be expressed in many different ways, and semantic roles, as labels of the argument structure, serve to interpret their contribution in the event in a congruent way across different syntactic expressions. This good correspondence has motivated NLP tasks such as Semantic role labeling, in which the goal is to identify and label the arguments of the sentence. In our approach we also rely on this good correspondence to approach event modelling.

2.3 SIMILARITY: USES AND LIMITATIONS

The notion of similarity refers to a specific relationship that is established between two objects on the basis of their common properties, features, behaviour or any other element that is used to characterize these objects. This notion is not unproblematic (Goodman, 1972), as it has been considered to be too broad or vague to be useful. Additionally it requires that the grounds on which two objects are considered similar be determined beforehand. Nevertheless, it is ubiquitous in many areas of knowledge and it is a powerful and useful concept that enables the discovery of correspondence patterns.

In **cognition studies**, similarity is a crucial notion that underpins a number of cognitive models. It has been often been given functional roles in processes such as categorization (Goldstone, 1994; Hampton, 1998), conceptualization (Lakoff, 1987), memory retrieval (Raaijmakers et al., 1981) and problem solving (Novick, 1988).

There are several approximations to the definition of similarity within cognition studies. Geometric models of similarity, one of the best-known among these approaches, characterize mental representations as points in a mental space configured by certain psychological dimensions. Therefore, the similarity between two representations can be calculated on the basis of their distance in this space: concepts that are close are more similar than those which are apart (Shepard, 1962a; Shepard, 1962b). Feature-set approaches (Tversky, 1977) represent concepts as feature ensembles. According to them, similarity between two concepts is measured by virtue of their common and distinct features. Structure-mapping models, also known as alignment-based models (Gentner and Arthur B Markman, 1997), are likewise based on the existence of features associated with concepts, but the basis on which objects are compared is the structure created by the relations between the features. Differently, transformational approaches (Garner, 1974;

Hahn et al., 2003) do not propose to study similarity using mediators such as points in space or feature sets. Instead they assume that any mental representation can be turn into another following a series of transformational operations. Based on this view, the number and type of transformations needed is what defines the similarity between two mental representations. Overall, although every approach has its own theoretical and operational definition of similarity, there is a common underlying conception of similarity as a linking process carried out on mental objects and based on their coincident (and divergent) aspects.

These different accounts of similarity in cognition have had a theoretical and practical impact on psycholinguistic models of the **mental lexicon** (cf. M. N. Jones et al. (2015) Ch. 11 for an overview). Furthermore, there is a body of research in psycholinguistics that highlights the evidence in favour of the role of similarity in constructing the mental lexicon, not only from priming experiments related with word recognition at several levels such as phonetic (Vitevitch et al., 1999; Luce et al., 1990; Vitevitch et al., 2016), semantic (Neely, 1991) and syntactic (Savage et al., 2003) but also from studies that deal with word formation (see Bybee (2010)).

If we move the focus to **Natural Language Processing (NLP)**, we find that the notion of similarity is crucial for the development of linguistic models and applications. Indeed, measuring the similarity between different linguistic units is important for many different tasks (evaluation metrics in machine translation, word sense disambiguation, textual entailment analysis, paraphrase detection, among others). Nevertheless, the formalization and measurement of similarity is affected by the broadness of this notion: there are many relationships between linguistic units that we can take into account in order to define similarity. These relationships can take place at different levels (Fellbaum, 2015), such as the lexical level (synonymy), the semantic-conceptual level (hyponymy, meronymy) and the syntagmatic level (selectional preferences).

Consequently, we also find a variety of strategies for measuring similarity between linguistic units in the literature, in most cases based on the semantic relation that holds between them. Among those that have to do with words and word senses, as in our case, we find models that represent linguistic units as nodes in a taxonomy connected by a variety of relations. Their similarity is measured on the basis of the path that links them (Resnik, 1995; Yang et al., 2006). Other accounts, such as distributional models, represent words as points in a space configured by their context. The fundamental idea in this approach is that words that share many contexts are similar (Schutze, 1992; Landauer et al., 1997; Mikolov et al., 2013). Besides, similarity has been calculated on the basis of the overlap between word definitions or glosses in dictionaries (Kozima et al., 1993; Banerjee et al., 2003; Patwardhan et al., 2003) and also taking into account the amount of overlap between feature sets associated with the words (Lin, 1998).

Finally, it should be mentioned that there are several collections of human ratings of word similarity. They consist of pairs of words that are rated by native speakers on a scale according to the perceived similarity between them. They are generally used to evaluate computational models of similarity (Faruqui et al., 2014). However, since most of these collections are limited to pairs of nouns, with the exceptions of Hill et al. (2015), Gerz et al. (2016), Yang et al. (2006) and S. Baker et al. (2014), the coverage of verbs (and other categories) is reduced if we compare it to the coverage of nouns. For Spanish there is a general scarcity of these resources: there are only ratings for nouns, either collected from native speakers (Moldovan et al., 2015), or translated from English (Finkelstein et al., 2001; Camacho-Collados et al., 2015). However, as noted by Hill et al. (2015), there is a major issue with human ratings of similarity because many times they are used to evaluate semantic similarity when, generally, human annotators do not rate just semantic similarity but also domain associations. That is why in many human ratings resources *cup* and *coffee* might be rated as being more similar than *train* and *car*.

2.4 VERBS AND THEIR SIMILARITY

With respect to the comparison between different perspectives of verb similarity, it is necessary to mention the work by Resnik and Diab (2000), which compares three models of verb similarity that differ in the type of lexical representation: semantic networks (Wordnet 1.5, Miller et al. (1990) and Fellbaum (1998)), distributional syntactic co-ocurrences obtained from corpus, and lexical conceptual structures (LCS), a type of enriched event decomposition representation (Dorr, 1993; Jackendoff, 1983) which constitutes a handcrafted resource.

To perform this comparison, they create a set of 48 verb pairs in which the members of each pair have the same subcategorization frame, aspectual class and are combined with the semantic role *theme*. For each approach, similarity scores between the members of each pair are obtained and then the correlation of these scores with human ratings of similarity is calculated using Pearson's ρ. Human ratings are collected independently for two conditions. In the first one, participants are asked to rate the perceived similarity between two verb lemmas presented in isolation. In the other condition, participants rate the perceived similarity between two verbs presented in the context of a sentence. The presence of context tends to yield lower similarity scores and the authors argue that this is due to more flexible interpretations of verb meaning in the absence of further data. The semantic network model of similarity is the one that correlates more with human ratings, particularly when the ratings are given for verbs in the context of a sentence (0.78). The model of similarity that correlates less with human ratings is LCS (0.313), maybe affected by the lack of differences of semantic structure in the verbs in the set. The distributional method, which they argue that it may be hindered by data sparsity, yields a correlation of 0.45 for the context condition. Finally, they combine these different models using a linear regression and they obtain a correlation coefficient of 0.87.

However, as we mentioned in the introduction, it is complicated to know what these figures mean, or to extract a conclusion about the qualities of the different resources since we cannot identify the specific basis on which humans rate two words as similar or dissimilar. Besides, coming up with an idea of what the combined model represents for verb similarity (from the human ratings point of view) results even more problematic.

2.5 VERB CLASSIFICATIONS

Classifications are a natural way to organize information. They are especially convenient when the number of elements to be classified is large and the aim is to capture relations in a comprehensive way. An additional advantage is that, once the classification is done, it can reveal relations between elements that were not known previously or that were not easily predictable. Thus, lexical classes are appealing from both the point of view of Linguistics and NLP. Related with what we just mentioned, Schulte Im Walde (2006) emphasizes their usefulness because they capture linguistic generalizations over the items being classified, and thus they are able to provide necessary information when abstracting away from the basic level is required. In this line, Korhonen (2009) argues that predictive power of classes can compensate for lack of data and that classifications can be seen as a tool for generalization, abstraction and prediction.

Handcrafted verb classifications have been successfully used for a wide range of tasks such as word sense disambiguation (Dorr and D. Jones, 1996; Kohomban et al., 2005; Brown et al., 2014), semantic role labelling (Swier et al., 2005; Pradet et al., 2013), document classification (Klavans et al., 1998), parsing (Carroll et al., 2004), query generalisation for information access (Navigli et al., 2003), question answering (Burke et al., 1997; Shen et al., 2007), machine translation (Dorr, 1997; Prescher et al., 2000; Koehn et al., 2007) and psycholinguistic modelling (Padó et al., 2006) among others (cf. Čulo et al., 2008 for a more exhaustive listing). Given their wide range of applications, the creation of automatic classifications of verbs has been explored to overcome the cost of manually developing one. However, the applications of automatically created verb classifications have been far less numerous, as we will see in the next section.

2.5.1 *Automatic verb classification*

We now turn to look at work done in automatic verb classification. The story of large scale automatic verb classifications starts in fact with a large scale manual classification, which is the work by Levin (1993). It represents one of the first and more comprehensive efforts towards creating a verb classification that had into account both semantic and syntactic information. This classification is based on the relation that can be established between verb meaning and verb behaviour (which here stands for the syntactic expression of verb arguments). As we already mentioned, the lexical classification of Levin is based on the hypothesis that the meaning of the verbs determines the expression and interpretation of its arguments. Here, the verb behaviour properties, controlled by verb semantics, are the diathesis alternations. These alternations constitute the different syntactic expressions of the arguments of a predicate. They may convey semantic or pragmatic contrast when comparing the two structures that form the diathesis (Lenci, 2009), but is generally acknowledged that they do not purport significant changes in the overall meaning. This hypothesis, namely, that it is possible to create semantically coherent classes of verbs by taking into account their *semantically relevant* syntactic properties, provided the criteria for the Levin's classification of around 3000 English verbs and also for many automatic verb classifications carried out afterwards.

As an additional note, it can be mentioned that Levin's classification also paved the way for several lexical resources such as VerbNet, which is a computational verb lexicon that grew out of Levin's work, modifying the hierarchy and expanding the information associated to the classes. As of today, VerbNet 3.2 has 6340 verbs, 273 total main classes and 214 total subclasses and it is linked with other resources in the Unified Verb Index: WordNet, PropBank and FrameNet (C. F. Baker et al., 1998) and also OntoNotes (Weischedel et al., 2011).

Automatic verb classifications share with Levin's classification the objective of achieving a semantic classification of verbs by characterising verb behavior, although different classification proposals have divergent views on which properties are semantically relevant. As we will see in the remainder of this section, one of the challenges in automatic verb classification is formalizing diathesis alternations. Therefore, most of these automatic classifications resorted to the use of more or less enriched subcategorization frames to capture verb behaviour. Subcategorization frames are regular patterns that specify the subcategorization behaviour of the verb. They commonly use the syntactic categories of the arguments associated with the verb (Noun Phrase, Prepositional Phrase, etc.).

A quite standard example of automatic verb classifications is the one developed by Schulte Im Walde and Brew (2002), who created a classification of 884 German verbs using the K-means algorithm to cluster them in classes. The features used to characterize verb behaviour are morphosyntactic subcategorization frames that specify the case of the arguments (nominative, accusative, dative) and, in the case of arguments that are subordinate clauses, the type of clause (e.g. indirect wh-questions or copula constructions). Additionally, this work experimented with enriched subcategorization frames by adding information related to prepositional preferences and selectional preferences, using the GermaNet (Hamp et al., 1997) top level category of the lexical head of the phrases to represent these preferences. Thus, each verb is described by a probability distribution over the enriched subcategorization frames that are obtained using this information. The evaluation of this automatic verb classification is carried out, as it is usually the case, by comparing the resulting classes with the ones in a manually constructed verb classification (gold standard). Other examples of works that follow this methodology are Schulte Im Walde (2006), Stevenson et al. (2003), Kingsbury et al. (2003) and Peterson (2019).

An automatic classification that differs from this common approach is the one carried out by Merlo and Stevenson (2001), who focus on the importance of argument structure for verb similarity. They classified 59 verbs into three classes on the basis of their argument structure (unergative, unaccusative and object-drop, which are classes that differ in the mapping of semantic roles to the syntactic category of the arguments). Their main hypothesis is that "verb class distinctions based in argument structure are reflected in statistics over corpora". Therefore, the membership of a verb to a class was determined using the statistic distribution of a set features (transitivity, causativity, animacy, verb voice, and use of past participle or simple past) that go beyond traditional frame information and are approximated using essentially PoS and lexical information extracted from the corpus. This work also differs from the previous ones in that they use a supervised algorithm (a decision tree algorithm) instead of clustering, which is unsupervised. They reach almost a 70 % of accuracy in the classification using all the features, although they determine that using information about the verb voice does not make any difference in the accuracy score.

2.5.2 *Features for automatic verb classification*

Throughout all the research carried out in this area, there is a constant concern about what are the linguistic features that better lead to a syntactic-semantic classification. Usually this concern is paired with a focus on the economy in the development of the feature set: classifications focus on being as comprehensive and adaptable as possible using as little human effort as it is viable.

A step towards this second goal is the work by Joanis et al. (2008), who aim at finding an inexpensive (in the sense of using only resources such as PoS tagging and chunking) and general set of features that allows for precise class distinctions across several criteria. To do so, they conceive a feature set inspired by Levin's work in alternations which includes syntactic information of the arguments, information about the overlaps between the nouns that fill the syntactic heads, animacy and verb features such as tense, voice and aspect. They evaluate this feature space on several tasks that consist in classifying verbs in subsets of Levin's classes that exhibit differences of semantic, syntactic and argumental nature. By using a supervised algorithm (Support Vector Machines) they achieve an average of 0.58 % of accuracy for the task of classifying 496 verbs in 14 of Levin 's classes. They also find out that most of this accuracy (0.57 %) can be accounted for using only frequencies of syntactic information (such as *direct object*, *indirect object*, PP_{for}, PP_{from}, PP_{into}, PP_{outof}) taken independently, not in subcategorization frames. Taking together these two accuracy results, we can conclude that the role of syntactic features is capital, but also limited (they do not achieve accuracies beyond 60%). Therefore, other types of features, more semantic in nature, may be necessary in order to achieve better results in verb classification. The specific type of this features and the way to combine them with syntactic features becomes then a relevant avenue for research.

Some works have made an effort in this direction, particularly regarding the optimal combination of lexical and syntactic features. Sun, Korhonen, and Krymolowski (2008) highlighted the importance of enriching subcategorization frames with prepositional preferences, achieving an accuracy of 0.64 in classifying 204 verbs into 17 Levin's classes. Li and Brew (2008) also investigate a range of feature sets relevant for classifying English verbs (word co-occurrences, subcategorization frames, enriched dependency relations) and determine that subcategorization frames alone have a limit in their performance. To overcome this limit, they combine subcategorization frames with co-occurrence information and obtain a better result in several classification experiments (accuracy of 0.53 in classifying 1300 verbs in 48 Levin's classes). The best performing feature sets from Sun and Korhonen (2009) also contain enriched subcategorization frames with selectional preferences. However, in this case they are not lexical preferences, but clusters of argument heads (a generalization of lexical preferences). They obtain an F_1 score of 0.80 for their clustering algorithm when classifying 204 verbs into 17 fine-grained Levin classes. Following the exploration of selectional preferences, Roberts et al. (2014) perform a thorough comparison of several selectional preference models that are used to enrich subcategorization frames for German verbs. The models consist of a lexical preference model, two types of noun clusterings (one based on the grammatical relation between verb and nouns and other based on distributional data), an LDA model, and GermaNet concepts. The best results are achieved with the noun clustering that is based on syntactic relations with verbs adapted from Sun and Korhonen (2009), that obtains an F_1 score of 0.40 when classifying 168 verbs into 43 classes. The results of this model are closely followed by the ones obtained using selectional preferences acquired using LDA techniques (F_1 score of 0.39). However, simple lexical preferences perform similarly (F_1 score of 0.38).

Most of these approaches aim at modelling the information that Levin deemed as relevant to adequately describe verb behaviour for a semantic classification i. e. diathesis alternations. Indeed, information about subcategorization frames seems to be a natural approximation in order to formalize them. However, there is a different approach that aims at capturing directly this type of information, followed by Sun, McCarthy, et al. (2013). They calculated the joint probability of two subcategorization frames on the assumption that a high such probability could entail a diathesis alternation. Although some of the alternations that they find are not straightaway interpretable with the classical typology of alternations, their work takes advantage of less frequent subcategorization frames that in other approaches are overlooked. Other works are not that heavily reliant on subcategorization frames, using instead tree kernels and LSA (Croce et al., 2012) or distributional information and knowledge bases such as NELL (Mitchell et al., 2015) in order to describe verb behavior (Sedoc et al., 2017; Wijaya, 2016).

2.5.3 *Verb classifications for languages other than English*

Although the majority of the efforts in automatic verb classification have been carried out for English, there are a number of works carried out for other languages: Spanish (Ferrer, 2004; Alonso Alemany et al., 2007), French (Falk et al., 2012), Persian (Aminian et al., 2013) and Italian (Lenci, 2014) among others not mentioned before. Additionally, cross-linguistic potential of verb classifications has been explored (cf. Sun, Korhonen, Poibeau, et al. (2010) and Merlo, Stevenson, et al. (2002) for a classification developed for English and transferred to French). Another work that explores multilingual data is the one developed by Tsang et al. (2002), which combines English and Chinese features from parallel corpora to take advantage of the information that each contribute. This is done in order to leverage smaller bilingual corpora and to alleviate the need of a large monolingual corpus for verb classification.

In the remainder of this section we get into more detail for the classifications developed for Spanish. The first classification that we mentioned, Alonso Alemany et al. (2007), explores the creation of an automatic verb classification based on syntactic information with the ultimate goal of acquiring subcategorization frames. Therefore, they build a classification of the verb senses in a corpus using three types of feature sets to characterize verb behaviour in an additive fashion: 1) syntactic category of the verb sense arguments, 2) syntactic category plus syntactic function, 3) syntactic category, syntactic function and semantic roles of the arguments. They find out that the best results are obtained when using a hierarchical clustering and syntactic categories plus syntactic function features.

Following with the work developed for Spanish, Ferrer (2004) applies automatic classification techniques already developed for English to Spanish. This work collects subcategorization frames from corpora, filtering out those that have low probability. The result is a set of 11 subcategorization frames with selectional preferences for prepositions. Besides, they add information for Named Entities to model meaning components (*no NE*, *people*, *locations*, and *institutions*). In order to create the automatic classification this work uses a bottom-up hierarchical clustering algorithm to group 514 verbs into classes. They compare their automatic classification against a manual classification developed by Vázquez et al. (2000), which constitutes their gold standard. They obtain an adjusted Rand measure of 0.07 for the 31 classes built without Named Entities and the same score for 15 classes created using Named Entities.

2.5.4 *Algorithms used to create the classifications and evaluation metrics*

As for the algorithms used to obtain the verb classifications, the literature shows a variety of methods to assign verbs to classes. Both supervised and unsupervised methods have been used. The most common unsupervised methods include K-means (introduced by Forgy (1965)) and used by Schulte Im Walde (2006); hierarchical clustering, applied by Stevenson et al. (2003), Reichart et al. (2013) and Sun and Korhonen (2011); spectral clustering used in Brew et al. (2002); Sun, McCarthy, et al. (2013) and Sun, Korhonen, Poibeau, et al. (2010); Expectation Maximization algorithm, applied by Rooth et al. (1999) and finally Dirichlet mixture models, used in Vlachos, Ghahramani, et al. (2008) and Vlachos, Korhonen, et al. (2009). The supervised methods include the classifier C5.0 (decision tree and rules), which is a modern version of C4.5 (Quinlan, 1992), used by Merlo and Stevenson (2001) and Tsang et al. (2002); support vector machines (SVMs) used by Joanis et al. (2008) and Croce et al. (2012) and Bayesian multinomial logistic regression, applied by Li and Brew (2008) and Li, K. Baker, et al. (2011).

The evidence about which algorithm performs better is limited because, as we mentioned before, the datasets and evaluations vary widely. However, there are some works that focus on comparing several methods for the same task and data: Sun, Korhonen, and Krymolowski (2008) experimented with four supervised methods (K-nearest neighbours, support vector machines, maximum entropy classifiers and Gaussian classifiers) and one unsupervised (cost-based pairwise clustering), finding that the best was the Gaussian classifier. Sun and Korhonen (2009) compared the results obtained using pairwise clustering and a variation of spectral clustering which exploits the MNCut algorithm, finding that the spectral clustering algorithm outperforms the pairwise clustering. Furthermore, Roberts et al. (2014) compared the hard and soft versions of Latent Dirichlet Allocation (Blei et al., 2003), finding that the hard version performed better with their data.

The comparison between different automatic classifications is also hindered by the variety of metrics used to evaluate the results of these algorithms. However, in the case of the evaluation of unsupervised algorithms (clustering) against a gold standard, there is a number of works that have used the harmonic mean of purity and accuracy for evaluation in a variety of languages (Scarton et al. (2014) for Portuguese, Sun (2013) for English, Sun, Korhonen, Poibeau, et al. (2010) for French).

To calculate purity, each cluster is paired with a golden class which is most frequent in the cluster (dominant class). The number of verbs in the cluster that are associated to the golden class are considered as correctly classified, whereas the verbs in the cluster that belong to other classes are considered as classification errors. This metric is biased towards small clusters and sometimes it is modified so it penalizes singleton clusters to minimize this bias. The method to compute this metric can be seen in equation 2.1, were C represents the automatic classification, G the gold standard, N is the number of verbs, and $|w_k \bigcap c_j|$ stands for the number of verbs common in cluster k and golden class j.

$$purity(C,G) = 1/N \sum^{k} max|c_k \cap g_j| \tag{2.1}$$

Accuracy is the proportion of verbs that are labelled correctly, i.e. that fall in the dominant class. The method to compute this metric can be seen in equation 2.2, where the equivalence of the cluster label c and the class label g of the ith verb is checked, adding 1 to the total count that is then divided by N, the number of verbs.

$$accuracy(C,G) = max(1/N \sum_{i} 1(c_i = g_i)) \tag{2.2}$$

To balance these two measures, their harmonic mean is calculated as in equation 2.3.

$$Hmean = \frac{2 \cdot purity \cdot accuracy}{purity + accuracy} \tag{2.3}$$

2.5.5 *Open issues in verb classification: domains, polysemy and applications*

Most automatic verb classifications use data extracted from newspaper corpora, assuming that the classification they obtain is representative of general language. However, one way in which verb classifications can be applied to practical tasks is to specialize them for a specific domain. This idea was pursued in Korhonen, Krymolowski, and Collier (2006) and Korhonen, Krymolowski, and Collier (2008), where an automatic classification for the biomedical domain was developed. However, these classifications were evaluated by comparing them against a manually created classification and not tested in BioNLP tasks.

Besides, it is important to note that the majority of the work done in this area assumes only one sense per verb lemma, this is, it does not differentiate between verb senses. However, different senses of the same lemma are likely to have different syntactic and semantic behaviour (Hare, McRae, et al., 2003). Additionally, Dorr and D. Jones (1996) highlight that verb sense distinctions are crucial for deriving semantic information from syntactic cues. Little research has put the focus on the effects of polysemy in verb classification. Among the exceptions, we can name the work of Korhonen, Krymolowski, and Marx (2003), who classified 110 polysemous English verbs into 34 classes using subcategorization frames. They found out that the impact of polysemy is substantial: senses belonging to polysemous verbs with a predominant sense and verbs that exhibited regular polysemy showed a strong tendency to be classified together, whereas senses that belonged to verbs that showed strong irregular polysemy were frequently assigned to singleton clusters.

The strategies that aim at tackling polysemy are varied: some follow a two step approach (Kawahara et al., 2014; Sedoc et al., 2017) where the first step consists in clustering the sentences within each verb lemma, obtaining classes that may be seen as different senses of the verb lemma. Then, they proceed to cluster these induced senses as in previous works. Other approaches use selectional preferences to distinguish between verb senses, such as Lapata et al. (1999), who use syntactic subcategorization information with prepositional preferences to disambiguate polysemous verbs. Finally, other approaches, such as the one by Gildea (2002), rely on the clustering algorithm for this task. Thus, instead of using a hard clustering algorithm where each verb is assigned to one cluster, they use soft clustering algorithms, where each verb has a probability of being assigned to each cluster. Therefore, if a verb has a probability higher than a threshold to being assigned to two or more clusters, it can be considered polysemous.

As for the evaluation, we have seen that most of the automatic verb classifications are validated using Levin's classification or a custom one. Among the few automatic classifications actually used in NLP tasks we can name the work by Lamirel, Falk, et al. (2015), which is applied to semantic role labeling, something that is linked with our objectives and that will be discussed in more detail next; the work by Shutova et al. (2010), applied to argumentative zoning, the work by Alonso Alemany et al. (2007), applied to subcategorization acquisition and, finally, Guo et al. (2011), who applied clustering to metaphor identification. Besides, automatically acquired classes have been also used for estimating a set of features from social media language (locus of control, sarcasm and sentiment) in the work of Sedoc et al. (2017).

2.5.6 *Verb classifications and predicate-argument structure*

As we have already seen, although most of the work in verb classification (either manually or automatically created) mentions the possible benefits and applications of the classes, few of them are designed with the goal of being applied in a task. This situation presents some significant exceptions related to our objectives that are described more detail in what follows. First, within manually constructed verb classifications we find the work of Swier et al. (2005), where VerbNet was used for Semantic Role Labeling (SRL), a task that is typically addressed using supervised methods, with interesting results. In this approach they propose a frame matching technique for argument identification and labelling. Using the 20 % of the British National Corpus (Burnard, 2000) as test data, they first proceed to chunk the sentences that contain verbs that are also in VerbNet. Then they match the syntactic-semantic frames associated to that verb in VerbNet to the sentence chunks. To do so, they consider each of the chunks an argument candidate and develop a metric that takes into account the overlap between that candidate and the arguments in the corresponding VerbNet frames. If there is a match and both segments are aligned, it means that the arguments have been found and the roles of the VerbNet frame are assigned to the matched sentence chunks. For ambiguous assignments, where more than one role may correspond to an argument candidate, a statistic model calculates the probability of each assignment by taking into account the verb, the syntactic function of the chunk and the noun head of the chunk head. If the probability obtained does not achieve a threshold, a backoff model based on more general categories is applied. Finally, the newly labelled arguments are then used to enrich the model in a bootstrapping fashion. The authors obtained a F1 of 0.65 on this task of argument identification and labelling using VerbNet 1.5. Later, Pradet et al. (2013) experimented with a similar methodology, although this time they used the VerbNet 3.2. and showed that handling specific constructions, such as the passive voice, could improve the results. They obtained a F1 of 0.70 on gold arguments (arguments whose boundaries are already determined), although the performance falls to 0.47 when the arguments are automatically identified.

Another work that uses manually created classifications for SRL is the one by Giuglea et al. (2006). Their strategy for the SRL task consists in creating a mapping across several lexical resources (FrameNet, VerbNet and PropBank). They first link FrameNet and

VerbNet through the ILC (Intersective Levin's classes (Dang et al., 1998), by taking into account the number of verbs shared between the VerbNet classes and the FrameNet frames. Once this is done, frame elements (e.g. *Addressee*) are mapped to VerbNet semantic roles (e.g. *Recipient*). They annotate FrameNet sentences with roles automatically, using an SVM classifier and then they use the mapped classes as features together with other standard features used for SRL from Gildea and Jurafsky (2002) and Pradhan et al. (2005), finding that there is virtually no difference in the accuracy when using these mapped classes as compared to using the gold frame label. They also perform experiments labelling the sentences in PropBank, finding that the accuracy increases notably when the feature for the classes is used (from 62% to 81%).

The work by Lamirel, Falk, et al. (2015) is particularly related to our objectives. They create an automatic verb classification for French using syntactic and semantic features from manually constructed resources (a syntactic lexicon for French verbs and the English VerbNet). Regarding the syntactic features, they use subcategorization frames and a set of additional features that indicate whether a verb accepts symmetric arguments (e.g. John met Mary/John and Mary met), has four or more arguments, combines with a predicative phrase, takes a sentential complement or an optional object or accepts the passive with the particle *se*. As for semantic features, they select features from the lexicon that indicate whether a verb takes a locative or an asset argument and whether it requires a concrete object (non-human role) or a plural role. Besides these features, they use the list of semantic roles associated to a VerbNet verb classes, obtaining it by linking the French verbs with VerbNet classes using translation techniques. Their automatic classification is created using IGNGF (Incremental Growing Neural Gas with Feature maximization), a clustering algorithm developed by Lamirel, Mall, et al. (2011) and contains 2183 verbs in 13 classes. A first evaluation is carried out by comparing the classes obtained automatically to the classes present in the translation of part of Levin's classification to French (116 verb in 12 classes). In this evaluation they obtain an F_1 score of 0.70. This algorithm also associates each cluster with a set of semantic and syntactic features used to create the classification. Therefore, each cluster has a set of associated subcategorization frames, semantic features and semantic roles. To evaluate whether this information associated to the class is correct they formulate several tasks in which this information is compared to information available in a manually annotated gold standard corpus.

The gold standard corpus is created by selecting sentences that contain the verbs used in the classification and labelling their arguments with roles. These sentences are taken from the Paris 7 Treebank (Abeillé et al., 2003). Then, in order to evaluate how representative were the semantic roles associated to the automatic classes by the algorithm, they compare the roleset for each verb in the gold standard and in the automatic classification. More specifically, the association between a verb and the roleset from its automatic class is considered correct if the roleset of the automatic class is a superset of the roles associated to the verb in the gold standard. This has the inconvenient of leaving the possible overgeneration of semantic roles unmeasured:

if automatic classes were associated with all the semantic roles, the comparison with the gold standard would yield a 100% of recall. This issue aside, they report a 0.48 of recall. They attribute this low result to the fact that many associations between verbs and semantic roles in VerbNet are lost in the translation to French, and therefore could not be used as features for the automatic classification. When they take into account verbs and role sets that are both in the automatic classification and the gold standard they obtain a 0.75 of recall. Another problem that they identify is polysemy: 21% of the verb instances in the gold standard have a different sense from the one that is present in the verb classification.

They further proceed to test the performance of the automatic classification in assigning roles to specific verb arguments using Swier and Stevenson's algorithm that we described above. In order to assign semantic roles to the syntactic arguments in the test sentences, they take the semantic roles associated to the classes in the automatic classification (e.g. *[Agent, Destination, Theme]*), considering all the different configurations of the arguments (e.g. *Agent-Theme, Agent-Destination, Agent-Destination-Theme* for the previous example). In this way they create a set of candidates consisting in semantic roles plus syntactic functions that are used to label the arguments in the sentences, obtaining a F1 of 0.71 (0.72 of precision and 0.7 recall) in this task. These results are better than a default baseline that always assigns the same roles to syntactic arguments (agents to subjects, patients to objects, etc.), for which they report a 0.65 F1.

They also compare their results with a baseline in which there are no classes (using directly the information associated to the corresponding verb in the lexicon), obtaining a F1 of 0.45 (0.77 precision and 0.32 recall). As we can see, the presence of classes lowers a bit the precision but improves significantly the recall. Wrapping it up, their approach to evaluate the association between verbs and semantic roles using automatic classes combines the use of lexical resources specific for French, the translation of VerbNet, the creation of an automatic classification and a semantic role expansion strategy. Therefore, although they demonstrate that automatic classifications can organize argumental information in a coherent way, the question of whether automatic classifications can capture information related to argument structure without being created with explicit information about semantic roles and the extent to which this information can be generalizable to verbs outside the classification remains unanswered.

In this chapter we aimed at providing a review of several issues that frame our research: first we presented the anchoring of verbs and argument structure in our mental representations of events and the linguistic materialization of these representations, and secondly we introduced an overview of the approaches to calculate verb similarity and to create automatic verb classifications, looking at the work done and what is yet to be accomplished. In the next section we will dive deeper in the specific resources that we used and the methodology followed to carry out our study.

Chapter 3

Methods and materials

The hypothesis that constitutes the backbone of this work, as we saw in section 1.3, is that the existence of regularities in predicate-argument structures that can be both captured and generalized with similarity based techniques. The two sub-hypothesis that stem from this idea elaborate on the specific properties of these regularities. The first sub-hypothesis is related to the nature of the regularities and establishes that even though there are different perspectives of study on predicate-argument structure that may emphasize different aspects, it is possible to discover some basic common features among them that reflect systematicity in argument structure. The second sub-hypothesis, related to the generalization, affirms that an automatic classification of predicates can reflect patterns of homogeneous behaviour in predicate-argument structure in a way that accounts for predicates not included originally in the classification.

Thus, the objectives of this dissertation revolve around analyzing and modelling verb similarity according to predicate-argument structure and, consequently, our work is organized around these two main pillars. The first one deals with the analysis of verb similarity, looking at it from different perspectives. The second focuses on the modelling of verb similarity, and more specifically, on the creation of a flexible automatic verb classification based on corpus data and a clustering algorithm. The details of the methodology of each experiment are explained in chapters 4 and 5 respectively, where the experiments are described in detail. This chapter focuses then on the procedures and data that are common to these studies.

3.1 METHODOLOGY

In relation to the methodology followed, it needs to be put forward that we deal with verb senses as the unit of analysis. In both studies verb senses are characterized by taking into account their statistical distribution over several sets of features, keeping track of how many times a certain verb sense co-occurs with a given feature. In the first experiment, where we aim at analyzing verb similarity from different perspectives, we take into account co-occurrences between these senses and features from three different sources of linguistic information related to argument structure and event representation:

- Verb constructions from linguistic theory.
- Semantic roles from corpus annotations.
- Psycholinguistic data extracted from a word association experiment.

In order to carry out the analysis we select 20 verb senses from SenSem corpus (Fernández-Montraveta et al., 2014) and we characterize each of them using relevant linguistic features obtained from the three mentioned perspectives. As a result, we end up with three different characterizations, one from each perspective. We measure the pairwise similarities between these 20 verbs according to each of these characterizations, obtaining similarity coefficients for all the possible verb sense pairs (a total of 190 verb sense pairs). In other words, each perspective defines similarity between verb senses in a different way and in this step we capture these three definitions in a quantitative way, obtaining three parallel sets of similarity coefficients. These perspectives and their different definitions of similarity are then compared. First, we look at the correlations between the three sets coefficients obtained from each perspective. After that, we analyze the linguistic characteristics of the most representative pairs of verb senses for each perspective. This design enables us to study coincidences and divergences between the three approaches with regard to verb similarity within the argument structure domain.

As for the modelling part, the scope of the study of verb similarity builds on the previous one and expands beyond the pairwise framework, looking instead at the different configurations of similarity that arise when more verb senses are taken into account. In order to do this, we characterize more than 600 verb senses from the SenSem corpus using several types of linguistic information related to argument structure that are found relevant in the previous experiment. After that, we use an agglomerative clustering algorithm to create an automatic classification of verb senses. In order to test the coherence and generalization power of the classification two evaluations are performed:

- a comparison with other classification of reference, the gold standard, as it is usual in automatic classification approaches. This gold standard is also based on argument structure (semantic roles).
- an extrinsic evaluation, where a set of verb senses not included in the classification (test verbs) are assigned to the obtained classes using a similarity metric.

The semantic roles associated to these test verb senses and those associated to the class are compared.

Besides, we perform a comparison between the verb similarities drawn by the construction and psycholinguistic data and the similarities found in the clustering. This allows us to test the flexibility of automatic classes regarding different formalizations of argument structure, and therefore, event information. Figures 3.1 and 3.2 display the flowcharts that summarize the steps carried out in each part.

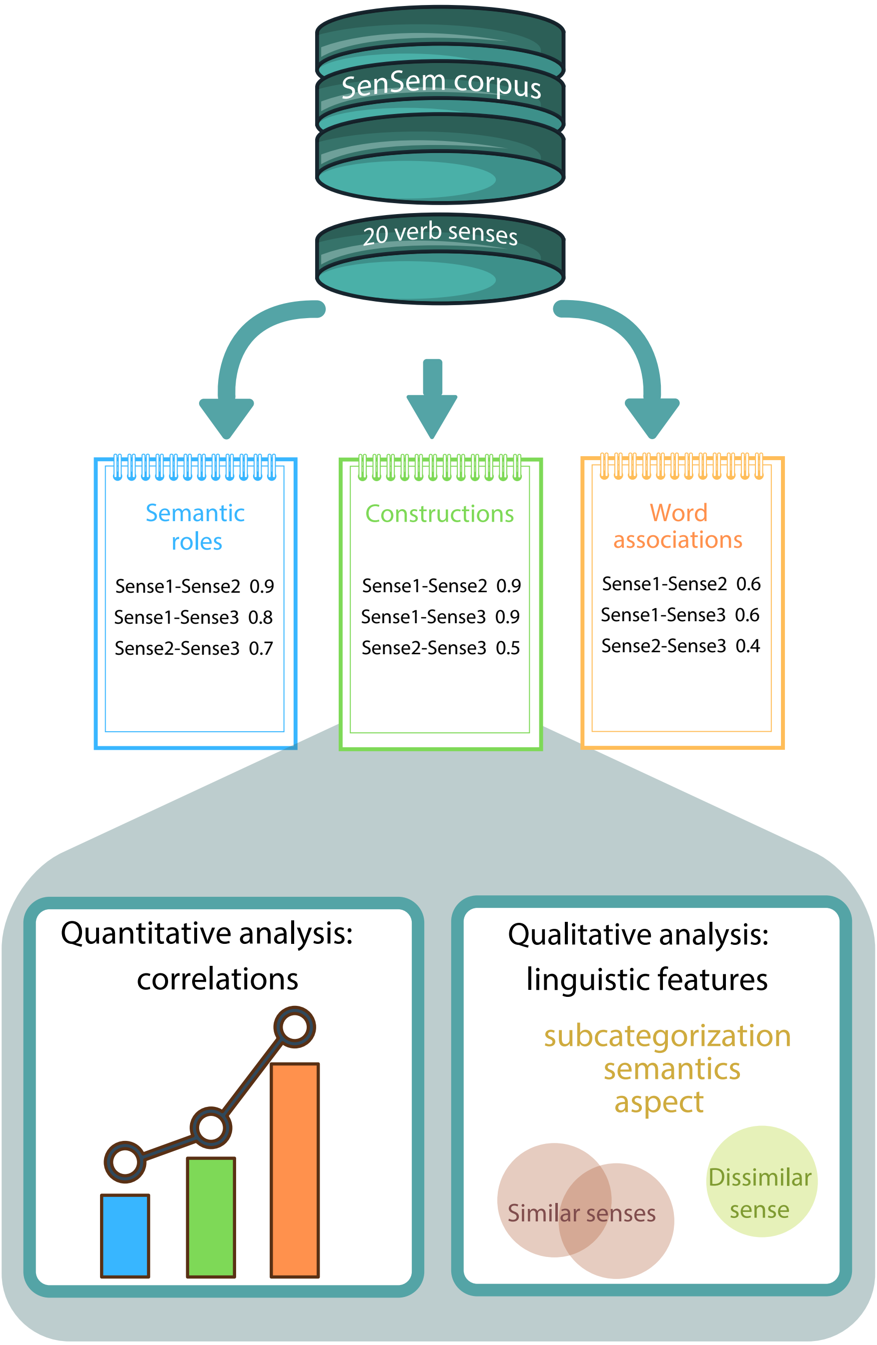


Figure 3.1: Steps carried out for analysis of verb similarity

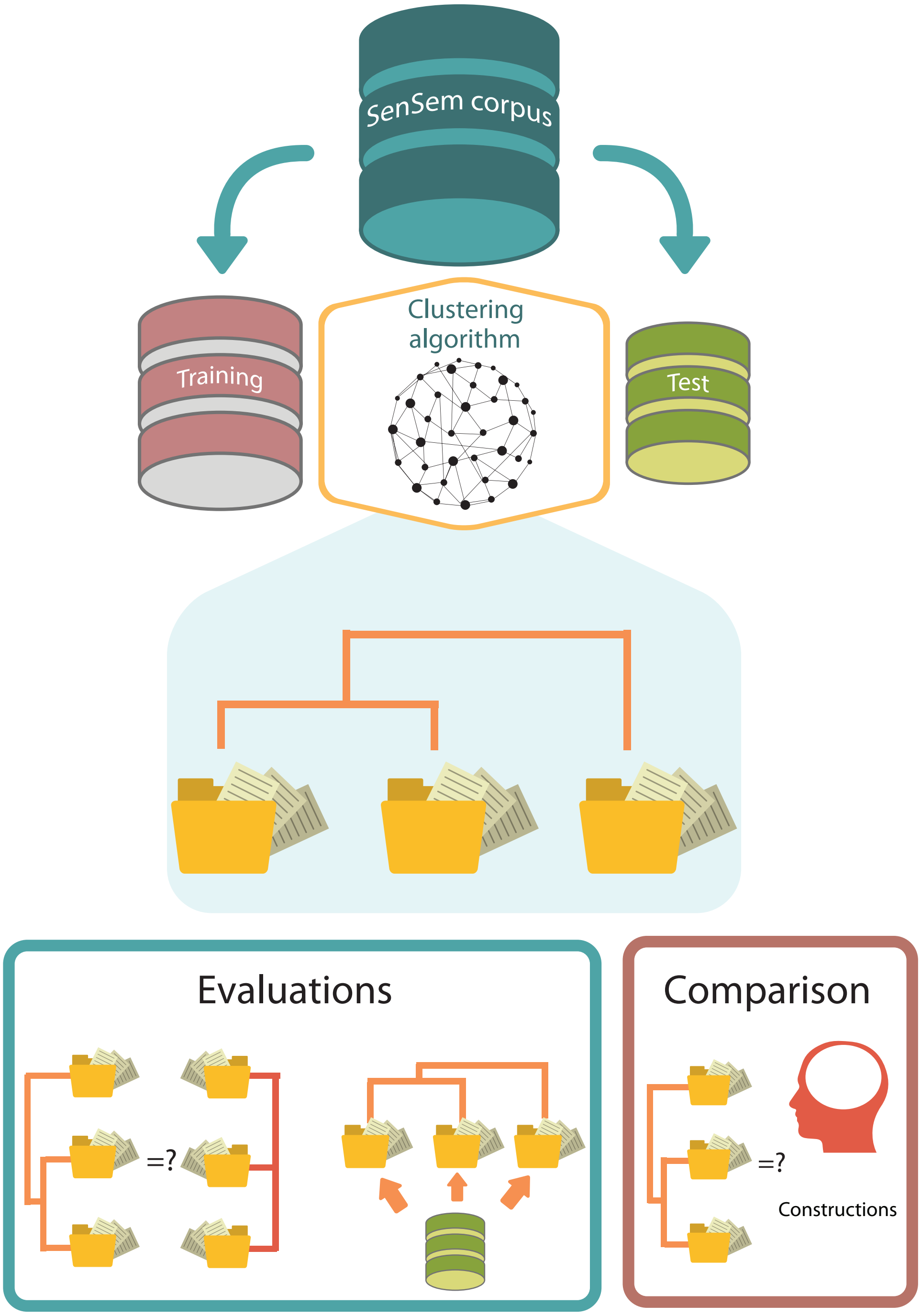


Figure 3.2: Steps carried out for the modelling of verb similarity

3.2 MATERIALS

As for the materials, one of the sources that we will use to characterize the verb senses is the SenSem corpus (Fernández-Montraveta et al., 2014), which contains more than 100 sentences for each of the 250 verb lemmas that are included. These lemmas are among the most frequent verb lemmas for Spanish. Besides, the senses of these verb lemmas are distinguished. Summarizing, the corpus consists of a total of 30,365 sentences and 988 different verb senses. Figure 3.3 illustrates the distribution of the number of senses per lemma. This corpus is annotated at the syntactic and semantic level: it contains information about the constructions in which verbs participate and the syntactic category, function and semantic role of the arguments. Additionally, the nucleus of the arguments are marked.

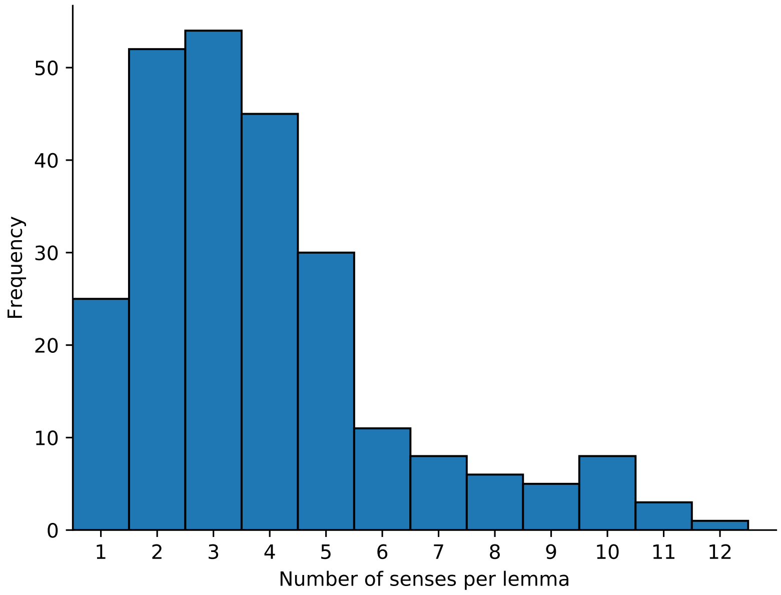


Figure 3.3: Number of senses per lemma

With respect to the semantic roles, SenSem contains 40 semantic roles. The Sensem role set is particularly detailed in the description of the agent and theme categories, for which their different subtypes are specified (agent-goal, agent-experimenter, affected-created-theme, affected-destroyed-theme, etc.). We think that this characteristic makes the SenSem role set particularly suitable for organizing its semantic roles in a hierarchy. A semantic role hierarchy has the advantage of offering flexibility in order to experiment with the different degrees of abstraction of semantic role information. Therefore we constructed a semantic role hierarchy based on the LIRICS approach (Bonial et al., 2011), that offers a hierarchical organization of semantic roles based on the ones present in

VerbNet.

Our manually built three-level hierarchical role set can be seen in figure 3.4. SenSem roles were directly used for the fine-grained, lower level (red in figure 3.4). In order to build up the next levels of the hierarchy, we follow a bottom up approach where Sensem roles are mapped to the third level roles in the LIRICS proposal to constitute the intermediate level of our hierarchy (blue in the figure 3.4). This mapping was created in two steps: an initial mapping was developed first and then further modifications were added to improve its correspondence to the SenSem roles. Afterwards, the resulting intermediate roles are mapped to the four basic roles in LIRICS: *actor*, *undergoer*, *place* and *time*, which constitute the more coarse-grained level in the hierarchy (orange in the figure 3.4).Besides, we keep the Sensem role *Circumstance* separated, because it is more general and does not fit in LIRICS categories. Once this final step is done, we obtained a hierarchy of three levels of semantic role granularity that enables us to formalize argument structure in a layered and flexible way.

A complete description of the SenSem semantic roles can be found at http://grial.edu.es/sensem/apps/admin/public/dwn/10_en_Descripcion%20etiquetas%20constituyentes%20SenSem.pdf. The definitions of the LIRICS semantic roles that we use in our hierarchy are detailed next.

Abstract level of semantic roles

- Actor: Participant that is the instigator of an event.
- Undergoer: Participant in a state or event that is not an instigator of the event or state.
- Place: Participant that represents the state in which an entity exists.
- Time: Participant that indicates an instant or an interval of time during which a state exists or an event took place.
- Circumstance: This role comes from the SenSem roleset. It is user to refer to a circumstance that modifies the event.

Intermediate level of semantic roles

- Actor#Actor: The participant which is responsible of initiating the event, not necessarily intentional (e.g. *This implies large changes*)
- Actor#Cause: Actor in an event (that may be animate or inanimate) that initiates the event, but that does not act with any intentionality or consciousness; it exists independently of the event.
- Actor#Agent: Actor in an event who initiates and carries out the event intentionally or consciously, and who exists independently of the event.
- Undergoer#Theme: Undergoer that is central to an event or state that does not have control over the way the event occurs, is not structurally changed by the event, or is characterized as being in a certain position or condition throughout

the state.

- Undergoer#Attribute: Undergoer that is a property of an entity or entities, as opposed to the entity itself.
- Undergoer#Patient: Undergoer in an event that experiences a change of state, location or condition, that is causally involved or directly affected by other participants, and exists independently of the event.
- Undergoer#Instrument: Undergoer in an event that is manipulated by an agent, and with which an intentional act is performed; it exists independently of the event.
- Place#Goal: Place that is the end point of action and exists independently of the event.
- Place#Location: Participant that represents the state in which an entity exists.
- Place#Place: Participant that refers to the path that is followed to accomplish an action.
- Place#Source: Place that is the starting point of action; exists independently of the event.
- Time#Initial time: Time that indicates when an event begins or a state becomes true.
- Time#Final time: Time that indicates when an event ends or a state becomes false.
- Time#Time: It indicates the moment in which the event occurs.

Regarding our unit of analysis, as we already mentioned, in this study we focus on verb senses rather than on lemmas to obtain more precise models of similarity, because different senses of the same lemma are likely to have different syntactic and semantic behaviour. This can be exemplified with the lemma *valer* from the Sensem corpus. This lemma has several senses listed in the Sensem lexicon: sense 1 of *valer* means 'to cost', its predominant subcategorization frame is *NP V NP* and its typical semantic roles are *theme* and *measure*. However, with sense number 2, *valer* means 'for something to be the cause of a result', its predominant subcategorization frame is *NP V NP PP* and its typical semantic roles are *cause*, *theme* and *goal*. Therefore, it is advisable to use specific senses of verbs in order to obtain suitable data for an issue as sensitive to polysemy as it is similarity.

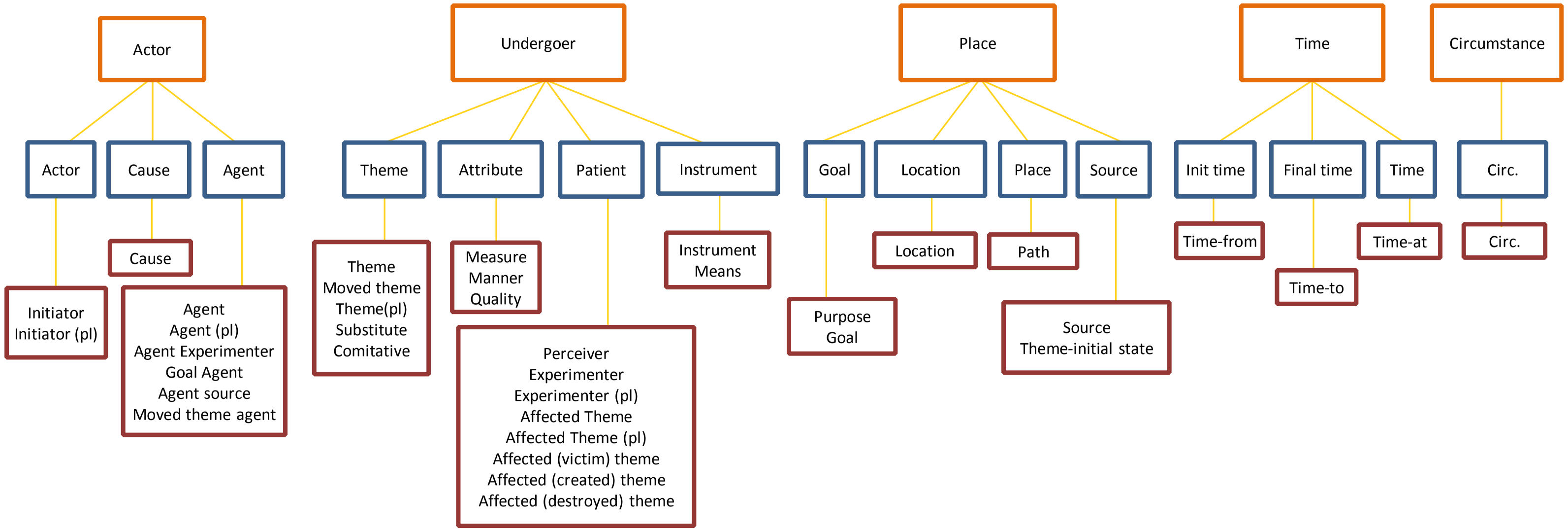


Figure 3.4: Hierarchy of semantic roles

Besides the corpus SenSem, we also use several ontologies as feature sources in order to represent the information associated to the verb senses:

1. WordNet (Fellbaum, 1998): WordNet is a lexical database created initially for English and extended afterwards to other languages.[1] WordNet includes nouns, verbs adjectives and adverbs, which are grouped in sets of synonyms called synsets. It also specifies their definition and the semantic relationships between these synsets. This lexical resource is one of the largest of its kind for many languages and also one of the most widely used for all types of applications, particularly for word sense disambiguation and semantic similarity related tasks. Among its limitations, it is often mentioned the fact that it does not cover domain-specific vocabulary and that the conceptual distances between the more abstract synsets are larger than the distances between more specific nodes. From WordNet we select two levels of the taxonomy: the hypernym, which is the parent of the node of interest in the *is-a* relationship, and the supersense, which is a much more abstract node. Supersenses are the *lexicographer class* labels used in WordNet. They act as broad semantic fields and they were used as an aid to organize more specific categories. Some examples of WordNet supersenses are *person*, *communication*, *artifact* or *act*. We illustrate the different positions and scope of those two categories in the WordNet taxonomy in image 3.5[2]: *motorcar* is the hypernym of *hatch-back*, *compact* and *gas guzzler*. *Artefact* is the supersense of all the categories present in this image, including those three.

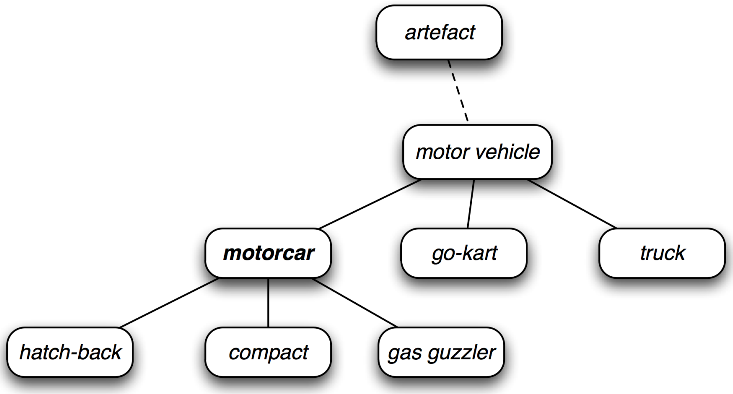


Figure 3.5: WordNet hierarchy

2. TCO ontology (Álvez, Atserias, et al., 2008): This ontology was born alongside with the development of EuroWordnet (Vossen, 1998) as a tool to organize the common Base Concepts shared among WordNets from different languages. It consists of 63 fundamental semantic distinctions (e.g. static/dynamic) and it

[1] A list of all available WordNets can be found at http://globalwordnet.org/resources/wordnets-in-the-world/

[2] Image from Bird et al. (2009), shareable under CC license.

is organized following Lyons'approach (Lyons, 1977). Entities in this ontology fall into three different sets: 1st order entities (physical things such as *vehicle*, *animal* or *substance*), 2nd order entities (situations: *begin*, *continue* or *be*) and 3rd order entities (unobservable entities: *idea*, *information*, *theory*, *plan*). These sets are associated to basic features such as the following:

- For 1st order entities we find features as *form*, *function* and *composition*, among others
- For 2nd order entities there are features such as *situation type* and *situation component*.
- 3rd order entities do not have further feature subdivisions.

The distribution of these features for each set can be seen in table 3.1.

As a result, a concept may be described using one or more features, which allows for a great level of flexibility when characterizing concepts. For example, a concept such as *factory* can be characterized in different ways:

- Building+Group+Object+Artifact, if we look at the function, composition, form and origin features
- Any of the former features in isolation
- The generic set: 1st order entity

Top		
1stOrderEntity	2ndOrderEntity	3rdOrderEntity
Origin	**SituationType**	
Natural	Dynamic	
Living	BoundedEvent	
Plant	UnboundedEvent	
Human	Static	
Creature	Property	
Animal	Relation	
Artifact		
Form	**SituationComponent**	
Substance	Cause	
Solid	Agentive	
Liquid	Phenomenal	
Gas	Stimulating	
Object	Communication	
	Condition	
Composition	Existence	
Part	Experience	
Group	Location	
	Manner	
Function	Mental	
Vehicle	Modal	
Representation	Physical	
MoneyRepresentation	Possesion	
LanguageRepresentation	Purpose	
ImageRepresentation	Quantity	
Software	Social	
Place	Time	
Occupation	Usage	
Instrument		
Garment		
Furniture		
Covering		
Container		
Comestible		
Building		

Table 3.1: TCO features

3. SUMO ontology (Suggested Upper Merged Ontology), by Niles et al. (2001): This ontology merges and combines other publicly available ontologies into one. It comprises around 25,000 terms and around 80,000 axioms when all the merged resources are taken into account. It contains an upper ontology connected to several domain-specific ontologies. Its upper levels consist of general terms and support semantic inferences of broad nature and also interoperability between the connected domain-specific ontologies. It has a hierarchical structure, with Entity as its top node, subsuming physical concepts (entities that have a position in space and time) and abstract concepts (everything else). The following hierarchy represents the top level concepts in SUMO.

   - entity
     - Physical
       - Object
         - SelfConnectedObject
         - Collection
       - Project
     - Abstract
       - SetClass
         - Relation
       - Proposition
       - Quantity
         - Number
         - PhysicalQuantity
       - Attribute

   In this work we us a version of the SUMO ontology, called the Adimen-SUMO ontology, which transformed the format of the ontology to first order language (Álvez, Lucio, et al., 2012).

Each of these ontologies are divergent world views that organize knowledge in different ways, according to the criteria applied when designing them. Each exhibits different categories, properties and relations. By using them in our experiments, we aim at determining whether there is a specific world view formulation (this is, a set of categories, properties and relations given by an ontology) that best captures relevant information in order to deal with verb similarity. All of these ontologies are linked through the MCR [3] (Multilingual Central Repository, Atserias et al. (2004), and therefore the equivalents in all ontologies can be used.

In this chapter we provided an overview of the data used (corpus and ontologies) and the methodology followed in order to achieve our goals. Firstly, we laid out the steps for an study of the qualitative and quantitative pairwise similarity of verbs from three

[3] The MCR connects the EuroWordNets from six different languages, including Spanish, using WordNet 3.0 synsets as linking indexes (ILI, interlingual index) and integrates and connects the TCO and SUMO ontologies.

different perspectives. Secondly, we presented an automatic verb clustering experiment and its evaluation. These two pieces of research are explained in detail in the following chapters 4 and 5 respectively.

II

Verb similarity: analysis, modelling and evaluation

CHAPTER 4

Perspectives on the analysis of verb similarity

Similarity is a rather abstract and fuzzy concept. It is hard to precise exactly what it means for an object A to be similar to an object B because similarity is not something totally objective and external to us. However, if we want to use it as a mechanism that enables us to gain knowledge about the elements engaging in the similarity relationship, the respects in which the similarity between two elements is being defined should be carefully specified. This is not a trivial task because, narrowing this issue down to language, there are many aspects that can be taken into account when describing and quantifying the similarity between two given concepts, words or sentences (cf. De Deyne, Peirsman, et al. (2009) for a discussion in the semantic domain).

In particular, for lexical units there are many relationships that can be taken into account in order to define two such units as similar: synonymy (*car*, *automobile*), antonymy (*happy*, *sad*), hyponymy (*horse*, *animal*), meronymy (*handle*, *mug*), paradigmatically related (*Spanish*, *Italian*), syntagmatically related (*dog*, *bark*) and the dictionary definitions among others (Hontoria, 2003). Of course, the selection of the aspects that are relevant in order to determine the similarity between two lexical units depends on the goals of the comparison.

In our case, the specific goal is to study verb similarity as defined by argument structure, which in turn is connected with the linguistic expression of events. Since there are several views on the nature and properties of argument structure, in this chapter we select three of such views that come from different areas and look at their coincident and divergent aspects when defining verb similarity. As we mentioned in the previous chapter, this comparative approach for the analysis of verb similarity enables us to discover coincidences and divergences between the approaches (both quantitative and qualitative), revealing the systematicities related to predicate-argument structures.

Particularly, in our work, the different data and perspectives used to characterize verbs from the point of view of argument structure are the following:

- The constructions in which verbs participate, from the theoretical linguistics perspective.
- The semantic characterization (semantic roles) of the arguments that combine with the verbs, from the corpus linguistics perspective.
- The words that are frequently associated with the verbs, from the psycholinguistic perspective.

We consider that, in order to properly asses the systematicity in the definition of similarity of these different approaches, we need to provide both a quantitative and qualitative account of their effects on the calculation of verb similarity. The quantitative account will measure the amount of convergence between the different perspectives, whereas the qualitative account will aim at gaining some insights in the sources of the convergence or divergence.

With respect to the materials of the experiment, as we explained in the section 3.2, this experiment is carried out using a sample of the verb senses available in the SenSem corpus. Regarding the selection of the verb senses, we chose 20 senses that met several requirements: 1) having a frequency of more than 10 sentences in the corpus; 2) belonging to a variety of semantic fields, as defined by WordNet supersenses ('lexicographer file'labels); 3) showing a variety of subcategorization frames and semantic roles associated with them. These criteria were designed and applied in an effort to maximize the representation of diverse linguistic features in the selected sample, which constitutes a controlled data set suitable to be explored in detail. In section A.1 from appendix A there is a list of the verb senses chosen, along with their definitions and their frequency in the corpus; in section A.2 their semantic fields are specified. Furthermore, since none of the selected senses are related to coincident lemmas, sense identifiers have been omitted in the remainder of this work to improve legibility.

In order to be able to compute pairwise similarity between all the selected senses, we first obtained all possible sense pairs. To do so we avoided combinations of repeated elements (such as *abrir-abrir* 'open-open') and we did not take into account the order of the elements (e. g. *cerrar-abrir* 'close-open'was considered to be the same pair as *abrir-cerrar* 'open-close'). As a result, we obtained 190 verb sense pairs. Therefore, by calculating the similarity between the senses in each pair, we are able to measure the similarity among all the verb senses in our sample.

Calculating the similarity coefficient between the members of each pair involves first characterizing each verb according to a feature set. Therefore, we create a vector for each verb that records the co-occurrence of this particular sense with each feature of the feature set. Thus, the vectors hold information about the distribution of verbs over a collection of features. In this way, the similarity coefficient between any two senses (the members of the pair) can be computed by applying a metric to their corresponding vectors.

The remainder of this chapter is organized in the following way: we first describe each perspective of analysis, explain how it is formalized through feature sets and we detail relevant elements in its definition of similarity between verb senses. After that, we explore the correlations between the similarity coefficients calculated for each perspective. Finally, we investigate the linguistic features that underlie the similarity configurations in each perspective.

4.1 PERSPECTIVE FROM PSYCHOLINGUISTICS. WORD ASSOCIATIONS

Word associations (WA) refer to the word or words that first come to mind in response to a stimulus, typically a word or a sequence of words. Researchers from different areas have long pursued the idea of WA as a suitable mechanism for the study of the mental lexicon. Thus, WA have been present in studies that tap into a wide variety of (psycho)linguistic phenomena such as language disorders (Eustache et al., 1990), language acquisition (Namei, 2004) and semantic cognition (De Deyne, Navarro, et al., 2013) among others. Nevertheless, the usage of WA originated with quite different goals. The first major collections of WA can be traced back to the beginnings of the 20th century, when Jung and collaborators (Jung, 1910) and Kent and Rosanoff (Kent et al., 1910) established them as a diagnostic tool in order to discriminate normal behaviour from pathological behaviour on the basis of the type of responses given by the patients (Fitzpatrick, Playfoot, et al., 2013). WA later moved beyond its initial role as a psychiatric screening tool to become a device to research semantic memory and answer questions such how words are stored, represented and retrieved in memory (Mollin, 2009; Fitzpatrick and Izura, 2011; Nordquist, 2009). In this regard, WA have been used to predict the results of several memory tasks: they show semantic similarity effects in free recall and cued recall, they correlate with human similarity ratings (Steyvers et al., 2004) and they also exhibit semantic clustering effects in recall experiments (Manning et al., 2012).

However, due to the unbounded nature of responses, there is no straightforward or agreed-upon interpretation of the relationship between stimuli and responses (De Deyne and Storms, 2015), although it is generally acknowledged that these relations have a broad semantic basis (Nelson et al., 2000; Roediger et al., 2001; McRae, Khalkhali, et al., 2012; Brainerd et al., 2008). Evidence suggests that we might find typical relations such as hyponymy, synonymy and antonymy (Clark, 1970), although other types of relations such as cause-effect or part-whole (Herández Muñoz et al., 2014) and thematic relationships such as *cheese-yellow* (Peirsman et al., 2009) are frequent. Also, other types of relations, such clang associations (*click-clock*) have been mentioned in the literature (Fitzpatrick, Munby, et al., 2014). All in all, the landscape drawn by this diversity points towards a complex network of relationships that can not be accounted for using a single source. This complexity finds a more suitable explanation in embodied multimodal accounts of cognition. For example, Barsalou et al. (2008) claims that processes in which the body engages, such as perception and action, are crucial in the generation of conceptual representations. Under this framework, relationships such as those elicited in WA reflect not only linguistic phenomena but are also dependent on sensory inputs and context. Works such as the one done by Guida et al. (2007) explore the implications that this complexity has when the stimuli are verbs. In this work they found out that nouns elicited using verbs as stimuli referred to core arguments of verbs in around the 50% of the cases, extending also to prototypical causes, instruments, locations, etc. They related this phenomenon to the importance of the contextual setting of the event expressed by the verb:

> The distribution of associate types can therefore be explained by assuming that when participants produce associations in a verb association task they access a highly contextualized representation of the event or situation expressed by the stimulus, possibly through its virtual re-enactment in a typical setting and with typical participants.

From this point of view, the verb would be inserted in a enriched mental representation of an event in which WA are able to recover part of this mental representation in such a way that they not only reveal core arguments of verbs (traditional predicate-argument structures) but also recover other type of event participants. A priori, this would constitute a marked difference with the other approaches to argument structure that we include here, that only take into account core arguments.

Also focusing on verbs, Schulte im Walde and Melinger (2005) carried out a study of the properties and features that could be found in responses elicited by German verbs in a word association task. They looked at several sources of information: semantic relations in a taxonomy, morpho-syntactic information (part of speech of the responses, syntactic function of the noun responses with respect to the verb) and distributional information. Regarding the taxonomical relations, they found that more than 50% of the responses did not have a relationship with its verb stimuli that could be found in GermaNet, with hypernymy and hyponymy being the most frequent relations from those that could be identified in this resource. As for the morpho-syntactic information, they found that almost half of the responses elicited were nouns but that this proportion varied according to the semantic class of the stimulus verb (e.g. creation verbs such as *bake* received more proportion of noun responses than aspectual verbs such as *stop*). Additionally less than 30% of the noun responses were fillers of syntactic functions associated with the stimuli verbs. In relation to co-occurrence information, taking into account all responses, 77% of them were found to co-occur at least once with the stimulus verb in a window of 20 words to the left or to the right to the target stimulus verb. For nouns that were not syntactic slot fillers in the grammar model, 70% of them could be found in the co-occurrence window. This could be seen as further evidence pointing towards the need of a more general, scene-like context that goes beyond subcategorization in order be able to account for the responses in a WA task.

Moving now to the methodologies followed in word association experiments, it has to be mentioned that the traditional strategy when studying WA consisted determining the associative strength between the response and the stimulus by taking into account the frequency of association between responses and stimuli. The notion of associative strength was thus the primary tool to research structures and processes in the mind using WA until the 1960s, when the focus was moved to the response distribution. Consequently, the relevant issue ceased to be the connections between stimulus and response, and changed to the overlap between common associations (De Deyne and Storms, 2015). This last approach is the one that we follow in the experiment that we present next.

4.1.1 *Collecting word associations*

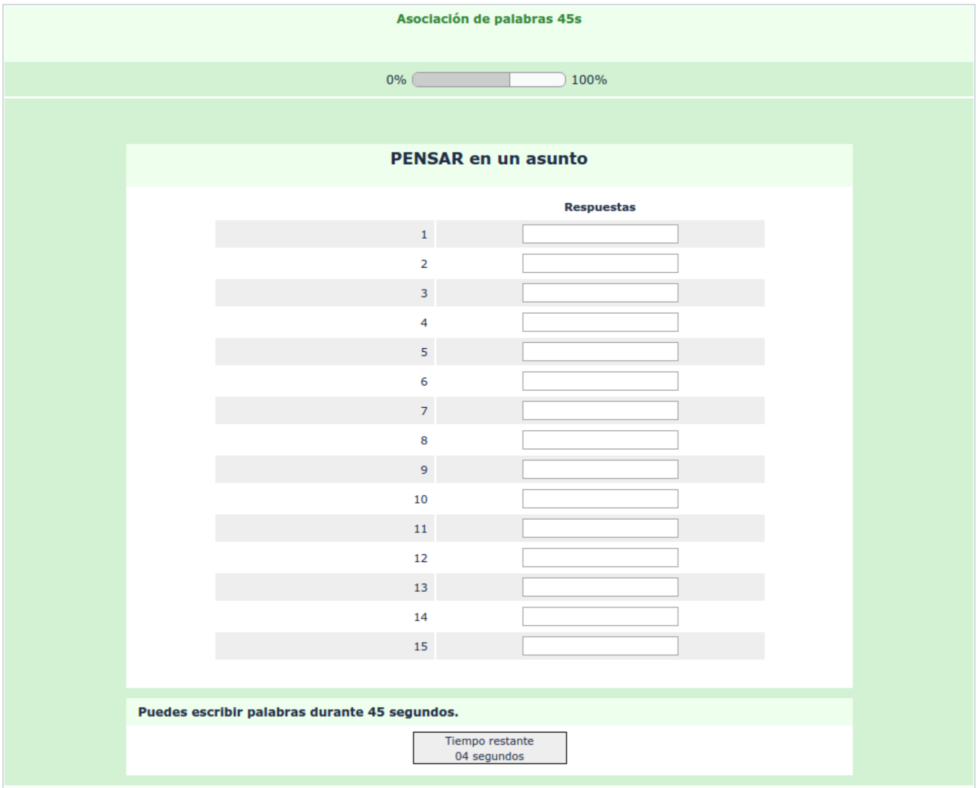
In order to characterize verb senses from a psycholinguistic perspective, we use the output of a word association task. This task is designed so as to enable us to obtain responses associated to each verb sense. These responses will be the features used to characterize the verb senses. This methodology follows a distributional assumption that entails that a large overlap of responses for two verb senses indicates that these senses are similar. The converse also holds: the lack of overlap of responses for two given senses points towards the lack of similarity between them.

The stimuli used in order to gather word association responses were the 20 verb senses presented already. These senses were embedded in a phrase that provides context (*ABRIR una ventana*, 'to open a window'). The aim of this type of contextualized presentation is to ensure that the collected associate responses were related to the specific senses that are the object of this study. Therefore, for each of the selected senses, we created a contextualized phrase stimulus that intended to achieve a balance between neutrality (keeping the responses from being biased due to the words in the phrase) and disambiguation (containing enough context to disambiguate the sense). To ensure the first requisite, we collected frequent selectional preferences for each of the senses and we looked for a generic alternative for the words in the set of selectional preferences (e.g. *person* vs. *thief* for *chase*; *means of transportation* vs. *train* for *travel*). To accomplish the second objective (disambiguation), we checked that none of the other senses with the same lemma in the corpus fitted in the sentence stimulus. The stimulus phrases contain the verb in infinitive, which is presented in the first position, in boldface and in capital letters. The other words in the phrase, no more than four, were presented in normal font. The list of stimuli can be found in appendix A in section A.3. The experiment was carried out using LimeSurvey, an open source tool that allows the researcher to administer it over the Internet and to save the data gradually as the experiment progresses. In addition, it supports timing the duration of the stimuli presentation to the participants.

In the first part of the experiment, we gathered information about the training of the participants in linguistics or related areas (e.g. translation) and the languages that the participants spoke as natives, besides Spanish. Moreover, participants were also asked to list the variety of Spanish that they spoke. In the second part, the participants were presented with the instructions. The presentation of the stimuli and the expected format of the answer (a single word) was explained. Participants were also encouraged to rely on their intuition. Each sentence stimulus was presented to the participants during 45 seconds, during which time they were asked to write up to 15 words that came to their mind. When the allotted time for each stimulus had passed, the program advanced automatically to the next stimulus. Each of these possible responses had a space allocated, as shown in figure 4.1. We chose to collect several responses to characterize each stimulus because Nelson et al. (2000) and De Deyne, Navarro, et al. (2013) suggest that these non-primary responses contain useful information and have proven to perform better in lexical access tasks (Balota et al., 1984). In addition to these advantages, from the point of view of similarity calculation between verb senses, having more than one response per stimulus and participant helps in reducing data sparsity (De Deyne, Navarro, et al., 2013).

A total of 102 native Spanish speakers collaborated in the experiment. Regarding the geographical variety, 85 were speakers from Spanish varieties from Spain and 17 participants listed themselves as speakers of Spanish varieties spoken outside of Spain (4 participants were from Chile, 4 from Argentina, 4 from Colombia, 2 from Venezuela, and the remaining 3 were from Peru, Uruguay and Ecuador). As for their linguistic training, out of the total amount of participants, 8 acknowledged a high level of formal linguistic training and 6 a high level of professional linguistic training (translators or teachers in secondary education). The total amount of responses collected was 11,617. On average, the participants responded with 5.7 words for each stimulus, this is, 113.9 words in total for the 20 stimuli, with 24 being the minimum number of words (a bit more than one word per stimulus) and 263 the maximum (around 13 words per stimulus). Tallying up all participants, the set of senses received between 454 and 730 responses. The responses given by the participants were looked up in the dictionary provided by Freeling (Padró et al., 2012)[1]. Those which were not present in the dictionary were disregarded in our experiments. This was done in an effort to minimise the noise produced by incorrectly written or partial words. From the total amount of 11,617 words collected, 11,504 were found in the Freeling dictionary. We also used Freeling to lemmatize the obtained responses and to tag them automatically with their part of speech. In cases of ambiguity,

[1] This dictionary contains 555,000 forms corresponding to more than 76,000 lemma-PoS combinations



Asociación de palabras 45s

PENSAR en un asunto

Respuestas

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Puedes escribir palabras durante 45 segundos.

Figure 4.1: Experiment window

the Freeling tagger assigned the most frequent part of speech to the lemma. Figure 4.2 shows the frequency of the most common PoS categories per verb sense.

We see that most of the associated responses were nouns, closely followed by verbs. Adjectives and adverbs are less common, although for certain stimuli, such as *valer* and *parecer*, adjectives have a more prominent weight than for others. However, the difference of frequency between noun and verb responses is not as large as it is in other word association studies for verbs, such as in Schulte im Walde (2008).

After the collection and cleaning processes, our final dataset contains a set of the lemmatized responses together with their frequency for each verb sense. In table 4.1 we show a small example of what this dataset looks like, with target verb senses placed in rows and the responses in columns.

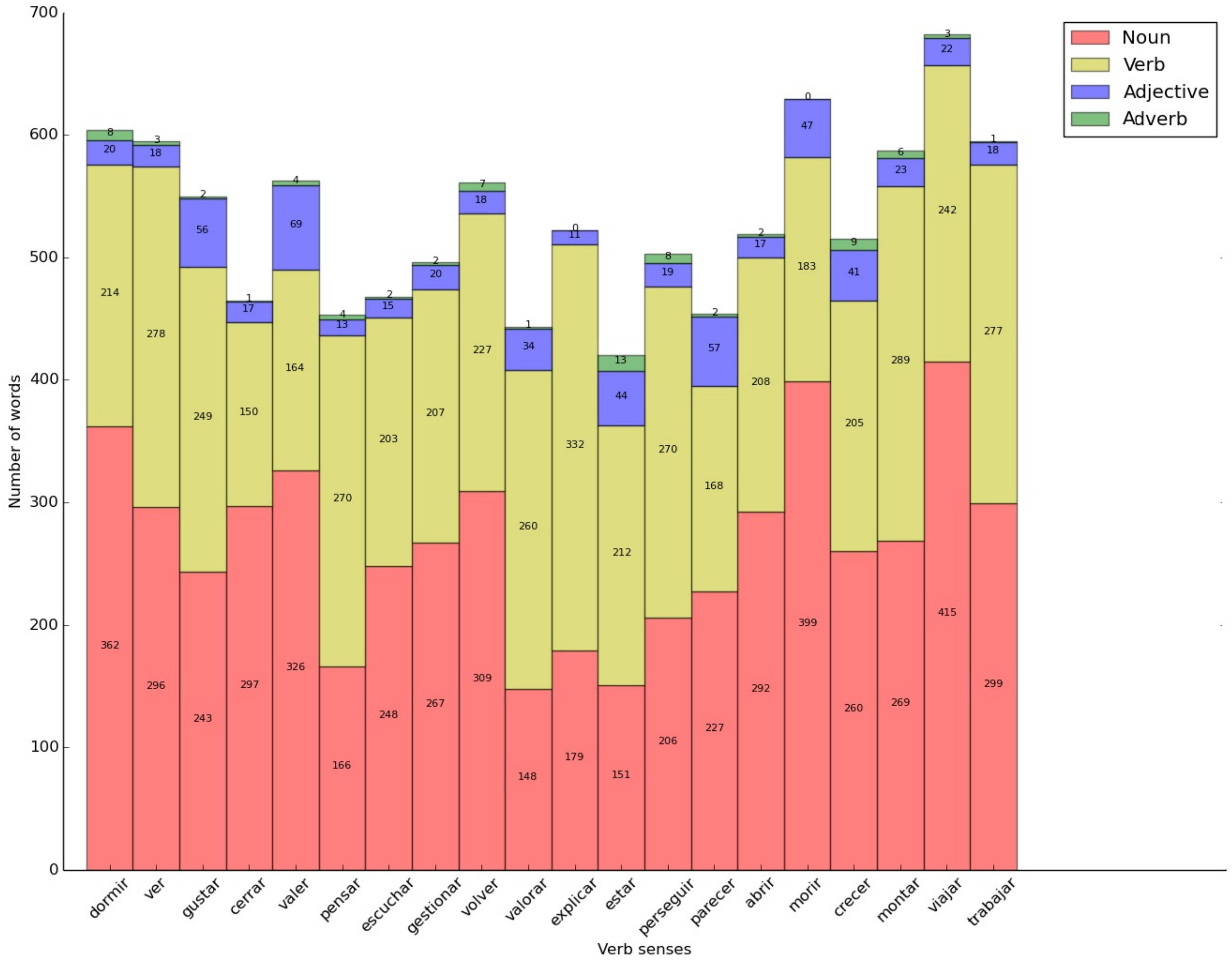


Figure 4.2: Frequency of the morphosyntactic category of the responses per verb

Verb sense	Entrar (enter)	Niño (boy)	Frío (cold)	Noche (night)
Abrir (open)	24	0	15	0
Cerrar (close)	1	0	0	5
Crecer (grow)	0	17	0	0

Table 4.1: Word association dataset example

4.1.2 *Extending word association data*

As we can see in the sample presented in table 4.1, the frequency matrix obtained is quite sparse: there are many responses and only a few of them co-occur with several stimuli. This fact affects the measurement, which is based on overlap, by skewing it towards dissimilarity. Thus, in order to alleviate this problem and also to experiment with different degrees of granularity and scope in the representation of the responses, we associated the responses with categories from several ontologies. To do this, we first annotated each lemmatized response with its corresponding WordNet synset (the identifier of a set of synonyms that map to a concept in WordNet) using Freeling, which assigned the most frequent synset to each lemma.

After that, the synset was used as a proxy to obtain the corresponding categories of the lemma in several ontologies. This was done through the MCR (Atserias et al., 2004), which provides mappings between synsets and such ontologies. Therefore, besides having the lemma of each response, we were able to use the equivalent categories from the following ontologies: hypernyms and supersenses from WordNet, and SUMO and TCO. A visualization of the process followed in order to link the responses obtained in the experiment to their corresponding lemmas and ontologies is shown in figure 4.3, exemplified with the response *saber* ('to know'). In table 4.2, the number of different categories used per resource is presented. It is important to note that some of the responses could not be represented in these ontologies: there were cases were their syntactic category (such as adverb) was not covered by some ontologies; in other cases there were responses whose synset was not mapped to one of the ontologies. In these events the responses not linked with a category in an ontology were not taken into account for that specific ontology.

Lemma	Hypernym	SUMO	TCO	Supersenses
2,691	1,275	595	328	40

Table 4.2: Number of different categories used

4.1.3 *Results*

As a result of the process presented in the previous section, we obtained, for each verb sense, a frequency distribution over all the responses in their several variations (lemmas, hypernyms, SUMO categories, TCO categories and supersenses). This entails that verb senses are characterized in five different ways, using five different formalizations of the psycholinguistic data. These formalizations, represented by five different co-occurrences matrices, are used independently to calculate the similarity coefficient between the verb senses of each pair (pairwise similarities). These coefficients are obtained by measuring the cosine similarity between the frequency distributions associated to each of the verb senses that compose the pair. This process is repeated for each of the five formalizations,

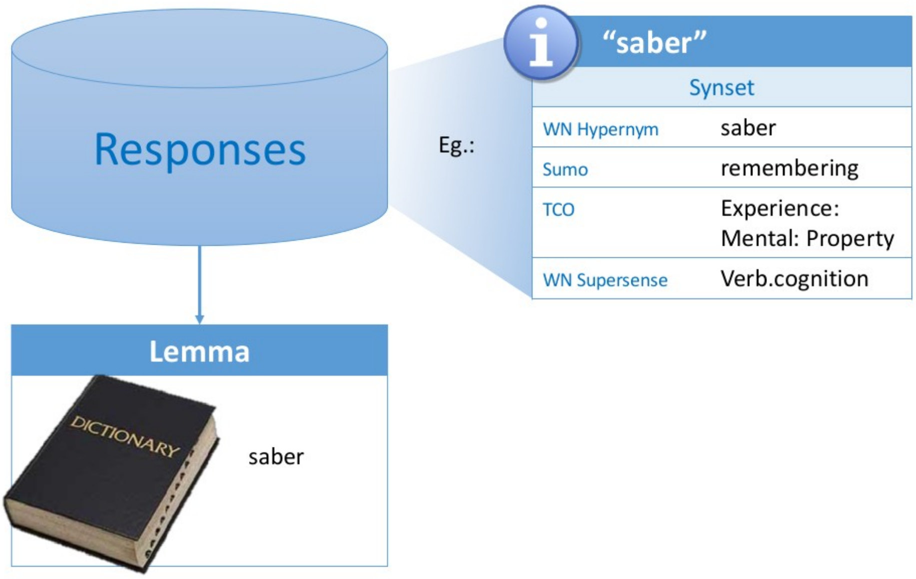


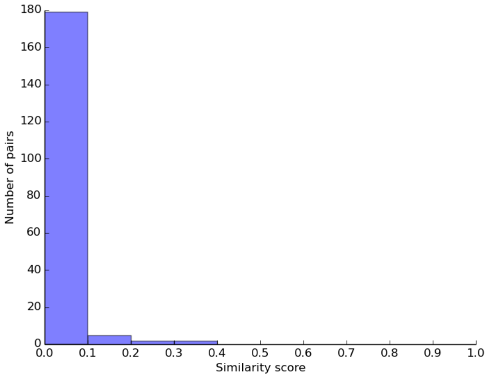
Figure 4.3: Obtaining categories from ontologies

as each one entails different frequency distributions. The similarity coefficients obtained range from 0 (not similar at all) to 1 (identical).

After obtaining the pairwise similarities for the verb senses according to each of the formalizations, we generated separated similarity rankings by ordering the pairs from more to less similar. In figures 4.4 a summary of these rankings containing the verb sense pairs and their corresponding similarity coefficients is displayed. These tables are accompanied by histograms in which the frequencies of the similarity coefficients for each formalization is displayed.

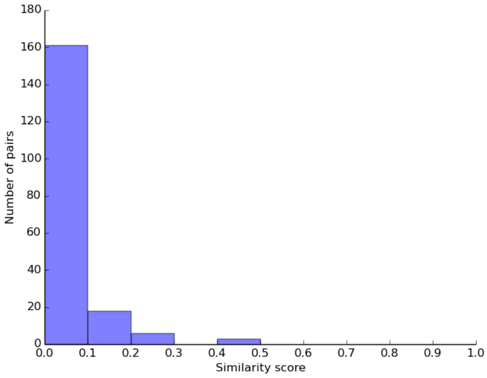
There are some noticeable differences in the distribution of similarity coefficients depending on the resource used to categorize the responses. First, there are two formalizations, lemmas and the hypernyms, that clearly skew the distribution towards dissimilarity. Using them to calculate the similarity between verb senses has the consequence of obtaining similarity coefficients that range between 0 and 0.1 for the overwhelming majority of pairs (around 90%). This is due to the low overlap of lemmas and hypernyms associated to the responses obtained in the WA task. Secondly, the formalizations that use SUMO and TCO categories produce a distribution that is also skewed towards the dissimilarity but with a much more gradual slope. The majority of the similarity coefficients fall within the interval 0.1-0.3, but we find some values in higher intervals, such as 0.6-0.7. Thirdly, using Supersense categories yields a distribution that closer to a flattened normal distribution than those obtained in the previous formalizations, with the majority of the pairs with values that range between 0.3 and 0.8, and no more than 40 pairs within the same interval.

Thus, we can conclude that each formalization of psycholinguistic data yields different distributions of similarity coefficients. These differences are not only given by the amount of categories used in each ontology, since using twice as many categories do not always double the amount of dissimilarity (see the case of lemmas and hypernyms), but also by the design and organization of each ontology.



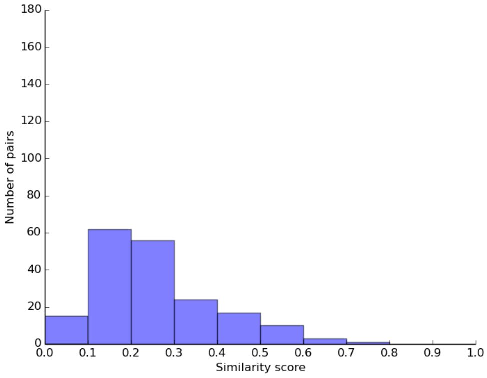
Verbs	Similarity score
abrir_cerrar	0,342
montar_viajar	0.332
gestionar_trabajar	0.265
...	...
dormir_escuchar	0
cerrar_explicar	0
cerrar_parecer	0

(a) Lemmas



Verbs	Similarity score
abrir_cerrar	0,7
gestionar_trabajar	0,7
montar_viajar	0,7
...	...
montar_pensar	0.002
dormir_valorar	0.002
cerrar_explicar	0.001

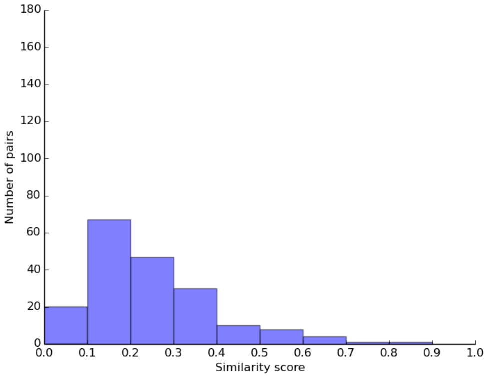
(b) Hypernyms



Verbs	Similarity score
gestionar_trabajar	0,7
gustar_pensar	0,602
pensar_trabajar	0,635
...	...
morir_ver	0,047
dormir_montar	0,046
valer_ver	0,045

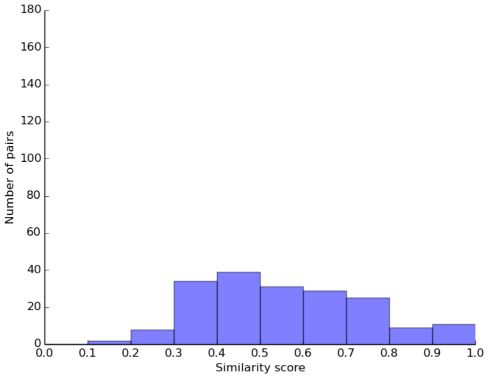
(c) SUMO categories

Figure 4.4: Histograms and similarity coefficients for verb pairs according to WA data



Verbs	Similarity score
pensar_valorar	0,85
gestionar_trabajar	0,733
montar_viajar	0,655
...	...
valer_ver	0.048
dormir_pensar	0.048
dormir_parecer	0.041

(d) TCO categories



Verbs	Similarity score
montar_viajar	0,976
pensar_valorar	0,963
viajar_volver	0,953
...	...
explicar_valer	0.213
pensar_valer	0.192
dormir_explicar	0.176

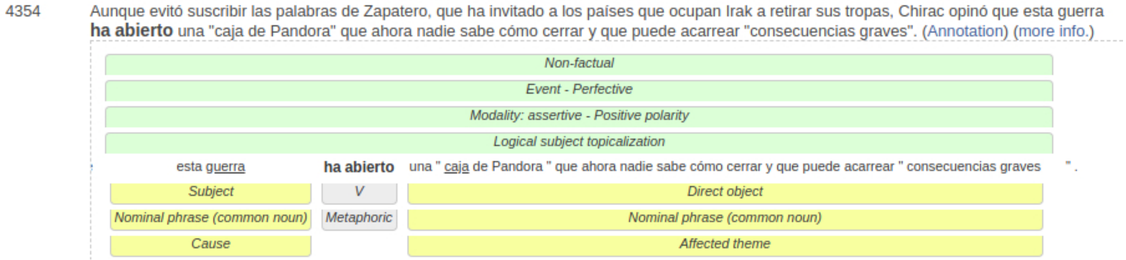
(e) Supersenses

Figure 4.4: Histograms and similarity coefficients for verb pairs according to WA data (cont.)

4.2 PERSPECTIVE FROM CORPUS: SEMANTIC ROLES

In this section we detail how we approach verb similarity from the corpus perspective. To do so, we use the data provided by the SenSem corpus to capture and formalize information related to the argument structure of the verbs. Each of the verb senses is associated to a number of sentences that are labeled with semantic and syntactic information as detailed in section 3.2. The features used to characterize the verb senses describe the arguments that combine with those senses. Regarding the distinction between argument and adjunts in SenSen corpus, the criteria is specified in Fernández-Montraveta et al. (2007), where arguments are defined as those verb complements that have subject, attribute and object positions or those that are kept when the verb is converted into a noun or a participle.

Thus, the feature set that describes the verb senses is based on the semantic roles associated to the arguments present in the corpus SenSem. Specifically, we use the three-level hierarchy of semantic roles explained in section 3.2 (see figure 3.4). This hierarchy allows us to experiment with different degrees of abstraction in terms of the information associated to the arguments. These different degrees of abstraction in turn generate different feature spaces for the representation of the argumental information associated to the verbs. We can see an example of this in figures 4.5 and 4.6, where two different sentences and their associated semantic roles obtained using each of the levels of the hierarchy are displayed.



4354 Aunque evitó suscribir las palabras de Zapatero, que ha invitado a los países que ocupan Irak a retirar sus tropas, Chirac opinó que esta guerra **ha abierto** una "caja de Pandora" que ahora nadie sabe cómo cerrar y que puede acarrear "consecuencias graves". (Annotation) (more info.)

Non-factual

Event - Perfective

Modality: assertive - Positive polarity

Logical subject topicalization

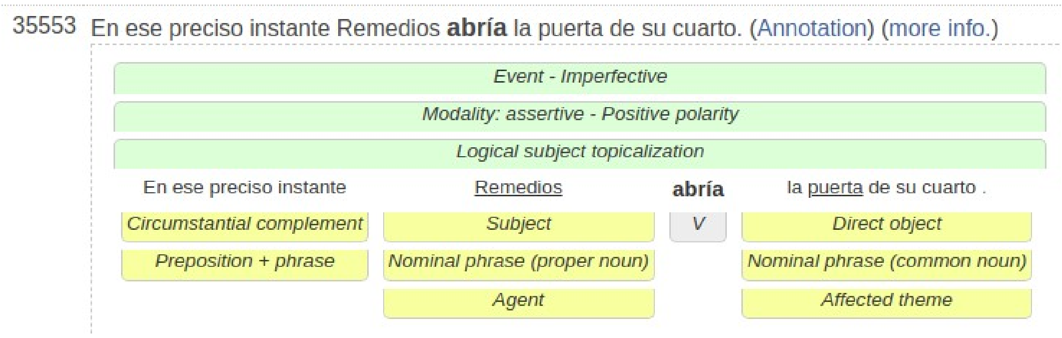
esta guerra	**ha abierto**	una " caja de Pandora " que ahora nadie sabe cómo cerrar y que puede acarrear " consecuencias graves " .
Subject	V	Direct object
Nominal phrase (common noun)	Metaphoric	Nominal phrase (common noun)
Cause		Affected theme

Coarse-grained level	Intermediate level	Fine-grained level
Actor, Undergoer	Cause, Theme	Cause, Affected theme

Figure 4.5: Sentence 1 and its semantic roles

Sentences shown in figures 4.5 and 4.6 display a variation in the semantic roles associated to the sentence when we move across the different levels of the hierarchy: both sentences have different semantic roles associated when taking into account the intermediate and fine-grained levels (for those levels, the sentence in example 4.5 has *cause* and *theme/affected theme* whereas the sentence in example 4.6 has *agent* and *theme/affected theme*), but the same semantic roles when taking into account the more coarse-grained level (both sentences in 4.5 and 4.6 have *actor* and *undergoer*). Therefore, by moving across the levels of the hierarchy when characterising argument structure, we can examine the effects that the different degrees of semantic abstraction have in the computation of similarity.

Besides using these three different levels from the hierarchy of semantic roles, we also use the syntactic functions of the arguments in order to investigate the results of adding explicit syntactic information when capturing argument structure. Therefore, we use two types of features independently to characterize verb senses: 1) semantic roles and 2) semantic roles combined with the corresponding syntactic function of the argument. In addition, the information extracted from Sensem about semantic roles and



35553 En ese preciso instante Remedios **abría** la puerta de su cuarto. (Annotation) (more info.)

Event - Imperfective

Modality: assertive - Positive polarity

Logical subject topicalization

En ese preciso instante	Remedios	**abría**	la puerta de su cuarto .
Circumstantial complement	*Subject*	*V*	*Direct object*
Preposition + phrase	*Nominal phrase (proper noun)*		*Nominal phrase (common noun)*
	Agent		*Affected theme*

Coarse-grained level	Intermediate level	Fine-grained level
Actor, Undergoer	Agent, Theme	Agent, Affected theme

Figure 4.6: Sentence 2 and its semantic roles

semantic roles and syntax was considered in two formats: **single arguments** (e.g. *agent* or *patient* if we just take into account semantic roles; *agent-subject* or *patient-object* if we take into account semantic roles plus syntactic information) and **complete structures**, which are the sum of the arguments of a sentence (e.g. *agent+patient* if we just take into account semantic roles; *agent-subject+patient-object* if we take into account semantic roles plus syntactic information). We illustrate this feature set creation and representation process with an example in table 4.4 based on the annotated sentence in figure 4.6. The rows of the table specify the three levels of abstraction of semantic roles that we are taking into account and the columns detail all the possible combinations taking into account the *single arguments/complete structures* and the *semantic roles/semantic roles plus syntactic functions* options. The resulting features of the combinations of these options are specified in italic font. When there are two features we separate them with semicolons ('Af_theme'stands for *Affected theme* and 'DObj'stands for *Direct object*)

	S. Arguments		C. Structures	
	Roles	Roles + Syntax	Roles	Roles + Syntax
Coarse-grained	*Actor*; *Undergoer*	*Actor-Subject*; *Undergoer-DObj*	*Actor+Undergoer*	*Actor-Subject+Undergoer-DObj*
Intermed.	*Agent*; *Theme*	*Agent-Subject*; *Theme-DObj*	*Agent+Theme*	*Agent-Subject+Theme-DObj*
Fine-grained	*Agent*; *Af_theme*	*Agent-Subject*; *Af_theme-DObj*	*Agent+Af_theme*	*Agent-Subject+Af_theme-DObj*

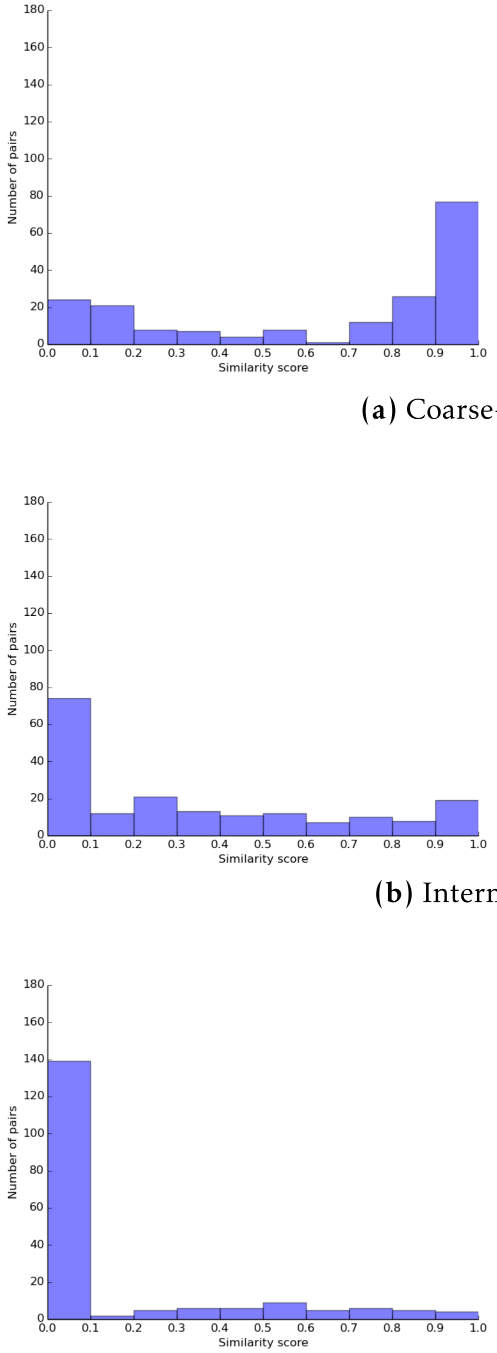
Table 4.4: Example of features obtained from the corpus

Summarizing, each verb sense is characterized by its frequency distribution over 12 different feature sets, which are obtained by taking into account different ways in which argument structure can be captured according to the three mentioned factors: the presence or absence of explicit syntactic information, the formalization in complete structures or single arguments and the different levels of granularity of semantic roles. The frequency distributions are obtained using the information from the annotated sentences associated to each of the verb senses in the corpus.

4.2.1 *Results*

Obtaining pairwise similarities for all senses follows essentially the same procedure described in section 4.1.3: the cosine similarity of any two given verb senses is calculated on the basis of the their co-ocurrence with the features from the above defined feature sets. This process is repeated for all the different formalizations (12 in the case of the corpus perspective). As in the previous case, the coefficients obtained range from 0 (not similar at all) to 1 (identical).

Recapitulating, this setting allows us to obtain similarity coefficients between all verb senses according to 12 different formalizations of argument structure. As a result, we generated 12 similarity rankings from corpus data, one for each formalization. In these rankings, sense pairs are ordered from most to least similar, according to the coefficients obtained in the previous step. In figure 4.7 there are displayed three of these similarity rankings corresponding to the three levels of granularity with *single arguments* and *only semantic roles* as parameters. Additionally, each of them is associated to a histogram that illustrates the frequency of the similarity coefficients. The granularity of semantic roles is one of the main sources of variation in the coefficients obtained, therefore these three rankings and their associated histograms are representative of the differences between the different formalizations.



Verbs	Similarity score
pensar_ver	1
estar_parecer	1
crecer_valer	1
...	...
montar_parecer	0
montar_ver	0
trabajar_valer	0

(a) Coarse-grained level

Verbs	Similarity score
gestionar_perseguir	0,99
parecer_valer	0,99
explicar_gestionar	0,99
...	...
morir_valer	0
cerrar_parecer	0
trabajar_valer	0

(b) Intermediate level

Verbs	Similarity score
abrir_cerrar	0,93
montar_volver	0,93
abrir_crecer	0,85
...	...
cerrar_parecer	0,047
trabajar_valer	0,046
abrir_perseguir	0,045

(c) Fine-grained level

Figure 4.7: Histograms and similarity coefficients for verb pairs according to corpus data

The picture that emerges from this panorama reveals that also verb similarities obtained from corpus data are quite sensitive to the specific formalization chosen. Formalizations that contain fine-grained roles show similarity coefficient distributions heavily skewed towards dissimilarity, with the overwhelming majority of the pairs obtaining a similarity coefficient between 0 and 0.1. Formalizations that contain intermediate roles follow a similar tendency, although more smoothed, particularly in the case of the usage of the formalizations created taking into account single arguments. Finally, formalizations that include coarse-grained roles show the less skewed distributions towards dissimilar: coefficients that indicate high similarity have the predominance. Nevertheless, the usage of complete structures and syntactic information (the other two parameters that were not used for these figures) yields a distribution where the majority of the pairs are dissimilar, creating coefficient distributions that resemble those in b) and c) in figure 4.7.

4.3 PERSPECTIVE FROM A THEORETICAL APPROACH TO LANGUAGE: CONSTRUCTIONS

The final characterization of verb senses and their argument structure comes from linguistic theory and it is based on the notion of *construction*. In this section we present the general framework of this theoretical approach and the details of the strategy that we follow to define and calculate verb similarity.

The idea that syntax and semantics are intimately related lies at the heart of approaches as diverse as construction grammar (Goldberg, 1995) and lexicalist theories of meaning (Levin, 1993). It is precisely this last framework the one that has been typically used to define groups of similar verbs. We have already mentioned that the basic idea in Levin's approach is that the meaning of a verb determines its behaviour. In this approach, verb behaviour is materialized as diathesis alternations, which are alternates in the expression of the arguments that are sometimes accompanied by certain changes in the meaning of the sentence that do not imply changes in the meaning of the verb. Therefore, diathesis alternations are generally regarded as different ways of packaging information. The typical example of diathesis alternation is the causative/incoative alternation (examples 5 and 6):

(5) Joanna broke the glass.

(6) The glass broke.

However, from a operational perspective, diathesis alternations present some difficulties when being used to characterize verbs in a systematic way that we detail in what follows. The first type of difficulty is on the productivity level: there is not always a structure which is in a meaningful opposition with another one, according to the definition of diathesis alternation given above. For example, the work of Levin itself contains a number of single structures that do not enter any alternation but are nevertheless used to define classes (e.g. x's way). Something similar can be found for Spanish (Márquez, 1991: 501). Besides, there are diatheses and structures that apply to a small number of verbs (e.g. obligatory passive for *reincarnate*, *rumor* and *repute* in Levin's). Another type of difficulty lies in the definition of the opposition that connects the alternating structures. There are different criteria used to define what are the elements that should be taken into account in order to relate two structures in such a way that they exhibit a meaningful opposition. If we focus on the work done for Spanish, we can see some examples of this situation in the variety of criteria used to identify the principles that support the opposition (García-Miguel, Costas, et al., 2005; Vázquez et al., 2000; Cifuentes Honrubia, 2006 and Castellón et al., 1997, among others). Finally, it is important to consider some a priori limitations of the diathesis approach. Goldberg (2002) points out that single structures, as opposed to those combined in diathetic pairs, have more generalization power for linguistic analysis since they do not place initial restrictions on the types of combinations and oppositions that can occur. Therefore, for Goldberg, the discriminative capacity lies in the structures considered individually and not in the alternations.

We consider that these objections are relevant in the light of the design of our study: since we have a limited set of verb senses, we need a way to characterize them that is as general as possible and does not leave some senses undescribed. Similarly, and related to this, since we aim at developing a systematic analysis, we think that not having a predefined set of oppositions can help discover relations between structures that are not limited in number or type. For these reasons, our work in the theoretical approach to language rests on the notion of *construction*. According to Goldberg (1995), a construction is a linguistic sign, with form and meaning, not compositional, that combines syntactic structure and semantic information and is the basic unit of the grammar. Constructions, as linguistic units, can have different degrees of complexity, ranging from morphemes to more complex elements such as argument structures. Goldberg assumes that the last corresponds to the codification of basic experiences of speakers (transfer, movement and caused movement among others). It is this level of complexity the one that we consider suitable for our verb similarity study, since it is directly related with the level of the linguistic expression of events that we are exploring in this work.

As we have already seen, the semantic weight that supports the interpretation of the structures is placed on the opposition in the diathesis alternation framework. This semantic content is not lost in the construction approach. Rather, it is incorporated in the construction itself. More in detail, the construction grammar approach assumes that a verb and a construction can be combined when there is compatibility between the meanings of both elements (Goldberg, 1995). More specifically, each argumental construction has an abstract meaning that also defines a set of associated semantic roles. Likewise, verbs have a lexical meaning and a conceptual framework with a series of associated participants. The compatibility between verb and construction is then defined taking into account the compatibility between verb participants and construction roles.

Our proposal to determine verb sense similarity within this perspective is based on the ability of each verb sense to instantiate a series of constructions. In this way, it will be determined which verbs and constructions can be combined and this data will be collected and treated in a fashion similar to what we have already done for the other perspectives. The similarity between verb senses will depend on the number of shared constructions. Therefore, there are not a priori restrictions on the combinations of constructions. With respect to the selection of constructions for our study, we chose general constructions described in multiple grammars (e.g. Bosque et al., 1999): transitive, intransitive, ditransitive, predicative and attributive. Besides, we have adapted the proposals of Levin and Goldberg taking into account work from Cifuentes Honrubia (2006), Martí et al. (2002) and Vázquez et al. (2000) for Spanish. Given that the amount of verb senses is limited, we have prioritized constructions that have a more general character over those more verb specific. The constructions selected are detailed next.

1. Prototypical causative: it is a construction that conveys a complex event in which the cause of the event is made explicit through different linguistic resources, although in this work we limit ourselves to those in which the cause is expressed through a subject. This subject might be volitional (agent) or not (cause). The entity that has the syntactic status of object is generally modified or affected in different degrees.

    (7) La falta de lluvias secó el río. ('The lack of rain dried up the river'.)

2. Prototypical anticausative (with *se*): it is an intransitive construction where the affected entity occupies the syntactic position of object and the agent or cause is typically omitted.

    (8) El río se secó ('The river dried up'.)

3. Periphrastic causative: it is a causative structure in which the components of cause and result are expressed separately: the predicate that expresses the notion of causativity is the auxiliary *hacer* ('to make'), that appears in infinitive, and there is a second predicate that expresses the event and also the resulting action or state (*trembling* in example 11). The entity that occupies the position of subject is usually the cause. The subject of the intransitive construction in example 9 can not be the object of the transitive (example 10), but it can be the object in the periphrastic causative (example 11). We exclude from this typology the structures that are built with an indirect cause that augments the number of participants in the event (as in example 12).

    (9) La niña tembló. ('The girl trembled')

    (10) *La película tembló a la niña. ('*The film trembled the girl'

    (11) La película hizo temblar a la niña. ('The film caused the girl to tremble')

(12) Pedro hizo a los alumnos repetir el examen ('Pedro made the students repeat the exam')

4. Anticausative without *se* (or process anticausative, Vázquez et al. (2000)): in this construction the constituent that expresses the cause is elided. An entity that is not clearly affected occupies the position of subject, although the degree of involvement of the entity may vary according to the verb (example 13). The difference with the prototypical anticausative (example 15) is that in the anticausative without *se* it is not possible a state as a result of the process defined by the verb (example 14 versus example 16).

   (13) Las temperaturas han bajado ('The temperatures have dropped')

   (14) *Las temperaturas están bajadas ('The temperatures are lowered')

   (15) El río se secó ('The river dried')

   (16) El río está seco 'The river is dried')

5. Middle construction: it expresses an state or property of the subject without needing to combine with an attributive verb. It is generally associated with an adverbial complement that reinforces the stative or timeless reading. This reading can be also emphasized by modal verbs. It is different from the prototypical anticausative, which has a dynamic interpretation.

   (17) La pintura se esparce con facilidad ('Paint sprays easily')

6. Impersonal pronominal: this construction has no grammatical subject, either explicit or available through context. The verb typically appears in the 3rd person.

   (18) Se aconseja el uso del cinturón de seguridad. 'The use of safety belt is advised'.

7. Oblique subject: in this construction the initiator of the event appears embedded in a prepositional phrase. It is usually subdivided in several types according to the meaning conveyed by the preposition and the object (e.g. origin, time, instrument). However, given the limited amount of verb senses that we are analysing, we have not taken into account these subdivisions.

   (19) La gente se beneficia de las nuevas medidas, ('People benefit from new measures').

8. Reflexive: in this construction the action carried out by the subject falls back onto itself. Some authors consider that this implies that the referent (syntactic subject) has the semantic role of agent and patient simultaneously.

   (20) Maria se peina. ('Maria combs her hair').

9. Reciprocal: this construction expresses simultaneous events. The subject of this construction is plural. Each of the components of the subject exerts an action over the others, and is in turn the object of the action of the others.

(21) Juan y Pedro se desafiaron ('Juan and Pedro challenged each other').

10. Periphrastic passive: in this construction the object is in the topic position and the verb requires an auxiliary to carry the morphological features. In general, the agent of the action can be expressed through a prepositional phrase, although it is commonly omitted.

    (22) Los bizcochos fueron comidos por los niños. ('The biscuits were eaten by the children'.

11. Reflexive passive: this construction also a passive but it is built with the particle *se*. Typically the subject is proposed with respect to the particle *se* and it does not bear the role of the initiatior of the action.

    (23) Se pasaron los trabajos a ordenador ('The works were computerized').

12. Cognate object: in this construction a usually intransitive verb it is combined with an object with which it is etymologically related, or with which it shares some redundancy in terms of its meaning. Generally, the sentences in which this construction appears have tautological character.

    (24) Cantamos una canción ('We sang a song').

13. Resultative with *estar*: it is an stative construction that details the state that is the result of the event expressed by the verb. In other words, it expresses a property of an entity which is the result of the process undergone by that entity. The initiator of this process can be agentive or causative.

    (25) El pan está cortado. ('The bread is cut')

In order to calculate the similarity between verb senses within this perspective, the first step was to determine which constructions could be combined with each verb sense. To do that, three linguists searched in descriptive grammars, such as Bosque et al. (1999). In dubious cases or those not covered by the grammars, each linguist gathered examples for each verb sense and construction or marked the impossibility of finding any example. Starting from these data, and after discussion sessions in which the adequacy of the examples (or lack of thereof) was evaluated, a global agreement was reached. The outcome of this process is a matrix that registers co-occurrences of verb senses and constructions. Differently from the other perspectives, in which we gathered observed frequencies of co-occurrence, in this case we obtain binary data for each verb sense. These binary values capture whether a sense can be combined with a construction (1) or not (0). An example of this can be seen in table 4.5.

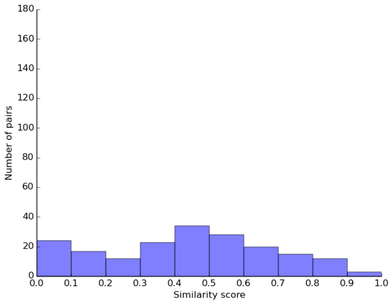
In order to determine the similarity between all senses, we calculate the pairwise similarity between them. As in previous perspectives, this entails measuring the similarity between the vectors that characterize the verb senses. However, in this case the vectors are binary and thus we use a metric that is adequate to this data, the Dice similarity.

Verb sense	Prototypical causative	Prototypical anticausative
Abrir (open)	1	1
Ver (see)	0	0

Table 4.5: Sample of the binary matrix

4.3.1 *Results*

As in the previous cases, the value of the obtained coefficients range between 0 (senses are totally different with respect to constructions they combine with) and 1 (senses are identical with respect to constructions they combine with). Verb sense pairs were ranked from more to less similar. In figure 4.8 we show a small sample of this ranking and the histogram that displays the frequency of the similarity coefficients.



Verbs	Similarity score
abrir_cerrar	1
trabajar_volver	1
escuchar_explicar	0,9
...	...
explicar_parecer	0
cerrar_parecer	0
trabajar_valer	0

Figure 4.8: Histograms and similarity coefficients for verb pairs according to construction data

We see in figure 4.8 that the similarity values are quite balanced across the intervals, with two peaks: one at the 0-0.1, which highlights the existence of a number of pairs with low similarity, and another more populated, at the interval 0.4-0.5, which points to a relatively high number of pairs that fall into a neutral zone: they are neither similar not dissimilar.

4.4 comparison of perspectives

In this section we present the comparison of the similarity coefficients obtained from corpus, constructions and psycholinguistic data. Firstly, we present a quantitative analysis of the relationship between these three approaches to verb similarity. Secondly, we carry out a qualitative analysis in which we look closely at the semantic, aspectual and subcategorization information that prevails in each type of data and analyse the coincidences and differences.

The quantitative study was performed through a Spearman correlation analysis. Spearman's correlation coefficient is a non parametric measure of correlation that quantifies the monotonic relationship between two variables. It is equivalent to the Pearson correlation between the ranked values of the two variables, and it is adequate in our case because, given the similarity coefficient distributions that we saw in figures 4.4-4.8, we can not suppose a normal distribution of the data and the relation between the perspectives of verb similarity may not be necessarily linear. The equation for the Spearman correlation coefficient (ρ) is shown in equation 4.1, where d represents the pairwise distances of the ranks of the variables and n is the number of samples.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{4.1}$$

We looked at the correlation between the similarity coefficients calculated using all the different formalizations for each of the three approaches that we have detailed in the previous sections. Summarizing, for corpus data we had three levels of granularity of semantic roles and two different structural combinations: complete structures and single arguments. In addition, we also experimented with the presence/absence of explicit syntactic functions. As for psycholinguistic data, we had five different formalizations for similarity: lemmas, hypernyms, SUMO categories, TCO categories and supersenses. Finally, for the theoretical approach, we had a collection of co-occurrences between verb senses and constructions.

4.4.1 *Correlations between corpus data and constructions*

In this section we display the correlations between the similarity coefficients obtained from corpus formalizations and those obtained from constructions data. Spearman's correlation coefficient for formalizations that contain explicit syntactic information besides semantic roles is shown in table 4.6, whereas correlation coefficients for formalizations that only contain semantic roles are shown in table 4.7. The coefficients shown are statistically significant ($p < 0.05$). Otherwise, we mark it with a hyphen.

Corpus data		Correlation with constructions
Fine-grained roles	S. Arguments	0.225
	C. Structures	0.3
Intermediate roles	S. Arguments	0.481
	C. Structures	0.528
Coarse-grained roles	S. Arguments	0.45
	C. Structures	0.582

Table 4.6: Features: semantic roles and syntactic functions

Corpus data		Correlation with constructions
Fine-grained roles	S. Arguments	-
	C. Structures	0.252
Intermediate roles	S. Arguments	0.261
	C. Structures	0.398
Coarse-grained roles	S. Arguments	-
	C. Structures	0.222

Table 4.7: Features: semantic roles

Overall, we see that correlations range between 0.222 and 0.582. The correlation strength is weak to medium and always positive. Besides these general considerations, there are several aspects of corpus data that may play a role in the correlation coefficients and that are analysed here:

1. Presence vs absence of syntactic functions in the features (table 4.6 vs table 4.7): the presence of syntax in corpus formalizations increases the correlation of the similarity coefficients from corpus data with similarity coefficients calculated from theoretical constructions.
2. Unit of argumental information (single arguments vs. complete structures): formalizations that contain complete structures tend to show higher correlations with constructions than those that contain single arguments, both for formalizations with or without syntax. This fact, taken together with the previously examined aspect, seems to suggest that the presence of structural information in corpora (either explicit syntax or constituent order captured through complete structures) is important in order to achieve similarity coefficients that are in line

with those obtained from constructions.

3. Semantic role granularity: data suggests that the intermediate and coarse-grained levels of semantic roles show higher correlations with constructions. More significant coefficients would help in reaffirming this tendency.

4.4.2 *Correlations of corpus data and word associations*

In this section we present correlations between corpus data formalizations and word associations (WA) formalizations from the psycholinguistics perspective. Table 4.8 displays the correlations between the corpus data that does not contain syntactic functions in its features (just semantic roles) and the WA data. Table 4.9 shows the correlations between the corpus data that contains syntactic functions besides semantic roles and the WA data.

		Word association data				
Corpus data		Lemma	Hypernym	SUMO	TCO	Supersense
Fine-grained roles	S. Arguments	-	-	-	-	-
	C. Structures	-	-	-	-	-
Intermediate roles	S. Arguments	-	0.235	0.262	0.278	0.153
	C. Structures	-	0,154	0,24	0.169	-
Coarse-grained roles	S. Arguments	-	-	-	-	-
	C. Structures	-	-	-	-	-

Table 4.8: Features: semantic roles

		Word association data				
Corpus data		Lemma	Hypernym	SUMO	TCO	Supersense
Fine-grained roles	S. Arguments	-	0.188	0.193	0.18	0.201
	C. Structures	-	-	-	-	-
Intermediate roles	S. Arguments	0.187	0.272	0.320	0.297	0.239
	C. Structures	-	0.160	0.236	0.208	-
Coarse-grained roles	S. Arguments	-	-	0.147	-	-
	C. Structures	-	-	-	-	-

Table 4.9: Features: semantic roles and syntactic functions

Overall, the correlation strength is low and weaker than the previous comparison, with the higher score being 0.32. Approximately half of the correlations are significant and all of them are positive. As previously, we examine several aspects of the correlation data that may be important in order to understand these results:

1. The presence/absence of syntax in corpus data (table 4.8 vs table 4.9): the highest correlation between corpus data and WA data (0.32) is obtained when using a corpus formalization that contains syntactic functions. In general, higher correlation scores are achieved when this information in corpus data. However, this tendency is not as strong and clear as in the previous case.
2. Unit of information (single arguments vs. complete structures): we see higher correlation values of corpus data with WA data when using single arguments regardless of the specific formalization within WA. However, more significant coefficients would help strengthen this tendency.
3. Semantic role granularity: Regarding the effects of varying the granularity of the elements that are being compared, we can not determine its effects on the correlation strength in this case since we do not have enough amount of significant data for comparison.
4. Type of category used for WA data: as in the previous aspects explored for this correlation, there are not many coefficients that would allow us to extract solid conclusions, but there is a tendency for WA data formalized using intermediate categories (hypernyms, SUMO categories and TCO categories, as opposed to specific categories such as lemmas or broad categories such as supersenses) to yield higher correlation values with corpus data formalized using intermediate semantic roles.

4.4.3 *Correlations constructions and word associations*

Finally, in this last section we explore the correlations between the similarity coefficients yielded by the different formalizations of WA data and those coefficients obtained using constructions to characterize the verb senses. We display in table 4.10 the correlation scores between these two approaches. The correlation strength values are slightly lower to those of corpus and WA, being the highest correlation value 0.22. As in the previous case, similarity coefficients obtained using data from constructions seem to correlate more with similarity coefficients from WA data when these are represented using ontologies that contain intermediate categories, although the evidence for this trend is not strong.

	Word association data				
	Lemma	Hypernym	SUMO	TCO	Supersenses
Constructions	0,187	-	0.197	0.225	0.155

Table 4.10: Construction correlations

4.4.4 *Synthesis of comparisons*

Overall, the highest single correlation coefficient is achieved when comparing the similarity coefficients obtained by using construction data and corpus data, when the last is formalized using coarse-grained roles plus syntactic functions using complete structures as unit (0.58). In general, correlation coefficients between these two perspectives are higher than between the other perspectives (corpus and WA, constructions and WA).

The presence of structural information in corpus data (in the form of syntax or in the form of complete structures, that gives information about the order or relations between the arguments) is relevant to obtain higher correlation coefficients with the other perspectives. However, whereas correlations between corpus and construction data require the presence of both syntactic categories and complete structures to reach its maximum, correlations between corpus and WA tend to be lower if complete structures are used to formalize corpus data. This fact suggests that information from complete structures is relevant in the corpus approach in order to establish relationships between verbs in a way that is similar to those that are established in data from constructions, but not to those that arise from psycholinguistic data. As for why this might be so for construction data, it can be argued that the definition of 'construction'implies an association with a set of semantic roles and that, since the formalization of corpus data that uses complete structures also contains information about the set of roles, there is a higher coincidence between the two, at least at the level of definition.

Besides, correlations between corpus and construction data seem to be sensitive to the granularity of the semantic information used: corpus data with coarse-grained and intermediate semantic roles tends to correlate more with constructions. Finally, although correlation coefficients between WA and constructions are low, we saw that there is a tendency for constructions to correlate more with intermediate WA formalizations.

The general panorama that arises is rich and complex: there is not a specific type of information that clearly and systematically varies correlation between the approaches. Besides, on some occasions, a larger amount of significant correlation coefficients would be desirable in order to confirm the partial tendencies that were found. Thus, to shed more light on this issue, in the next section we explore more in deep what lies at the heart of each approach by looking at the linguistic features associated to the verbs in the most similar and dissimilar pairs according to each perspective.

4.5 QUALITATIVE ANALYSIS

The objective of the qualitative analysis is to look more in depth into verb similarity taking into account basic linguistic features associated with verbs. More specifically, we look into semantic, aspectual and subcategorization features. In order to find out the particularities of each perspective regarding these three types of features, we compare their values for the most similar and dissimilar pairs of each perspective (the ones with the highest and lowest similarity coefficients respectively). These pairs are expected to show a rather systematic behaviour regarding the linguistic features associated to the verb senses: the verb senses that compose the similar pairs are expected to have common features whereas the members of dissimilar pairs are expected to show fewer or none common features.

In order to obtain the more and less similar pairs we follow the methodology detailed next. We already saw in the last section that for any given formalization we created a ranking of verb sense pairs, ordering them from most to least similar according to the coefficients obtained. Several instances of these rankings were shown in figures 4.4-4.8. In order to carry out the qualitative analysis, we selected a sample of the most similar and dissimilar pairs. To do so, we took the first and last ten pairs of each ordered ranking, which amounted to a total of 20 pairs out of 190 (around 10% of the ranked data). In the case of a tie (when several pairs obtained the same similarity score as the cut-off pair), we took all the pairs that had the same score. Finally, for each formalization, we went on to analyse the semantic, aspectual and subcategorization coherence of the pairs of these two selected groups. Specifically, we looked at semantic, aspectual and subcategorization features considering the following:

1. For the semantic analysis we took into account how many of the selected pairs contained verb senses that had the same semantic field category according to Adesse macro-classes (García-Miguel and Albertuz, 2005). Adesse is a resource that contains a lexical database and a corpus for Spanish verbs. It includes semantic and syntactic information for Spanish verbs. Additionally, each verb is associated to a class. These classes are in turn grouped under 12 macro-class labels, that were created following the types of processes definded by Halliday et al. (2004). Verb senses that are members of a similar pair are expected to belong to the same semantic field. Conversely, senses that are members of a dissimilar pair are expected to be associated with different semantic fields.
2. For the aspectual analysis we counted how many of the selected pairs contained verb senses that had the same broad aspectual category (static or dynamic). Verbs that are members of a similar pair are expected to have the same aspectual category. Conversely, verbs that are members of a dissimilar pair are expected to have different aspectual categories.
3. For the analysis of the subcategorization we counted how many subcategorization frames were shared between the members of the pairs. Verbs that are members of a similar pair are expected to share a high number of subcategorization frames. Conversely, verbs that are members of a dissimilar pair are expected to share a low number of subcategorization frames.

The objective of this analysis is to detect the linguistic features that could be considered markers of similarity relations between the verbs in our study for the several perspectives covered.

4.5.1 *Semantic analysis*

To provide an overview of the results obtained, we show in figure 4.9 the ratio of the sampled similar and dissimilar pairs whose verb senses share the semantic field, according to the Adesse categories. On the X-axis we enumerate the different formalizations that are being compared. The names of the formalizations that correspond to each perspective are represented with different colours on the X-axis (blue for corpus data, red for WA data and black for constructions). On the Y-axis we specify the percentage of pairs whose senses share the semantic field, ranging from 0 to 1. Similar pairs are represented with yellow bars and dissimilar pairs are represented with green bars.

In order to analyze these data we look at several aspects: first, the influence of the different factors taken into account when creating the corpus data formalizations (complete structures vs single arguments, syntax vs lack of thereof, and semantic role granularity); second, the different categories used in the formalization of WA data; and third, the overall differences between the three approaches.

In the case of the semantic field analysis, the most noticeable observation is that, in almost every case, there are more similar pairs that share the semantic field than dissimilar pairs, which corresponds to our expectations and indicates that these perspectives are, at least up to some point, semantically structured. However, there are some differences between the formalizations and perspectives worth to be noted: regarding the corpus formalizations, we see that there is a tendency for configurations that use complete structures to show a larger percentage of shared semantic category both for similar and dissimilar pairs than formalizations that use single arguments. Additionally, the incorporation of syntactic functions decreases the percentage of similar pairs with shared semantic categories, except in the case of the level that uses coarse-grained roles. Corpus formalizations that contain only coarse-grained roles as single arguments become more semantically driven when syntactic information is added. This is illustrated in figure 4.10, where we contrast the percentage of pairs that share the semantic field (Y axis) with the similarity coefficients (X axis) for the complete set of 190 pairs. We can see that there are noticeable differences between the graph in 4.10a, which has no structural information (coarse-grained roles in single arguments), and the graphs in 4.10b (coarse-grained roles plus syntactic information, also in single arguments). In 4.10a the pairs that share the semantic field are neither similar nor dissimilar (their coefficients range between 0.3 and 0.7) and the pairs which have a high similarity coefficient do not contain senses that have the same semantic field in the majority of the cases (around 70 %). [2]

As for the levels of abstraction in semantic roles, we see that when no syntactic information is added, the presence of fine-grained or intermediate roles increase the

[2] The differences between the highest percentages on the Y axis in figures 4.9 and 4.10 are due to the fact that in 4.9 we are only considering the most and least similar pairs according to their rank and therefore pairs with varying similarity coefficients may be conflated.

percentage of similar pairs with shared semantic category. As for the differences within WA formalizations, the more semantically coherent formalizations (those where the percentage of similar pairs with shared semantic fields is maximal whereas the percentage of dissimilar pairs with shared semantic fields is minimal) are lemmas and supersenses, the more and less specific categories respectively.

Overall, WA are the most semantically structured approach, followed by fine-grained and intermediate levels of semantic roles (without syntax) from the corpus approach. Differences between similar and dissimilar pairs in terms of shared semantic category are smaller for constructions and other corpus approaches. In fact, data from constructions and data from the corpus approach with coarse-grained roles and syntax in single arguments behave similarly.

4.5.2 *Aspect analysis*

As for the aspect, the results are shown in figure 4.11. As in figure 4.9, on the X-axis, the different formalizations taken into account are specified, and on the Y-axis, the percentage of pairs whose members have the same coarse aspect category, static or dynamic, is shown. Regarding the use of these two categories for aspect, it should be mentioned that there is a rich variety of accounts of aspect that differ in the number and types of aspectual categories. Therefore, we adapt the aspect labels assigned to the verb senses in the Sensem lexicon to contemplate only two general categories, as in D. R. Dowty (1979) and Jackendoff (1983): dynamic events and states. These are two broad categories that are proposed as essential by these two authors and which constitute a opposition kept in different aspectual categorization proposals.

Regarding the differences in aspectual coherence between the different formalizations, we can see in figure 4.11 that, in general, the percentage of senses with identical aspect is higher for similar pairs, except for the corpus formalization that includes coarse-grained roles organized in single arguments. If we look only at corpus data, focusing on the different levels of granularity in semantic roles, we can see that the number of dissimilar pairs with shared aspect decreases slightly when abstraction increases. WA data percentages are similar to those of corpus, although there is a slighly higher tendency for dissimilar pairs to share the same aspectual category. The greater difference between similar pairs and dissimilar pairs occurs within the constructions approach. Members of similar pairs share the aspectual category in all cases whereas only 12% of the most dissimilar pairs share the aspectual category. Therefore, the opposition between dynamic and static events is a powerful backbone of verb similarity as defined by constructions and differently from the other approaches to verb similarity. Regarding this difference, it is important to note that constructions have an inherent dynamic or static interpretation. Therefore, the aspect of the construction and the lexical aspect of the verb that it is combined with need to be made compatible. These results suggest that this compatibility has an important role in defining verb similarity according to this perspective. In figure 4.12 it is contrasted the evolution of the percentage of pairs with shared aspect (y axis) as the similarity increases (x axis) for all 190 pairs. On the left part, figure 4.12a displays this evolution for the formalization that shows the less coherent behaviour according to aspect (dissimilar pairs tend to share the aspectual category more frequently than similar pairs), which is coarse-grained roles organized in single arguments. On the right side, in figure 4.12b, it is shown the corresponding figure for constructions. As it can be observed, they show a quite opposed behaviour: while the percentage of pairs with coherent aspect augments as similarity increases in 4.12b, this percentage is high for pairs that have intermediate similarity and then decreases in figure 4.12a.

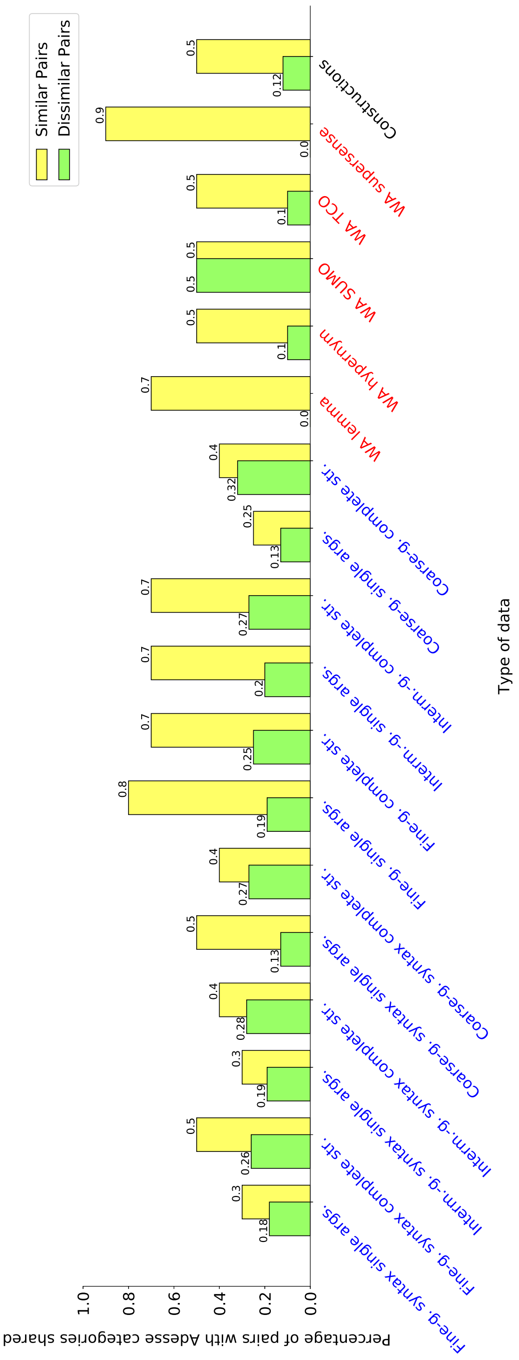
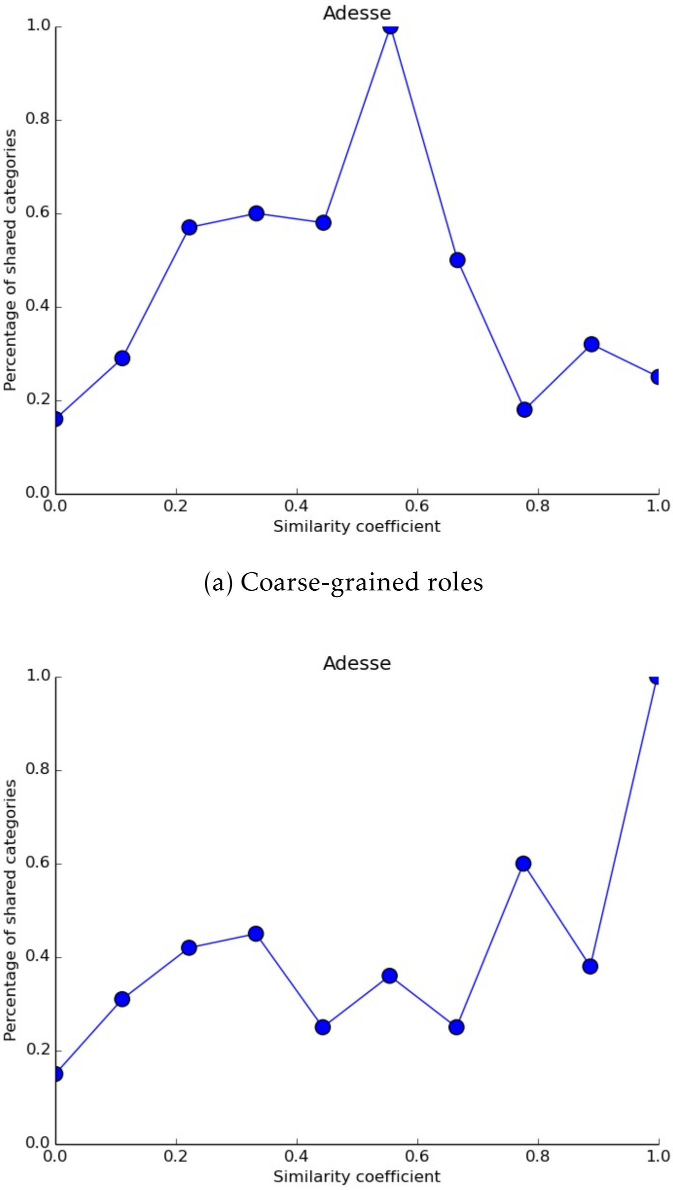


Figure 4.9: Semantic fields



(a) Coarse-grained roles

(b) Coarse-grained roles and syntax

Figure 4.10: Adesse categories for different formulations

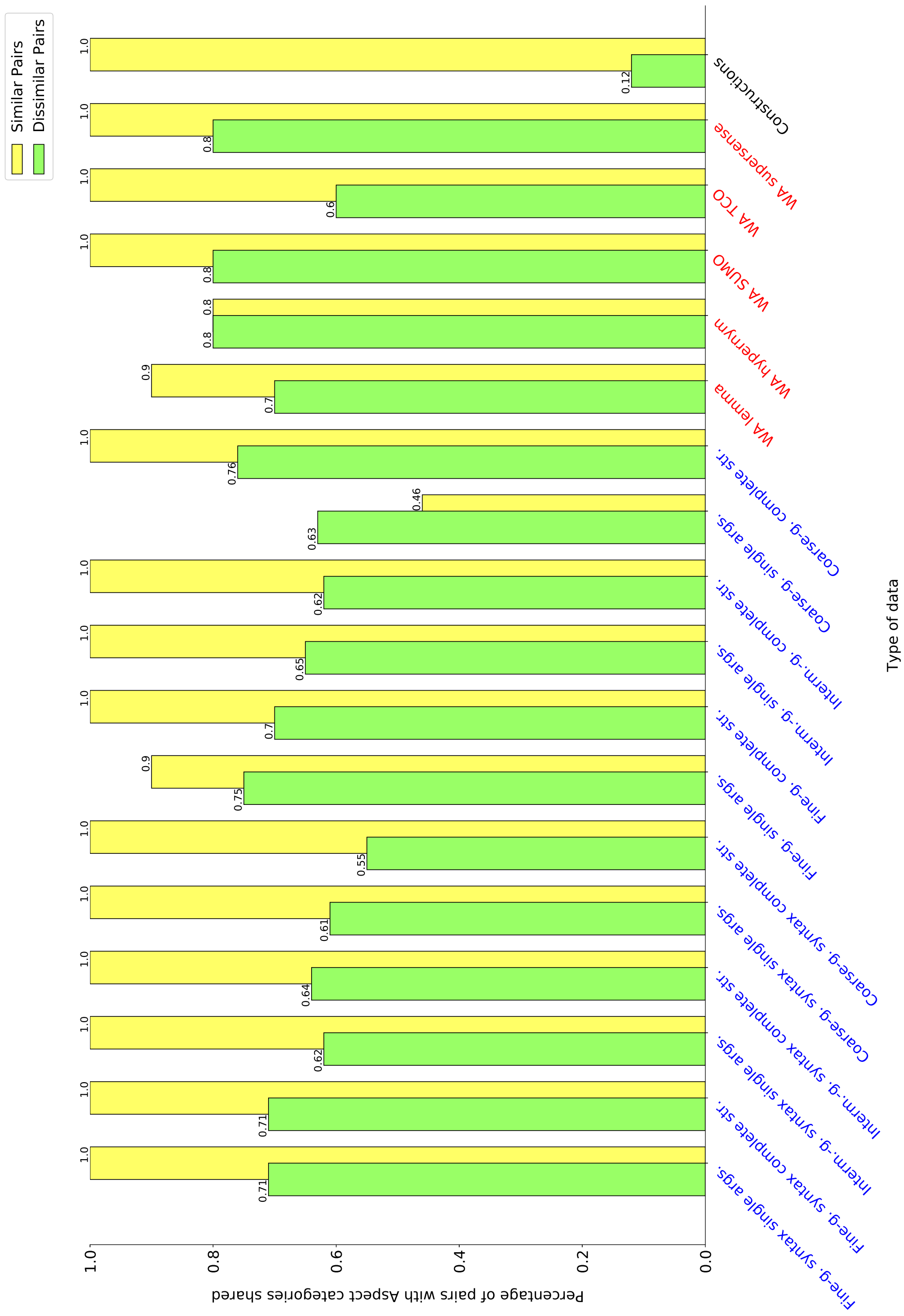
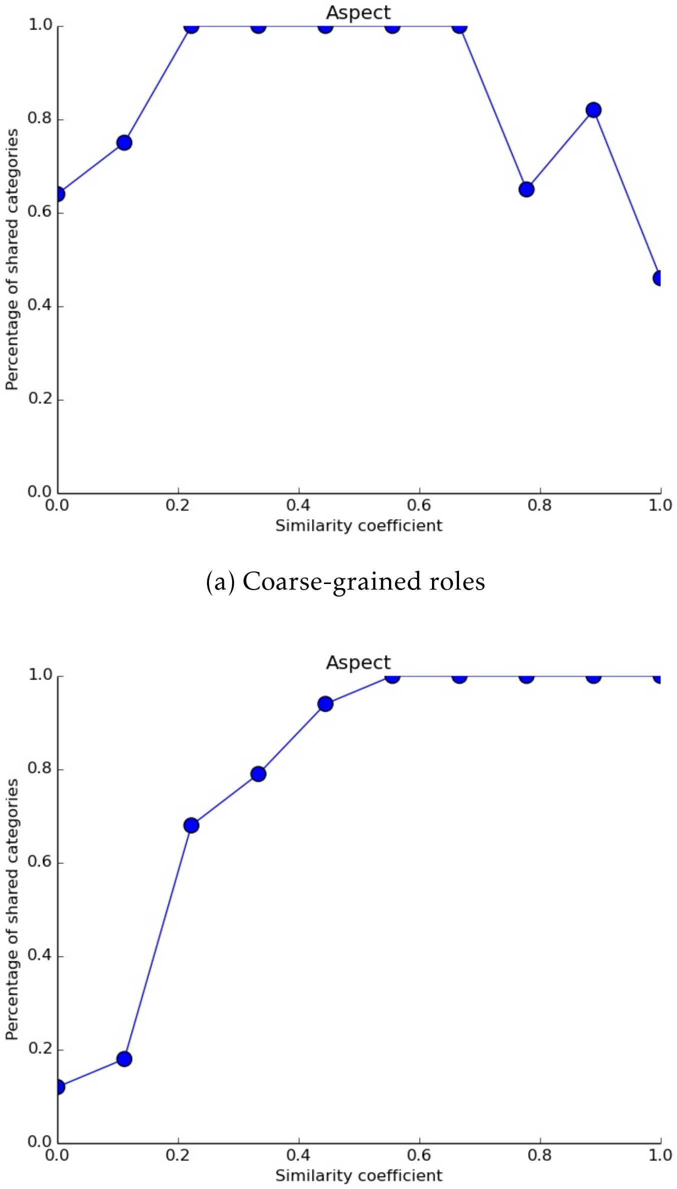


Figure 4.11: Aspectual information



(a) Coarse-grained roles

(b) Theoretical constructions

Figure 4.12: Aspectual information

4.5.3 *Subcategorization analysis*

Finally, in figures 4.13 and 4.14 we show the percentage of subcategorization frames that are shared by the members of the pairs. This constitutes a slight change of focus with respect to the previous two linguistic features analysed. Before we looked at whether the verb senses that formed the pairs shared certain features and we gathered the percentages of pairs where this was the case. By contrast, in this analysis we take into account the percentage of shared subcategorization, not of pairs sharing subcategorization, as it would be difficult to find verb senses that combine with exactly the same subcategorization frames when our sample contains a controlled and reduced number of verb senses. The information about the amount of shared subcategorization is further subdivided into two cases according to the frequency of the subcategorization frames. In figure (4.13) we present the differences between similar and dissimilar pairs when taking into account all subcategorization frames, while in figure (4.14) we only take into account the subcategorization frames that are less frequent, this is, those that are not *NP-V* or *NP-V-NP*. As in previous figures, on the X-axis there are the different formalizations of the three perspectives researched. On the Y-axis, the percentage of subcategorization frames shared is displayed. This percentage is below 40% in all cases.

If we take into account all the subcategorization frames (figure 4.13), we can see that there is a clear difference between WA data and the other two approaches. While in corpus and construction data, senses in similar pairs share more subcategorization frames than senses in dissimilar pairs, this situation is reversed for WA data, most notably when TCO categories and supersenses are used. This fact is along the lines of the findings of Schulte im Walde (2008), who suggested that WA go beyond subcategorization information. However, if we look at the less frequent subcategorization frames (figure 4.14), the former situation no longer holds: in almost all cases, for corpus, constructions and WA data, senses in similar pairs share more subcategorization frames than senses in dissimilar pairs. This suggests that verb sense similarities, as obtained from WA data, are sensitive only to less frequent subcategorization frames and may indicate that these types of frames, and not the frequent ones, are the ones that are relevant in shaping similarity relationships between verbs in a way that is coherent with the concepts associated to verbs in the minds of the speakers. In any case, their role is rather small, given the small percentage of subcategorization frames shared in general and the slight difference between similar and dissimilar pairs across all perspectives.

As an additional note, we can mention that this contrasts with most of the work done in automatic verb classification, which relies heavily on subcategorization frames to create verb classifications that contain semantically coherent classes, following Levin's insight (Schulte Im Walde, 2006; Sun, Korhonen, and Krymolowski, 2008). However, the verb pairs that share the semantic fields (according to Adesse categories) share up to 36% of the subcategorization frames at most. This fact, together with the small difference between similar and dissimilar pairs, suggests that subcategorization frames need to be supplemented with additional information in order to gain discriminative power.

In figures 4.15 and 4.16 we illustrate the behaviour of the three perspectives under the light of subcategorization for all 190 pairs. Each figure contains three plots, one per perspective. In the case of corpus and WA perspectives, we selected the formalizations that best represent the differences in the percentages of shared subcategorization when going from all subcategorization frames to low frequency subcategorization frames. Figure 4.15 corresponds to all subcategorization frames and figure 4.16 illustrates low frequency subcategorization frames

Figures 4.15a and 4.16a, that correspond to the corpus perspective, display a quite flat behaviour in the two scenarios considered. Figures 4.15b and 4.16b, that correspond to WA data, go from a decreasing tendency, in which less subcategorization is shared when the similarity increases, to a neutral one. As for figures 4.15c and 4.16c, which illustrate the behaviour of constructions, we see that they are the most coherent with respect to subcategorization (the percentage of shared subcategorization increases as the similarity increases) and that this tendency is maintained for all subcategorization and for the less frequent subcategorization.

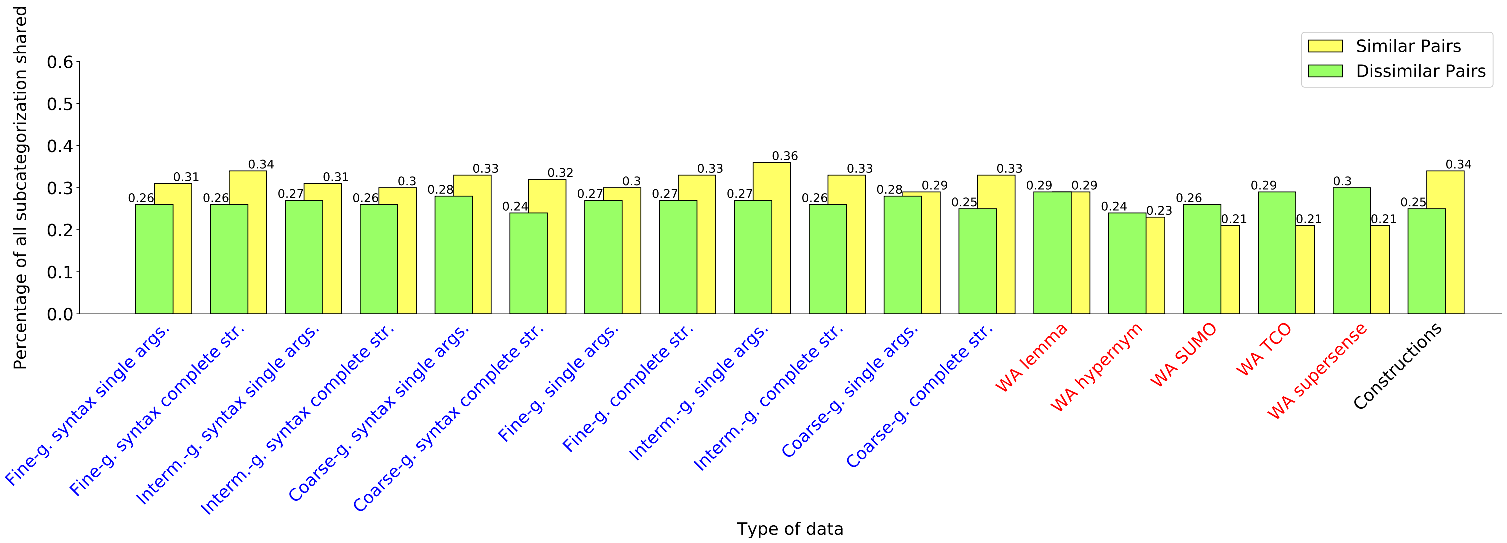


Figure 4.13: All subcategorization

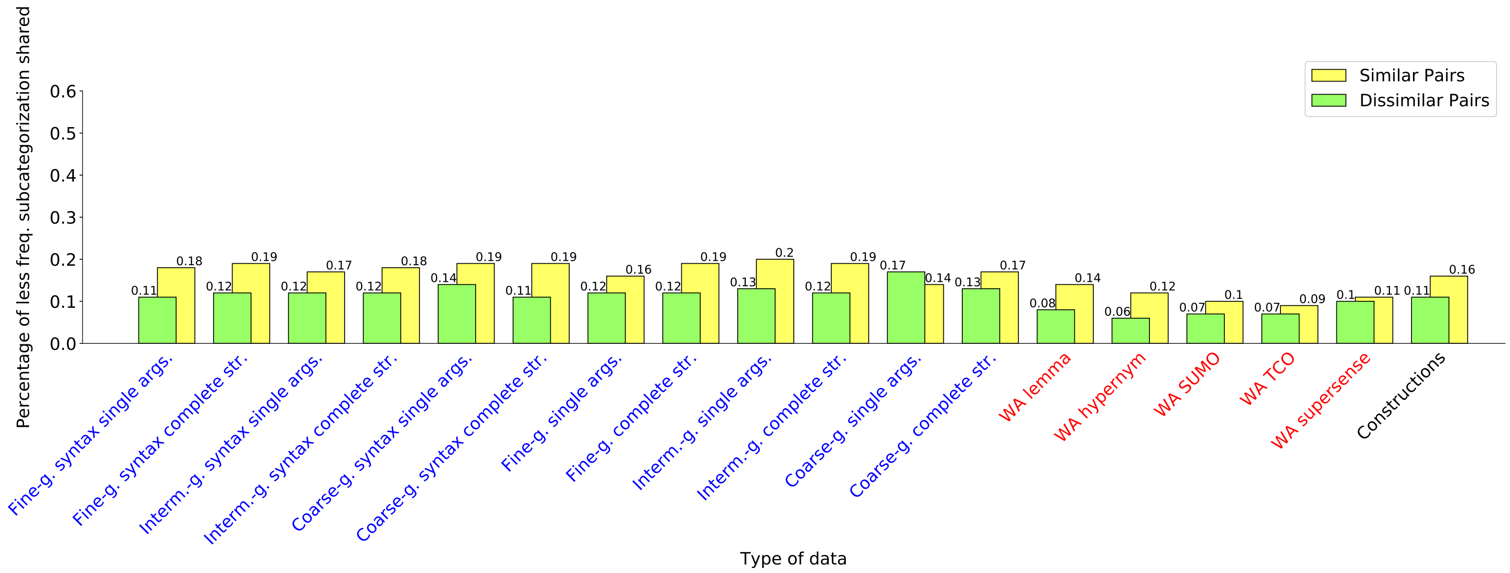
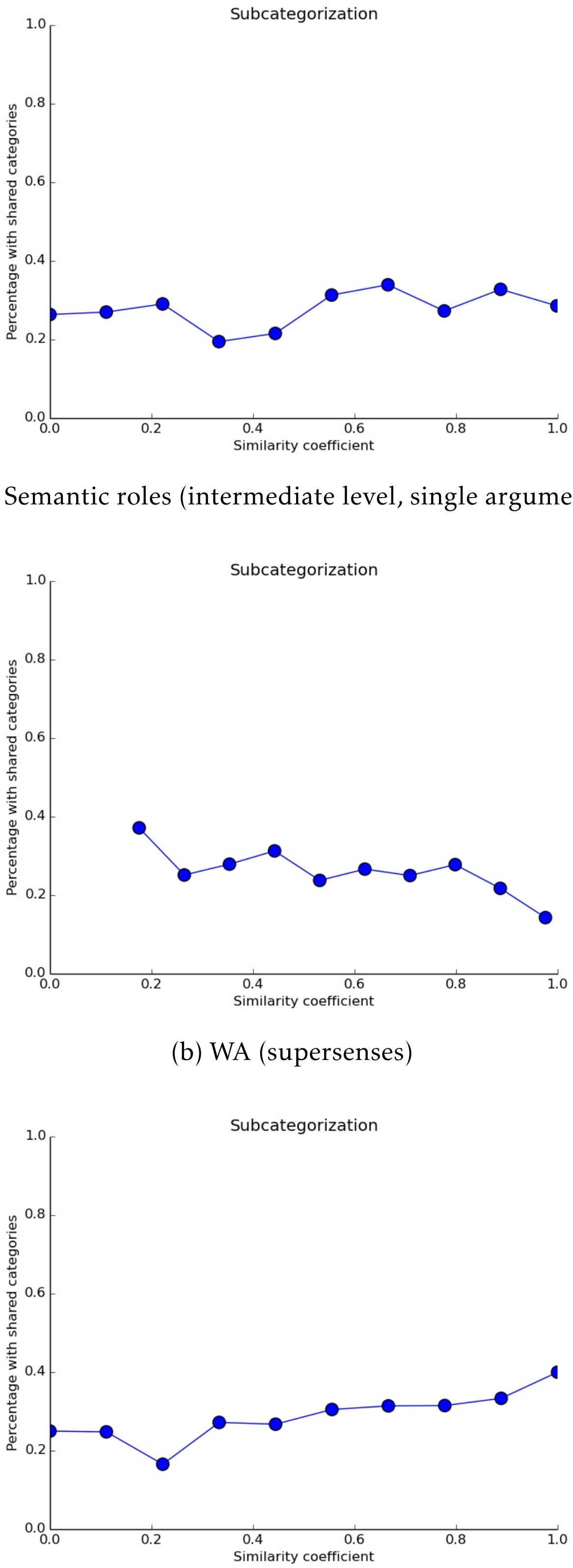


Figure 4.14: Low frequency subcategorization

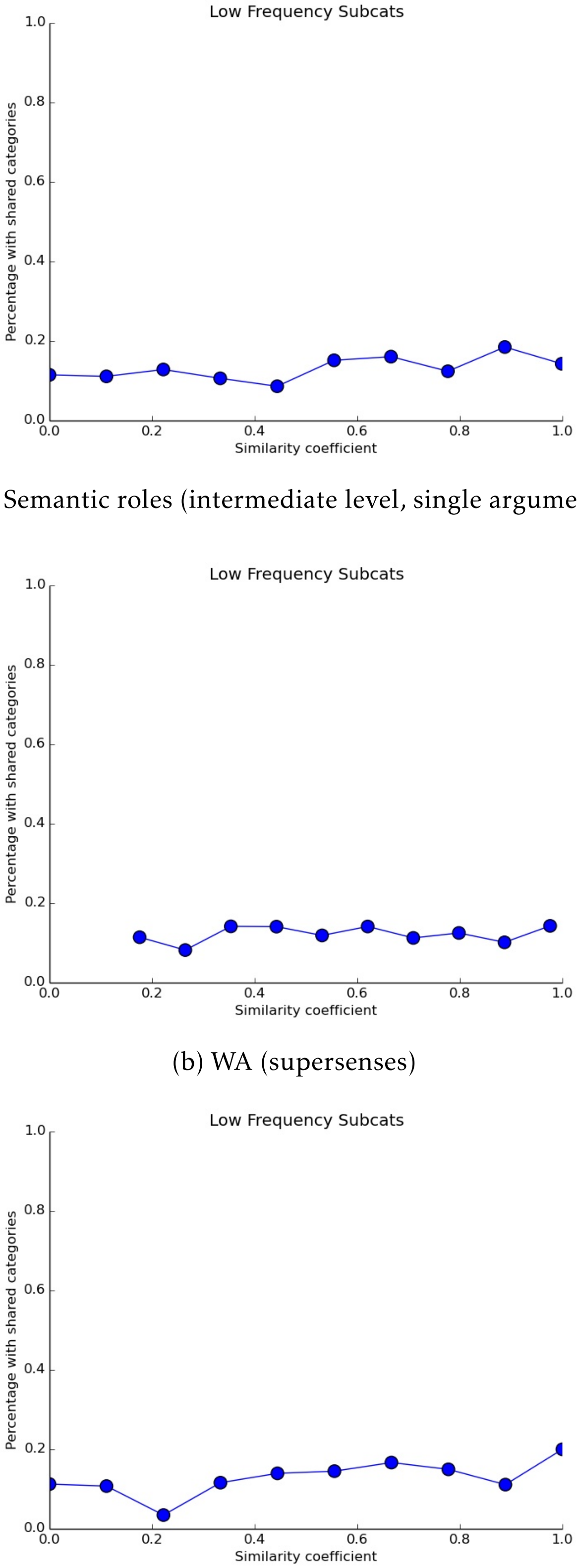


(a) Semantic roles (intermediate level, single arguments)

(b) WA (supersenses)

(c) Constructions

Figure 4.15: Subcategorization information



(a) Semantic roles (intermediate level, single arguments)

(b) WA (supersenses)

(c) Constructions

Figure 4.16: Subcategorization information (low frequency)

4.5.4 *Synthesis of the qualitative analysis*

Overall, we can conclude that the three approaches share some common ground when it comes to determining similarity between verb senses, but each perspective puts the focus on diverse linguistic aspects. If we take into account semantic field coherence, we see that verb sense similarity as defined by WA data and by some types of corpus data is more sensitive to semantic relations in terms of broad semantic fields. More specifically, we find that formalizations based on WA and on semantic roles without syntax are more semantically driven than other formalizations. As for aspectual cohesion, the major difference between the perspectives analysed is due to constructions, where similarity/dissimilarity is clearly articulated along the opposition dynamic/static. Finally, concerning the subcategorization information, differences between similar and dissimilar pairs across different formalizations are less prominent. Nevertheless, evidence suggests that similarity drawn from WA data is not sensitive to subcategorization information, at least when taking into account all subcategorization frames and not only the less frequent ones.

4.6 CONCLUSIONS OF THE CHAPTER

In this chapter we have compared three different perspectives of verb sense similarity which had not been previously related or comparatively studied: similarity defined by the semantic roles associated to the verb senses, obtained from corpus; similarity obtained from psycholinguistic data gathered through WA and similarity given by verb co-occurrences with constructions drawn from theoretical linguistics. Unlike other approaches to similarity, the comparison has been established using verb senses instead of lemmas, as ambiguity can be an obstacle to determine similarity at a detailed level. We have compared the similarity coefficients obtained when characterising verbs senses using different formalizations of these perspectives, finding low to moderate correlation in the similarity coefficients. The correlation strength tended to vary substantially depending on the specific formalizations explored within each perspective. After that, in order to further investigate the specific characteristics of each perspective, we looked in detail at the most and least similar verb senses, trying to find if the differences between them were articulated in terms of some linguistic features.

In this regard, there are a couple of findings that are of particular relevance because both their unexpectedness and the strength of the trend they draw: first, we find that verb similarities according to construction data are well articulated along the aspectual categories of static/dynamic, and more coherent than other perspectives regarding subcategorization. Secondly, we find that the similarity landscape drawn by WA data implies that it is more connected to the semantic component of language than the other types of data. Besides, in this landscape there is little influence of subcategorization information, at least of the more frequent subcategorization frames.

Moreover, we have also touched on the issue of the maleability of corpus data under the argument structure focus. Depending on the specific characteristics included in the formalization of corpus data, different linguistic features will be stressed and divergent similarity relationships between verb senses will be created.

Overall, while we find some coincidences in the information provided by the different approaches to verb similarity, both quantitative and qualitative, we have also uncovered some relevant sources of variation that need to be taken into account when comparing those approaches or when using them individually.

CHAPTER 5

Verb classification

Similarity relations between verbs are not necessarily restricted to interactions between two verb senses. Therefore, while in the previous chapter we focused on the analysis of the behaviour of specific pairs of verbs from the point of view of argument structure, in this chapter we go beyond the restrictions of the pairwise based analysis and we explore the relations of similarity that are established within a sample of 635 verbs, also from the point of view of predicate-argument structures. More specifically, in this chapter we aim at, firstly, **representing verb similarity using a clustering algorithm** to create an automatic verb classification for Spanish, applying the findings from the previous chapter to do so, and, secondly, **evaluating** this representation with respect to the argument structure information associated to the verb senses.

Our goal is not only to examine the performance of the automatic classification in the light of the results obtained, but also to establish the usefulness, extensibility and limits of the methodology. Taking this into account, we will pay particular attention to the evaluation of the models obtained, assessing through different means their adequacy in capturing information relevant for the expression of the events. The evaluation will be carried out using a range of methods that we describe more in detail in the following sections.

A comprehensive evaluation is necessary and relevant in order to address two fundamental shortcomings that were pointed out in the literature review (section 2.5). The first one refers to the evaluation methodology that is generally applied in automatic classifications: while inducing verb classes automatically has the advantage of being cost-effective as compared to manually crafting them, their effectiveness and applicability has yet to be thoroughly assessed. In section 2.5 we listed a reduced number of works that did use automatically created verb classifications for some tasks, although even in this case some of them (Shutova et al., 2010 and Guo et al., 2011) actually use the classes as another feature in a supervised classifier system and, therefore, the evaluation of the actual contribution of the classification is done by indirect means. Thus, a detailed and direct evaluation of the resulting verb classification is needed in order to overcome this lack of information about the performance of automatic classifications.

The second shortcoming in automatic classification is that most of the automatic classifications, and also the work that applies them in a task, is focused on English. Therefore, the work on modellization and evaluation carried out in this research will also provide an assessment of the applicability of automatic verb classifications to Spanish.

5.1 FEATURES FOR AUTOMATIC VERB CLASSIFICATION

In order to create the verb classification, verb senses are characterized on the basis of the arguments they co-occur with. Therefore, each verb sense is associated to numeric vector that captures the co-ocurrences of that verb with different arguments within the corpus. The arguments, in turn, are characterized by a set of features that we define next.

We have already seen that Levin's proposal of representing verb behaviour through diathesis alternations has deeply influenced the approach followed when characterizing verbs for automatic verb classifications. However, the impact of this proposal has been somewhat indirect because capturing diathesis alternations is particularly difficult. Thus, in practice alternations are approximated using subcategorization frames, which are sometimes enriched with selectional preferences.

We share with the work done previously the idea of using subcategorization information. However, we add a number of modifications due to the conclusions reached in the previous chapter. These modifications are detailed in the remainder of this section. Besides, since we are modelling verb senses instead of lemmas, we directly extract information associated to the senses from the SenSem corpus, in which verb senses are already disambiguated.[1]

As we just mentioned, subcategorization information is the main source of verb characterization in verb classification tasks. However, we saw in the former chapter that frames that used syntactic categories (such as NP-V-NP) were not very helpful in order to distinguish between similar and dissimilar verb senses in terms of their argument structure. Therefore here we consider three strategies in order to alleviate this shortcoming: firstly we enrich the syntactic information with several types of semantic information (ontologies and semantic clusters) related to verb arguments and we use syntactic functions as well as syntactic categories; secondly, we add *aspect*

[1] The complete list of categories that are used to annotate this corpus can be found in http://grial.edu.es/sensem/doc?idioma=in

as a feature to characterize verbs, as it proved to have discriminative power between similar and dissimilar verb senses; thirdly, we decompose the subcategorization frames into meaningful parts to explore the effects of giving more relevance to having a less constrained representation of the subcategorized information.

Therefore, the features that are used to characterize the arguments include the following types of information:

Syntactic information

1. Syntactic function of the argument (direct object, subject, etc.)
2. Syntactic category of the argument (noun phrase, prepositional phrase, etc.)

Semantic information

For the semantic information we characterize the argument heads using lemmas, categories from several ontologies, and automatically induced categories from semantic clusters. Besides, we include aspectual categories (static or dynamic aspect). Here we must clarify that the information under the tag *aspect* in the corpus corresponds to the aspect of the sentence. Therefore, although the scope of this type of information goes beyond the argument unit, aspect is related to the argument structure (typically the quantification of the direct object). This circumstance, together with the fact that it proved to be relevant for the definition of verb similarity with respect to argument structure, encourages us to include it in the set of features used to characterize verbs. Summarizing, we include the following pieces of information:

1. Lemma of the argument head
2. Sumo category of the argument head
3. Supersense category of the argument head
4. TCO category of the argument head (linked features such as *building+object+artifact*)
5. Automatic category of the argument head, obtained from semantic clusters created using Word2Vec (Mikolov et al., 2013)
6. Aspect of the verb sense (equivalent to lexical aspect but specific for each sense): static or dynamic

In SenSem corpus the arguments are manually annotated with their syntactic information. Semantic information, on the other side, is just partially defined manually. More specifically, only the head of the argument is marked. Therefore, the ontological information is obtained through the synset of the argument head, which serves as a handler to map the word marked as head to a semantic category from the different ontologies listed above. The synset was obtained using Freeling, which includes the UKB disambiguator Agirre et al., 2009, a PageRank-based word sense disambiguation algorithm. Thus, argument heads are automatically disambiguated and assigned a synset (when possible), that in turn is used to obtain an ontological category. Besides, as we have seen, the lemma of the head of the argument is also used to characterize it.

Ontologies are manually built resources, and as such, besides having many advantages such as their quality and precision, they are costly to create and inherently limited. Therefore we think that an interesting addition to the characterization of verb arguments from a semantic point of view is the usage of an automatic resource. To achieve such characterization and obtain automatic categories for verb arguments (the 5^{th} item in the above list of *Semantic information*), we created semantic clusters using the Word2Vec algorithm. In order to build them we followed these steps:

- First we took a corpus of 3 million words collected using data from *El periódico*, a Spanish journal. We normalized it, converting the text to lowercase and substituting the numbers with the label <digit>).
- Then the corpus was used as input to create low dimensional word vectors (embeddings) using the Word2Vec implementation available in the Gensim library (Řehůřek et al., 2010). As parameters for the Word2Vec algorithm we selected 100 dimensions, a minimun of 5 occurrences for tokens and a sliding window of 5 words. These parameters were selected on the basis of an ad-hoc validation of the most similar words (nearest neighbours) produced by the Word2Vec model to a given set of words.
- After this step was performed, we obtained a model of word embeddings, where each word was associated to a vector that contains information about the contexts in which a word occurs. Thus, each word is characterized by the words that it co-occurs with, following the proposal of the distributional semantics paradigm.
- The last step consisted in creating the clustering of words using the embeddings as input. To do so, we used the spectral clustering algorithm implementation available in the scikit-learn library (Pedregosa et al., 2011), using as parameters 5 for the number of neighbours and *cosine* as the affinity metric. Again, the algorithm and parameters are established through a validation of the contents of the clusters, by observing the semantic congruence between the meanings of the words wich are members of the same cluster.
- Once the clusters were defined, in order to create the features based on the *semantic cluster labels*, for each argument head defined in the corpus we identified its cluster ID and we assigned it to the argument. Therefore, arguments whose head belong to the same cluster may be represented by the same feature. For argument heads that were not included in the cluster we used the generic label <OOV> (out of vocabulary) instead of the cluster identifier.

Finally, as mentioned in the motivation for the features to be included in the automatic classification, none of the perspectives of verb similarity had subcategorization frames as a strong foundation on which they were able to distinguish between similar and dissimilar verb senses. Thus, besides using the traditional approach based on subcategorization frames, we experimented with the effects of breaking down these frames into their meaningful components, in order to test whether these types of information are more helpful in order to create a verb classification that captures the relevant information of the arguments. Therefore, we designed two different configurations of features (subcategorization frames and subcategorized constituents, that we will abbreviate as *constituents*). To better understand what each type of information represents we can look at example 26:

(26) Jeff [$_{\text{SUBJECT,PERSON}}$] drives [$_{\text{PREDICATE}}$] cars [$_{\text{OBJECT,ARTIFACT}}$]

- Subcategorization frames → each feature corresponds to the argument structure of the sentence:
  - ◇ subject-person+object-artifact
- Constituents → each feature corresponds to an argument:
  - ◇ subject-person
  - ◇ object-artifact

Besides, in order to collect co-occurrence information we use both probabilities and binary data (co-occurrence or not co-occurrence). Probabilities are a very common way of quantifying co-occurrence, but infrequent features might be overlooked in comparison with more frequent ones, and this is problematic if they carry important information. Binary data only considers whether two elements co-occur, disregarding the strength of the concurrence, which might alleviate the problem above mentioned but also increase the importance of noisy features. In this context, we explore these two ways of gathering quantitative information in order to determine which one is more adequate.

Since we have a considerable pool of features, both syntactic and semantic and two possible configurations (subcategorization frames and constituents), we set up the following criteria in order to keep the possible combinations within a reasonable range. Therefore, we generate feature sets following these guidelines:

- At least a type of syntactic information is included. Since in the previous chapter syntactic information was found to be a useful and robust feature in order to obtain higher correlations between different approaches to verb similarity, we decided to include it as a main component.
- The syntactic information can be used alone or combined with one of the following semantic features: categories from the ontologies, semantic clusters or lemmas (e.g. syntactic categories+lemmas; syntactic categories+clusters; syntactic functions+lemmas, etc.). In this way we can test the impact of using different types of selectional preferences for the arguments.
- Additionally, besides using syntactic information alone and syntactic informa-

tion combined with semantic features, we create a collection of feature sets in which this information is supplemented with aspectual information available in the corpus in the form of two broad categories (static or dynamic). The corresponding versions of the examples in the previous example with added aspectual information would be the following: aspect+syntactic categories+lemmas; aspect+syntactic categories+clusters; aspect+syntactic functions+lemmas.

These configurations allow us to create several feature sets that are used separately to characterize verb senses. Features that have a frequency lower than 2 are removed in order to diminish the possible noise due to labelling errors.

5.2 VERB SENSES SELECTED

Senses that occur less than 5 times in the corpus (have less than 5 sentences associated) are disregarded so as to ensure that verb senses have at least a minimal characterization. Therefore, we are left with 705 senses out of the initial 988, which amounts to a total of 29,961 sentences. These data are split in training and test sets as follows:

- training: 635 verb senses
- test: 70 verb senses

The complete list of the verbs that fall in the training and test sets can be consulted in appendices A.4 and A.5, respectively. The senses for the test were randomly chosen from the pool of verbs available after the first pruning step. Table 5.1 summarizes the number of senses per frequency interval in the corpus (second column) and the number of senses selected for the test from each given interval (third column).

Frequency (number of sentences)	Number of total senses	Number of senses selected for testing
5	33	5
6-19	264	34
20-39	117	12
40-59	91	7
60-79	54	5
80-99	59	4
100-122	87	3

Table 5.1: Number of senses

5.3 CLUSTERING ALGORITHM

There are a number of clustering algorithms that can be used for the task of automatic verb classification. They can be grouped in families of models, such as density models or connectivity models, among others, according to their definition of similarity. For example, in density models two items are similar if they belong to a dense region whereas in connectivity models two items are similar if they are close in the space that has been defined. Each of the clustering algorithms has its own advantages and weaknesses. In our case, we performed some preliminary experiments with spectral clustering, K-means, agglomerative clustering and HDBSCAN and we decided to use agglomerative clustering, which belongs to the family of connectivity models, because it presents a series of advantages: is not as computationally expensive as other algorithms, such as spectral clustering, it is not as highly dependant on the distribution of the data, as it is K-means, and it is able to cluster all data points, differently from HDBSCAN.

The agglomerative clustering starts with each element in a single clustering. It then proceeds to recursively merge the clusterings according to a given measure of similarity and a linkage criterion. The linkage criterion determines whether two clusters should be merged according to the distance between the elements that compose them. We experimented with several types of linkage methods. Specifically, we report results for the following linkages:

- Single linkage: two clusters are merged if the distance between their closest data points (one from each cluster) is the minimum over all the clusters.
- Complete linkage: two clusters are merged if the distance between their most distant data points (one from each cluster) is the minimum over all the clusters.
- Average linkage: two clusters are merged if the average distance between all their respective data points is the minimum over all the clusters.

We also experimented with the number of classes, ranging from 2 to 200 in steps of 3 (2, 5, 8, etc.).

5.4 EVALUATION AND RESULTS

In order to evaluate the resulting automatic classifications, we carry out two types of evaluation. First, we present an evaluation of the similarity between our automatic classification and a gold standard, following the traditional evaluation methodology in automatic verb classification. This gold standard is created by classifying together the verbs of SenSem that share the same set of semantic roles. Its creation is detailed in the following section.

Secondly, we present an evaluation that looks into the generalization power of automatic verb classifications. More specifically, we test the compatibility between the argumental information that is brought together in the automatic classes and the argumental information associated to verbs that are outside of the classification (test verbs) when those verbs are assigned to the automatic classes. Besides, we perform a comparison with the word association data and the theoretical data that were defined in the previous chapter, assessing how well the pairwise similarity relations defined in the psycholinguistic experiment and in the theoretical data are kept in the automatic classification.

5.4.1 *Evaluation against a gold classification*

As we already saw in chapter 2, traditionally, automatic classifications are evaluated against a gold standard classification which is developed manually. Their degree of coincidence is measured using metrics such as purity, homogeneity, mutual information or F_1 score, which test whether the gold classes and the automatic classes organize the verbs in similar ways. In this section we perform an evaluation against a gold classification adapted from the SenSem corpus and we present the results obtained using the F_1 score.

In order to generate the gold standard for this evaluation, using the data present in SenSem, we follow these steps:

1. For each verb, we collect the coarse-grained semantic roles present in the sentences in which it appears. The output of this step is a set of semantic roles for each verb.
2. The roles in each set are ordered alphabetically. This ordered set is used as class identifier (e.g. Actor+Place+Undergoer).
3. Verb senses that occur with the same set of roles are assigned to the same class, with that set of roles as class identifier.

The outcome of the gold standard generation are 13 classes. Within each class, all the verbs combine with the same set of semantic roles.

Next, we compare the automatic classes that we generated with the gold standard and we provide a quantitative index of how much overlap there is between them. In relation to the evaluation measure used in this section, we mentioned in section 2.5 that there was a measure, the harmonic mean of purity and accuracy, that had been used in a number of works. However, we do not use it because, as Sun (2013) explains, it is biased towards classifications with larger number of clusters and, therefore, it is not advised in cases where no number of classes is selected a priory (as it is our case). Therefore, we select another measure, the pairwise F_1 score, which does not have this problem.

We calculate the pairwise F_1 score as follows: for every possible pair of verb senses it is checked whether they fall in the same class or in different classes both for the gold classes and the automatic classes. More specifically:

- two verb senses are together both in an automatic cluster and in a gold class → true positive (TP)
- two verb senses are together in an automatic cluster but separated in a gold class → false positive (FP)
- two verb senses are separated in an automatic clustering and together in a gold class → false negative (FN)

These indicators serve to calculate pairwise recall (equation 5.1), pairwise precision (equation 5.2) and pairwise F_1 score (equation 5.3). Next we present in table 5.2 the automatic verb classification that obtains the best score in this evaluation.

$$pairwise\ recall = \frac{paired\ TP}{paired\ TP + paired\ FN} \tag{5.1}$$

$$pairwise\ precision = \frac{paired\ TP}{paired\ TP + paired\ FP} \tag{5.2}$$

$$pairwise\ F_1 = \frac{2 \cdot pairwise\ precision \cdot pairwise\ recall}{pairwise\ precision + pairwise\ recall} \tag{5.3}$$

Linguistic features	Configuration	Algorithm parameters	F_1
aspect, syntactic functions	constituents	14 classes, average linkage, dice distance	0.43

Table 5.2: Best classification for gold standard evaluation

A comparison between the structure of the automatic classification and the gold classification can be seen in figure 5.1[2]. The automatic classes (clusters) are represented on the left side in red, whereas the gold classes are represented on the right side in blue with their corresponding semantic role labels. We can observe that the cluster under the label 10 tends to conflate different gold classes that contain the *Undergoer* role (classes *Actor@Undergoer*, *Actor@Place@Undergoer* and part of *Undergoer*). The gold

[2] For better visualization and interaction visit the following url: https://plot.ly/~lgilva/73

class *Undergoer* is split into two clusters (cluster 5 and cluster 10). Another main class, *Actor@Place*, is split into three clusters (8, 9, 10). Cluster 8 contains verb senses that belong to different *Actor* classes. Therefore, although the 1-to-1 correspondence between gold classes and automatic classes is not perfect, we can see that the basic organization of the gold classes is preserved.

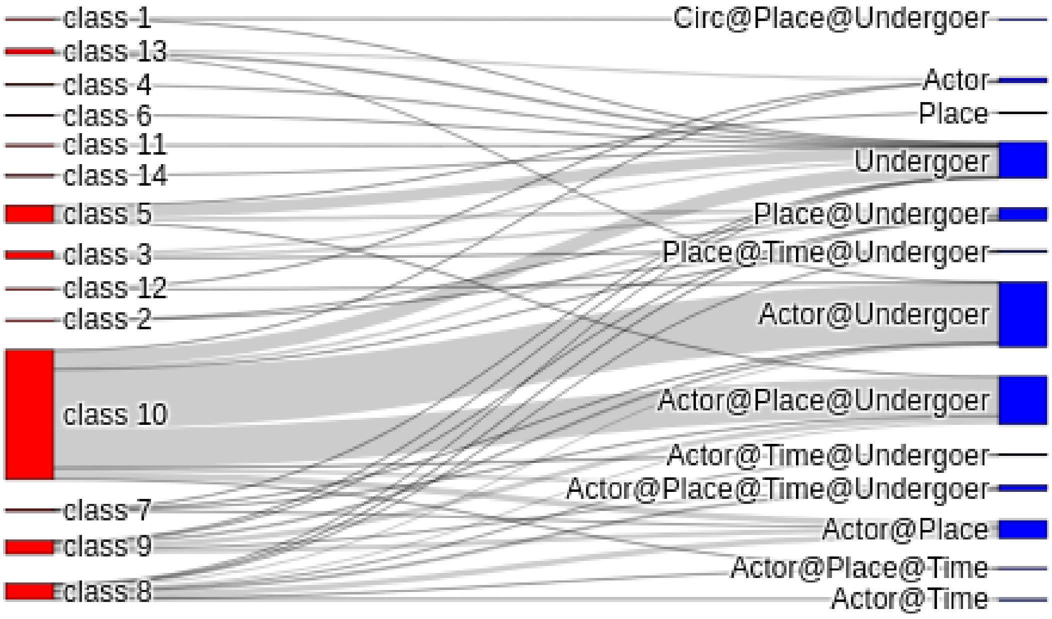


Figure 5.1: Comparison between gold classification and best automatic classification

Finally, in order to provide more context for the evaluation score obtained, we calculated the harmonic average of purity and accuracy of this classification, which is 0.47. This score is not directly comparable with the evaluations of the other classifications that we mentioned, as they use different gold standards. However, we can see that this score is in line with the scores obtained for those classifications. For example, Scarton et al. (2014) obtained a harmonic average of 0.42 for Brazilian Portuguese, with a gold standard of 540 verbs in 16 classes; Sun, Korhonen, Poibeau, et al. (2010) got a score of 0.55 for French with a gold standard of 171 verbs in 16 classes; Sun and Korhonen (2009) reached very different scores for two datasets for English, the first dataset contains 586 verbs into 14 classes, for with they obtained a score of 0.57, and the second dataset contains 204 verbs into 17 classes, for which they obtained a score of 0.80. All in all, the differences between the data and tools developed for English and the ones developed for other languages are reflected in the results. Besides, the results tend to be better when the ratio of verbs per class is low.

In the next evaluation we aim at overcoming the limitations of this type of evaluation, which does not take into account the generalization power of the classifications. Consequently, we go one step beyond in testing the ability of the automatic classes to organize verb senses in meaningful ways, classifying the test verbs in the automatic classes and looking at the compatibility between the role information associated to the test verbs and the role information associated to the verbs already in the classes.

5.4.2 *Generalization of the automatic classification to new verb senses.*

In the former section we covered the classical evaluation in automatic verb classifications. In this section we take a new angle on the evaluation of verb classifications and we look at the information existing in the classes. In particular, the goal for this evaluation is to examine the generalization power of the information contained in the automatically created verb classification that performed best in the former evaluation. This evaluation sheds light on the generalization power of the classification by identifying up to what point the linguistic information that it contains is extensible to elements that are outside it by using a definition of similarity. To do so, instead of using the classes obtained as features in an external task, as it is done in other approaches, we directly use the information included in the automatic classification. The setup for this evaluation is inspired on the Semantic Role Labelling (SRL) task because we deal with the semantic content of the arguments of the predicates. However, it is defined in a more general way: while the SRL task consists in the identification and classification of the arguments of a predicate that appear in a sentence, our goal of this evaluation is focused on specifying the type of semantic roles that typically combine with a predicate according to the sentences in which this predicate occurs. In other words, the task we use for this evaluation does not require to assign a semantic role to each argument, but to constrain the inventory of semantic role patterns (complete structures) that can occur with each predicate.

In order to carry out this evaluation, we take the automatic classification that performed best in the former evaluation and we assign each sense from the test to a cluster according to the distance and type of linkage used when creating the clustering. Then, we compare the role structures associated to the test verb senses (the roles labelled in the sentences in which those senses occur) and the role structures that are representative of their assigned cluster, using standard evaluation measures to quantify the degree to which they overlap.

In order to assign the test verb senses to clusters we formalize them in the same way that was used to create the clusterings, so that clusters and senses are comparable. Therefore, to assign a sense to a cluster the first step is to represent it according to the features used to create the clustering. The second step consists of selecting the metric and the linkage criterion that was used to build the clustering and apply those parameters to assign the test sense to a cluster. Once the sense has been assigned to a cluster (or several clusters if the distances obtained are equal), the set of roles of the sentences in which this sense occurs and the role structures that are representative of that cluster are compared.

The information used to evaluate the classification are the semantic roles associated to the verbs. We perform separated evaluations for the three levels of semantic roles (coarse-grained semantic roles, intermediate semantic roles and fine-grained semantic roles). More in detail, we asses the overlap between the complete role structures associated to the test verbs and those that are associate to the class in which this verb falls. Those structures are obtained from the sentences in the corpus associated to the relevant verbs. An example of this type of information is *Agent+Theme*, or *Actor+Undergoer+Place*

The evaluation measure used in order to perform the comparison is, as in the previous case, the F_1 score, which is a standard measure across NLP tasks. We already saw that it is defined as the harmonic average of precision and recall but in our case we adapt the definition of those two items to the task at hand.

- recall: number of correctly identified semantic roles structures of the test verb (structures that are both associated to the verb and to the cluster) divided by the total number of structures that are associated to the test verb.
- precision: number of correctly identified semantic roles structures of the test verb divided by the total number of structures present in the clustering.

Additionally, we set up two baselines in order to measure the impact of using classes and to assess the possible weaknesses of this methodology.

- The first baseline is created using the **closest sense from the training set** to the test sense. This entails selecting the closest sense in the training set (the collection of senses used to create the automatic classification) to a given test sense and comparing the role structures that are associated to each one. The motivation for creating this baseline is testing the usefulness of having classes: we test whether, under the same feature characterization, the information contained in a class is more precise and complete than the information of the most similar sense.
- The second baseline is created using a **purely unsupervised method**: the Doc2vec technique[3] (Le et al., 2014). This evaluation helps us assess the usefulness of manually defined features versus using an unsupervised characterization of verb senses based on neural networks. The Doc2vec technique is an extension of the Word2vec framework. Word2vec creates word embeddings, which are vectors that capture the linguistic context of a word, by representing this context in a low dimensional space. Doc2vec extends this methodology to documents by taking into account document labels. It is trained with the objective of predicting these document labels besides the words in context. Therefore, it is able to measure the similarity between documents taking into account the words that they contain.

[3] As implemented in the gensim package (https://radimrehurek.com/gensim/models/doc2vec.html)

> In order to use this technique for our baseline, we modelled each sense as a document composed by the sentences in which it occurs. Therefore, measuring the distances between these sense-documents is equivalent of measuring the distances between the verb senses. In order to evaluate this baseline, training and test verbs are characterized using Doc2vec and then each test verb is paired with the most similar training verb. The roles of both verbs are compared using the criteria previously explained.

In order to carry out the evaluation, the automatic classes are characterized by the semantic roles associated to the verbs that they contain. A first approach to evaluating the classification may be associating each automatic class with *all* the semantic roles that correspond to the training verbs that it contains. However, this approach may be problematic because it ensures a high recall but poses a problem in the precision of the system: since the classes will consist of a variety of semantic roles pertaining to several verb senses, it can happen that the majority of the semantic roles associated to the test verbs will be found but the classes will have more semantic roles than needed.

While this situation was not problematic for other approaches (Lamirel, Falk, et al., 2015) because they did not take precision into account, we consider that the precision should be taken into account as well when evaluating the performance of a classification that captures argument structure information. A recall of 100% can be achieved if the classes contain very heterogeneous verbs in terms of their semantic roles, but this does not ensure that the classes are coherent in terms of the argument structure information they contain. Therefore, it is necessary to pay attention to the precision of the classes.

In consequence, we explore several ways of selecting the relevant semantic roles for each automatic class. More specifically, we evaluate the following alternatives for selecting the relevant roles for the classes:

1. Select the roles with higher F score (Lamirel, Falk, et al., 2015): F score here works as an internal measure to decide which semantic role structures from the ones associated to a class are more representative. It is calculated on the basis of a particular definition of recall and precision: recall for role information selection is defined as the frequency of role structure (e.g. Actor+Undergoer) in a given class divided by its total frequency in all the clusters. Precision for role structure selection is defined as the frequency of a role structure in a given class divided by the frequency of all the role structures in the same class. Thus, a role structure is kept for a given class if its F score is higher than the average F score of all the structures and the average of that structure in all the classes.
2. Select the roles that fall within a threshold based on convexity (Gonzàlez et al., 2015): in figure 5.2 it is shown a example of how this threshold is established. The weights of the roles in each cluster are first normalized so they fall in the range [0-1]. At that point they are sorted in descending order (from more to less weight). The sorted weights are shown in blue and the red line marks

the convex distribution. Then a threshold can be set in order to separate roles that are relevant for the class from roles that are not so relevant. To calculate this threshold, the distance from each normalized weight to the origin (0,0) is measured. The threshold is established at the feature that minimizes the distance, marked in the figure with a solid arrow (as opposite to the dotted arrow) and the roles that fall within this threshold (roles whose weight is equal or higher than the feature at the threshold) are kept. They are marked in darker blue in the figure.

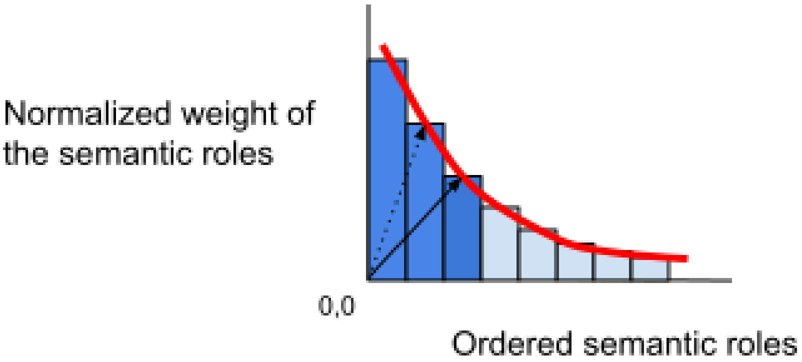


Figure 5.2: Convexity threshold

First we investigated the effectiveness of each method of role selection by evaluating the automatic classification for each of them, besides the default method of keeping all the semantic roles. In table 5.3 we present the scores obtained for each of these methods. The scores are averaged over the different levels of semantic roles used for the evaluation.

	All features	F selection	Convexity threshold selection
precision	0.1	0.05	0.46
recall	0.56	0.16	0.38
F_1 score	0.16	0.08	0.41

Table 5.3: Scores for each method of selecting roles relevant for the classes

As predicted, using all the information related to roles in the classes, when the number of classes is small in relation to the number of verbs classified, yields a high recall but a low precision. The method that gives the best results for role selection is the one based in the convexity threshold. Next we offer the results based on this methodology of role selection, specifying the scores obtained for each level of abstraction. After that, we also offer the scores obtained when using all the semantic roles, following the same division by level of abstraction, so that both approaches can be compared more in detail.

The results of the automatic classification that performed best in the gold standard evaluation are showed for this evaluation in table 5.4. This table shows the F_1 achieved by the classification for each of the abstraction levels of semantic roles, as well as the results obtained by the baselines (*bs* in the table).

	Automatic classes	most similar bs.	Dov2Vec bs.
Coarse-grained roles	0.46	0.44	0.28
Intermediate roles	0.24	0.22	0.15
Fine-grained roles	0.20	0.15	0.09

Table 5.4: F_1 scores for the three levels of abstraction of semantic roles

The F_1 scores achieved by the automatic classification range between 0.46 for role structures within the more coarse-grained level (such as *Actor+Undergoer*) and 0.20 for the more fine-grained level of information (structures such as *Cause+Affected theme*). Logically, the more specific the semantic roles used to evaluate the extensibility of the classification, the harder the task becomes.

As for the baselines, we can see that the information gathered in the classes outperforms the information that the most similar sense offers in all the levels of abstraction, with a wither margin in the fine-grained level. Regarding the doc2Vec baseline, it performs worse than the automatic classification and the most similar verb baselines. Thus, we can conclude that for this type of task, characterizing verbs using linguistic features proves to be more useful, either using classes or single verbs.

Besides, we present in table 5.5 the results obtained in the evaluation when all the roles of the classes are used, without applying any selection method. In this table we can see the scores achieved when using all the role structures in the automatic classes (left column) and the difference in the performance with the convexity threshold method for role selection (right column). As we anticipated, for this evaluation, grouping a large number of verb senses in a limited number of classes (14 in our case) may have negative effects on the precision of the classes that need to be resolved either by augmenting the number of classes or by selecting the relevant information of the class.

	Automatic classes	Difference with convexity scores (table 5.4)
Coarse-grained roles	0.28	-0.18
Intermediate roles	0.13	-0.11
Specific roles	0.09	-0.11

Table 5.5: Comparison of F_1 scores

5.5 COMPARISON WITH PSYCHOLINGUISTIC DATA

In chapter 4 we described the details of a comparison that deepened in the relationship between verb similarity as depicted by word associations (WA) and other approaches to argument structure (corpus data and constructions from linguistic theory). The goal of this section is to investigate whether the similarity relationships yielded by WA data, that we observed and analized in chapter 4, can be reproduced using a clustering algorithm and, if so, what are the linguistic features and the strategy when building the classes (number of classes, linkage for the clustering algorithm, etc.) that allows to create a clustering in which verb senses are classified in a way that is coherent with the similarity determined by WA. In particular, we look at whether verb senses that are considered highly similar or highly dissimilar according to WA data are also considered as similar or dissimilar by the clustering structure. Consequently, we asses here the ability of the clustering to correctly place similar and dissimilar verb senses according to their definition in the WA data.

It is important to note here that this is not considered a typical evaluation, at least in quantitative terms, as we look at a limited amount of pairs of verb senses. However, while we acknowledge that more data is needed in order to perform a thorough evaluation, we think that this approach will provide us with a first estimation of whether the characterization of similarity obtained from psycholinguistic data can be replicated using a clustering algorithm.

The methodology to carry out this evaluation starts by recovering the already created similarity-dissimilarity rankings for WA and the selected representative verb sense pairs from it. In chapter 4.5 we explained in detail how those verb sense pairs were selected. Summarizing the complete process, we created rankings of similarity by calculating the pairwise distances of a total of 190 pairs, that were ordered from more to less similar according to the similarity scores given by WA data. To create a reliable set of similar and dissimilar pairs, we selected the most and least similar (the extremes of the ranking), setting a threshold so that only the 10 most similar and 10 most dissimilar pairs were taken into account.

Once the sets of similar and dissimilar pairs have been defined, it should be established whether their status as similar or dissimilar is maintained in the clustering. From the point of view of the clustering, a simple and intuitive way to define similarity and dissimilarity is to consider that two senses are similar if they fall in the same cluster and dissimilar if they fall in different clusters. Thus, the process of evaluation is carried out in the following way: for pairs of senses that are similar according to WA, it is verified whether both are located in the same cluster. For pairs of dissimilar senses, it is verified whether both fall into different clusters. In order to obtain an balanced measure for similar an dissimilar pairs, we avoid using the traditional F_1 score, because when the clustering correctly places the dissimilar pairs and fails with the similar pairs the F_1 score is still 0.5, which is not representative of the performance of the clustering. Instead we calculate the harmonic average for the percentages of similar and dissimilar pairs correctly placed in the clustering classes.

The automatic classification achieves a score of 0.37, which is the average mean of the percentage of the pairs correctly classified as similar (0.28%) and of the pairs correctly classified as dissimilar (0.56%). Thus the classification proves to be better at classifying dissimilar pairs than similar pairs. In section 5.7 we provide more insights on the linguistic features and clustering parameters that have an impact on this result.

5.6 COMPARISON WITH CONSTRUCTION DATA

Similarly as we did with the WA data, in this section we compare the similarity relationships defined by construction data and those defined by the automatic clustering. We follow the same methodology that was explained in the previous section: we first recover the similarity rankings from the section 4 for the construction data, then we select the most similar and dissimilar sense pairs and finally we look at their status in the automatic verb classification.

As previously, we consider that similar verbs according to constructions are adequately represented in the clustering if the two members of the similar pair are located within the same automatic class and, conversely, we consider that the dissimilar verbs are adequately captured if the members of the dissimilar pairs fall in different automatic classes. As previously, we calculate the harmonic average of the percentages of similar and dissimilar verbs correctly placed in the clustering.

The automatic verb classification obtains a score of 0.93 (0.90% of similar pairs and 0.96% of dissimilar pairs are correctly classified). This score is notably better than the scores obtained when comparing the automatic classification with WA data. In section 5.7 we provide an analysis of the role of the linguistic features in relation to these results with the goal of understanding better the reasons behind this difference, besides other issues.

5.7 FEATURE SWAP STUDY

In this section we carry out a study with the aim of better understanding which features and parameters are important for the performance of the automatic classification. More specifically, this study looks into the role of the linguistic features used in the automatic classification with respect to the evaluation scores. Since we created a set of automatic classifications by exploring a number of systematic combinations within a closed collection of linguistic features and clustering parameters, we can compare the scores achieved by the automatic classifications when each of the different linguistic features and parameters are used.

In order to be able to assess the role of these different components of the clustering, we compare the performance of clusterings that are identical except for one feature, the one whose contribution is being studied at each moment. For example, we can compare the contribution of syntactic functions and syntactic categories, which are in complementary distribution within the clusterings. To do so, we set side by side the collection of clusterings that use syntactic functions and the collection of clusterings using syntactic categories, which mirrors the previous collection exactly in terms of the rest of features and clustering parameters. Particularly, we look at the following types of features:

- The type of semantic information used (lemmas vs TCO vs SUMO vs supersenses vs semantic clusters vs not using semantic information)
- The type of syntactic information (syntactic categories vs syntactic functions)
- Using aspect vs not using it
- The type of feature configurations (constituents, subcategorization frames)

For each of these 4 types of features we compare the scores obtained by clusterings that use the different information comprised within each type (e.g. scores of clusterings that include aspect as linguistic feature vs those that do not include aspect). We report results for the performance of each type of information in the evaluations (generalization power and gold standard) and comparisons (psycholinguistic and constructions), showing the differences in each case. These results are represented graphically in 6 bipartite figures (figure 5.3 to figure 5.18), each of which consists of a cumulative kernel density plot (upper part) and a box plot (lower part).

The kernel density plot is a variation of a histogram that uses a smoothing strategy to represent values while trying to mitigate the effects of noise (in our case we use a Gausian kernel). The kernel plots represent the distribution of the evaluation scores for a given feature in a cumulative fashion. Therefore, for each feature and score X, we obtain the probability for the clusterings that include this feature of getting a score value that is equal or less to X. The cumulative information is particularly useful when two or more variables are being compared.

A feature whose values are distributed following a flat shape at the beginning (indicating that it has a low probability for low scores) and presents a steep rise at the end (indicating that it has high probabilities of achieving higher scores) can be considered to contain information that is important in achieving good scores for a specific evaluation.

The box plots are another way of looking at the data. They divide the data sample in quartiles, indicating the median of the data, its dispersion and its skewness. The box contains the 50% of the data, those that are between the 2nd and the 3rd quartiles. This distance is the interquartile range (IQR). In our figures the fences or whiskers of the box plot correspond to 1.5 times the IQR.

Next, we explain in more detail the information that is present in the following figures. In the density plots, on the X axis there are represented the possible values for the scores achieved by the clusterings that contain a specific feature. On the Y axis, we can observe the values of the cumulative density function. These values indicate the probability for the clusterings that contain a specific feature of yielding a score that is equal or less than the score that corresponds to its coordinate on the X axis. As we explained, this implies that useful features will present a distribution that grows slowly at the beginning (low scores on the X axis) and their growth will be more marked at the end of the X axis (higher scores). Since it is cumulative, the distributions always grow until they reach a probability of 1.

The box plots show the median for each feature being compared and the two closest quartiles that hold the 50% of the data. When data is more spread the IQR is larger. Conversely, if the data is dense, concentrated in a small range of values, the IQR is smaller. Besides, they also show if the data is skewed (not evenly distributed), by presenting the median line closer to one of the ends of the IQR. There is a different box for each of categories used.

In the remainder of this section we examine the role of the different features/parameters for the different evaluations and comparisons presented in the previous sections of this chapter.

1. Aspect
   We start the analysis of the role of the different features used in the classification examining the impact of adding aspectual information versus not adding it. We remind here that the *aspect category* comprises two possible values, static or dynamic, that are used to characterize the sentences associated to each verb sense. As for the *gold standard evaluation*, we see in figure 5.3 that its presence tends to improve the scores, as configurations that include aspect have more probability of obtaining higher scores. Besides, we can see in the right upper corner of the figure that only classifications that include aspect reach the higher scores. Also, in the box plot we can observe that the median for aspect is slightly higher.

The impact of using aspect is negligible for the *extensibility evaluation* (figure 5.4). The distributions and medians are very similar for configurations that include aspect and for those that do not include it.

Regarding the effect of using aspect in the clusterings for the *comparison with psycholinguistic data*, it can be seen in figure 5.5 that the differences in the distribution of the scores between the approaches that use aspectual information and those that do not use it are not large, with configurations that use aspect having more probability of obtaining higher scores. This is also corroborated by the boxplot, in wich we see that the median for configurations that have aspect is higher. Data is skewed for configurations that do not contain aspect: the scores that are below the median have a smaller range, meaning that the scores in the 2nd quartile (lower scores) are more concentrated than those in the 3rd quartile (higher scores).

As for the presence of aspect in relation to the *comparison with data from constructions* (figure 5.6), we see that configurations with aspect tend to obtain higher F_1 scores. For example, 50% of the configurations that use aspect obtain scores above 0.6, whereas in the case of not using it, this percentage is reduced to 30%. Besides, the score median value is clearly higher for configurations that contain aspect as opposed to those that do not (0.58 vs 0.4).

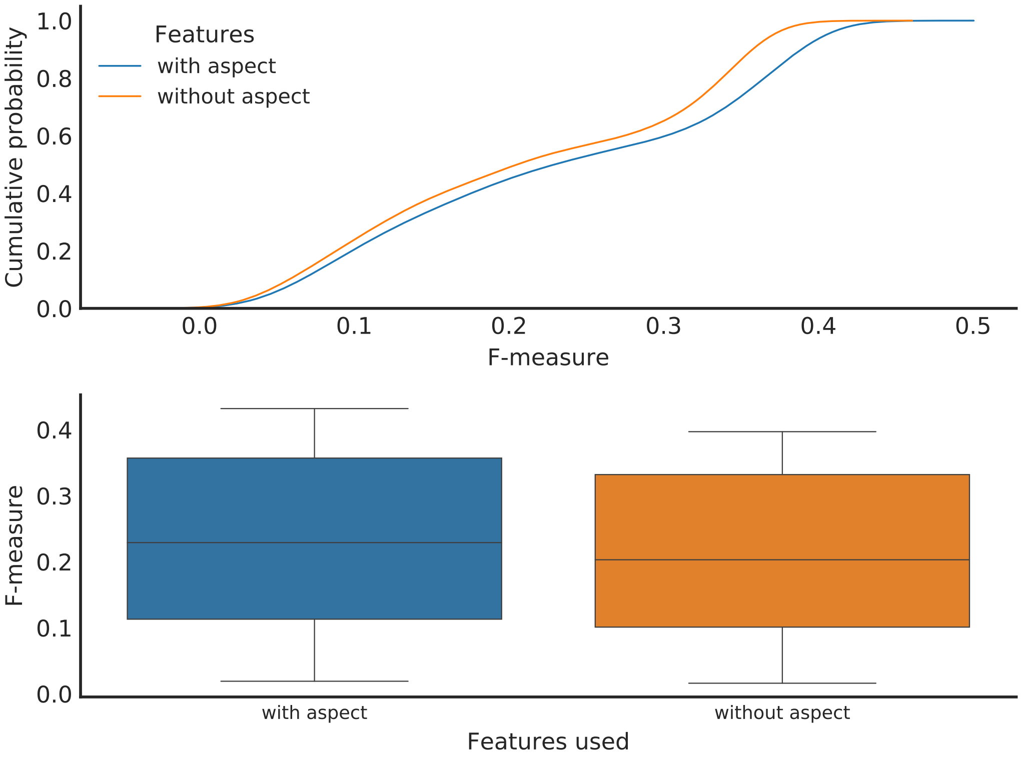


Figure 5.3: Gold standard evaluation: Aspect

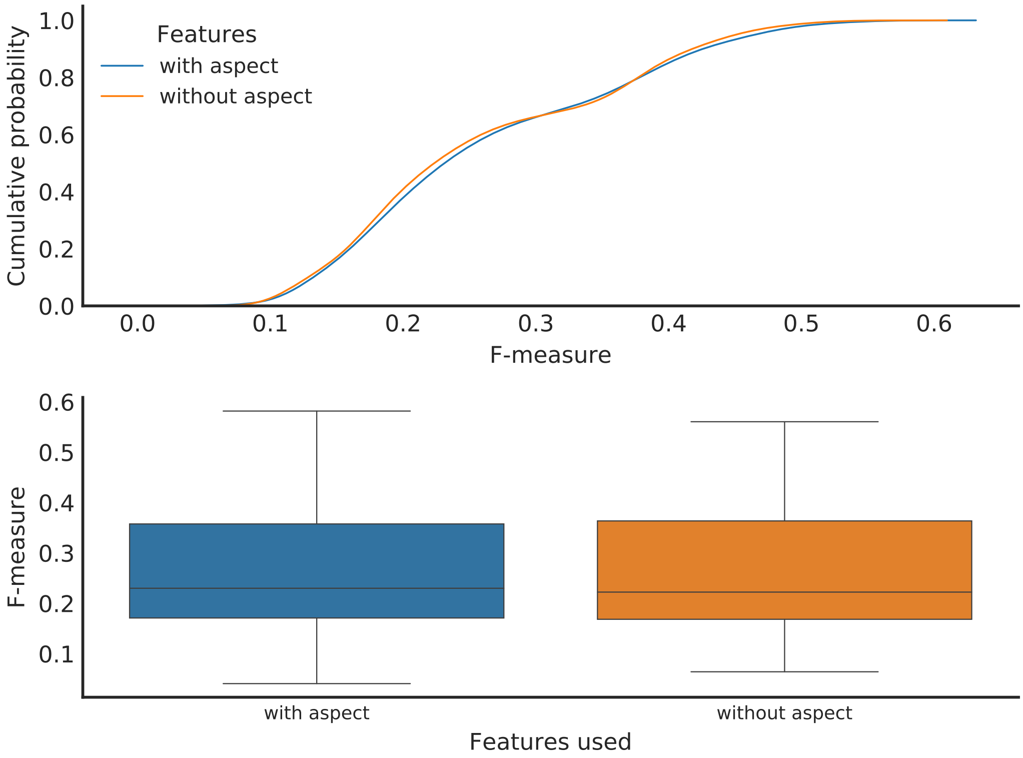


Figure 5.4: Extensibility evaluation: Aspect

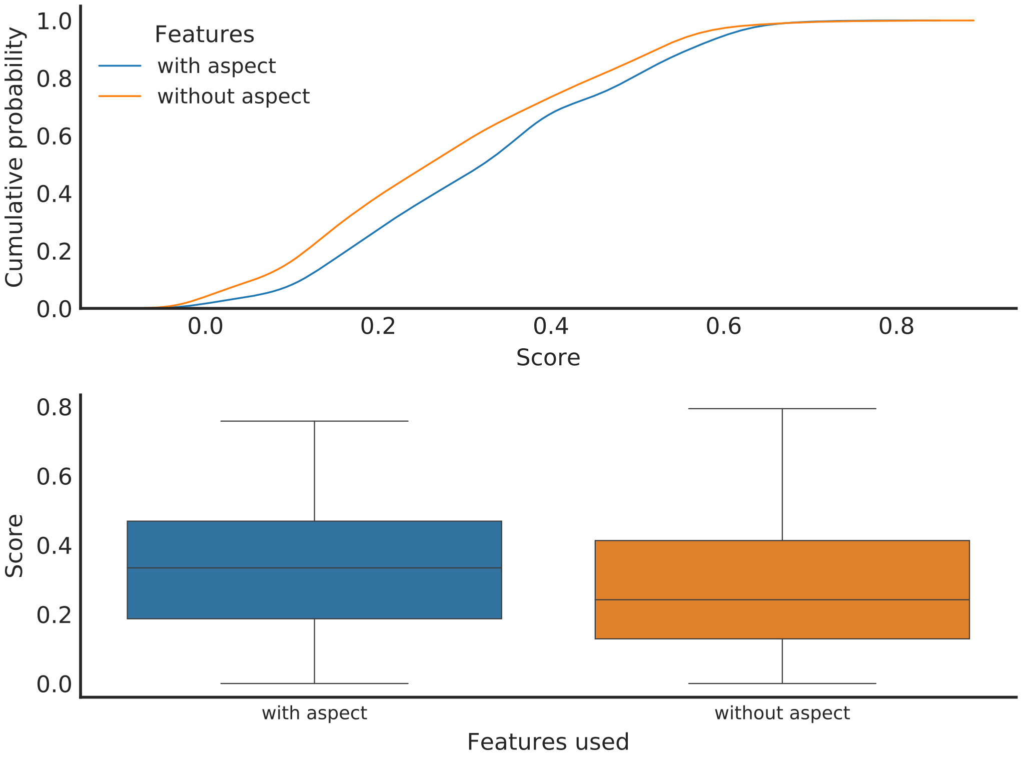


Figure 5.5: Psycholinguistic evaluation: Aspect

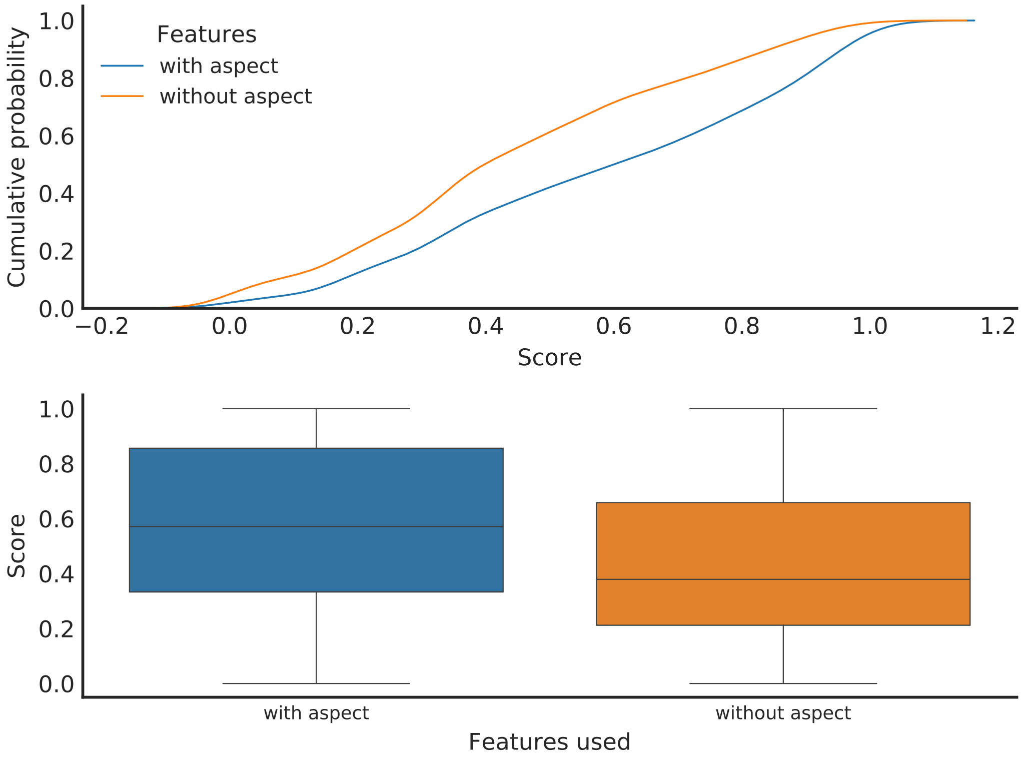


Figure 5.6: Theoretic data evaluation: Aspect

2. Semantic information

   As for semantic information included in the configurations, we should remind here that there are a number of categories: not including semantic information, lemmas, SUMO categories, TCO categories, Supersense categories and semantic clusters. As for the evaluation against a *gold standard* (figure 5.7), we see that the difference between these categories is not as easily interpretable as in the previous case. However, it can be seen that configurations that do not include semantic information are more likely to obtain higher F_1 scores. Fo example, there is a probability of 70% for scores over 0.2 as opposed to 50% of the semantic clusterings. Configurations that do not use semantic information also obtain the highest median, closely followed by the categories based on ontologies.

   As for the *extensibility evaluation* (figure 5.8), we find a similar landscape, with configurations that do not include semantic information being clearly more likely to obtain higher F_1 scores (30% for scores over 0.4 as opposed to 20% or less for the rest of the categories). This situation is also reflected in the boxplot, with the configuration that do not contain semantics having the higher median than the other configurations, which behave very much alike.

   In the *comparison with psycholinguistic data* (figure 5.9) this situation is reversed, with the lack of semantic information being the category that obtain lower scores with more probability and the one that has the lowest median. However, there is no clear winner in the rest of the categories, all of them show similar distributions, with SUMO and TCO categories having slightly more probabilities of obtaining higher F_1 scores.

   Regarding the *comparison with data from constructions* (figure 5.10), we find a situation in which is difficult to tell apart the contribution of each category, with lemmas having the lowest median and all of the rest exhibiting very similar distributions and medians.

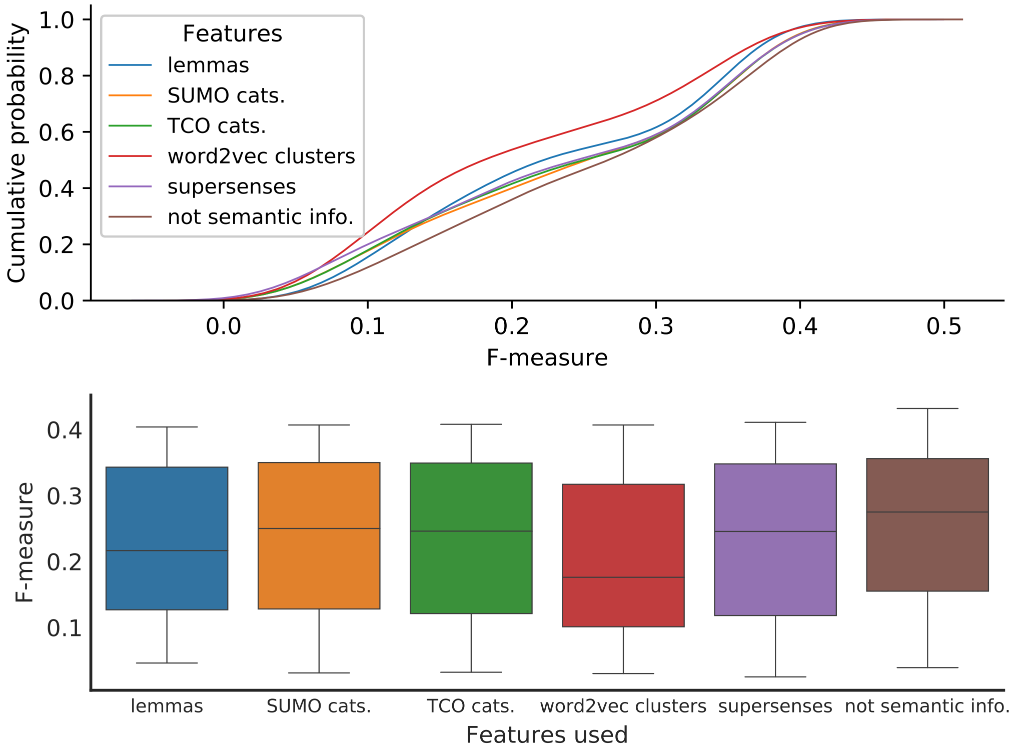


Figure 5.7: Gold standard evaluation: Semantics

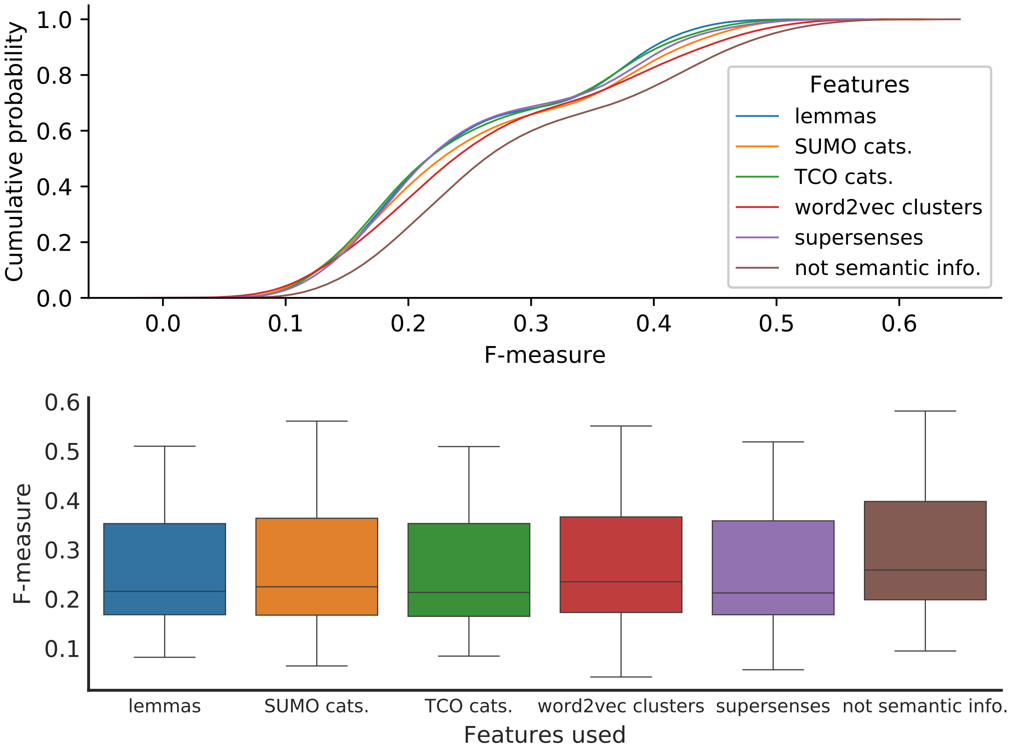


Figure 5.8: Extensibility evaluation: Semantics

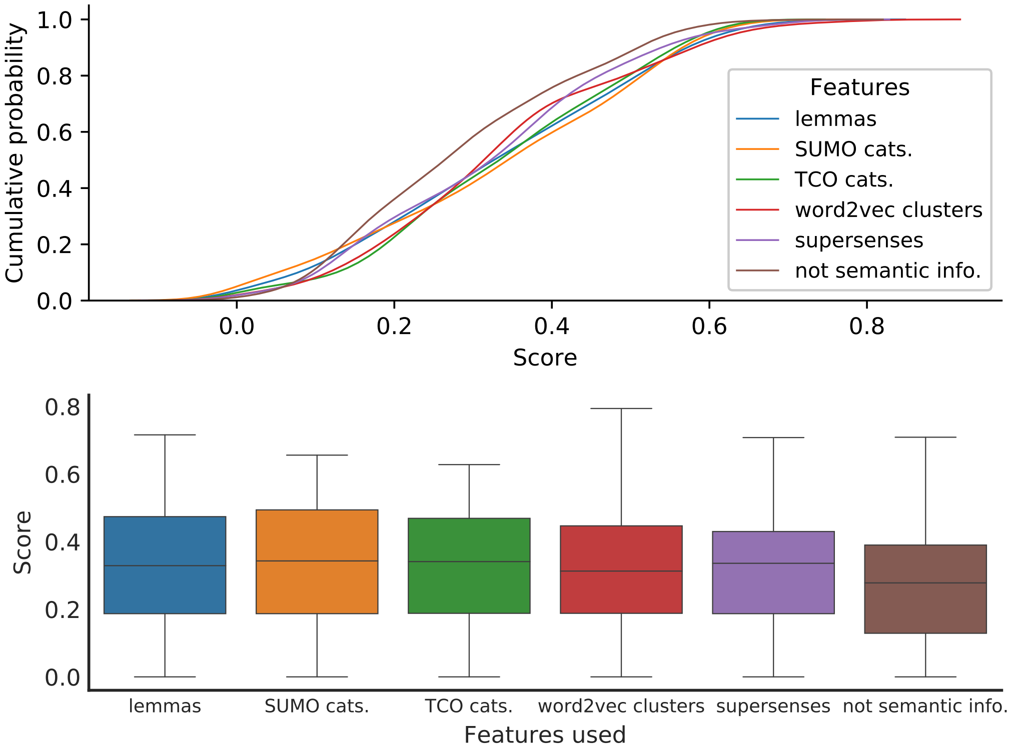


Figure 5.9: Psycholinguistic evaluation: Semantics

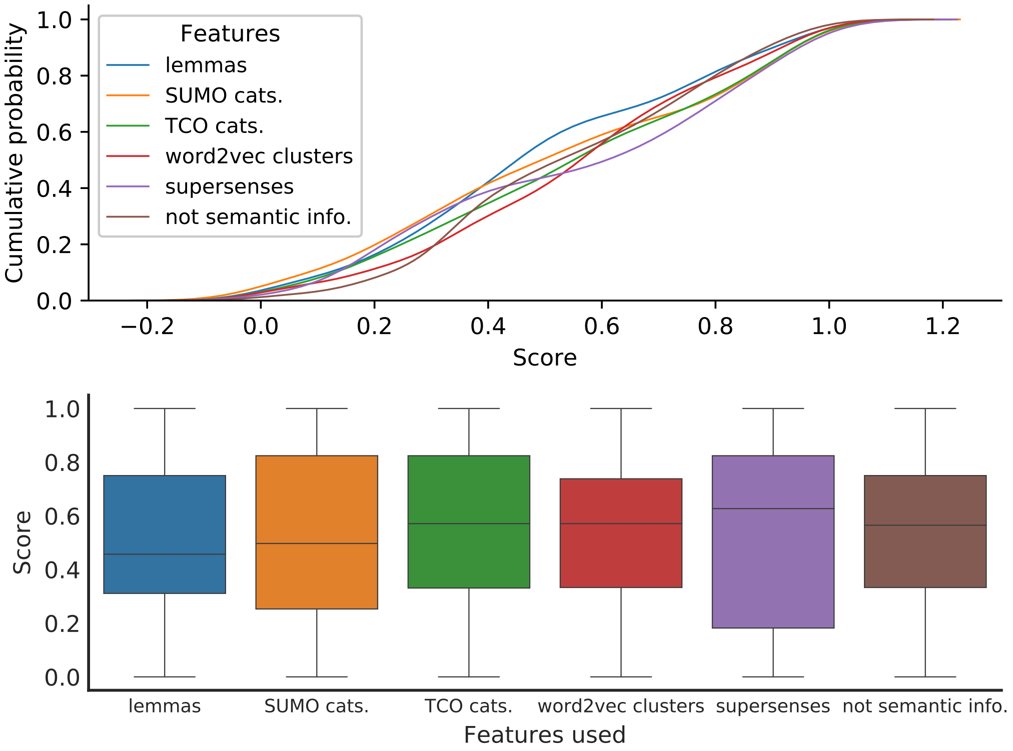


Figure 5.10: Theoretic data evaluation: Semantics

3. Syntax

   In terms of the role of the syntactic information used (syntactic categories such as *NP* versus synatctic functions such as *Subject*), we see that for the evaluation against the *gold standard* (figure 5.11), syntactic functions have a higher median and less probabilities of getting lower F_1 scores. Also, they are necessary in order to obtain the highest F_1 scores.

   This situation also holds in the case of the *extensibility evaluation* (figure 5.12), although with a less pronounced difference between the two characterizations.

   As for the *comparison with psycholinguistic data* (figure 5.13), the distribution of the syntactic functions and syntactic categories is reversed, with syntactic categories performing slightly better, and also their medians are very close, with syntactic categories having a slightly higher value.

   Regarding the *comparison with construction data* (figure 5.14), the distribution of the scores of syntactic functions and syntactic categories points to better performance of syntactic functions. For example, configurations with syntactic functions have a probability of 40% of obtaining a score under 0.6, whereas for configurations with syntactic categories this probability is 60%. Besides, in the boxplot we can see this difference reflected in the medians of each type of configuration.

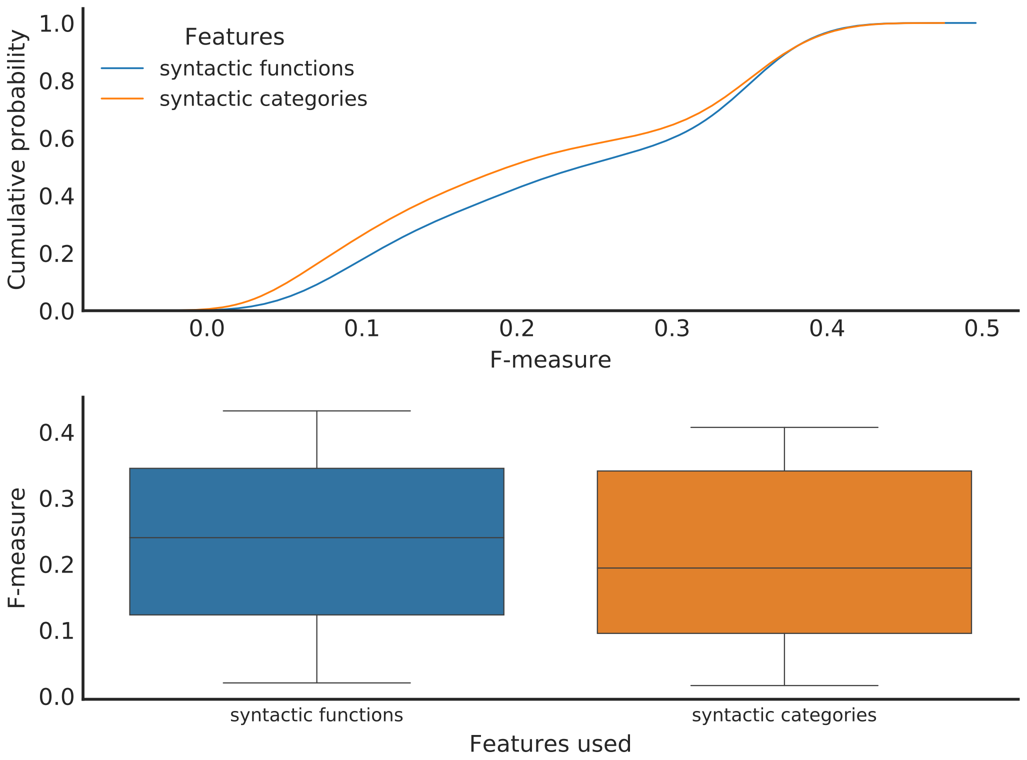


Figure 5.11: Gold standard evaluation: Syntax

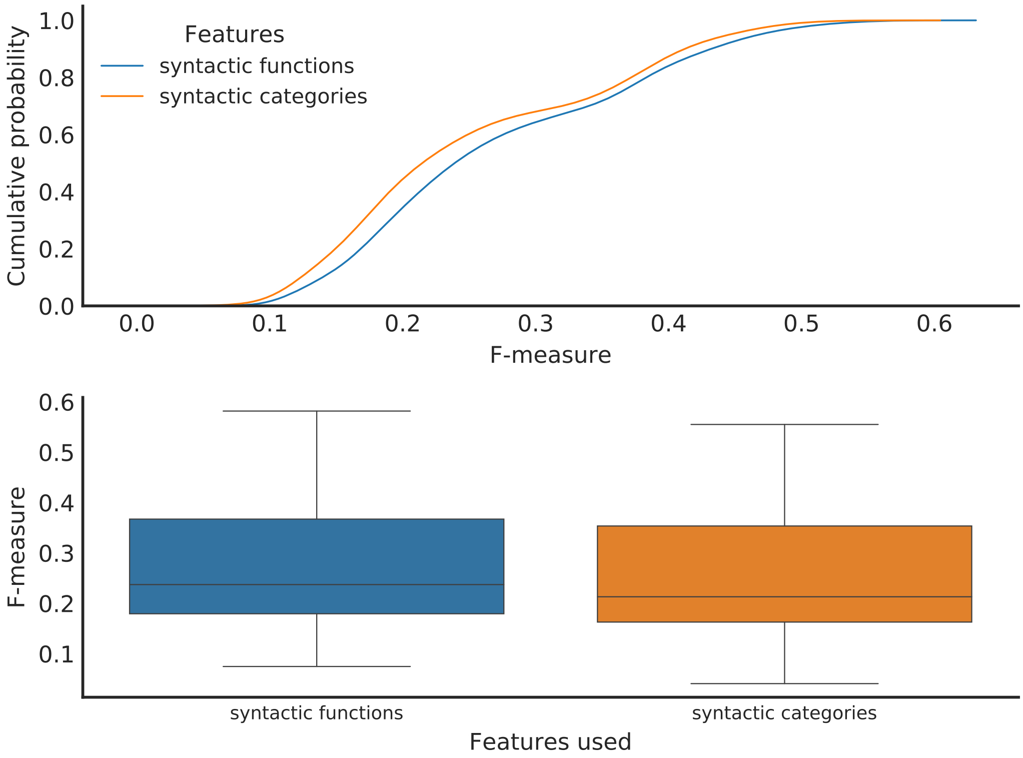


Figure 5.12: Extensibility evaluation: Syntax

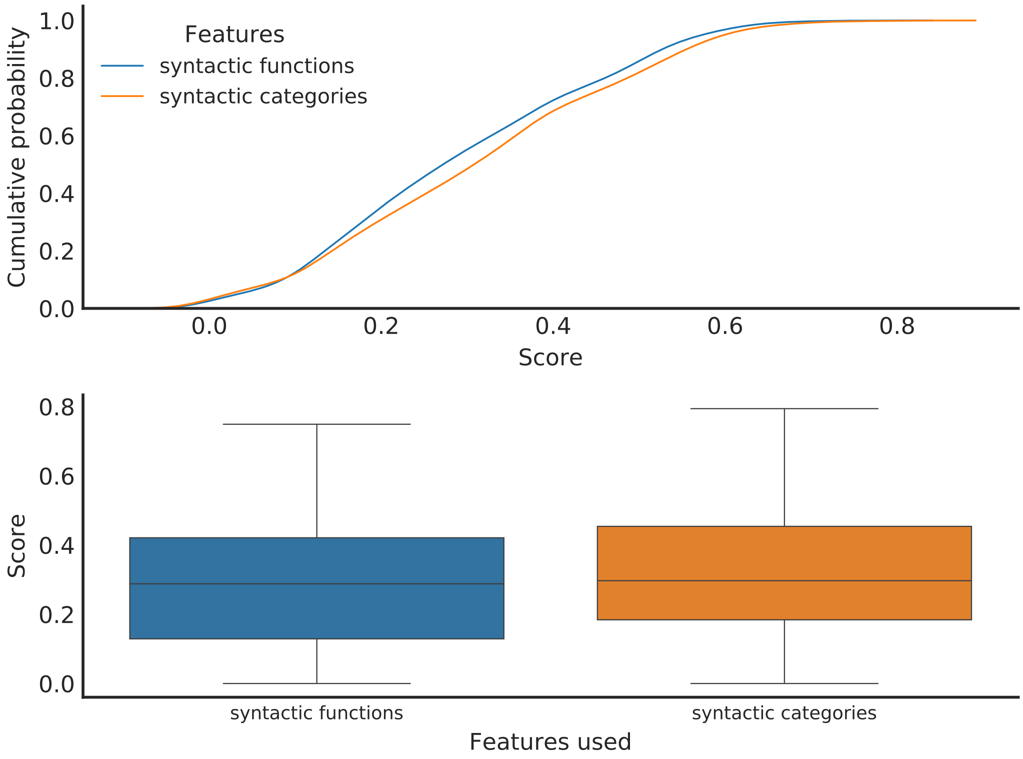


Figure 5.13: Psycholinguistic evaluation: Syntax

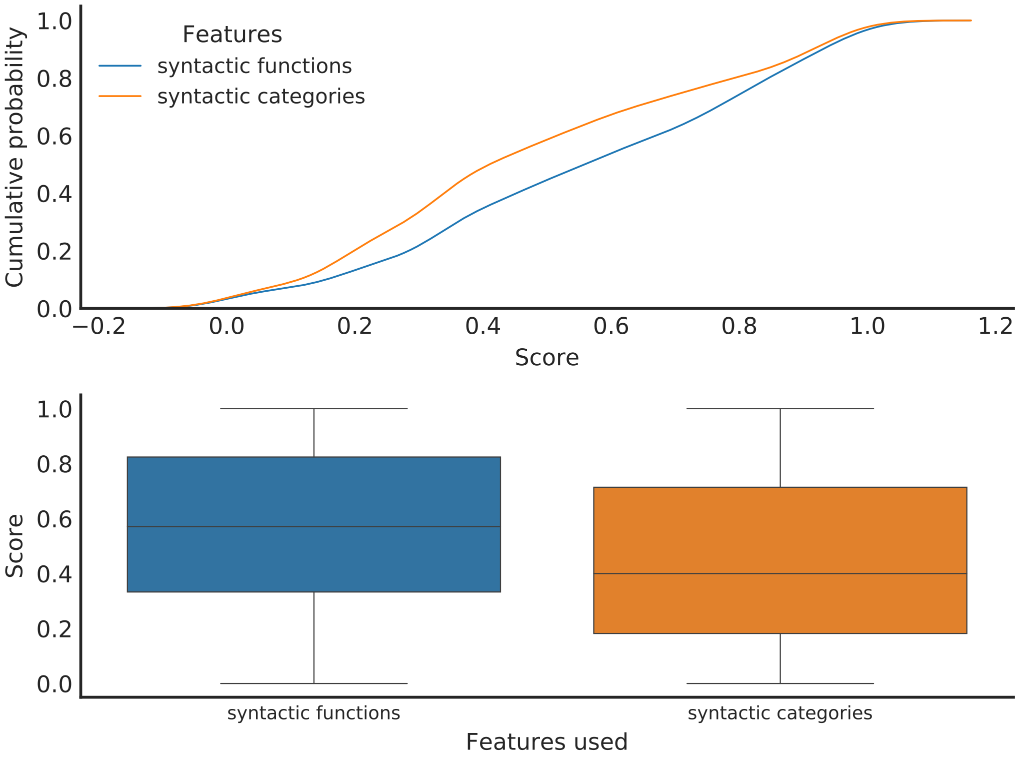


Figure 5.14: Theoretic data evaluation: Syntax

4. Configuration

In this section we examine the impact of the feature configuration in the scores obtained. There are two possible configurations that capture the argument structure in different ways: constituents (e.g. Subject_Human, Object_Human) and subcategorization frames (e.g. Subject_Human+Object_Human). In the case of the *gold standard evaluation*, *extensibility evaluation* and *the comparison with psycholinguistic data* (figures 5.15, 5.16 and 5.17 respectively) the two approaches produce very similar results, with the two configurations having very similar medians. On the contrary, in the *comparison with data from constructions* (figure 5.18) we see that configurations with subcategorization frames show a better general performance, with a 40% probability of obtaining scores over 0.7 as compared to less than 30% for constituents and a higher median.

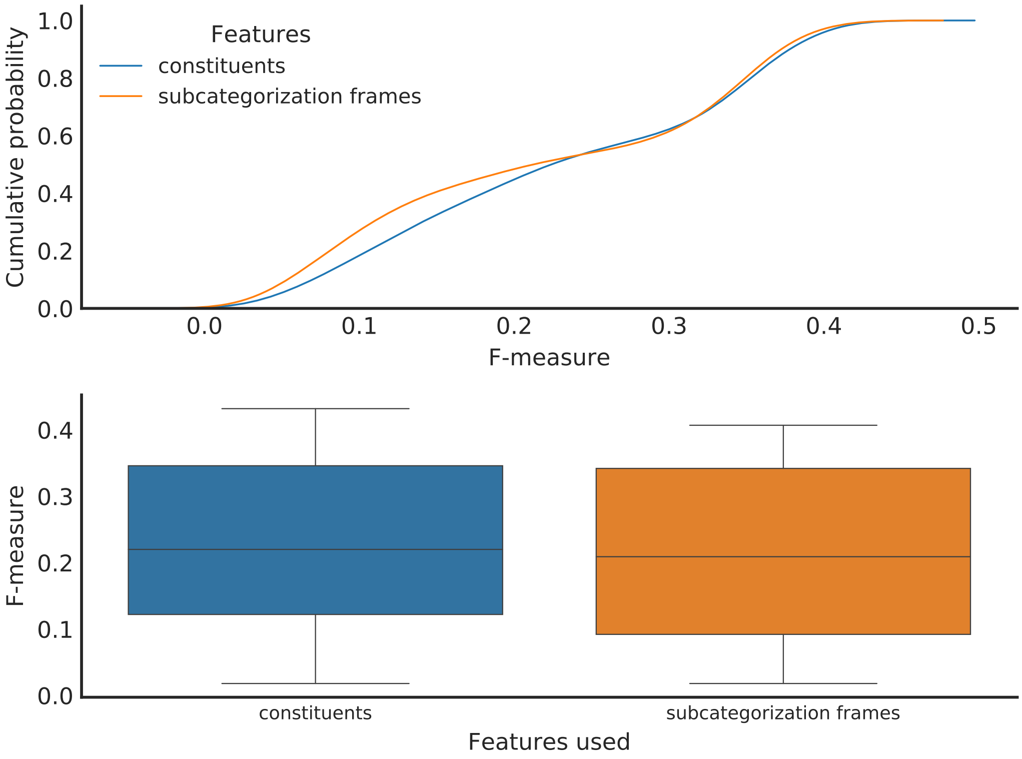


Figure 5.15: Gold standard evaluation: Configuration

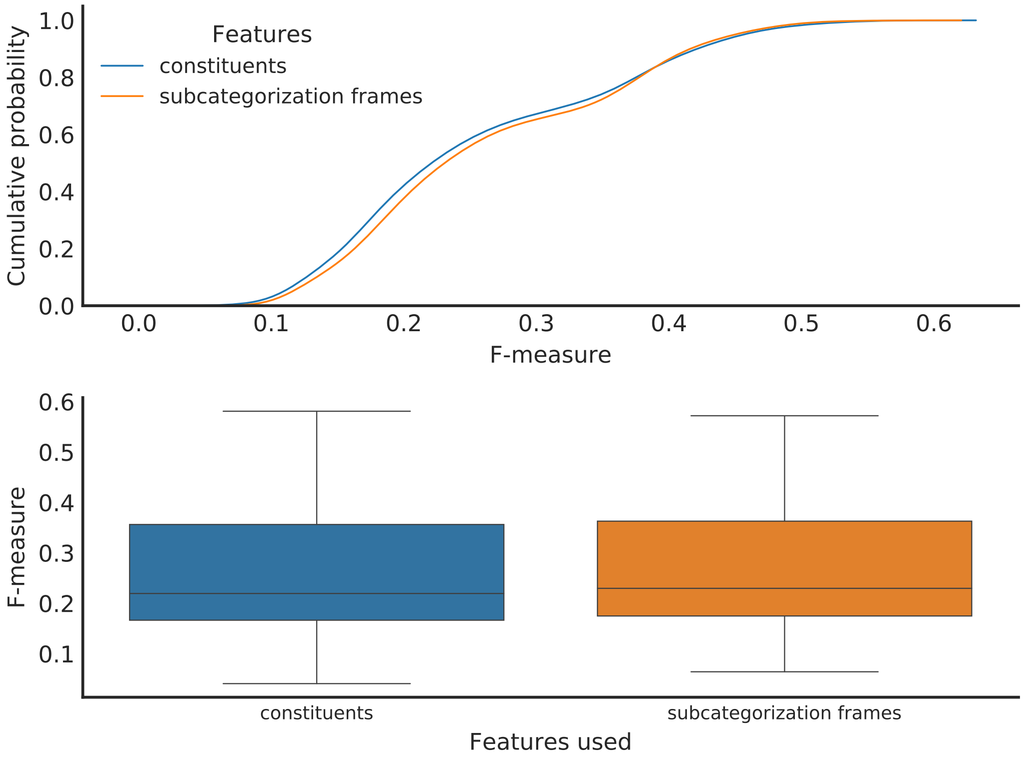


Figure 5.16: Extensibility evaluation: Configuration

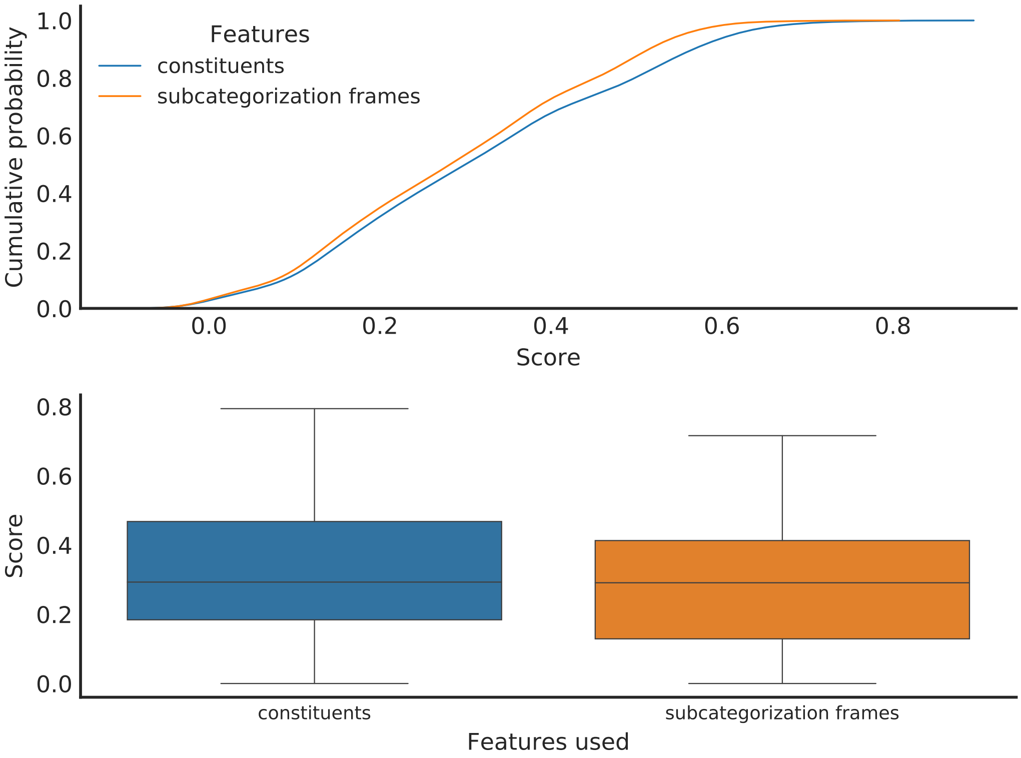


Figure 5.17: Psycholinguistic evaluation: Configuration

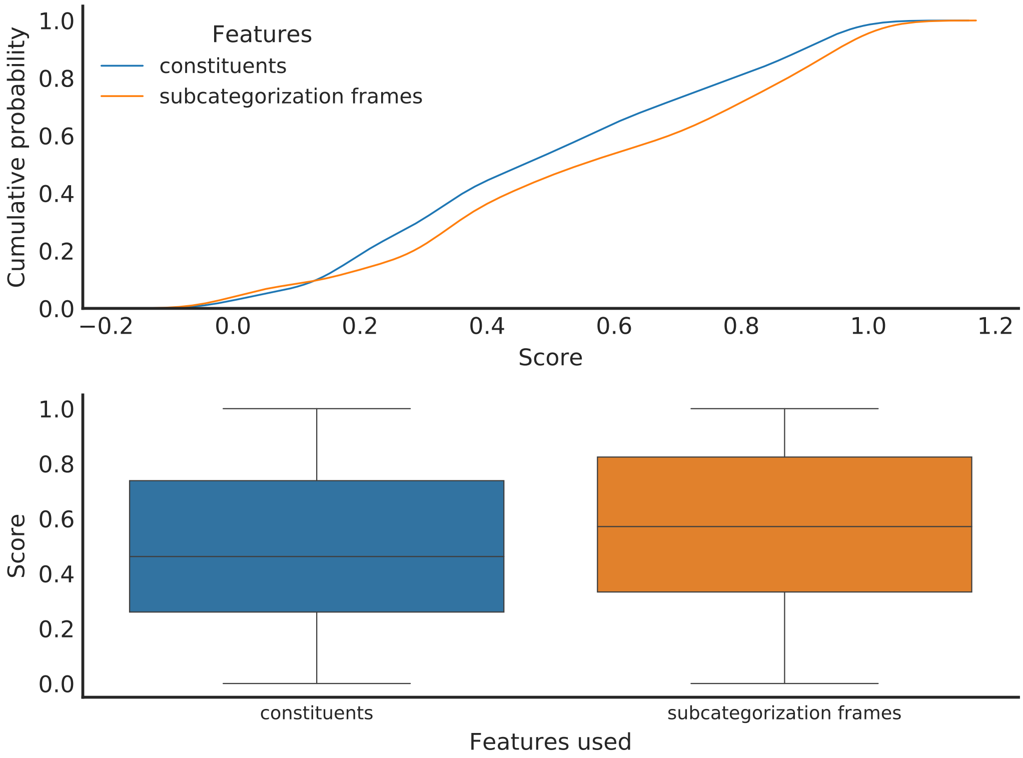


Figure 5.18: Theoretic data evaluation: Configuration

In table 5.6 we summarize the conclusions obtained in this study. With it we aimed at isolating the contribution of the individual linguistic features used in the creation of the clustering, looking at the relation between their presence and the scores obtained by the automatic classifications in the different evaluations and comparisons. The importance of this analysis lies its informativeness: taking an individual classification and looking at its scores in different evaluations and comparisons would not be enough to extract conclusions about which linguistic features produce classifications with better results because their role may be modulated by other factors such as the number of classes, the feature configuration, etc. Therefore, with this analysis, in which we take a collection of classifications and look at the influence of the individual linguistic features while controlling for all the other factors that influenced the classification, we aim at gaining a more global understanding on the contribution of the different pieces of linguistic information included in the classifications.

	Gold ev.	Extensibility ev.	Psycholinguistic comp.	Constructions comp.
Aspect	slighly positive	indifferent	slighly positive	positive
Syntax	functions	functions	categories	functions
Semantic cats.	not semantic info.	not semantic info.	semantic info.	indifferent
Configuration	indifferent.	indifferent	indifferent	s. frames

Table 5.6: Impact of the linguistic features in the classifications for each task

5.8 conclusions of the chapter

In this chapter we explained the process followed when creating a collection of automatic classifications. Afterwards, we evaluated them in two different ways: we compared them with a gold classification and we looked at their extensibility power of the best classification. Besides, we looked at how well the best classification in those evaluations kept the similarity relationships defined by word association data and construction data, which were defined in the former chapter.

As for the gold standard evaluation, an automatic classification of 14 classes created using aspect and syntactic functions, organized in constituents yielded an F_1 score of 0.43. It should be stressed here that there are no other classifications for Spanish to which we can compare directly, because they use different gold standards. Despite this, if we look at the previous experiments in automatic verb classification carried previously for languages other than English, we find that the performance of this classification is in line with their results.

In relation to the classification that looks at the extensibility of the argument structure information gathered in the classes obtained, we saw that results varied depending on the level of abstraction tested. The performance of the automatic classes for the highest level of abstraction is notably better than for the others. Nevertheless, the baselines do not improve over the classification in any of the levels. We also saw that establishing a threshold based on the distribution of the weights of the semantic role structures associated to the class helps in identifying the structures associated to the test verbs, particularly in finer grained levels of abstraction of semantic roles.

Finally, we saw that automatic classifications can organize verbs in a way that has some overlap with the output of a psycholinguistic experiment and a high overlap with the similarities defined by a theoretical linguistics approach based on constructions. It is relevant to underscore the convenience of verifying these results with a larger amount of verbs.

As for the role of the linguistic features involved in the modelization of the verbs, we saw that there were some differences according to each of the evaluations carried out, although a general behaviour can be highlighted: the presence of syntactic functions (except with psycholinguistic data) increases the probabilities of obtaining higher scores. As for the particularities, the presence of semantic categories related to the arguments is more important in order to automatically group verbs in a way which is coherent for psycholinguistic data, and the formalization in subcategorization frames and the presence of aspect is important to do so for construction data. These findings endorse the conclusions of the former chapter, namely that semantic information in corpus is relevant in order to create similarities that align with those given by psycholinguistic data and that the presence of aspect and subcategorization information in corpus is important in order to align with construction data.

III

Outcome

CHAPTER 6

Conclusions and future work

In this thesis we have explored the usefulness of verb similarity for the task of modelling events, more specifically, participants in those events. The ground for this study is set by a relevant property of verbs, namely, their relational nature, that enables them to be associated with an argument structure. This structure contains essential information about the participants in the event described by the verb. In relation to this, we have looked in detail how different perspectives of argument structure, that come from different linguistic approaches (construction theory from *linguistic theory*, word associations from *psycholinguistics* and semantic roles from *corpus linguistics*), define similarity between verb senses.

Building on these insights, we have designed an experiment with a set of automatic verb classifications, which explore a range of clustering parameters and linguistic features, with the goal of testing to what extent automatic verb classifications are able to resemble a gold standard based on argument structure. Besides, we examined the generalization power of the classification, by looking at its ability to expand the information relative to arguments that is kept within its classes to verb senses that are out the scope of the classification. This type of evaluation is crucial to prove the usefulness of the automatic classification and had not been carried out before. Finally, we have compared the similarity relationships given by the classification with the similarity defined by two of the perspectives examined in the analysis: linguistic theory (construction theory) and psycholinguistics (word associations).

Finally, after performing these comparisons and evaluations, we have delved into the analysis of the role of the linguistic features used to create the automatic classifications in relation to the results obtained. This analysis was carried out by comparing classifications that were identical except for the feature being analyzed at each point, with the goal of identifying systematic behaviours of the linguistic components of the automatic classification.

6.1 REVISITING THE INITIAL HYPOTHESIS

In section 1.3 we set the following general hypothesis:

- predicate-argument structure has regularities that can be captured and generalised with similarity-based techniques.

This hypothesis is materialized in two sub-hypothesis that direct the research towards specific goals. Next we recover those hypothesis, explaining in each case the conclusions achieved after studying the results.

- different perspectives of predicate-argument structure exhibit basic consistent behaviour in defining verb similarity.

On the quantitative side, we found significant correlations ranging from weak to medium between the similarity scores assigned by different perspectives to a group of 190 pairs of verb senses. More in detail, we found that the strength of these correlations was modulated by a series of factors such as the different granularities of semantic roles and the different word association characterizations. On the qualitative side, when we examined a range of linguistic features associated to the most similar and dissimilar pairs (semantic field, aspect and subcategorization), we found that different perspectives emphasized diverse features associated to the verbs. In particular, the more salient findings were that similarity relationships defined by psycholinguistic data were aligned with the semantic fields of the verb senses, and similarity relationships defined by construction data were aligned with the aspect of the verb senses. All in all, there is indeed basic consistent behaviour in defining verb similarity. However, it can also be observed that, when we look closely, there are differences at the quantitative and qualitative levels.

- A classification that is automatically created from corpus data can reveal patterns of homogeneous linguistic behaviour which is representative enough to represent predicate-argument structure successfully and, also, to generalize verb behaviour with respect to argument structure.

The results of the evaluation and comparisons carried out indicate that an automatic verb classification is able to organize the information in a way that aligns with the basic argument structure of the verbs, and that is able to extend the information in the classes to verbs that are outside of the classification (test verbs) more successfully than purely unsupervised methods or than the most similar sense to the test verbs.

To wrap up, if we recover the main hypothesis, which stated that *predicate-argument structures have regularities that can be captured and generalised in order to capture event participant information*, we can conclude that there are basic regularities that can be captured and extended to provide information for scenarios for which we do not have information. Besides, classifications have enough flexibility to accommodate the similarity relationships defined by construction theory and more than a third of those defined by word associations. Therefore, the hypotheses have been substantiated.

6.2 CONTRIBUTIONS

The analysis that we carried out in chapter 4 allowed us to, first, **quantify the common ground between the different perspectives on argument structure (linguistic theory, corpus data, psycholinguistic data)**. The other relevant finding in this chapter is on the qualitative side: we found that **each perspective tended to establish the axis similarity-dissimilarity onto different linguistic features**. Based on the idea that verbs in similar pairs are expected to show common linguistic features (same aspect, same semantic field, large amount of subcategorization) and verbs in dissimilar pairs are expected to exhibit different features (different aspect, different semantic field and a low amount of subcategorization), we examined up to which point this was actually the case for the different perspectives explored. We saw that **aspect was an important marker for similarity according to the linguistic theory based on constructions** (the verb senses in similar pairs tended to share aspect whereas senses in dissimilar pairs tended to have different aspect) and that **semantic field was coherent with similarity relations obtained from psycholinguistic data** (verb senses in similar pairs tended to share the semantic field whereas senses in dissimilar pairs tended to have different semantic fields). Corpus data presented an intermediate step between the other two perspectives: **similarities based on corpus are also structured by semantic fields, although less consistently than the case of psycholinguistic data, and by aspect, although less strongly than construction data**. Nevertheless, more in general, members of the similar pairs tended to share more categories than members of the dissimilar pairs, as it was expected. However, a relevant exception to this were subcategorization frames: **in general and more clearly in the psycholinguistic approach, subcategorization frames alone were not helpful in telling apart similar verbs from dissimilar verbs**, which supports the idea that, at least, they need to be enriched with other types of information. These findings were used as a guide when creating feature sets in order to characterize verb senses for the automatic classification.

As for the automatic classification described in chapter 5, this work makes a contribution to the area of the **applications of verb classification in Spanish**, a language that is not nearly as explored for this topic as English. Besides creating a classification and comparing it to a gold classification based on semantic roles, **we focus on applying the classification to gathering information about the semantic role structures that combine with verb senses, which has direct connections with detecting participants**

in events. This last area that has been mostly overlooked in other languages and has not been treated in Spanish for verb classifications.

In relation with the comparisons with psycholinguistic data and construction data, **we verified that the automatic classification that best resembled the structure of a semantic role-based classification was also able to account for similarity relations found in other perspectives**: construction theory and, to a lesser extent, those found in psycholinguistic data. Besides, a study of the role of the linguistic features in the collection of classifications created showed that **the linguistic ingredients that were found relevant in the pairwise-based perspective study (chapter 4) were also important in order to recreate those relations within the automatic classifications**. An additional finding of chapter 4 that was endorsed by the feature study (section 5.7) is that **the information supplied by subcategorization frames is on pair with that provided by single arguments**, except from the case of similarities obtained by construction theory.

6.3 LIMITATIONS AND FUTURE WORK

The two main limitations identified have to do with the amount of data available and the resources used. As for the first issue, in the context of the study of the perspectives of verb similarity, having more verbs included in the study would be desirable in order to refrend the results obtained. In the context of the automatic classification, it would also be informative to replicate the extensibility experiment and the comparisons with psycholinguistic and construction data with more verb senses, to better asses the reliability of the results.

In relation to the second issue, the resources used, we think that in the case of the automatic classification, it would be interesting to experiment more extensively with unsupervised information, such as embeddings obtained from distributional data, as it would make the results less dependant on the resources available. This type of information may add richness in the representation of linguistic phenomena in languages with less handcrafted resources and it can alleviate the cost of porting information between languages in the case of crosslinguistic tasks.

On the opposite end in terms of the cost of the resource we find syntactic dependencies. Few automatic classifications make use of dependency information, as it is costly to obtain and not straightforward to integrate. However, given the reliability of this type of information for other tasks we think that it would be informative to assess its performance in automatic classification.

Another element of concern is the semantic information associated to the arguments that we used in the automatic classification, because it did not prove to be helpful except in the case of the comparison with psycholinguistic data. In future work we would like to look more in deep at this issue, and to analize other ways of integrating this information because there is still room for improvement in Spanish and purely syntactic information has proven its limitations for other languages.

Another issue open for experimentation in this work is expanding the amount of automatic annotation used. We relied on manually annotated information in some cases (delimitation of arguments, their syntactic category and function, the aspect of the verbs). We think that an interesting future avenue for research would entail substituting this manually annotation for automatically annotated one and re-running the evaluations to asses the impact of switching to automatic annotation.

Finally, we think that the automatic modellization of event information could be applied to other tasks that deal with this type of content, such as event type recognition.

On a different note, in the context of the multiperspective approach we saw that there was consistent behaviour across the various perspectives (linguistic theory, constructions and psycholinguistics) regarding the status of *similar* or *dissimilar* of verb senses. Therefore, a question that may arise, beyond the qualitative and quantitative analysis is the following: are there specific verb senses that are considered as similar or dissimilar consistently across perspectives?

To answer this, we can consider that each perspective votes a verb sense pair as similar or dissimilar if it is among the most similar or dissimilar pairs for that perspective.[1] According to this criterion, there are four pairs that obtain votes as similar from all three perspectives in a consistent manner (see between parenthesis their specific weight of their vote, over a maximum value of 3): *volver-viajar* (1.85), *abrir-cerrar* (2.83), *valorar-pensar* (2.21) and *explicar-escuchar* (1.56). Conversely, there are pairs that obtain votes as dissimilar from all three perspectives also in a consistent manner: *dormir-parecer* (1.99) and *parecer-cerrar* (1.95).

As we see, the perspectives agree on the status of specific verb pairs: the similar pairs that we just listed combine with similar semantic roles, constructions and received comparable responses in the word association experiment, and the opposite situation also holds for the dissimilar pairs. However, the motivation behind this situation needs to be more thoroughly explored. For example, there is not a specific semantic relation (synonymy, antonymy, troponymy, etc) that systematically holds between the members of all these pairs and that can be clearly identified. Thus, more research into the continuum similarity-dissimilarity, the features of the elements that are at the edges of this continuum and the robustness of their status is needed. In this regard, we think that theories that take into account the context in which linguistic exchanges occur, such as multimodal accounts of linguistic knowledge (Barsalou et al., 2008) may be an interesting avenue for research.

[1] Since perspectives may have different formalizations, the weight of this vote should be proportional to the agreement between the different formalizations of that perspective: the weight is 1 if all of them agree, and it decreases in proportion to the number of formalizations that do not consider the verb pair among the most similar or dissimilar.

6.4 RESOURCES AND PUBLICATIONS

Two resources for the study of verb similarity were created:

- A collection of linguistic constructions, defined for Spanish according to construction theory and freely available, associated to a set of Spanish verb senses. It can be downloaded from the following link: https://zenodo.org/record/3603387#.Xhd_xZhKjrc
- A resource that contains a total of 11,617 word association responses to a set of verb senses. The main source for word associations are nouns and therefore this resource, also freely available, contributes to provide information for the less studied category of verbs. It can be downloaded from the following link: https://zenodo.org/record/3603398#.XheEG5hKjrc

Additionally, the code developed to perform this study is freely available at https://github.com/aralittle/verb-classification

The main publications related to this thesis are listed next.

- Gil-Vallejo L., I. Castellón, M. Coll-Florit (2015). "Hacia una definición de la similitud verbal para la extracción de eventos", *e-AESLA. Revista digital de Lingüística Aplicada, Sección Lingüística de corpus, computacional e ingeniería lingüística*, 1, p. 1-15. ISSN: 2444-197X
- Gil-Vallejo, L. (2015). "Exploiting verb similarity for event extraction", *Proceedings of the Spanish Society for Natural Language Processing (SEPLN)*. Doctoral Symposium.
- Gil-Vallejo, L., I. Castellón, M. Coll-Florit, J. Turmo (2015). "Hacia una clasificación verbal automática para el espaÃśol: estudio sobre la relevancia de los diferentes tipos y configuraciones de información sintáctico-semántica", *Linguamática*, 7:1, p. 419-452. ISSN: 1647-0818
- Gil-Vallejo, L., I. Castellón, M. Coll-Florit (2018). "Similitud verbal: Análisis comparativo entre lingüística teórica y datos empíricos extraídos de corpus", *Revista Signos: Estudios de Lingüística*, 51:98, p. 310-332. ISSN: 0035-0451
- Gil-Vallejo, L., M. Coll-Florit, I. Castellón, J. Turmo (2018). "Verb similarity: Comparing corpus and psycholinguistic data", *Corpus Linguistics and Linguistic Theory*, 14:2, p. 275-307. ISSN: 1613-7035

IV

Appendix and bibliography

APPENDIX A

Appendix

A.1 Verb senses, definitions and frequency

In this section we present the 20 verb senses selected for the experiments described in 4, together with their definitions and frequencies

- **Abrir** 18: Move the latch, unlock, unclick any piece that closes something. (15)
- **Cerrar** 19: Secure with a lock, a latch or other instrument a door, window, lid, etc. to stop it from opening. (14)
- **Crecer** 1: Increase the amount or the importance of something, develop. (116)
- **Dormir** 1: Remain in a state in which there are no voluntary movements, usually in order to rest. (99)
- **Escuchar** 1: Listen, pay attention to what is being heard. (107)
- **Estar** 14: To be something or someone in a specific state. (101)
- **Explicar** 1: Explain, give information about a specific issue. (106)
- **Gestionar** 1: Manage, go through a procedure to achieve an objective. (36)
- **Gustar** 1: Like, find somebody or something appealing. (117)
- **Montar** 2: Get in a vehicle or get on top of an animal. (26)
- **Morir** 1: Die; cease to exist (somebody or something). (115)
- **Parecer** 1: To pretend to be something without necessarily being it. (51)
- **Pensar** 2: Think, reason, examine an idea. (25)
- **Perseguir** 1: Chase somebody or pursue something in order to reach it. (53)
- **Trabajar** 1: Work, do a specific task or job. (80)

- **Valorar** 2: Value, recognize the importance of a fact, thing or action. (70)
- **Valer** 1: For something to have a specific value. (45)
- **Ver** 1: See, perceive through the eyes. (86)
- **Viajar** 1: Travel, go from one place to another distant one, usually in a means of transportation. (111)
- **Volver** 1: Return, go back to a place where one has already been. (84)

A.2 SEMANTIC FIELDS OF THE VERB SENSES

Verb senses and their semantic fields (WordNet supersenses).

Verb sense	Semantic field
Abrir (open)	change
Cerrar (close)	change
Crecer (grow)	change
Dormir (sleep)	activity (bodily)
Escuchar (listen)	perception
Estar (to be)	state
Explicar (explain)	communication
Gestionar (manage)	activity (social)
Gustar (like)	cognition
Crecer (grow)	change
Montar (get in/on)	movement
Morir (die)	change
Parecer (seem)	state
Pensar (think)	cognition
Perseguir (chase)	movement
Trabajar (work)	activity
Valorar (assess)	communication
Valer (cost)	state
Ver (see)	perception
Viajar (travel)	movement
Volver (turn back)	movement

Table A.1: Selected senses and their semantic fields

A.3 LIST OF STIMULI USED IN THE PSYCHOLINGUISTIC EXPERIMENT

- ABRIR una puerta. /TO OPEN a door.
- CERRAR una ventana. /TO CLOSE a window
- CRECER a cierto ritmo. /TO GROW at a certain rate
- DORMIR durante un rato. /TO SLEEP for a while
- ESCUCHAR atentamente. /TO LISTEN carefully

- ESTAR en una determinada condici on. /TO BE in a specific state
- EXPLICAR una cuesti on. /TO EXPLAIN an issue
- GESTIONAR un tr amite. /TO HANDLE a procedure
- GUSTAR mucho. /TO LIKE a lot
- MONTAR en un vehículo. /TO GET IN a car
- MORIR alguien. /TO DIE somebody
- PARECER fuerte. /TO SEEM strong
- PENSAR en un asunto. /TO THINK about an issue
- PERSEGUIR a una persona. /TO CHASE a person
- TRABAJAR en algo. /TO WORK in something
- VALORAR la importancia de algo. /TO ASSESS the importance of something
- VALER dinero. /TO COST money
- VER una imagen. /TO SEE an image
- VIAJAR en un medio de transporte. /TO TRAVEL in a means of transportation
- VOLVER a un lugar. /TO GO BACK to a place

A.4 VERB SENSES IN TRAINING PARTITION

abrir_1, abrir_17, abrir_18, abrir_7, abrir_9, acabar_1, acabar_3, acabar_4, acabar_7, acabar_9, acceder_1, acceder_2, acceder_3, aceptar_1, aceptar_2, acercar_2, acercar_3, acercar_5, aclarar_2, aclarar_7, acompañar_1, acompañar_2, acordar_1, acordar_2, actuar_2, actuar_3, actuar_4, actuar_5, actuar_6, acudir_1, acudir_3, adquirir_1, adquirir_2, afectar_1, afectar_4, afirmar_2, agradecer_1, agradecer_3, alcanzar_3, alcanzar_6, alcanzar_8, analizar_1, anunciar_1, aparecer_1, aparecer_2, aparecer_3, aparecer_4, aparecer_5, aparecer_6, apostar_2, apostar_3, aprovechar_1, aprovechar_2, apuntar_1, apuntar_10, apuntar_2, apuntar_3, apuntar_4, apuntar_6, apuntar_8, arreglar_2, arreglar_3, arreglar_6, arreglar_7, asegurar_2, asegurar_4, añadir_1, añadir_2, bajar_2, bajar_4, bajar_5, bajar_7, bajar_9, beneficiar_1, beneficiar_2, buscar_1, buscar_3, caber_1, caber_2, calificar_1, cambiar_1, cambiar_7, casar_1, casar_3, ceder_1, ceder_2, ceder_3, ceder_4, celebrar_1, celebrar_2, celebrar_3, cerrar_13, cerrar_19, cerrar_2, cerrar_3, cerrar_6, cerrar_7, citar_1, citar_2, citar_3, citar_4, coincidir_1, coincidir_2, coincidir_3, coincidir_4, comentar_1, comenzar_1, comenzar_3, comer_1, comer_6, compartir_1, compartir_2, compartir_3, completar_1, comprobar_1, comprometer_2, comprometer_3, comprometer_5, conceder_1, conceder_3, concentrar_1, concentrar_3, concentrar_4, concentrar_5, conducir_1, conducir_2, conducir_3, conducir_5, conducir_6, confesar_1, confesar_2, confesar_3, conocer_1, conocer_3, conocer_5, conocer_7, conocer_8, conseguir_1, considerar_1, considerar_4, consistir_1, consistir_2, constituir_1, constituir_2, constituir_3, construir_1, construir_2, contar_1, contar_3, contar_4, contar_6, contener_1, contener_2, continuar_1, continuar_2, continuar_3, continuar_4, continuar_5, contratar_1, contratar_2, contribuir_1, contribuir_2, controlar_1, convertir_1, convocar_1, convocar_2, corresponder_1, corresponder_2, corresponder_3, costar_1, costar_2, crear_1, crecer_1, creer_1, creer_2, creer_4, cumplir_1, cumplir_2, cumplir_4, cumplir_5, dar_1, dar_2, dar_6, dar_7, dar_8, deber_1, deber_3, decidir_1, decir_1, dedicar_1, dedicar_2, dedicar_3, defender_1, defender_3, definir_1, definir_3, definir_4, dejar_1, dejar_4, dejar_6, dejar_7, dejar_8, demostrar_1, demostrar_2, denunciar_1, denunciar_2, depender_1, depender_2, desaparecer_1, desaparecer_2, desaparecer_3, desarrollar_1, desarrollar_3, descartar_1, descubrir_2, descubrir_3, descubrir_4, desear_1, desplazar_1, desplazar_2, desplazar_3, destacar_1, destacar_2, destinar_1, destinar_2, destinar_3, detectar_1, detener_1, detener_2, detener_4, devolver_2, devolver_4, dirigir_1, dirigir_2, dirigir_3, dirigir_4, disponer_4, dormir_1, dormir_4, dormir_8, durar_1, durar_2, echar_1, echar_15, echar_2, echar_3, efectuar_1, elaborar_1, elaborar_2, empezar_1, encargar_1, encargar_2, encargar_3, encontrar_1, encontrar_3, entender_1, entender_2, entender_4, entender_5, entender_7, entrar_1, entrar_3, entrar_8, escribir_1, escuchar_1, escuchar_2, esperar_1, esperar_2, esperar_3, esperar_4, estar_14, estar_9, estudiar_1, estudiar_2, evitar_1, evitar_2, existir_1, existir_2, explicar_1, facilitar_2, faltar_1, fijar_2, fijar_3, fijar_4, financiar_1, formar_1, formar_2, formar_4, ganar_1, ganar_2, ganar_5, ganar_8, gastar_1, gastar_3, gestionar_1, gestionar_2, gustar_1, hablar_1, hablar_5, hacer_1, hacer_14, hacer_15, hacer_16, hallar_1, hallar_2, hallar_3, hallar_5, identificar_1, identificar_3, identificar_4, identificar_5, impedir_1, implicar_1, implicar_2, implicar_3, impulsar_1, impulsar_3, inaugurar_1, incluir_1, incluir_2, incorporar_2, incorporar_3, indicar_1, indicar_2, iniciar_3, instalar_1, instalar_2, instalar_3, integrar_1, integrar_2, integrar_3, intentar_1, interpretar_1, interpretar_3, intervenir_1, intervenir_2, invertir_1, invertir_2, invertir_3, ir_1, ir_13, ir_14, ir_2, ir_6, jugar_1, jugar_2, jugar_3, jugar_5, jugar_6, juzgar_1, juzgar_2, lamentar_1, lamentar_2, limitar_1, limitar_2, limitar_4, llamar_2, llamar_3, llamar_4, llamar_5, llamar_6, llegar_3, llegar_6, llegar_7, llenar_3, llevar_1, llevar_11, llevar_12, llevar_2, llevar_3, llevar_4, lograr_1, manifestar_1, manifestar_2, manifestar_3, manifestar_4, manifestar_5, mantener_2, mantener_3, mantener_5, mantener_6, marcar_2, marcar_8, marcar_9, merecer_1, montar_1, montar_2, montar_4, montar_5, morir_1, morir_2, mostrar_2, mostrar_3, mover_1, mover_2, mover_6, mover_7, nacer_1, nacer_2, nacer_3, necesitar_1, negociar_1, obligar_1, observar_1, observar_3, ocupar_1, ocupar_5, ocurrir_1, ocurrir_2, ofrecer_1, ofrecer_5, ordenar_1, ordenar_2, oír_1, oír_2, oír_3, pagar_1, pagar_2, parar_1, parar_2, parar_5, parar_6, parecer_1, parecer_2, parecer_3, parecer_4, partir_1, partir_3, partir_5, pasar_10, pasar_2, pasar_3, pasar_6, pedir_1, pensar_2, pensar_4, pensar_5, perder_1, perder_3, perder_4, perder_6, permitir_1, permitir_2, perseguir_1, perseguir_2, pertenecer_1, pertenecer_2, plantear_1, plantear_2, plantear_3, poner_1, poner_14, poner_2, poner_3, poseer_1, practicar_3, precisar_1, precisar_2, preferir_1, presentar_1, presentar_2, presentar_3, presentar_7, pretender_1, pretender_2, prever_1, prever_2, proceder_1, proceder_2, proceder_3, producir_2, producir_4, prometer_2, proponer_1, provocar_1, quedar_1, quedar_2, quedar_3, quedar_4, quedar_8, quejar_1, querer_1, realizar_1, recibir_1, recibir_3, reclamar_1, reclamar_2, reclamar_3, recoger_1, recoger_2, recoger_4, recoger_8, reconocer_2, reconocer_3, reconocer_5, recordar_1, recordar_2, recordar_3, recorrer_1, recorrer_2, recuperar_1, recuperar_2, recuperar_6, reducir_2, reducir_5, rendir_1, rendir_2, rendir_3, rendir_4, renovar_1, renovar_2, renovar_3, repartir_1, repartir_2, repartir_3, repartir_5, repetir_1, repetir_5, residir_1, residir_2, resolver_1, resolver_3, resultar_1, resultar_2, reunir_1, reunir_2, reunir_3, saber_2, saber_3, saber_4, salir_1, salir_11, salir_13, salir_16, salir_2, salir_3, salir_5, salvar_1, salvar_3, salvar_5, seguir_1, seguir_10, seguir_2, seguir_3, seguir_7, seguir_8, sentar_1, sentar_2, sentar_6, sentir_1, sentir_5, servir_1, servir_4, señalar_2, señalar_3, significar_1, situar_1, situar_4, solicitar_1, solucionar_1, sorprender_1, sorprender_2, sorprender_3, sorprender_4, sostener_1, sostener_2, sostener_5, subrayar_1, subrayar_2, suceder_2, sufrir_1, sufrir_2, sufrir_3, superar_1, superar_2, superar_3, suponer_1, suspender_2, suspender_3, suspender_4, tardar_1, tener_1, tener_10, tener_9, terminar_1, terminar_3, terminar_4, terminar_9, tocar_1, tocar_11, tocar_3, tocar_5, tocar_8, tomar_1, tomar_10, tomar_11, tomar_12, tomar_13, tomar_3, tomar_4, tomar_6, trabajar_1, trabajar_6, trabajar_8, traer_1, traer_2, traer_3, tratar_10, tratar_8, tratar_9, usar_2, utilizar_1, valer_1, valer_4, valer_8, valorar_1, valorar_2, valorar_3, vender_1, venir_1, venir_4, ver_1, ver_5, ver_6, ver_7, viajar_1, viajar_3, vivir_1, vivir_2, vivir_3, vivir_4, volver_1, volver_2, volver_6, votar_1

a.5 verb senses in test partition

abrir_16, abrir_5, acercar_1, acercar_9, acompañar_4, acompañar_7, afectar_2, alcanzar_5, alcanzar_9, apostar_5, arreglar_1, bajar_3, bajar_8, cerrar_1, compartir_4, completar_2, conocer_6, dar_3, desarrollar_4, devolver_1, disponer_1, disponer_3, empezar_3, encontrar_2, encontrar_6, explicar_4, facilitar_1, faltar_2, hablar_3, identificar_2, impulsar_2, incorporar_1, integrar_4, interpretar_2, ir_11, llamar_1, llegar_4, mostrar_1, mover_3, ocupar_3, ofrecer_2, participar_1, partir_2, pasar_16, pasar_4, perder_2, plantear_4, practicar_2, practicar_4, presentar_6, prometer_1, proponer_3, realizar_2, recoger_5, reivindicar_1, resolver_2, resolver_4, salir_6, seguir_5, sentir_3, significar_3, subrayar_3, suceder_1, suponer_2, tomar_8, tratar_1, usar_1, valer_2, venir_2, volver_5

Bibliography

Abeillé, Anne et al. (2003). "Building a treebank for French". In: *Treebanks*. Springer, pp. 165–187.

Agirre, Eneko et al. (2009). "Personalizing pagerank for word sense disambiguation". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 33–41.

Alonso Alemany, Laura et al. (2007). "Inducción de clases de comportamiento verbal a partir del corpus SENSEM". In: *Procesamiento del Lenguaje Natural, nº 39 (sept. 2007), pp. 123-130.*

Altmann, Gerry T. M. et al. (2007). "The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing". In: *Journal of Memory and Language* 57.4. Language-Vision Interaction, pp. 502–518. issn: 0749-596X.

Álvez, Javier, Jordi Atserias, et al. (2008). "Consistent annotation of eurowordnet with the top concept ontology". In: *Proceedings of Fourth International WordNet Conference (GWC-08).*

Álvez, Javier, Paqui Lucio, et al. (2012). "Adimen-SUMO: Reengineering an ontology for first-order reasoning". In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 8.4, pp. 80–116.

Aminian, Maryam et al. (2013). "Unsupervised Induction of Persian Semantic Verb Classes Based on Syntactic Information". In: *Language Processing and Intelligent Information Systems*. Springer, pp. 112–124.

Atserias, Jordi et al. (2004). "The meaning multilingual central repository". In: *Proceedings of the 2nd International Global WordNet Conference*. Brno: Masaryk University, pp. 80–201.

Baker, Collin F. et al. (1998). "The Berkeley Framenet project". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 86–90.

Baker, Simon et al. (2014). "An Unsupervised Model for Instance Level Subcategorization Acquisition." In: *EMNLP*, pp. 278–289.

Balota, David A. et al. (1984). “Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage.” In: *Journal of Experimental Psychology: Human perception and performance* 10.3, p. 340.

Banerjee, Satanjeev et al. (2003). “Extended gloss overlaps as a measure of semantic relatedness”. In: *Ijcai*. Vol. 3, pp. 805–810.

Barsalou, Lawrence W. et al. (2008). “Language and simulation in conceptual processing”. In: *Symbols, embodiment, and meaning*, pp. 245–283.

Bicknell, Klinton et al. (2010). “Effects of event knowledge in processing verbal arguments”. In: *Journal of memory and language* 63.4, pp. 489–505.

Bird, Steven et al. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O’Reilly Media, Inc."

Blei, David M. et al. (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Bonial, Claire et al. (2011). “A hierarchical unification of LIRICS and VerbNet semantic roles”. In: *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*. IEEE, pp. 483–489.

Bosque, Ignacio et al. (1999). *Gramática descriptiva de la lengua española*. Madrid: Espasa.

Brainerd, C. J. et al. (2008). “Semantic processing in “associative” false memory”. In: *Psychonomic Bulletin & Review* 15.6, pp. 1035–1053.

Brew, Chris et al. (2002). “Spectral clustering for german verbs”. In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, pp. 117–124.

Brown, Susan Windisch et al. (2014). “VerbNet class assignment as a WSD task”. In: *Computing Meaning*. Springer, pp. 203–216.

Burke, Robin D. et al. (1997). “Question answering from frequently asked question files: Experiences with the faq finder system”. In: *AI magazine* 18.2, p. 57.

Burnard, Lou (2000). *Reference guide for the british national corpus (world edition)*.

Bybee, Joan (2010). “Analogy and similarity”. In: *Language, usage and cognition*. Cambridge University Press. Chap. 4, pp. 57–75.

Camacho-Collados, José et al. (2015). “A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets.” In: *ACL (2)*, pp. 1–7.

Carroll, John et al. (2004). “The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser”. In: *International Conference on Natural Language Processing*. Springer, pp. 646–654.

Castellón, Irene et al. (1997). “Propuesta de alternancias de diátesis verbales para el español y el catalán”. In: *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural* 21, pp. 31–48.

Christensen, Janara et al. (2010). “Semantic role labeling for open information extraction”. In: *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*. Association for Computational Linguistics, pp. 52–60.

Cifuentes Honrubia, José Luis (2006). “Alternancias verbales en español”. In: *Revista portuguesa de Humanidades* 10.2, pp. 107–132.

Clark, Herbert H (1970). “Word associations and linguistic theory”. In: *New horizons in linguistics* 1, pp. 271–286.

Croce, Danilo et al. (2012). “Verb classification using distributional similarity in syntactic and semantic structures”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pp. 263–272.

Croft, William (1998). *Event structure in argument linking. The projection of arguments, ed. by Miriam Butt & Wilhelm Geuder, 21–63.*

Čulo, Oliver et al. (2008). “Comparing and combining semantic verb classifications”. In: *Language Resources and Evaluation* 42.3, pp. 265–291.

Dang, Hoa Trang et al. (1998). “Investigating regular sense extensions based on intersective Levin classes”. In: *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 293–299.

De Deyne, Simon, Daniel J. Navarro, et al. (2013). “Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations”. In: *Behavior Research Methods* 45.2, pp. 480–498.

De Deyne, Simon, Yves Peirsman, et al. (2009). “Sources of semantic similarity”. In: *Proceedings of the Cognitive Science Society*. Vol. 31. 31.

De Deyne, Simon and Gert Storms (2015). “Word associations”. In: *The Oxford handbook of the word*.

Dorr, Bonnie (1993). *Machine translation: a view from the Lexicon*. MIT press.

— (1997). “Large-scale dictionary construction for foreign language tutoring and interlingual machine translation”. In: *Machine Translation* 12.4, pp. 271–322.

Dorr, Bonnie and Doug Jones (1996). “Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues”. In: *Proceedings of the 16th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 322–327.

Dowty, David (1991). “Thematic proto-roles and argument selection”. In: *language*, pp. 547–619.

Dowty, David R (1979). *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague’s PTO*. Reidel.

Eustache, Francis et al. (1990). “Word-association responses and severity of dementia in Alzheimer Disease”. In: *Psychological reports* 66.3, pp. 1315–1322.

Falk, Ingrid et al. (2012). “Classifying french verbs using french and english lexical resources”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pp. 854–863.

Faruqui, Manaal et al. (2014). “Improving vector space word representations using multilingual correlation”. In: Association for Computational Linguistics.

Fellbaum, Christiane (1998). *WordNet*. Wiley Online Library.

Fellbaum, Christiane (2015). “Lexical Relations”. In: *The Oxford handbook of the word*. Ed. by John R Taylor. OUP Oxford, pp. 350–363.

Fernández-Montraveta, Ana et al. (2007). “Problemas sobre la distinción entre argumentos y adjuntos en el corpus SenSem”. In: *Perspectivas de análisis de la unidad verbal. Seres.* Ed. by Irene Castellón Masalles et al., pp. 35–48.

— (2014). “The SenSem Corpus: An annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality”. In: *Corpus Linguistics and Linguistic Theory* 10.2, pp. 273–288.

Ferrer, Eva Esteve (2004). “Towards a semantic classification of Spanish verbs based on subcategorisation information”. In: *Proceedings of the ACL 2004 workshop on Student research*. Association for Computational Linguistics, p. 13.

Ferretti, Todd R, Marta Kutas, et al. (2007). “Verb aspect and the activation of event knowledge.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33.1, p. 182.

Ferretti, Todd R, Ken McRae, et al. (2001). “Integrating verbs, situation schemas, and thematic role concepts”. In: *Journal of Memory and Language* 44.4, pp. 516–547.

Fillmore, Charles J. (1967). “The case for case.” In:

Fillmore, Charles J. et al. (2003). “Background to framenet”. In: *International journal of lexicography* 16.3, pp. 235–250.

Finkelstein, Lev et al. (2001). “Placing search in context: The concept revisited”. In: *Proceedings of the 10th international conference on World Wide Web*. ACM, pp. 406–414.

Fitzpatrick, Tess and Cristina Izura (2011). “Word association in L1 and L2: An exploratory study of response types, response times, and interlingual mediation”. In: *Studies in Second Language Acquisition* 33.3, pp. 373–398.

Fitzpatrick, Tess, I Munby, et al. (2014). “Word associations and the L2 lexicon”. In: *Dimensions of vocabulary knowledge*, pp. 92–105.

Fitzpatrick, Tess, David Playfoot, et al. (2013). “Establishing the reliability of word association data for investigating individual and group differences”. In: *Applied Linguistics* 36.1, pp. 23–50.

Forgy, Edward W. (1965). “Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications”. In: *Biometrics* 21, pp. 768–769.

García-Miguel, José M and Francisco J Albertuz (2005). “Verbs, semantic classes and semantic roles in the ADESSE project”. In: *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pp. 50–55.

García-Miguel, José M, Lourdes Costas, et al. (2005). “Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintácticosemánticos en el proyecto ADESSE”. In: *Entre semántica léxica, teoría del léxico y sintaxis. Frankfurt am Main: Peter Lang*, pp. 373–384.

Garner, Wendell R (1974). *The processing of information and structure*. Psychology Press.

Gentner, Dedre and Kenneth J Kurtz (2005). “Relational categories”. In: *Categorization inside and outside the lab*, pp. 151–175.

Gentner, Dedre and Arthur B Markman (1997). “Structure mapping in analogy and similarity.” In: *American psychologist* 52.1, p. 45.

Gerz, Daniela et al. (2016). “SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity”. In: *EMNLP*.

Gildea, Daniel (2002). “Probabilistic models of verb-argument structure”. In: *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 1–7.

Gildea, Daniel and Daniel Jurafsky (2002). “Automatic labeling of semantic roles”. In: *Computational linguistics* 28.3, pp. 245–288.

Giuglea, Ana-Maria et al. (2006). “Semantic role labeling via framenet, verbnet and propbank”. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 929–936.

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

— (2002). “Surface generalizations: An alternative to alternations”. In: *Cognitive Linguistics* 13.4, pp. 327–356.

Goldstone, Robert L. (1994). “The role of similarity in categorization: providing a groundwork”. In: *Cognition* 52.2, pp. 125–157. ISSN: 0010-0277.

Gonzàlez, Edgar et al. (2015). “Unsupervised ensemble minority clustering”. In: *Machine learning* 98.1-2, pp. 217–268.

Goodman, Nelson (1972). “Seven strictures on similarity”. In:

Grishman, Ralph (2003). “Information extraction”. In: *The Handbook of Computational Linguistics and Natural Language Processing*, pp. 515–530.

Gruber, Jeffrey Steven (1965). “Studies in lexical relations.” PhD thesis. Massachusetts Institute of Technology.

Guida, Annamaria et al. (2007). “Semantic properties of word associations to Italian verbs”. In: *Italian Journal of Linguistics* 19.2, pp. 293–326.

Guo, Yufan et al. (2011). “A weakly-supervised approach to argumentative zoning of scientific documents”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 273–283.

Hahn, Ulrike et al. (2003). “Similarity as transformation”. In: *Cognition* 87.1, pp. 1–32.

Halliday, Michael Alexander Kirkwood et al. (2004). *An introduction to functional grammar*. Routledge.

Hamp, Birgit et al. (1997). “Germanet-a lexical-semantic net for german”. In: *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pp. 9–15.

Hampton, James A. (1998). “Similarity-based categorization and fuzziness of natural categories”. In: *Cognition* 65.2, pp. 137–165. ISSN: 0010-0277.

Hare, Mary, Jeffrey L. Elman, et al. (2009). "The wind chilled the spectators, but the wine just chilled: Sense, structure, and sentence comprehension". In: *Cognitive Science* 33.4, pp. 610–628.

Hare, Mary, Ken McRae, et al. (2003). "Sense and structure: Meaning as a determinant of verb subcategorization preferences". In: *Journal of Memory and Language* 48.2, pp. 281–303.

Herández Muñoz, Natividad et al. (2014). "Análisis de las relaciones semánticas a través de una tarea de libre asociación en español con mapas auto-organizados". In: *RLA. Revista de lingüística teórica y aplicada* 52.2, pp. 189–212.

Hill, Felix et al. (2015). "SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation". In: *Computational Linguistics* 41.4, pp. 665–695.

Hontoria, Horacio Rodríguez (2003). "Similitud semántica". In: *Lexicografía computacional y semántica* 64, p. 119.

Jackendoff, Ray (1972). "Semantic interpretation in generative grammar." In:

— (1983). *Semantics and cognition*. Vol. 8. MIT press.

— (1992). *Semantic structures*. Vol. 18. MIT press.

Joanis, Eric et al. (2008). "A general feature space for automatic verb classification". In: *Natural Language Engineering* 14.3, pp. 337–367.

Jones, Michael N. et al. (2015). "Models of semantic memory". In: *Oxford handbook of mathematical and computational psychology*. Ed. by Busemeyer et al. Chap. 11, pp. 232–254.

Jung, Carl G. (1910). "The association method". In: *The American journal of psychology* 21.2, pp. 219–269.

Kamide, Yuki et al. (2003). "The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements". In: *Journal of Memory and language* 49.1, pp. 133–156.

Kawahara, Daisuke et al. (2014). "A Step-wise Usage-based Method for Inducing Polysemy-aware Verb Classes." In: *ACL (1)*, pp. 1030–1040.

Kent, Grace Helen et al. (1910). "A study of association in insanity". In: *American Journal of Psychiatry* 67.1, pp. 37–96.

Kingsbury, Paul et al. (2003). "Deriving verb-meaning clusters from syntactic structure". In: *Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9*. Association for Computational Linguistics, pp. 70–77.

Klavans, Judith et al. (1998). "Role of Verbs in Document Analysis". In: *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*. COLING '98. Montreal, Quebec, Canada: Association for Computational Linguistics, pp. 680–686. DOI: 10.3115/980451.980959.

Koehn, Philipp et al. (2007). "Factored Translation Models." In: *EMNLP-CoNLL*, pp. 868–876.

Kohomban, Upali S. et al. (2005). "Learning semantic classes for word sense disambiguation". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 34–41.

Korhonen, Anna (2009). "Automatic Lexical Classification–Balancing between Machine Learning and Linguistics." In: *PACLIC*, pp. 19–28.

Korhonen, Anna, Yuval Krymolowski, and Nigel Collier (2006). "Automatic classification of verbs in biomedical texts". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 345–352.

— (2008). "The choice of features for classification of verbs in biomedical texts". In: *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 449–456.

Korhonen, Anna, Yuval Krymolowski, and Zvika Marx (2003). "Clustering polysemic subcategorization frame distributions semantically". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 64–71.

Kozima, Hideki et al. (1993). "Similarity between words computed by spreading activation on an English dictionary". In: *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 232–239.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. University of Chicago Press. ISBN: 9780226468044.

Lamirel, Jean-Charles, Ingrid Falk, et al. (2015). "Federating clustering and cluster labelling capabilities with a single approach based on feature maximization: French verb classes identification with IGNGF neural clustering". In: *Neurocomputing* 147, pp. 136–146.

Lamirel, Jean-Charles, Raghvendra Mall, et al. (2011). "Variations to incremental growing neural gas algorithm based on label maximization". In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, pp. 956–965.

Landauer, Thomas K et al. (1997). "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." In: *Psychological review* 104.2, p. 211.

Lapata, Maria et al. (1999). "Using subcategorization to resolve verb class ambiguity". In: *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 266–274.

Le, Quoc et al. (2014). "Distributed representations of sentences and documents". In: *International Conference on Machine Learning*, pp. 1188–1196.

Lenci, Alessandro (2009). "Argument alternations in italian verbs: a computational study". In: *Atti del XLII Congresso Internazionale di Studi della Societa di Linguistica Italiana*.

— (2014). "Carving verb classes from corpora". In: *Word Classes: Nature, Typology, and Representations*, pp. 17–36.

Levin, Beth (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Levin, Beth and Malka Rappaport Hovav (2005). *Argument realization*. Cambridge University Press.

Li, Jianguo, Kirk Baker, et al. (2011). “Which are the Best Features for Automatic Verb Classification: a Corpus Study of Levin Verb Classification”. In:

Li, Jianguo and Chris Brew (2008). “Which Are the Best Features for Automatic Verb Classification.” In: *ACL*, pp. 434–442.

Lin, Dekang (1998). “An Information-Theoretic Definition of Similarity”. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML ’98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 296–304. ISBN: 1-55860-556-8.

Liu, Ding et al. (2010). “Semantic role features for machine translation”. In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 716–724.

Luce, Paul A. et al. (1990). “Similarity neighborhoods of spoken words.” In: *ACL-MIT Press series in natural language processing. Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Ed. by Gerry T. M. Altmann. The MIT Press, pp. 122–147.

Lyons, John (1977). “Semantics (vols i & ii)”. In: *Cambridge CUP*.

Manning, Jeremy R. et al. (2012). “Interpreting semantic clustering effects in free recall”. In: *Memory* 20.5, pp. 511–517.

Markman, Arthur B. et al. (2013). “The nature of mental concepts”. In: *The Oxford Handbook of Cognitive Psychology*. Oxford University Press New York, NY, pp. 321–345.

Márquez, Pedro Pablo Devís (1991). “Esquemas sintáctico-semánticos: el problema de las diátesis en español”. PhD thesis. Universidad de Cádiz.

Martí, M Antònia et al. (2002). *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita*. Vol. 53. Edicions Universitat Barcelona, pp. 50–57.

McRae, Ken, Mary Hare, et al. (2005). “A basis for generating expectancies for verbs from nouns”. In: *Memory & Cognition* 33.7, pp. 1174–1184.

McRae, Ken, Saman Khalkhali, et al. (2012). “Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy.” In:

McRae, Ken and Kazunaga Matsuki (2009). “People use their knowledge of common events to understand language, and do so as quickly as possible”. In: *Language and linguistics compass* 3.6, pp. 1417–1429.

Merlo, Paola and Suzanne Stevenson (2001). “Automatic verb classification based on statistical distributions of argument structure”. In: *Computational Linguistics* 27.3, pp. 373–408.

Merlo, Paola, Suzanne Stevenson, et al. (2002). “A multilingual paradigm for automatic verb classification”. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 207–214.

Mikolov, Tomas et al. (2013). “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*, pp. 3111–3119.

Miller, George A. et al. (1990). "Introduction to WordNet: An on-line lexical database". In: *International journal of lexicography* 3.4, pp. 235–244.

Mitchell, T. et al. (2015). "Never-Ending Learning". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15).*

Moldovan, Cornelia D. et al. (2015). "Semantic similarity: normative ratings for 185 Spanish noun triplets". In: *Behavior research methods* 47.3, pp. 788–799.

Mollin, Sandra (2009). "Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations". In: *Corpus Linguistics and Linguistic Theory* 5.2, pp. 175–200.

Namei, Shidrokh (2004). "Bilingual lexical development: A Persian–Swedish word association study". In: *International Journal of Applied Linguistics* 14.3, pp. 363–388.

Navigli, Roberto et al. (2003). "An analysis of ontology-based query expansion strategies". In: *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia*, pp. 42–49.

Neely, James H (1991). "Semantic priming effects in visual word recognition: A selective review of current findings and theories". In: *Basic processes in reading: Visual word recognition* 11, pp. 264–336.

Nelson, Douglas L et al. (2000). "What is free association and what does it measure?" In: *Memory & cognition* 28.6, pp. 887–899.

Niles, Ian et al. (2001). "Towards a standard upper ontology". In: *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*. ACM, pp. 2–9.

Nordquist, Dawn (2009). "Investigating elicited data from a usage-based perspective". In: *Corpus Linguistics and Linguistic Theory* 5.1, pp. 105–130.

Novick, Laura R (1988). "Analogical transfer, problem similarity, and expertise." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14.3, p. 510.

Padó, Ulrike et al. (2006). "Modelling Semantic Role Pausibility in Human Sentence Processing." In: *EACL*.

Padró, Lluís et al. (2012). "Freeling 3.0: Towards wider multilinguality". In: *LREC2012*.

Palmer, Martha et al. (2005). "The proposition bank: An annotated corpus of semantic roles". In: *Computational linguistics* 31.1, pp. 71–106.

Patwardhan, Siddharth et al. (2003). "Using measures of semantic relatedness for word sense disambiguation". In: *CICLing*. Vol. 2588. Springer, pp. 241–257.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peirsman, Yves et al. (2009). "Predicting strong associations on the basis of corpus data". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 648–656.

Peterson, Daniel Wyde (2019). "Probabilistic Modeling of VerbNet Clusters". PhD thesis. University of Colorado at Boulder.

Pradet, Quentin et al. (2013). "Revisiting knowledge-based Semantic Role Labeling". In: *LTC'13*, p. 71.

Pradhan, Sameer et al. (2005). "Support vector learning for semantic argument classification". In: *Machine Learning* 60.1-3, pp. 11–39.

Prescher, Detlef et al. (2000). "Using a probabilistic class-based lexicon for lexical ambiguity resolution". In: *Proceedings of the 18th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, pp. 649–655.

Quinlan, Ross (1992). *C4. 5 Release 8*.

Raaijmakers, Jeroen G. et al. (1981). "Search of associative memory." In: *Psychological review* 88.2, p. 93.

Řehůřek, Radim et al. (2010). "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. `http://is.muni.cz/publication/884893/en`. Valletta, Malta: ELRA, pp. 45–50.

Reichart, Roi et al. (2013). "Improved Lexical Acquisition through DPP-based Verb Clustering." In: *ACL (1)*, pp. 862–872.

Resnik, Philip (1995). "Using Information Content to Evaluate Semantic Similarity in a Taxonomy". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., pp. 448–453. ISBN: 1-55860-363-8, 978-1-558-60363-9.

Resnik, Philip and Mona Diab (2000). "Measuring verb similarity". In: *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*. Philadelphia, PA, pp. 399–404.

Roberts, Will et al. (2014). "A Comparison of Selectional Preference Models for Automatic Verb Classification." In: *EMNLP*, pp. 511–522.

Roediger, Henry L et al. (2001). "Factors that determine false recall: A multiple regression analysis". In: *Psychonomic Bulletin & Review* 8.3, pp. 385–407.

Rooth, Mats et al. (1999). "Inducing a semantically annotated lexicon via EM-based clustering". In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, pp. 104–111.

Sanz-Torrent, Mònica et al. (2017). "Auditory word recognition of verbs: Effects of verb argument structure on referent identification". In: *PloS one* 12.12, e0188728.

Savage, Ceri et al. (2003). "Testing the abstractness of children's linguistic representations: lexical and structural priming of syntactic constructions in young children". In: *Developmental Science* 6.5, pp. 557–567.

Scarton, Carolina et al. (2014). "Verb clustering for brazilian portuguese". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 25–39.

Schuler, Karin Kipper (2005). "VerbNet: A broad-coverage, comprehensive verb lexicon". In:

Schulte Im Walde, Sabine (2006). "Experiments on the automatic induction of German semantic verb classes". In: *Computational Linguistics* 32.2, pp. 159–194.

Schulte im Walde, Sabine (2008). "Human associations and the choice of features for semantic verb classification". In: *Research on Language and Computation* 6.1, pp. 79–111.

Schulte Im Walde, Sabine and Chris Brew (2002). "Inducing German semantic verb classes from purely syntactic subcategorisation information". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 223–230.

Schulte im Walde, Sabine and Alissa Melinger (2005). "Identifying semantic relations and functional properties of human verb associations". In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 612–619.

Schutze, Hinrich (1992). "Dimensions of meaning". In: *Supercomputing'92., Proceedings*. IEEE, pp. 787–796.

Sedoc, Joao et al. (2017). "Deriving Verb Predicates By Clustering Verbs with Arguments". In: *arXiv preprint arXiv:1708.00416*.

Shen, Dan et al. (2007). "Using Semantic Roles to Improve Question Answering." In: *Emnlp-conll*, pp. 12–21.

Shepard, Roger N. (1962a). "The analysis of proximities: multidimensional scaling with an unknown distance function. I." In: *Psychometrika* 27.2, pp. 125–140.

— (1962b). "The analysis of proximities: Multidimensional scaling with an unknown distance function. II". In: *Psychometrika* 27.3, pp. 219–246.

Shutova, Ekaterina et al. (2010). "Metaphor identification using verb and noun clustering". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 1002–1010.

Stevenson, Suzanne et al. (2003). "Semi-supervised verb class discovery using noisy features". In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, pp. 71–78.

Steyvers, Mark et al. (2004). "Word association spaces for predicting semantic similarity effects in episodic memory". In: *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, pp. 237–249.

Sun, Lin (2013). "Automatic induction of verb classes using clustering". PhD thesis. University of Cambridge.

Sun, Lin and Anna Korhonen (2009). "Improving verb clustering with automatically acquired selectional preferences". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, pp. 638–647.

— (2011). "Hierarchical verb clustering using graph factorization". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1023–1033.

Sun, Lin, Anna Korhonen, and Yuval Krymolowski (2008). "Verb class discovery from rich syntactic data". In: *Lecture Notes in Computer Science* 4919, p. 16.

Sun, Lin, Anna Korhonen, Thierry Poibeau, et al. (2010). “Investigating the cross-linguistic potential of verbnet: style classification”. In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 1056–1064.

Sun, Lin, Diana McCarthy, et al. (2013). “Diathesis alternation approximation for verb clustering.” In: *ACL (2)*, pp. 736–741.

Surdeanu, Mihai et al. (2003). “Using predicate-argument structures for information extraction”. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 8–15.

Swier, Robert S. et al. (2005). “Exploiting a verb lexicon in automatic semantic role labelling”. In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 883–890.

Tsang, Vivian et al. (2002). “Crosslinguistic transfer in automatic verb classification”. In: *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 1–7.

Tversky, Amos (1977). “Features of similarity.” In: *Psychological review* 84.4, p. 327.

Vázquez, Gloria et al. (2000). “Clasificación verbal(alternancias de diátesis)”. In: *Quaderns de sintagma*.

Vendler, Zeno (1957). “Verbs and times”. In: *The philosophical review* 66.2, pp. 143–160.

Vitevitch, Michael S et al. (1999). “Probabilistic phonotactics and neighborhood activation in spoken word recognition”. In: *Journal of Memory and Language* 40.3, pp. 374–408.

— (2016). “Phonological neighborhood effects in spoken word perception and production”. In: *Annual Review of Linguistics* 2, pp. 75–94.

Vlachos, Andreas, Zoubin Ghahramani, et al. (2008). “Dirichlet process mixture models for verb clustering”. In: *Proceedings of the ICML workshop on Prior Knowledge for Text and Language*.

Vlachos, Andreas, Anna Korhonen, et al. (2009). “Unsupervised and constrained Dirichlet process mixture models for verb clustering”. In: *Proceedings of the workshop on geometrical models of natural language semantics*. Association for Computational Linguistics, pp. 74–82.

Vossen, Piek (1998). “A multilingual database with lexical semantic networks”. In: *Dordrecht: Kluwer Academic Publishers. doi* 10, pp. 978–94.

Weischedel, Ralph et al. (2011). “OntoNotes Release 4.0”. In: *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Wijaya, Derry Tanti (2016). “VerbKB: A Knowledge Base of Verbs for Natural Language Understanding”. PhD thesis. Carnegie Mellon University.

Yang, Dongqiang et al. (2006). “Verb similarity on the taxonomy of WordNet”. In: *The Third International WordNet Conference: GWC 2006*. Masaryk University.