This is a preprint of the paper: Gil-Vallejo, L., M. Coll-Florit, I. Castellón, J. Turmo (2017). "Verb similarity: Comparing corpus and psycholinguistic data", *Corpus Linguistics and Linguistic Theory*. ISSN (Online) 1613-7035, ISSN (Print) 1613-7027, DOI: https://doi.org/10.1515/cllt-2016-0045, January 2017.

Verb similarity: Comparing corpus and psycholinguistic data

Lara Gil-Vallejo, Marta Coll-Florit, Irene Castellón and Jordi Turmo

Abstract

Similarity, which plays a key role in fields like cognitive science, psycholinguistics and natural language processing, is a broad and multifaceted concept. In this work we analyse how two approaches that belong to different perspectives, the corpus view and the psycholinguistic view, articulate similarity between verb senses in Spanish. Specifically, we compare the similarity between verb senses based on their argument structure, which is captured through semantic roles, with their similarity defined by word associations. We address the question of whether verb argument structure, which reflects the expression of the events, and word associations, which are related to the speakers' organization of the mental lexicon, shape similarity between verbs in a congruent manner, a topic which has not been explored previously. While we find significant correlations between verb sense similarities obtained from these two approaches, our findings also highlight some discrepancies between them and the importance of the degree of abstraction of the corpus annotation and psycholinguistic representations.

Keywords: similarity, semantic roles, word associations

1. Introduction

Similarity is a crucial notion that underpins a number of cognitive models and psycholinguistic accounts of the lexicon. Furthermore, it is also important for natural language processing (NLP) models and a wide range of related NLP tasks. However, despite its usefulness, there is a fundamental problem in the definition of similarity. Bybee (2010) voiced such a concern in the context of models of novel utterance processing and production, but it is applicable to any similarity-based model: "The problem [...], however, is in specifying the relevant features upon which similarity is measured. This is a pressing empirical problem". There are many non-trivial aspects that can be taken into account when describing and quantifying the similarity between two given concepts, words, sentences, etc. (cf. De Deyne et al. 2009 for a discussion in the semantic domain). Moreover, it has been argued that these different aspects tap into different elements of linguistic knowledge (Maki and Buchanan 2008; Riordan and Jones 2007).

In this work we intend to gain further understanding of the particularities of two different approaches that come from the corpus and the psycholinguistic perspectives when defining verb sense similarity in Spanish as well as the relationship between them. The first type of knowledge can lead to a more advised use of each perspective regarding the linguistic features that they emphasize, whereas a better knowledge of their points of convergence and divergence can help in the development of a combined model that represents a more robust characterization of verb similarity. In particular we examine similarity between verb

senses[1] in Spanish. For our experiments we characterize it in the following way: on the one side, we use speakers' characterizations of verb senses obtained through word associations to embody the psycholinguistic perspective; on the other side we select argument structure of the sentences, the bearer of event information, to materialize the corpus perspective.

Several comparisons between corpus data and word associations have been carried out previously (Mollin 2009; Nordquist 2009), but they have been mostly based on collocational information, aiming to test the psycholinguistic reality or relevance of collocations extracted from corpora. Indeed, from a usage-based perspective, the hypothesis that conventionalized linguistic structures should appear both in corpora and elicited data is a compelling one (Nordquist 2009). In addition, several computational models of language have aimed to predict word associations by using word co-occurrence information from diverse corpora (Church et al. 1989; Wettler and Rapp 1993; Peirsman and Geeraerts 2009; Sahlgren 2006; Michelbacher et al. 2007). In relation to this, a number of psycholinguistic studies remark that associative strength between a stimulus and a response correlates with word co-occurrence probabilities in a corpus (Spence and Owens 1990; McKoon and Ratcliff 1992; Plaut 1995), although this correlation is not straightforward, as there are some linguistic properties mediating it (De Deyne, Verheyen and Storms 2015).

To the best of our knowledge, the specific comparison between word associations and corpus data that we propose here, focused on similarity, has not been performed before. Furthermore, we should note that most of the previous comparative studies that deal with similarity modelled through different methods, such as De Deyne et al. 2009, have been focused on the semantic component. Since our comparison involves argument structure, which is assigned typically to the interface between semantics and syntax, we also explicitly explore the role of the structural component in shaping similarity. In particular, our objective is twofold: firstly, we aim to look into the relationship between the similarity of verb senses as defined by corpus data (argument structure) and by psycholinguistic data (word associations); secondly, we want to investigate the role that the degree of abstraction of the data representation has in shaping this relationship.

The paper is organised as follows. The next section provides background on the notion of similarity in the different fields where it has a relevant role: cognitive science, psycholinguistics (with emphasis on word associations) and natural language processing. Section 3 details the methodology followed to establish the comparison between the corpus and psycholinguistic perspectives. The following two sections, 4 and 5, delve into the formalization of each of the aforementioned perspectives respectively. The comparisons and results are discussed in Section 6. Finally, we summarise the most relevant findings in the light of their possible applications.

2. Background on similarity

With regard to cognition studies, similarity turns out to be a pervasive topic: many accounts of inference generation, problem solving, memory, prediction and categorization rely on it as a basic working mechanism. However, there is no unified account of similarity. On the

[1] We provide the definitions for *sense*, *lemma* and *word* as used in this article. *Sense* is defined in line with Croft and Cruse (2004): a sense is a unit of meaning of a word, delimited from the rest of the meaning potential of this word. This delimitation is operated by conventional, cognitive and contextual constraints. In our work the senses are already defined in the corpus: they have a number that identifies them and a definition that delimits their boundaries. *Lemma* is the canonical form used to refer to a set of inflected forms. *Word* is the inflected form of a lexical unit.

contrary, there are a number of formal accounts that also define how similarity should be empirically measured. Goldstone and Son (2005) summarize them in four types of models: geometric, feature set-based, structure mapping-based and transformational. Geometric models characterize mental representations as points in the representational space. Therefore, the similarity between two concepts can be calculated on the basis of their distance: concepts that are close are more similar than those which are apart (Shepard 1962). Feature-set approaches (Tversky 1977) represent concepts as feature lists. According to this approach, similarity between two concepts is measured in virtue of their common and distinctive features. Structure-mapping models, also known as alignment-based models, (Gentner and Markman, 1997) are also based on the existence of features associated with concepts, but they place the emphasis on the role that each feature has in the representations regarding the feature-matching process. To the contrary, transformational approaches (Garner 1974; Hahn et al. 2003) do not propose to study similarity using proxies such as points in space or feature sets. Instead they assume that any mental representation can be transformed into another following a series of transformational operations. Based on this view, the number and type of transformations needed define the degree of 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 aspects.

These different accounts of similarity in cognition have had a theoretical and practical impact on psycholinguistic models of the mental lexicon (see Jones et al. 2015: Ch. 11 for an overview). Furthermore, there is a body of research in psycholinguistics that brings to light 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 and Luce 1999; Luce et al. 1990; Vitevitch and Luce 2016), semantic (Neely 1991) and syntactic (Savage et al. 2003), but also from studies that deal with word formation (see Bybee 2010: Ch. 4).

Within this field, word associations are also considered a useful tool to reveal information about the organization of the lexicon (Mollin 2009, Fitzpatrick et al. 2011) and the lexical retrieval process (Nordquist 2009). The term ‘word association’ originated in psychology to refer to the word or words that first come to mind in response to a subject being presented with a stimulus word. The first collections of word associations can be traced back to the beginnings of the 20th century, when Jung established them as a diagnostic tool to discriminate normal behaviour from pathological behaviour on the basis of the type of responses given by the patients. The introduction and usage of word association in psycholinguistics is more recent and happens in close connection with lexicon studies, which deal with how words are stored, represented and retrieved in memory. In word association experiments, the responses associated to each stimulus word are used to characterize it, and their frequency determines the associative strength between the response and the stimulus. As for the importance of word associations for the organization of the lexicon, some studies claim that word associations are crucial in the early conceptual processing that takes place when a word has been perceived (Barsalou et al. 2008: 249). In addition to this, word associations have been shown to correlate with semantic similarity effects in free recall and cued recall, as well as with human similarity ratings (Steyvers et al. 2004). In addition, they are argued to be able to account for semantic clustering effects in recall experiments (Manning 2012).

A number of studies have focused on the type of information that motivates associations between the stimulus and the response. It is generally acknowledged that these relations

have a broad semantic basis (Nelson et al. 2000; Roediger et al. 2001; McRae 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 (Hernández et al. 2014) and thematic relationships (Peirsman and Geeraerts 2009) are more frequent. Additionally, some studies suggest that the type of relation between responses elicited and stimuli can be influenced by the part of speech of the stimulus (Deese 1962; Goldfarb and Halpern 1984).

If we move the focus to natural language processing, 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, etc.). Nevertheless, the formalization and measurement of similarity is affected by the same problem of broadness that we mentioned in the introduction: 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): the lexical level (synonymy and antonymy), the conceptual level (hyponymy, meronymy), the syntagmatic level, etc. These relations are the building blocks of relational models such as WordNet (Fellbaum 1998), which are grounded in psycholinguistic data from child language acquisition (Keil, 1989), word association norms (De Deyne and Storms 2015) and co-occurrence information (Fellbaum 2015). Consequently, we also find a variety of strategies for measuring similarity in the literature, in most cases based on the semantic relation of the two linguistic units. Among those that have to do with words and concepts, as in our case, we find models that represent linguistic units as nodes in a taxonomy linked by a variety of relations, whose similarity is measured on the basis of the path that connects them (Resnik 1995; Yang and Powers 2006). Other accounts represent words as points in a space configured by their context, in which words that share many contexts are similar (Schutze 1992; Landauer and Dumais 1997; Mikolov et al. 2013). Besides, similarity has been calculated on the basis of the overlap between word definitions or glosses in dictionaries (Kozima and Furugori 1993; Banerjee and Pedersen 2003; Pathwardan et al. 2003) and taking into account the amount of overlap between feature sets associated with the words (Lin 1998).

Finally, it is important to mention that there are several collections of human ratings of word similarity, mostly based on semantics. 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 and Dyer 2014). However, since most of these collections are limited to pairs of nouns, with the exception of Hill et al. (2015), Yang and Powers (2006) and Baker et al. (2014), the coverage of verbs (and other categories) is reduced. For Spanish there is a general scarcity of these resources: there are only ratings for nouns, either collected from native speakers (Moldovan 2015), or translated from English (Finkelstein et al. 2001; Camacho-Collados et al. 2015).

3. Methodology

In this study we focus on verb senses rather than lemmas to obtain more precise models of similarity, because different senses of the same lemma are likely to have different syntactic and semantic behaviour (Hare et al. 2003). We exemplify this with the lemma *valer* from the Sensem Spanish corpus (Fernández-Montraveta and Vázquez 2014). 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[2] 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. In addition to this corpus example, there is experimental evidence that suggests that the different senses of a word have an impact on word recognition (Rodd and Marslen-Wilson 2002; Rayner and Frazier 1989) and similarity perception (Klein and Murphy 2002). Therefore, it is necessary to use specific senses of verbs in order to obtain data valid for an issue as sensitive to polysemy as it is similarity.

Regarding the selection of the verb senses used for the present study, 20 senses were chosen from among those of the Sensem corpus. The selected senses 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[3] ('lexicographer file' labels, Fellbaum 1998) and by Adesse corpus macro-classes (Albertuz 2007); and 3) showing a variety of subcategorization frames and semantic roles associated with them[4]. 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 set of data suitable to be explored in detail. In Appendix 1 there is a list of the verb senses chosen, along with their definitions and their frequency in the corpus; in Appendix 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 article to improve legibility.

In order to be able to compute pairwise similarity between all the senses, we obtained all possible sense pairings. For that, 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 were able to measure the similarity among all the verb senses in our sample.

In the following two sections (4 and 5) we detail the formalization of the corpus data and the psycholinguistic data respectively, specifying the features that we used in each perspective. The results of the comparison of both types of data are shown in Section 6.

4. Corpus perspective: argument structure and event representation

In this section we explain how we approach similarity from the corpus data perspective. Our corpus of reference is the Sensem corpus, which consists of 999,712 tokens and contains 30,365 sentences associated to 250 different verb lemmas. The lemmas are manually disambiguated, making a total of 980 different verb senses. The sentences are annotated at the syntactic and semantic levels (semantic roles and verb synsets). We use the data provided by this corpus to capture and formalize the argument structure of the verbs.

We consider that argument structure is relevant for verb similarity for two reasons. On the one hand, linguistic accounts look at argument structure as a key element associated with verbs (and also other units such as deverbal nouns or prepositions) that lives at the syntax-

[2] NP stands for 'noun phrase', V stands for 'verb' and PP stands for 'prepositional phrase'.
[3] The correspondence with WordNet supersenses is done through the synsets associated with the Sensem verb senses using the Multilingual Central Repository (Gonzalez-Agirre and Rigau 2013).
[4] Their subcategorization frames and their semantic roles can be found in the Sensem lexicon (http://grial.uab.es/sensem/lexico?idioma=en).

semantics interface of the linguistic system. The linguistic notion of argument structure refers to how many and what type of arguments are related to a predicate. There are diverse (and sometimes opposing) views of the nature of argument structure. To comment briefly on two influential examples, we can mention the Levin (1993) and Goldberg (1995) approaches. Levin's work is based on the hypothesis that "the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extend determined by its meaning". She used information about the common diathesis (shared argument alternations in the expressions of the arguments of a verb) in which verbs participate to model verb behaviour and classify around 3,000 English verbs. A quite different view is that of Goldberg. This author considers that argument structure is a type of construction and remarks that, as with every construction, it has independent meaning. Whether the relation between the argument structure and the verb 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. On the other hand, argument structure represents the linguistic codification of the events expressed in the sentence. From the point of view of the organization of the lexicon, information related to events has been gaining importance in its role in grounding linguistic knowledge and mental representations (Jones 2015; McRae et al. 2005; Ferretti et al. 2001; Chwilla and Kolk 2005).

A related work to the one presented in this article is the one by Merlo and Stevenson (2001), who showed that argument structure can be used to classify verbs using information from corpora. They classified 59 verbs into three classes according to their argument structure (unergative, unaccusative and object-drop). They hypothesized that the verb's argument structure could be determined on the basis of the distribution of key characteristics in corpus. These characteristics were modelled through a set of lexical features (transitivity, causativity, animacy, verb voice, and use of past participle or simple past). They used a decision tree algorithm, reaching almost 70 % of accuracy in the classification.

With regard to how to capture argument structure in an empirical manner, we use the semantic role annotations in the Sensem corpus. Semantic roles were first introduced in modern linguistics by the works of Gruber (1965), Fillmore (1968) 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 and Lapata 2007), information extraction (Christensen 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 corpus annotation at this level 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*, *painter*, etc.). The pioneering work carried out by Fillmore (1968) introduced a set of semantic roles based on interpretative categories: *agent, patient, theme, beneficiary*, etc., 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.). The numbering criterion

remains constant for a specific verb, but changes across different verbs. A more fine-grained account of semantic roles can be found in Frame Semantics theory and FrameNet (Fillmore et al. 2003), in which 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 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 (Proto-Agent and 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 (coarse-grained roles, e.g. *actor*, subsume related fine-grained roles: *cause*, *agent*, etc.).

We considered that this last approach was suitable for our experiments and research goals because of its flexibility, which allowed us to experiment with the different degrees of granularity of semantic roles. Therefore, we manually created a three-level hierarchical roleset. This was done mapping the roles annotated in the corpus with the LIRICS roles in a bottom up fashion: firstly, the Sensem semantic roles (38 in total) were used directly for the fine-grained level; secondly, these roles were mapped to the medium grained roles from the LIRICS proposal comparing the definitions of both role sets; next, according to this mapping, we used the LIRICS roles that had links with the Sensem roles for the medium-grained level (14 roles); finally, the coarse-grained roles are the abstract LIRICS roles that have links with our medium-grained roles (4 roles). We show the generated role hierarchy in Figure 1: the first level corresponds to coarse-grained roles, the second to medium-grained roles and the third to the fine-grained semantic roles. As a result, we obtained three levels of semantic role granularity that allowed us to formalize argument structure in a flexible way.

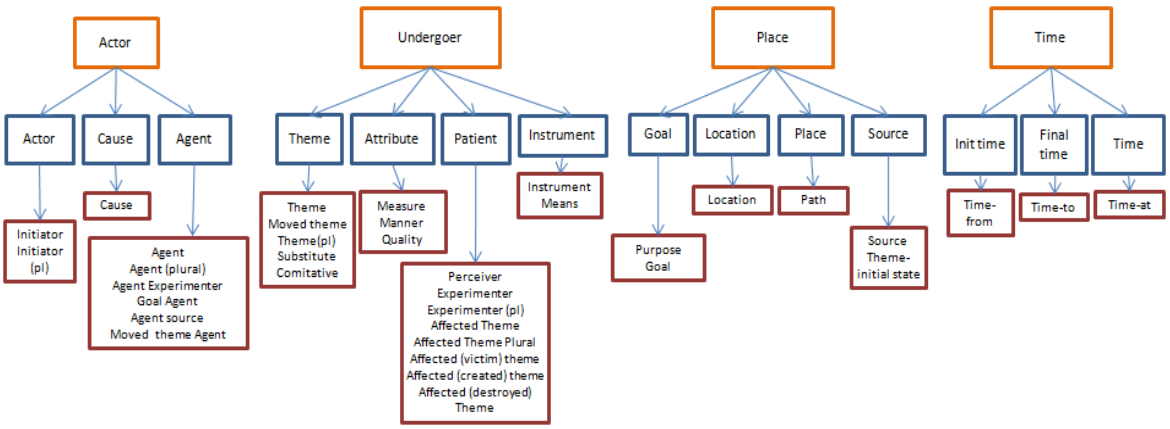

Figure 1: Hierarchy of semantic roles

The specific features from the corpus that we used to characterize the verb senses were: 1) each of the three levels of semantic roles and 2) these semantic roles combined with the corresponding syntactic function of the argument, forming a unit. We selected these two formalizations because they best captured the argument structure of the sentences and enabled us to explore the effects of adding extra syntactic information in the calculation of the similarity.

In addition, the information extracted from Sensem about semantic roles and semantic roles plus syntax was considered in two formats: constituents, which are single arguments (e.g. agent, patient; agent-subject, patient-object) and patterns, which are the combination of the arguments of a sentence (e.g. agent+patient; agent-subject+patient-object). We illustrate

this feature extraction and representation process with an example in Table 1 based on the annotated sentence *Remedios abrió su bolso* (‘Remedios opened her handbag’) in Figure 2.

Figure 2: Annotation from Sensem

Table 1: Example of features obtained from the sentence

	Constituents		Patterns	
Semantic roles	Roles	roles + syntax	Roles	roles + syntax
Fine-grained level	Agent; Affected theme	Agent-Subject; Affected theme-DirectObject	Agent+Affected theme	Agent-Subject+Affected theme-DirectObject
Medium-grained level	Agent; Theme	Agent-Subject; Theme-DirectObject	Agent+Theme	Agent-Subject+Theme-DirectObject
Coarse-grained level	Actor; Undergoer	Actor-Subject; Undergoer-DirectObject	Actor+Undergoer	Actor-Subject+Undergoer-DirectObject

Thus, for each sentence we obtained 12 different formalizations of its argument structure, according to the mentioned factors: the usage of either roles or roles combined with syntax to characterize the arguments of the verbs, the formalization in patterns or constituents and the different levels of granularity of semantic roles. Correspondingly, we generated 12 vector representations for every verb sense. The vectors contained the probability of co-occurrence of a verb sense with the features associated with the different formalizations, according to the data in the corpus. Therefore, for each sense we obtained a probability distribution over the designed features. We calculated pairwise similarity between all the verb senses according to each of the formalizations. The similarity values were obtained calculating the cosine distance of the vector representations of the senses. This distance was suitable for our purposes because it has the advantage of being independent of the magnitude and therefore the similarity value was not affected by the frequency of the sense. Summarizing, this setting allowed us to obtain pairwise similarity measures between all verb senses according to the 12 different ways of capturing argument structure.

As a result, we generated 12 similarity rankings from corpus data. In these rankings, sense pairs were ordered from most to least similar according to the values obtained in the previous step. These rankings were then compared to rankings obtained from psycholinguistic data. The results of this comparison are presented in Section 6.

5. Psycholinguistic perspective

In this section we present a psycholinguistic experiment carried out with the objective of capturing the way in which native speakers characterize verb senses. Similarity between verb senses is calculated on the basis of these characterizations. We avoid asking participants to rate the perceived similarity between two senses directly, since this method

is more suitable for lemmas and has been used primarily to rate similarity between concrete nouns. Instead, we calculate the similarity between verb senses using word associations.

As for existing resources of word associations, there are several collections for English which contain a high number of stimuli and responses, such as Kiss et al. (1973) (8,400 stimuli) and Nelson et al. (1998) (5,019 stimuli). Regarding word association resources for Spanish, there are several small collections,[5] all below 700 stimuli. However, very few include word associations for verbs (Coll-Florit and Gennari 2011) and none of them has been applied in researching verb similarity.

An in-depth examination of word associations for verbs was carried by Schulte im Walde (2008). This author explored the applications of the information obtained from word associations in German for automatic verb classification. The aim was to test whether word associations could provide useful insights into key linguistic features for this task. The author collected word associations for 330 German verbs and performed an analysis of the collected words at three levels: their part of speech (PoS), the syntactic function of the nouns with respect to the verb and their co-occurrence in corpus with the stimulus verb. As for the PoS analysis, it was found that, on average, nouns accounted for 62% of word associations, followed by verbs (19.86%), although this percentage varied across the semantic verb classes. Regarding the syntactic function of the associated nouns with respect to the verbs, only 28% of the nouns filled a position in the syntactic frames obtained using previously developed grammar. Of particular note, it was discovered that 50% of the association nouns were present in the grammar model but did not fill syntactic frame slots. Taking this fact together with the results of an analysis that showed that 77% of the nouns co-occurred at least once with the stimuli verb in a window of 20 words, it was concluded that the nouns that people associated with verbs were not restricted to subcategorization positions or selection preferences. In a related work, Schulte im Walde et al. (2008) analysed the semantic relation between the stimulus verb and the verbs obtained in the experiment as associated responses: they found that 47% of the responses had a relation in GermaNet (the German equivalent of WordNet). Other types of relations existed in the associations, but were not represented in GermaNet (cause, consequence, result, etc.).

For our experiment we also focused on the verb category to design the materials. However, unlike in Schulte im Walde (2008), the stimuli used for this experiment were the already presented 20 verb senses, which were embedded in a phrase that provides context. The aim of this type of contextualized presentation was 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 aimed 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, afterwards looking 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 contained the verb in infinitive, which was 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 3.

[5] See http://campus.usal.es/~gimc/normas/nipejcyl.htm for a compilation.

The experiment was carried out using LimeSurvey, an open source software 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 the timing of the duration of the presentation of stimulus to the participants. In the first part of the experiment, we gathered information about the formal knowledge 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 and their linguistic training, if any. In the second part, the participants were presented with the instructions. Firstly, how the stimuli were going to be presented and the expected format of the answer (a single word) was explained. Participants were also encouraged to rely on their intuition. Secondly, the participants were informed that they would be shown a sentence. The instructions specifically asked the participants to write down the words that came to their minds in relation to the word that was presented in capital letters in the sentence. 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 3. We chose to collect several responses to characterize each stimulus, as not limiting the responses to the first that came to mind has shown higher correlation with human similarity ratings (Steyvers et al. 2004) and it has proven to perform better in lexical access tasks (Chumbley and Balota 1984). Besides, Nelson et al. (2000) and De Deyne et al. (2013) suggest that these non-primary responses contain useful information. 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 et al. 2013).

Figure 3: Word association task screen

A total of 102 participants collaborated in the experiment. Regarding the geographical variety, 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 a total of 3 from Peru, Uruguay and Ecuador, respectively). 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, senses received between 454 and 730 responses.

The responses given by the participants were looked up in the dictionary provided by Freeling (Padró and Stanilovsky 2012). 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 morphosyntactic category. In cases of morphosyntactic ambiguity, the Freeling tagger assigned the most frequent part of speech to the lemma. Figure 4 presents the distribution of these categories per verb sense.

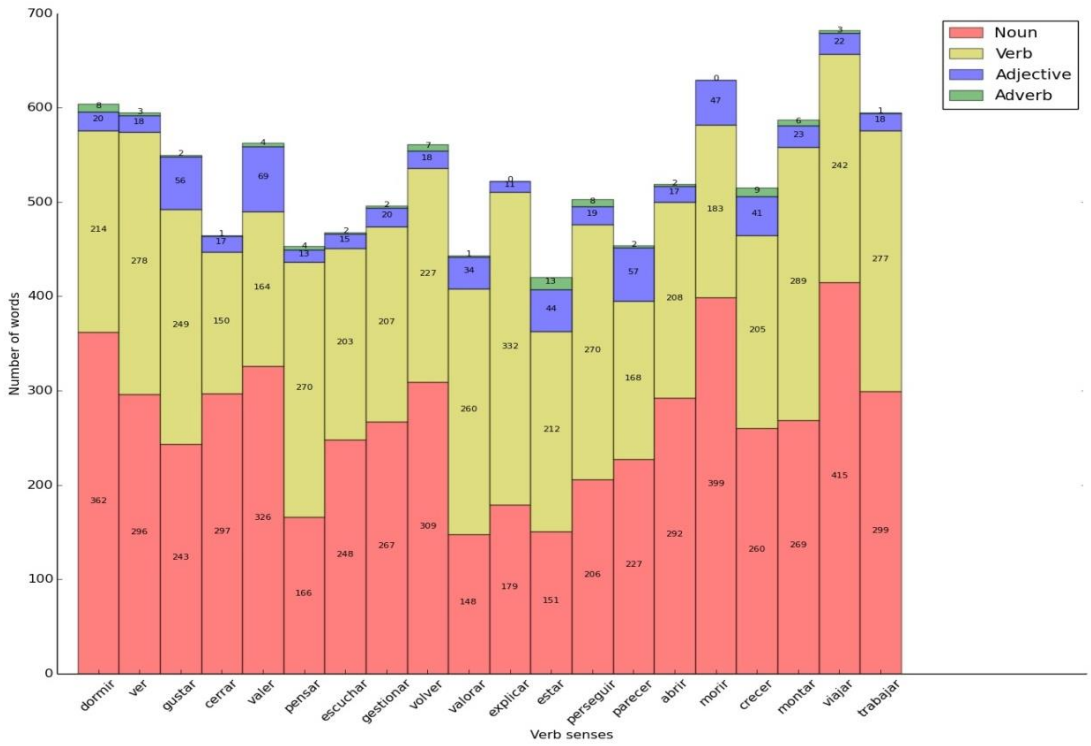

Figure 4: Frequency of the morphosyntactic category of the responses per verb

We see that most of the associated responses were nouns, closely followed by verbs. However, the difference between the two was not as large as it is in other word association studies for verbs, such as in Schulte im Walde (2008).

The frequency of the lemmatized responses obtained in the experiment was registered. In Table 2 we show a small example of the dataset, with target verb senses placed in rows and the responses in columns. As a first step, we obtained a frequency distribution over all these responses for each verb sense. The second step, as in the corpus perspective, was to obtain the similarity value of the verb senses in each pair, calculating the cosine distance of their corresponding probability distributions. Therefore, a large amount of highly probable shared responses for two verb senses yielded a large similarity value, and vice versa.

Table 2: Word association dataset example

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

As we can see in the small sample presented in Table 2, the frequency matrix obtained is quite sparse: there are many responses and only a few of them co-occur with each stimulus. This fact affects the overlap measure by skewing it towards dissimilarity. In order to reduce this effect and to experiment with different degrees of granularity and scope in the representation of the responses, we associated each response 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 correspond to a concept in WordNet) using Freeling, which assigned the most frequent synset to each word form. In turn, the synset was used as a proxy to obtain the corresponding category in a collection of ontologies. This was done through the MCR (Gonzalez-Agirre et al. 2012), which provides mappings between synsets and such ontologies. Therefore, besides having the equivalent lemma of each response, we were able to use categories from the following ontologies: hypernyms from WordNet, SUMO (Suggested Upper Merged Ontology) (Niles and Pease 2001), TCO (Top Concept Ontology) (Alvez et al. 2008) and supersenses from WordNet.

Hypernyms are the immediate, more abstract node (parent) in the is-a relationship in WordNet. SUMO is an ontology linked to WordNet which contains around 25,000 terms and 80,000 axioms. TCO is a collection of 64 features organized in three groups according to the work done by Lyons (1977). TCO features belong to first-order entities (physical things: *vehicle*, *animal*, etc.), second-order entities (situations: *happen*, *be*, etc.) and third-order entities (unobservable entities: *idea*, *theory*, etc.) and are used in a hierarchical combination to describe around 60,000 concepts from WordNet. Supersenses are 45 abstract concepts used as a first categorization level in WordNet, including labels such as *body*, *artefact*, *location*, *event*, *cognition*, etc. An example of the result of the process of linking our lemmatized responses to ontologies for the response *saber* (to know) is shown in Figure 5.

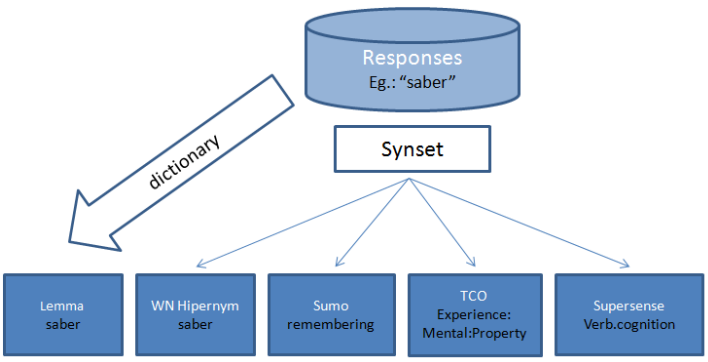

Figure 5: Generation of the diverse ontology categories

It is important to note that some of the responses could not be represented in these ontologies: there were categories (such as adverbs) that were not covered by the ontologies; or words whose synset was not mapped to an ontology. In these cases, the responses that were not linked with a category in an ontology were not taken into account for that specific ontology.

Each of these four types of ontological categorization (hypernyms, SUMO, TCO and supersenses) was used to calculate pairwise similarity in the same way as for the lemmas, substituting the lemma of the response with its corresponding category. In Table 3, the number of different categories used per resource is presented. Each of these different categories represents concepts with a different degree of granularity, as can be seen by looking at the number of categories.

Table 3: Number of different categories used in the data

lemma	2,691
hypernym	1,275
SUMO	595
TCO	328
Supersenses	40

As in the corpus perspective, we generated one similarity ranking for each formalization (to a total of 5 in this case, one for each type of resource) for all the pairs. In the next section we present the results of the comparison of corpus similarity rankings and psycholinguistic similarity rankings.

6. Comparison of both approaches

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

The quantitative study was performed through a Pearson correlation analysis. We studied the correspondence between verb sense similarities calculated using all the different formalizations for corpus and psycholinguistic data that we have explained in the previous sections. Summarizing, for corpus data we had three levels of granularity of semantic roles and two different structural combinations: patterns and constituents. In addition, we also experimented with the presence/absence of explicit syntactic information. As for psycholinguistic data, we had five different formalizations for similarity: lemmas, hypernyms, SUMO, TCO and supersenses. Moreover, we experimented with three versions of word associations: 1) similarities calculated using only the first response of each participant; 2) similarities obtained using the first three responses at most, as in De Deyne et al. (2013); and 3) all the responses obtained. We observed that taking into account the first three and all responses yielded higher correlations with corpus data than taking only the first response, in line with the findings of Nelson et al. (2000) and De Deyne et al. (2013). Overall, higher correlations were obtained when taking into account just the first three responses. Therefore, these are the results that we present and analyse here. These correlations were obtained comparing the similarity scores of every verb sense pair according to the corpus formalizations with its corresponding scores obtained using word association data.

The correlations' values can be seen in Table 4 (correlations of word associations with semantic roles) and Table 5 (correlations of word associations with semantic roles and syntax). Overall, the correlation strength ranges from medium to low. However, all of the correlations that we present here are significant ($p < 0.05$). When the correlation is not significant we indicate it with a dash.

Table 4: Correlations of semantic roles and word associations

			Psycholinguistic data				
Corpus data	Type of information	Type of format	lemma	hypernym	SUMO	TCO	supersense
	Fine-grained roles	patterns	**0.34**	0.31	0.18	0.30	0.24
		constituents	0.33	0.33	0.22	0.24	0.14
	Medium-grained roles	patterns	0.30	0.28	0.31	**0.34**	0.30
		constituents	0.30	0.28	0.30	0.33	0.20
	Coarse-grained roles	patterns	-	-	-	-	-
		constituents	-	-	-	-	-

Table 5: Correlations of semantic roles + syntax and word associations

			Psycholinguistic data				
Corpus data	Type of information	Type of format	lemma	hypernym	SUMO	TCO	supersense
	Fine-grained roles + syntax	patterns	0.32	0.29	0.16	0.24	0.19
		constituents	**0.40**	0.38	0.25	0.29	0.22
	Medium-grained roles + syntax	patterns	0.29	0.26	0.30	0.31	0.26
		constituents	0.35	0.32	0.38	0.35	0.28
	Coarse-grained roles + syntax	patterns	0.16	-	-	0.16	-
		constituents	0.20	-	0.16	-	-

If we look at the interplay of the presence/absence of syntax and the formalizations in constituents/patterns by comparing each row of Table 4 to the corresponding row in Table 5, we see that patterns that contain only semantic roles, or constituents that contain semantic roles with additional syntactic information correlate more with psycholinguistic data than the inverse combinations (semantic roles alone in constituents or semantic roles plus syntax in patterns). This suggests that information that deals with structure (either by expressing it through linear order or with syntactic categories) is relevant in the retrieval (and possibly configuration) of similarity relationships in the minds of the speakers. The lower correlation values of psycholinguistic data with corpus data formalized in patterns which include syntactic information show that the combination of these two types of information does not add up from a psycholinguistic viewpoint and in fact hinders the correlation.

As for the possible effects of including specific syntactic information, we see that, overall, roles plus syntax (Table 5) correlates the most with psycholinguistic data and also achieves the highest single correlation (fine-grained roles plus syntax in constituents yields a correlation of 0.40 with lemmas).

Regarding the effects of varying the granularity of the elements that are being compared, we find that, globally, fine-grained roles and fine-grained roles + syntax correlate more with fine-grained levels of word associations (lemmas and hypernyms), and that medium-grained roles correlate more with intermediate levels of word association (TCO for roles and TCO and SUMO for roles + syntax). However, coarse-grained roles obtain no significant

correlation or low correlation with lemmas and SUMO when syntax is present. Thus, the degree of specificity of the information related to the event (the argument information captured through semantic roles) is coherent with the degree of abstraction when representing verb senses through psycholinguistic data obtained from word associations. At the most abstract level of semantic roles, significant evidence for or against this connection cannot be found.

As for the qualitative analysis, we looked into the semantic, aspectual and subcategorization features of each of the similarity formalizations presented in the previous part. For any given formalization we first created a similarity ranking of verb sense pairs, ordering them from most to least similar according to the similarity values obtained. Next, 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, we went on to analyse the semantic, aspectual and subcategorizational coherence of these two groups of pairs. Specifically, we took into account how many of these pairs contained verb senses that had the same semantic field category according to Adesse macro-class and WordNet supersenses, and the same broad aspect category (static or dynamic). Finally, we counted how many subcategorization frames were shared between the members of the pairs. This analysis allowed us to detect the linguistic features that played a relevant role in establishing similarity relations in the different types of data.

To provide an overview of the results obtained in the semantic facet, we show in Figure 6 the ratio of the sampled similar and dissimilar pairs whose verb senses share the semantic field, according to the Adesse (upper plot) and WordNet supersense categories (lower plot). On the X axis we enumerate the different formalizations that are being compared in an abbreviated manner: FGR stands for 'fine-grained roles', MGR stands for 'medium-grained roles', CGR stands for 'coarse-grained roles' and WA stands for 'word associations'. On the Y axis we specify the ratio of pairs whose senses share the semantic feature, ranging from 0 to 1.

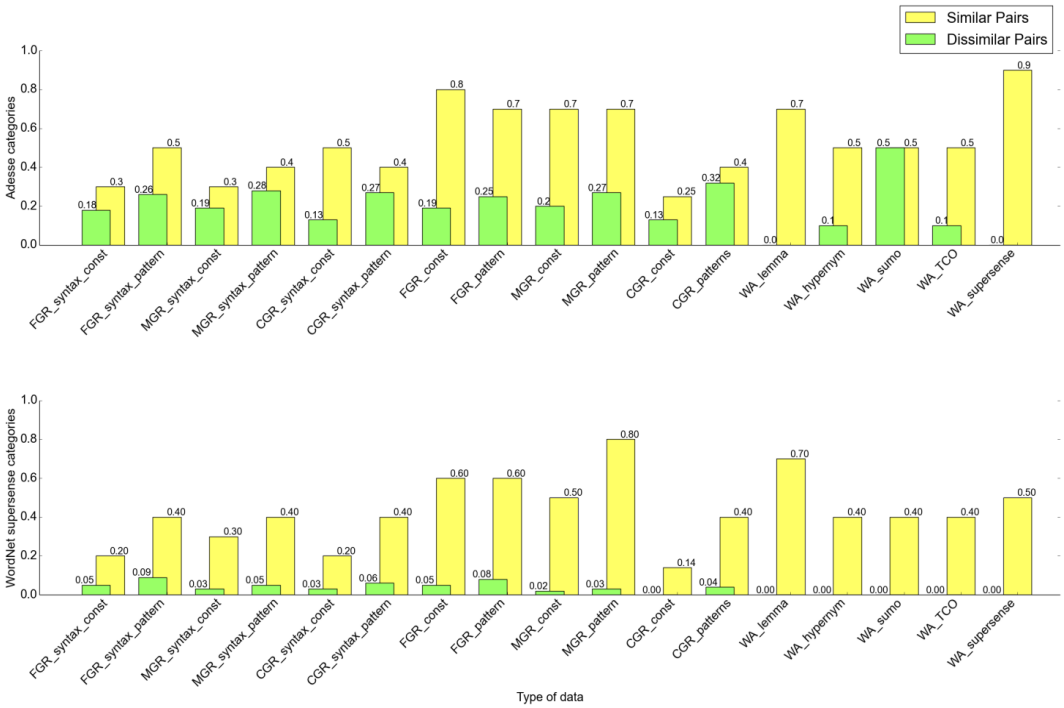

Figure 6: Ratios of pairs with the same semantic category

In the case of the semantic field analysis, there are some commonalities in the behaviour of the data according to the two types of semantic fields used for the analysis (Adesse and supersense categories). In almost every case, the similar pairs defined by the two perspectives contain senses that share the semantic field more frequently than dissimilar pairs, which is to be expected. However, there are some differences between the formalizations. Regarding corpus data, the presence of explicit syntactic information decreases the ratio of similar pairs with senses related to the same semantic field. In addition, the presence of patterns has an effect: it increases the ratio of dissimilar pairs that contain senses with a common semantic field. Focusing on the difference in the ratio of shared semantic field between similar and dissimilar pairs, this contrast is more striking for corpus data that contain semantic roles in isolation and for psycholinguistic data based on lemmas and on supersenses, the more and less specific categorizations, respectively. Indeed, in these formalizations, the number of similar pairs with senses that share the same semantic field is maximal, whereas the number of dissimilar pairs with senses related to the same semantic field is minimal. Therefore, these specific formalizations can be considered as more semantically structured than others.

In addition, it is worth mentioning the different results obtained using the Adesse and supersense semantic fields: the difference in ratio of shared semantic field between senses from similar and dissimilar pairs is more extreme using supersenses, whereas senses in dissimilar pairs usually have a higher ratio of shared semantic category according to Adesse and, therefore, the contrast between similar and dissimilar pairs is smaller. This can be related to the number and type of categories that each resource proposes to capture semantic fields. Supersenses have 15 coarse semantic domains for verbs (e.g. verbs of change, cognition, communication, etc.). From Adesse, we take the 6 macro-classes, that is, the upper level of the hierarchy. These classes are based on large semantic domains (mental, relational, material, etc.), with the difference that the classes are hierarchical and associated to examples in a corpus (Garcia-Miguel 2009).

As for the aspect facet of verb sense similarity, the results are shown in Figure 7. As in Figure 6, on the X axis the different formalizations taken into account are specified, and on the Y axis the ratio of pairs that have the same coarse aspect category, static or dynamic, is detailed for the given formalizations. 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 assigned to the verb senses in the Sensem lexicon to contemplate only two general categories, as in Dowty (1979) and Jackendoff (1983): dynamic events and states.

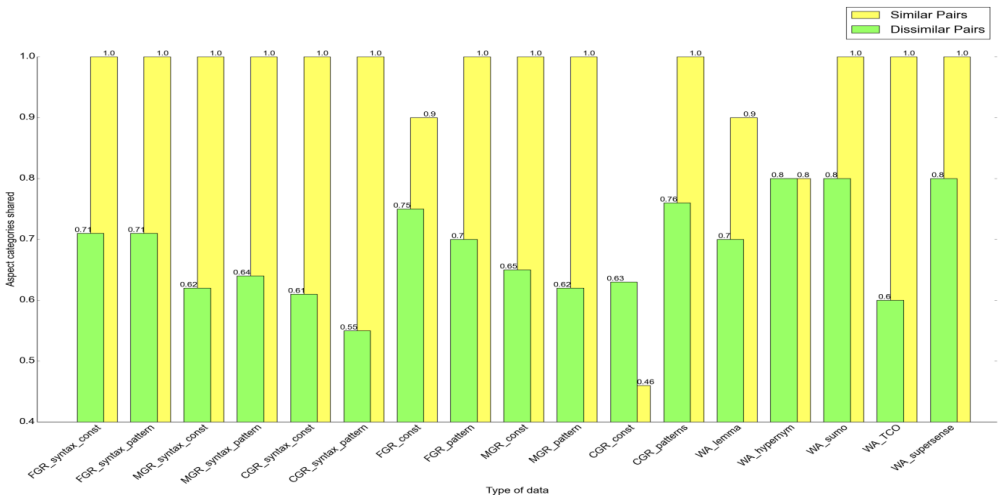

Figure 7: Ratio of sense pairs with shared aspect

Regarding the difference in aspectual coherence between the different formalizations, we can see in Figure 7 that, in general, similar pairs have a greater ratio of senses with identical aspect, except for the corpus formalization that includes coarse-grained roles organized in constituents. Moreover, it is important to note that this difference is more systematic in corpus data that includes syntax.

Finally, regarding the subcategorization facet of verb sense similarity, we show in Figure 8 the ratio of subcategorization frames that are shared by the members of the pairs. This information is further subdivided into two cases according to the frequency of the subcategorization frames. In the upper plot we present the difference between similar and dissimilar pairs when taking into account all subcategorization frames, while in the lower plot 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 Figures 6 and 7, on the X axis there are the different formalizations. On the Y axis, the ratio of subcategorization frames shared is displayed. This ratio is below 40% in all cases.

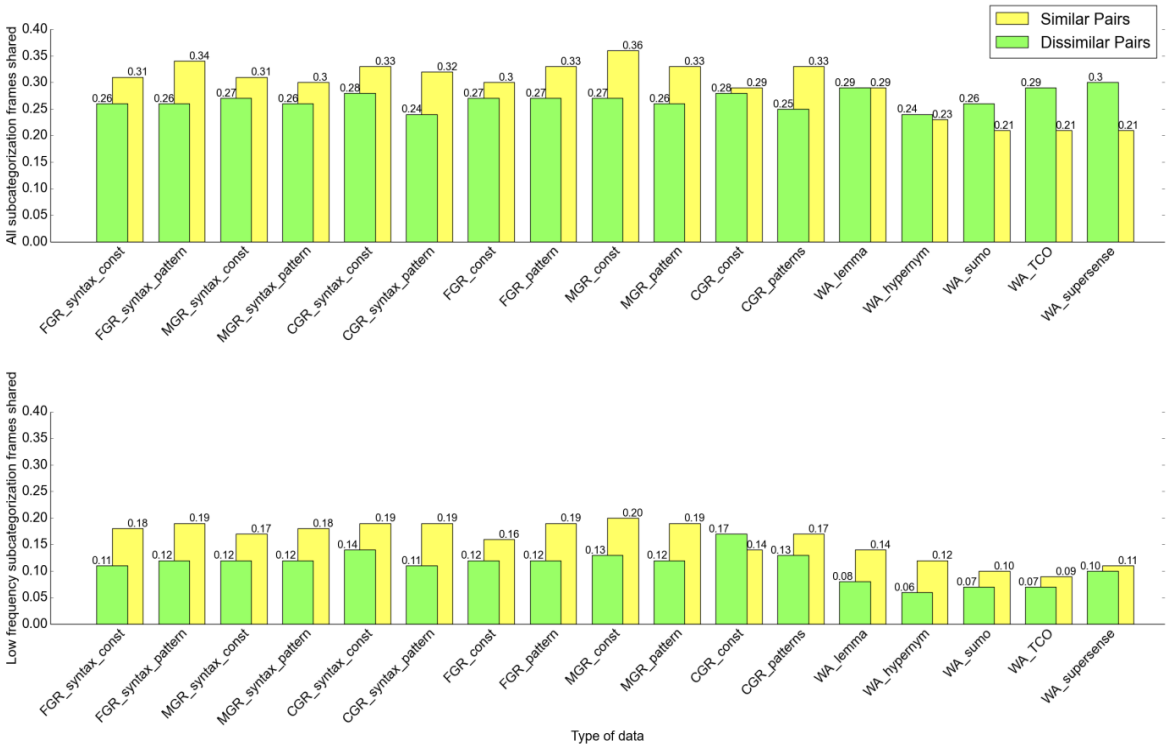

Figure 8: Ratios of shared subcategorization frames

If we take into account all the subcategorization frames (upper plot), we can see that there is a clear difference between corpus data and psycholinguistic data. While in corpus data, senses in similar pairs share more subcategorization frames than senses in dissimilar pairs; this situation is reversed for psycholinguistic data. This fact is along the lines of the findings of Schulte im Walde (2008), who suggested that word associations go beyond subcategorization information. However, if we look at the less frequent subcategorization frames (lower plot), the former situation no longer holds: in almost all cases, for corpus and psycholinguistic data, senses in similar pairs share more subcategorization frames than senses in dissimilar pairs. This suggests that verb sense similarities, as obtained from psycholinguistic data, are sensitive only to less frequent subcategorization frames and may

indicate that these types of frames, and not the frequent ones, have a role in shaping similarities in verbs in the lexicon. In any case, their role is small, given the small ratio of subcategorization frames shared in general. This diverges with most of the work done in automatic verb classification, which relies on subcategorization frames to create verb classifications that contain semantically coherent classes, following Levin's insight (Schulte im Walde 2006; Sun and Korhonen 2009). In contrast to this, we see that the pairs of senses that are semantically similar according to the Adesse and WordNet semantic fields 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 on their own do not have a role as crucial as previously thought in calculating similarity between verbs.

7. Conclusions

In this work we have compared two different perspectives of verb sense similarity which had not been previously related or comparatively studied: similarity as defined by the verb argument structure obtained from corpus and similarity as depicted by psycholinguistic data obtained through word associations. Unlike other approaches to similarity, the comparison has been established using verb senses instead of lemmas, as ambiguity is an obstacle to determine similarity at a detailed level. Using a set of 190 pairs created with 20 different verb senses, we have examined the relationship between different formalizations of these two perspectives, finding a low to moderate correlation in the similarity values that both assign to verb senses. We have also found that the correlation values are influenced by the degree of granularity of the different formalizations in such a way that finer granularity of semantic roles in corpus data correlates more with fine-grained formalizations of word associations. Conversely, medium-grained semantic roles correlate more with less specific formulations of word associations. Therefore, we can conclude that eventive information captured by semantic roles in corpus correlates to how verb senses are represented with respect to each other in the psycholinguistic dimension, also across different degrees of abstraction of the event information and the psycholinguistic representation. Additionally, we have found that certain linguistic aspects are crucial for strengthening the relationship between the two types of data: a structural component in corpus (syntactic function or linear order) is necessary to achieve higher correlation with psycholinguistic data.

Overall, we can conclude that word associations and argument structure data share common ground when it comes to determining similarity between verb senses, but each perspective puts the focus on diverse linguistic aspects. Globally, verb sense similarity as defined by psycholinguistic data is more sensitive to semantic relations in terms of broad semantic fields. More specifically, we find that corpus formalizations based on semantic roles alone are more semantically driven than formalizations that include specific syntactic information. As for aspectual cohesion, generally, similarity in corpus data is slightly more articulated by aspect than in psycholinguistic data. Finally, concerning the subcategorization information, differences between similar and dissimilar pairs across different formalizations are less relevant. Nevertheless, evidence suggests that similarity drawn from psycholinguistic data is not sensitive to subcategorization information; at least when taking into account all subcategorization frames and not only the less frequent ones. As a side observation, we can add that the lack of discriminating power of subcategorization frames in terms of differentiating similar from dissimilar verb senses points to the need to enrich this kind of information for NLP tasks that aim to model similarity for verbs.

It should be noted that, whereas the methodology yields interesting results, an analysis with more data could be carried out to further refine them. In future research we plan to take advantage of the commonalities as well as the specificities of each perspective to build a robust ranking of verb similarity that can be useful for corpus and NLP applications.

References

Albertuz, F. J. 2007. Sintaxis, semántica y clases de verbos: clasificación verbal en el proyecto ADESSE. *Actas del VI Congreso de Lingüística General, Santiago de Compostela, 3-7 de mayo de 2004.* 2015-2030.

Alvez, J., Atserias, J., Carrera, J., Climent, S., Oliver, A., and Rigau, G. 2008. Consistent annotation of eurowordnet with the top concept ontology. *Proceedings of Fourth International WordNet Conference (GWC'08)*.

Baker, S., Reichart, R., and Korhonen, A. 2014. An Unsupervised Model for Instance Level Subcategorization Acquisition. P*roceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 278-289.

Banerjee, S., & Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Ijcai* 3. 805-810.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. 2008. Language and simulation in conceptual processing. In *Symbols, embodiment, and meaning*. 245-283.

Bonial, C., Corvey, W., Palmer, M., Petukhova, V. V., and Bunt, H. 2011. A hierarchical unification of LIRICS and VerbNet semantic roles. *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference*. 483-489.

Brainerd, C. J., Yang, Y., Reyna, V. F., Howe, M. L., and Mills, B. A. 2008. Semantic processing in "associative" false memory. *Psychonomic Bulletin & Review*, 15. 1035-1053.

Bybee, J. 2010. *Language, usage and cognition*. Cambridge University Press.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. 2015. A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 1-7.

Christensen, J., Soderland, S., and Etzioni, O. 2010. Semantic role labeling for open information extraction. *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*. 52-60.

Chumbley, J. I., and Balota, D. A. 1984. A word's meaning affects the decision in lexical decision. *Memory & Cognition*, 12. 590–606.

Church, K., Gale, W., Hanks, P., and Hindle, D. 1989. Word Associations and Typical Predicate-Argument Relations. *Proceedings of the International Workshop on Parsing Technologies*.

Chwilla, D. J., and Kolk, H. H. 2005. Accessing world knowledge: evidence from N400 and reaction time priming. *Cognitive Brain Research*, *25*(3). 589-606.

Clark, H. H. 1970. Word associations and linguistic theory. *New horizons in linguistics*, *1*.

Coll-Florit, M. and S. Gennari. 2011. Time in language: Event duration in language comprehension, *Cognitive Psychology*, 62, 41-79.

Croft, W. and Cruse, A. 2004. *Cognitive linguistics*. Cambridge University Press.

De Deyne S. and Storms G. 2015. Word associations. In Taylor, J. R. (Ed.). (2015). *The Oxford handbook of the word*. OUP Oxford.

De Deyne, S., Navarro, D. J., and Storms, G. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*(2). 480-498.

De Deyne, S., Peirsman, Y., and Storms, G 2009. Sources of semantic similarity. *Proceedings of the 31th annual conference of the cognitive science society* (pp. 1834-1839).

De Deyne, S., Verheyen, S., and Storms, G. 2015. The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *The Quarterly Journal of Experimental Psychology*, *68*(8). 1643-1664.

Deese, J. 1962. Form class and the determinants of association. *Journal of verbal learning and verbal behavior*, *1*(2). 79-84.

Dowty, D. 1979. *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Dowty, D. 1991. Thematic proto-roles and argument selection. *Language*, 547-619.

Faruqui, M., and Dyer, C. 2014. Community evaluation and exchange of word vectors at wordvectors. Org. *ACL: System Demonstrations*.

Fellbaum, C. 2015. Lexical Relations. In *The Oxford Handbook of the Word*. Oxford University Press.

Fellbaum, C. 1998. *WordNet*. Blackwell Publishing Ltd.

Fernández-Montraveta, A., and Vázquez, G. 2014. The SenSem Corpus: an annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, *10*(2). 273-288.

Ferretti, T. R., McRae, K., & Hatherell, A. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory & Language*, 44. 516-547.

Fillmore, C. J. 1968. The case for case. *Universals in Linguistic Theory*. Holt, Rinehart and Winston. 1-88.

Fillmore, C. J., Johnson, C. R., & Petruck, M. R. 2003. Background to framenet. *International journal of lexicography*, *16*(3). 235-250.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman and E. Ruppin. 2001. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1). 116-131.

Fitzpatrick, T., & Izura, C. 2011. Word association in L1 and L2. *Studies in Second Language Acquisition*, *33*(03), 373-398.

García-Miguel, J M. 2009 A semantic classification of Spanish verbs. Verb Typologies Revisited: *Proceedings of A Cross-Linguistic Reflection on Verbs and Verb Classes*, February 5-7, Ghent University, Ghent, Belgium.

Garner, W. R. 1974. *The processing of information and structure*. New York: Wiley.

Gentner, D., and Markman, A. B. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1), 45-56.

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

Goldfarb, R., and Halpern, H. 1984. Word association responses in normal adult subjects. *Journal of Psycholinguistic Research*, *13*(1). 37-55.

Goldstone, Robert L.; Son, Ji Yun. 2005 Similarity. In Holyoak, Keith J.; Morrison, Robert G. *The Cambridge handbook of thinking and reasoning*. New York, NY, US: Cambridge University Press. 13-36.

Gonzalez-Agirre, A., Laparra, E., and Rigau, G. 2012. Multilingual Central Repository version 3.0. In *LREC*. 2525-2529.

Gruber, Jeffrey S. 1965. *Studies in Lexical Relations*. Diss. MIT. Published as *Lexical Structures in Syntax and Semantics*. Amsterdam: North Holland, 1976.

Hahn, U., Chater, N., and Richardson, L.B. 2003. Similarity as transformation. *Cognition,* 87. 1–32.

Hare, M., McRae, K., and Elman, J. L. 2003. Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, *48*(2). 281-303.

Hernández Muñoz, N. and López García, M. 2014. Análisis de las relaciones semánticas a través de una tarea de libre asociación en español con mapas auto-organizados. *RLA. Revista de lingüística teórica y aplicada*, *52*(2). 189-212.

Hill, F., Reichart, R., and Korhonen, A. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4). 665-695.

Jackendoff, Ray S. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, MA: The MIT Press.

Jackendoff, R. S. 1983. Semantics and Cognition. Cambridge, MA: MIT Press.

Jones, MN, J Willits, S Dennis, M Jones. 2015. Models of semantic memory. *Oxford Handbook of Mathematical and Computational Psychology*. 232-254.

Keil, F. C. 1989. *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. 1973. An associative thesaurus of English and its computer analysis. *The computer and literary studies*. 153-165.

Klein, D.E. and Murphy, G.L. 2002 Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language*, 47. 548-570.

Kozima, H., and Furugori, T. 1993. Similarity between words computed by spreading activation on an English dictionary. *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*. 232-239.

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

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

Lin, D. 1998. An information-theoretic definition of similarity. *ICML* 98. 296-304.

Liu, D., & Gildea, D. 2010. Semantic role features for machine translation. *Proceedings of the 23rd International Conference on Computational Linguistics*. 716-724.

Luce, P. A., Pisoni, D. B., and Goldinger, S. D. 1990. Similarity neighborhoods of spoken words. In Altmann, G. T. M (ed.). *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. *ACL–MIT Press series in natural language processing*. 122-147.

Lyons, J. 1977. *Semantics* (vols I & II). *Cambridge CUP*.

Maki, W. S., and Buchanan, E. 2008. Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*(3). 598-603.

Manning, J. R., & Kahana, M. J. 2012. Interpreting semantic clustering effects in free recall. *Memory*, *20*(5). 511-517.

McKoon, G., and Ratcliff, R. 1992. Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(6), 1155.

McRae, K., Hare, M., Elman, J. L., and Ferretti, T. R. 2005. A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33. 1174-1184.

McRae, K., Khalkhali, S., and Hare, M. 2012. Semantic and associative relations: Examining a tenuous dichotomy. In V. F. Reyna, S. Chapman, M. Dougherty, & J. Confrey (Eds.), T*he adolescent brain: Learning, reasoning, and decision making*. Washington.

Merlo, P., & Stevenson, S. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, *27*(3), 373-408.

Michelbacher, L., Evert, S., and Schütze, H. 2007. Asymmetric association measures. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007)*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space. *ICLR Workshop*.

Moldovan, C. D., Ferré, P., Demestre, J., & Sánchez-Casas, R. 2015. Semantic similarity: normative ratings for 185 Spanish noun triplets. *Behavior research methods*, 47(3), 788-799.

Mollin, S. 2009. Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory*, *5*(2). 175-200.

Neely, J. H. 1991. Semantic priming effects in visual word recognition: A selective review of current findings and theories. *Basic processes in reading: Visual word recognition*, *11*. 264-336.

Nelson, D. L., McEvoy, C. L., and Dennis, S. 2000. What is free association and what does it measure? *Memory & Cognition*, 28. 887-899.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. 1998. *The University of South Florida word association, rhyme, and word fragment norms*. http://www.usf.edu/FreeAssociation/ (16 April, 2016).

Niles, I., and Pease, A. 2001. Towards a standard upper ontology. In*Proceedings of the international conference on Formal Ontology in Information Systems*, Volume 2001. 2-9.

Nordquist, D. 2009. Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory*, *5*(1). 105-130.

Padró, L., and Stanilovsky, E. 2012. Freeling 3.0: Towards wider multilinguality. *LREC 2012*.

Palmer, M., Gildea, D., & Kingsbury, P. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, *31*(1). 71-106.

Patwardhan, S., Banerjee, S., and Pedersen, T. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Computational linguistics and intelligent text processing*. 241-257.

Peirsman, Y., and Geeraerts, D. 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*. 648-656.

Plaut, D. C. 1995. Semantic and associative priming in a distributed attractor network. *Proceedings of the 17th annual conference of the cognitive science society* Vol. 17, No. 2. 37-42.

Rayner, K. and Frazier, L. 1989 Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5). 779-790.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of IJCAI-95*. 448–453.

Riordan, B., and Jones, M. N. 2007. Comparing semantic space models using child-directed speech. In D. S. MacNamara & J. G. Trafton, *Proceedings of the 29th Annual Cognitive Science Society*. 599-604.

Rodd, J., Gaskell, G. and Marslen-Wilson, W. 2002 Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access. *Journal of Memory and Language*, 46. 245-266.

Roediger, H. L., III, Watson, J. M., McDermott, K. B., & Gallo, D. A. 2001. Factors that determine false recall: A multiple regression analysis. Psychonomic Bulletin & Review, 8. 385-407.

Sahlgren, M. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm: Stockholm University publishing service, Ph.D. thesis.

Savage, C., Lieven, E., Theakston, A., and Tomasello, M. 2003. Testing the abstractness of children's linguistic representations: Lexical and structural priming of syntactic constructions in young children. *Developmental Science*,*6*(5). 557-567.
Schuler, K. K. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Philadelphia, PA: University of Pennsylvania, Ph.D. thesis.
Schulte im Walde, S. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32, 159-194.
Schulte im Walde, S. 2008. Human associations and the choice of features for semantic verb classification. *Research on Language and Computation*, *6*(1).79-111.
Schulte im Walde, S., Melinger, A., Roth, M., and Weber, A. 2008. An empirical characterisation of response types in German association norms. *Research on Language and Computation*, *6*(2). 205-238.
Schütze, H. 1992. Dimensions of meaning. In *Supercomputing'92 Proceedings*. 787-796.
Shen, D., and Lapata, M. 2007. Using Semantic Roles to Improve Question Answering. In *EMNLP-CoNLL*. 12-21.
Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, 27.
Spence, D. P., and Owens, K. C. 1990. Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, *19*(5). 317-330.
Steyvers, M., Shiffrin, R. M., and Nelson, D. L. 2004. Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. 237-249.
Sun, L. and Korhonen, A. 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. EMNLP.
Tversky, A. 1977. Features of similarity. *Psychological Review, 84* (4). 327–352.
Vitevitch, M. S., and Luce, P. A. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3). 374-408.
Vitevitch, M. S., and Luce, P. A. 2016. Phonological Neighborhood Effects in Spoken Word Perception and Production. *Annual Review of Linguistics*, *2*. 75-94.
Wettler, M., and Rapp, R. 1993. Computation of word associations based on the co-occurrences of words in large corpora. In *Proceedings of the 1st Workshop on Very Large Corpora*. 84-93.
Yang, D., and Powers, D. M. 2006. Verb similarity on the taxonomy of WordNet. *Proceedings of GWC-06*. 121–128.

Appendices

Appendix 1. Verb senses, definitions and frequency (between brackets)

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. (18)
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)

Appendix 2. Semantic fields of the verb senses

Verb sense	WordNet supersense	Adesse macro-class
abrir 18	change	material
cerrar 19	change	material
crecer 1	change	material
dormir 1	activity (bodily)	material
escuchar 1	perception	mental
estar 14	state	relational
explicar 1	communication	verbal
gestionar 1	activity (social)	material
gustar 1	cognition	mental
montar 2	movement	material
morir 1	change	existential
parecer 1	state	relational
pensar 2	cognition	mental
perseguir 1	movement	material
trabajar 1	activity	material
valorar 2	communication	relational
valer 1	state	relational
ver 1	perception	mental
viajar 1	movement	material
volver 1	movement	material

Appendix 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 condición. / TO BE in a specific state
EXPLICAR una cuestión. / TO EXPLAIN an issue
GESTIONAR un trámite. / 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