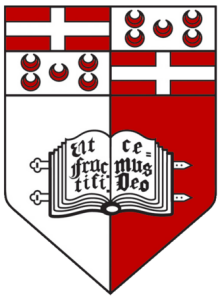
Automatic Correction of Multilingual Proposition Banks for Semantic Role Labelling

Carlos Fernando Diez Sánchez

MSc. Dissertation



Department of Intelligent Computer Systems

Faculty of Information and Communication Technology

University of Malta

2017

Supervisors:

Lonneke van der Plas, Institute of Linguistics and Language Technology, University of Malta

German Rigau i Claramunt, Faculty of Informatics, University of the Basque Country

Submitted in partial fulfilment of the requirements for the Degree of European Master of Science in Human Language Science and Technology

M.Sc. (HLST)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY UNIVERSITY OF MALTA

Declaration

Plagiarism is defined as "the unacknowledged use, as one's own work, of work of another person, whether or not such work has been published" (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I, the undersigned, declare that the Master's dissertation submitted is my own work, except where acknowledged and referenced.

I understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Student Name: Carlos Fernando Diez Sánchez

Course Code: CSA5310 HLST Dissertation

Title of work: Automatic Correction of Multilingual Proposition Banks for Semantic Role Labelling

Signature of Student:

Date: January 4th, 2018

Supervisors

Lonneke van der Plas

Institute of Linguistics and Language Technology

University of Malta

and

German Rigau i Claramunt

Faculty of Informatics

University of the Basque Country

Local Advisor

Michael Rosner

Department of Intelligent Computer Systems

University of Malta

Acknowledgements

Success is never the result of one person.

Gracias a mi familia, a mi mamá, a mi papá y a mi hermano. Todo lo que hago, incluso en la distancia, lo hago pensando en ustedes.

Thank you very much to everyone who joined me and helped me during this extraordinary, life-changing, LCT experience, especially to the coordinating authorities at Saarland University (Saarbrücken, Germany), the University of the Basque Country (UPV-EHU), and the University of Malta (UoM), to the professors and classmates I had the privilege to meet in the Basque Country, in Malta, and during the annual meetings.

Thank you to my supervisors, Dr Lonneke van der Plas (UoM) and Dr German Rigau i Claramunt (UPV-EHU), for their support, invaluable advice and feedback on this research.

Lonneke, words are not enough to thank you for all you have done for this work, a project that, despite all my evident limitations, I am very proud to have shared with such an incredible woman.

Germá n, muchas gracias, pues tus contribuciones desde el primer año y hasta este momento han sido fuente de ese tipo de inspiración y motivación que considero esencial para toda investigación digna de serlo.

Thank you to the people working at IBM Research Group (California, USA), especially to Yunyao Li and Chenguang Wang. The way this work has been enriched since we first met would have been unimaginable otherwise.

This research project was funded by the European Commission in the framework of the Erasmus Mundus Joint Master Degrees Programme, and also part of a joint study agreement between the UoM and the IBM Research Group in Almaden (California, USA).

Abstract

Semantic Role Labelling (SRL), the task of automatically identifying and labelling a predicate-argument structure at the sentence level, has been shown to be important for a broad spectrum of natural language processing (NLP) applications, such as information extraction, summarization, plagiarism detection, question answering, and machine translation. Due to the high costs of manual annotation for SRL, Akbik *et al.* (2016b) proposed a method to generate Proposition Banks (PBs) for novel languages by means of annotation projection in parallel corpora, followed by a manual correction step in order to filter and merge the created semantic frames. In this project, we propose a method to perform the correction process semi-automatically by using a multilingual distributional semantic model and a learning algorithm for classification. Although the project is aimed at creating a Corrected Spanish PB, the method is language-independent and will be used to correct PBs in other languages as well. The method was evaluated on the manually Curated French, German and Chinese PB, and obtained promising results in the Projected Spanish PB, which are expected to help speed up the manual correction process overall.

Keywords: Language Independent Semantic Role Labelling, Automatic Correction Method, Multilingual Proposition Banks, and Multilingual Distributional Semantics

Contents

List of Abbreviations

NLP	Natural Language Processing
NLU	Natural Language Understanding
SRL	Semantic Role Labelling
PB or PropBank	Proposition Bank
DS	Distributional Semantics
DSM	Distributional Semantic Models
BTA	Back-Translation Assumption
I-DR	Increased Dimensionality Reduction
D-DR	Decreased Dimensionality Reduction

List of Tables, Samples and Figures

Chapter 1

Introduction

1.1 Semantic Role Labelling

The task of semantic role labelling (SRL) refers to the automatic analysis of the predicate-argument structure at the sentence level in un-annotated text, that is to say, identifying and tagging the predicate[1] and its arguments (constituents) with semantic labels in a given corpus, indicating the role they play (kind of relation with the predicate) within a semantic frame.

This shallow analysis by a trained statistical system allows determining 'who did what to whom', plus optional roles and adjuncts ('how, when and where') [2] (See Figure 1.1) in the sentence or discourse, and therefore characterizes the participants (entities) and the events established by the predicate in a more stable or consistent representation across syntactically different sentences (that is to say, it does not vary if there are syntactic alternations) or even synonymous verbs, as semantic roles encode certain aspects of our conceptualization of the world. As Palmer, Gildea & Xue (2010, p. 1) put it: "For computers to make effective use of information encoded in text, it is essential that they be able to detect the events that are being described and the event participants".

[1] The term 'predicate' usally refers to verbal predicates, but it can also be applied, for

[2] Although nowadays there are different labelling schemes, they all should cover these original 'labels' (i.e. 'who', 'what', 'to whom'), along with a set of features (i.e. +/- ANIMATE, etc.). As stated by Fowler (1996): "[…] agency, state, process and so on, seem to be the basic categories in terms of which human beings present the world to themselves through language."

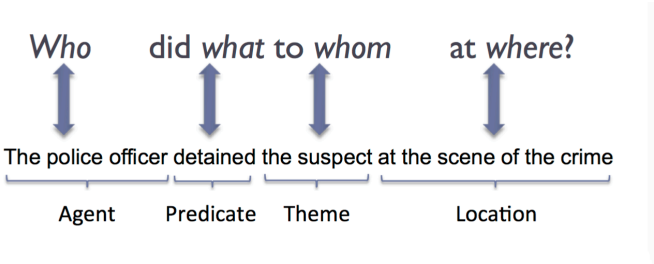


Figure 1.1 Semantic role labelling by Jurafsky & Martin (2017)[3].

Recognizing these event structures has been shown to be crucial for a broad spectrum of natural language understanding (NLU) applications, such as information extraction (Fader *et al.*, 2011), summarization (Khan *et al.*, 2015), plagiarism detection (Osman *et al.*, 2012; Paul & Jamal, 2015), question answering (Shen & Lapata, 2007; Maqsud *et al.*, 2014), and machine translation (Aziz *et al.*, 2011; Xiong *et al.*, 2012; Lo *et al.*, 2013), among others. The potential for semantic generalizations (namely, fewer types are needed for describing the entire lexicon) in NLU can be seen, for example, when describing inference rules, which could be written with respect to a finite set of cases rather than thousands of individual lexical items.

The substantial progress in the task of automatic semantic annotation in the last decades is based on the availability of annotated corpora, as they facilitate the development of semantic role labelling systems based on supervised machine learning techniques. Currently, there are three frameworks proposed for annotating corpora which provide an explicit predicate-argument structure: FrameNet (Baker, Fillmore, & Lowe, 1998), VerbNet (Schuler, 2005; Kipper *et al.*, 2006), and PropBank (or PB, Kingsbury & Palmer, 2002; Palmer, Gildea & Kingsbury, 2005).

All of them have been developed on the basis of English data, and they are compatible, although they differ in the granularity of the labels, among other things. As stated by Samardžic *et al.* (2010), although these frameworks are implementations of different linguistic theories, they have been developed to account for universal phenomena, so they should be suitable to be applied to other languages as well. In fact, Akbik & Li argue English PropBank labels have the potential to become a basis

[3] Draft of the 3rd ed. Website: *https://web.stanford.edu/~jurafsky/slp3/slides/22_SRL.pdf*

of universal semantic labels: “Such a unified representation of shallow semantics [...] may facilitate applications such as multilingual information extraction and question answering, much in the same way that universal dependencies facilitate tasks such as cross-lingual learning and the development and evaluation of multilingual syntactic parsers (Nivre, 2015). The key questions, however, are (1) to what degree English PropBank frame and role labels are appropriate for different target languages; and (2) how far this approach can handle language-specific phenomena or semantic concepts.” (2016a, p.2) Lexical resources similar to the English PB have been created for languages such as Chinese (Xue & Palmer, 2005) and Hindi (Bhatt *et al.*, 2009).

Unfortunately, the resources required to create SRL models are costly and not always available for most languages. Projects such as the ones mentioned above require corpora manually annotated with semantic labels, in order to produce statistical SRL parsers for the language, a factor hindering the expansion of SRL systems to new target languages.

As a promising alternative, previous approaches (such as Pado, 2007; Pado & Lapata, 2009; Van der Plas *et al.*, 2011) have investigated the direct annotation projection of semantic labels from English to other target languages via parallel corpora, in order to create the language specific PBs and thus enabling the training of SRL systems for other languages (Figure 1.2).

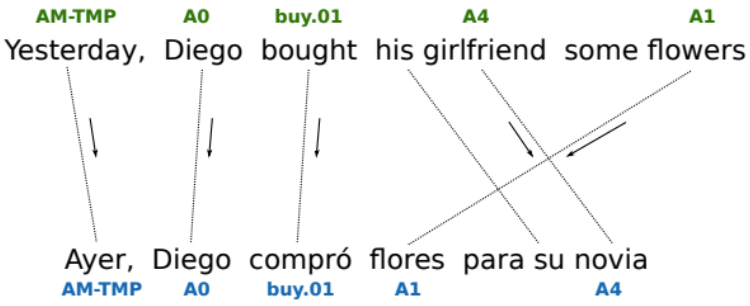


Figure 1.2. Annotation projection for a word-aligned English Spanish sentence pair by Akbik & Li (2016a).

The underlying assumption of this approach is the semantic equivalence of the original and translated sentences, where the semantic labels can be projected onto the aligned target corpus (Pado & Lapata, 2009). With this method, however, only a

subset of all the semantic labels is correctly projected due to translation shifts and non-literal translations (Akbik *et al.*, 2015). Previous works have proposed lexical and syntactic constraints in order to increase the quality of the projection (Pado & Lapata, 2009; Van der Plas *et al.*, 2011; Akbik *et al.*, 2015).

Under this assumption, Akbik *et al.* (2015) proposed a method for creating a SRL model in two stages for languages lacking the appropriate resources, by using monolingual (English) SRL models and multilingual parallel data. They generated Proposition Banks for 7 languages in three different language families, which included Spanish, although it still presents inconsistencies.

Akbik, Guan, & Li (2016) present a manual method to address lexicon-level inconsistencies in three Proposition Banks (PBs for French, German and Chinese) that were automatically generated using annotation projection. The manual method is performed in two steps: filtering to identify incorrect frames, and merging to reduce redundancy.

In this work, I propose an automatic alternative to the second step in the manual frame correction method proposed by Akbik *et al.* (2016). I make use of the multilingual, distributional semantics (DS) method, and prove that it can be leveraged for this task, within the framework of semantic role labeling (SRL).

The main problem with the manual method is the number of lexical items that need to be evaluated for each resource. That manual review has to be done by a language expert, which it is costly, time-consuming, and can lead to inconsistencies if there are variations in the annotators' criteria. Besides, we need to take into account that, although this work focuses on correcting the Projected Spanish PB, my method is language-independent, and could be also used to correct all the multiple PBs generated with this method in other languages.

Although my method has not reached the stage where it leads to the same decisions as made by the language experts and correct every case, it can function in a semi-automatic manner, that is to say, the method can detect strong and weak redundant candidates, which helps to speed up and improve the criteria of the correction process. Finally, even when the method does not get to the same results, it reveals the

thought process and the other linguistic parameters the human 'curators' used when performing such task.

1.2 Research Questions, Aims and Objectives

The main research questions that this thesis aims to answer are the following:

- Is it possible to partially automate the correction of a Projected Proposition Bank for Spanish (created by annotation projection) using multilingual DS as main method?

- If so, what are the most important parameters to improve the performance of a distributional semantic model (DSM) and how can these parameters be combined?

- Can this method be applied across languages, that is to say, for a PB in any language created with the same process?

Therefore, the aim of the thesis is to create an automatic alternative to the second step of the manual correction method proposed by Akbik *et al.* (2016), which identifies redundant semantic frames (shown as verb pair candidates) in Projected Proposition Banks (PropBanks or PBs) and merge them into one when appropriate. The creation of high-quality PBs has shown to be an important step to facilitate the development of semantic role labelling systems based on supervised machine learning techniques for languages lacking such resources.

The creation of the proposed method is based mainly on the combination of several multilingual DSMs, whose outputs (similarity scores), along with a role comparison between semantic frames, serve as features in a machine learning setting. To do this, a number of DSMs with different data and different parameters need to be built and tested to evaluate their usefulness when solving the proposed task. Although it is already possible to set a decision threshold (merging / not merging) based on the similarity scores of each individual model, a better and more consistent performace is achieved among languages when using a suitable machine learning algorithm to

combine the output of such models. Finally, role comparison can be added as an extra feature to further improve its performance. It is possible to develop and test the proposed method due to the existence of the already manually 'Curated' (or rather Corrected) PBs in French, German and Chinese by Akbik *et al.* (2016), which serve as a kind of 'gold standard', and a manually annotated sample of the Projected Spanish PB (in Akbik & Li, 2016a).

1.3 Thesis structure

The thesis is organized as follows: In Chapter 2, I talk about the complexity to build automatic and efficient SRL systems from the linguistic point of view (2.1); I also talk about the lexical resources available for training SRL systems (2.2), with a focus on the complex Spanish AnCora, which served during the first stage of the development of the whole method. In Section 2.3, I refer to the multilingual approaches using transfer methods, with a focus on the manual correction method for Projected PBs presented by Akbik *et al.* (2016) (2.3.1), which I try to make more efficient; and, since our automatic method is based on distributional semantic models (DSMs), I review some of the applications that have been built under these hypothesis (2.4).

In Chapter 3, after an outline of the proposed automatic method (3.1), the benefits that entail (3.2), and describe a small feasibility study I carried out when I was looking for other forms of evaluation (3.3), I talk about the design of the proposed method and the methodology I followed, with a focus on the collection of lexical resources (3.4), and extraction of verb pair candidates from the different PBs (3.5-6) as a first step. In Section 3.7, I talk about the architecture of the main component of the correction process, the DSMs, and how it was applied (3.8). I also talk about a partial evaluation I did with a gold standard resource for semantic similarity (3.9) and how I devised a combined model using machine learning algorithms to improve the system overall (3.5).

In Chapter 4, I report the results and evaluation of the semantic model on the Curated PBs in French, German and Chinese (4.1), and the results and evaluation on

the Projected Spanish PB (4.2), along with an error analysis (4.2.1). Finally, in Chapter 5, I discuss the conclusions of the thesis (5.1) and some directions for future work (5.2).

Chapter 2

Background & Related Work

In this chapter, I present the background to our research and some of the related works on this topic. In Section 2.1, I present one of the most comprehensive works on automatic SRL by Gildea & Jurafsky (2002), which shows the complexity of the semantic task when all the linguistic phenomena involved are taken into account, and the need to look for other alternatives. In Section 2.2, I focus on the importance of the lexical resources or annotated corpora to train SRL systems, and analyse the case of AnCora, the only lexical resource available for Spanish that existed before the Projected Spanish PB for SRL, which required a lot of work at different linguistic levels in order to include semantic annotation. In Section 2.4, I mention the most relevant works that have used cross-lingual projection, and specifically, in Section 2.5, the manual method to correct Projected PBs proposed by Akbik *et al.* (2016). Finally, I describe the applications for which Distributional Semantic Models (DSMs) have been used before this work (2.6).

2.1 An early automatic SRL system

In their seminal and comprehensive work on SRL, Gildea & Jurafsky (2002) presented a statistical algorithm for automatically learning to identify all the abstract or domain-specific semantic roles in a wide variety of predicates (unrestricted text).

Their statistical algorithm is trained on the manually annotated FrameNet database (Baker, Fillmore, & Lowe, 1998), with roughly 50,000 sentences from the British National Corpus. They presented a basic and extended version of their statistically trained system. The system is trained with 36,995 sentences, and extracts the next features, based on the assumption that syntactic realization can be generated from semantics (linking theory): *Phrase Type*, most commonly expressed as noun phrases (NPs, 47% of frame elements in the training set), and prepositional phrases (PPs, 22%); *Governing Category*, with two values, subjects (S) and objects of verbs (VP), which allows rules such as 'if there is and underlying Agent, it becomes the syntactic Subject of the sentence'; *Parse Tree Path*, indicating upward or downward movement

(through the parse tree) from the target word to the constituent in question (the 'PP argument / adjunct' as the most frequent path, with 14.2%), this feature is important later, when the constituents must be identified as frame elements for a given target word; *Position*, which simply indicates whether the constituent to be labelled occurs before or after the predicate defining the semantic frame; *Voice*, because the distinction between active and passive verbs plays an important role in the connection between semantic role and grammatical function; and *Head Word*, indicating the head of any phrase type (for example, head words of noun phrases can be used to express selectional restrictions on the possible semantic roles).

For their experiments, they divided one-tenth of the annotated sentences as a test set, and another one-tenth was set aside as a tuning set. To label the semantic role of a constituent automatically, they estimated a probability distribution indicating how likely the constituent is to fill a possible role, given all the features and the target word (predicate). Due to the sparsity of the data, the classifier is built by combining probabilities from different subsets of the features, because there is a trade-off between more-specific distributions (with high accuracy but low coverage), and less-specific distributions (with low accuracy but high coverage). To combine the strengths of the various distributions, they compared against the baseline three methods: linear interpolation, geometric mean, and back-off combination, where a lattice was constructed (more-specific conditioning events to less-specific).

Besides, in order to generalize the information of the head words in the noun phrases (50% of the constituents) seen in their training data, the researchers evaluated three different approaches: automatic clustering, semantic hierarchy (using WordNet), and bootstrapping. The automatic clustering technique is based on the expectation that words with similar semantics will tend to co-occur with the same sets of words. A total of 2,610,946 verb-object pairs from the British National Corpus were used as the training data for the clustering method, with a further 290,105 pairs used as a cross-validation set. In the semantic hierarchy method, if a headword had not been seen in the training examples, they try to scale any level in the hierarchy for which some training data were available. The unannotated data used for the bootstrapping method (that is, labelling the unannotated data with their own automatic system) consisted of 156,590 sentences, increasing roughly six times the

total amount of data available for the experiment (36,995 annotated training sentences).

A step further was to integrate the automatic syntactic parser described in Collins (1999), a form of chart parsing, and the semantic-role probability model, which computes the best frame element, and then choose the highest probability overall. In this way, they expected to improve the accuracy of the whole system, although the results show a not statistically significant increase in recall of frame elements. Finally, they try to generalize the information of the head words seen in the training data to the unseen predicates, frames, and domains, using the information in the FrameNet database. For that purpose, they collapsed the 67 FrameNet roles into a set of 18 abstract thematic roles, as abstract roles should be more useful for generalizing when the frame information cannot be found in the training data.

In order to see how much can be accomplished with as simple a SRL system as possible, the authors constructed a minimal back-off system, with just two distributions. This system classified 76.3% of the frame elements correctly. The full system performed with a 80.4% accuracy, which can be compared to the 40.9% by always choosing the most probable role for each target word. The model considers separately the question of locating the (boundaries of) frame element in a sentence. The features used are path, target word and head word. As an example, the path VB VP NP (direct object of a verb as the target) has a high probability of being a frame element. In this case, 79.6% of the constituents that had been previously identified as such were assigned this time the correct role. When clustered statistics are used with the full system, the performance on NP constituents increases from 83.4% to 85.0% (statistically significant at $p < .05$). Over the entire test set, it increases from 80.4% to 81.2%. The accuracy of the WordNet technique is roughly the same as the one shown in the automatic clustering (84.3%). As for the bootstrapping method, the accuracy is of 81.0%, reasonably close to the 87.0% by the system trained on the annotated data. The main difference between the three methods is the estimated coverage, where the automatic-clustering method performed better (only 2.1% of unseen cases). As for the integrated system, the frame element identification task obtained 71.3% in precision, and 67.6% in recall, whereas the frame element labelling task obtained 60.8% in precision and 57.6% in recall.

Finally, for unseen predicates they obtained an overall performance of 82.1% for thematic roles (compared to 80.4% for frame-specific roles). For unseen frames, the system achieves performance of 51.0%, compared to the 82.1% of the original system, clearly a difficult task even within a domain. For unseen domains, taking into account that the FrameNet database covers only a small portion of the language, the system classified 39.8% of the frame elements correctly. All this later results were expected, because making successively broader generalizations to more distant predicates degrades the performance, although this was be useful to help future statistical systems to generalize the labelling from similar words when training data is available.

Their comprehensive research left many topics unexplored. In their first original system, they did not use the maximum-entropy technique to combine the strengths of the various distributions for each feature. When they were trying to generalize the statistical information of the head words in the noun phrases to other head words using a semantic hierarchy (WordNet), they only used the first sense that was listed in the database. A plausible hypothesis is that a word sense disambiguation module capable of distinguishing sense could improve the results. Other aspects that could be included are a dictionary of proper nouns, indefinite and non animate pronouns, and a module for anaphora resolution. The authors did not try to combine the three methods they used (automatic clustering, semantic hierarchy, and bootstrapping). An in-depth study could find if the three systems are complementary or if they overlap in some way.

2.2 Lexical Resources for SRL

The substantial progress in the task of automatic semantic annotation in the last decades has been based on the availability of annotated corpora, as this facilitates the development of semantic role labelling systems based on supervised machine learning techniques.

Currently, there are three frameworks proposed for annotating corpora which provide an explicit predicate-argument structure: FrameNet (Baker, Fillmore, &

Lowe, 1998), VerbNet (Schuler, 2005; Kipper *et al.*, 2006), and PropBank (PB, Kingsbury & Palmer, 2002; Gildea & Palmer, 2002; Palmer, Gildea & Kingsbury, 2005).

All of them have been developed on the basis of English data, and they are compatible, although they differ in the granularity of the labels. As stated by Samardžic *et al.* (2010), although these frameworks are implementations of different linguistic theories, they have been developed to account for universal phenomena, so they should be suitable to be applied to other languages as well.

Akbik & Li argue that the English PropBank labels have the potential to become a basis of universal semantic labels: "Such a unified representation of shallow semantics [...] may facilitate applications such as multilingual information extraction and question answering, much in the same way that universal dependencies facilitate tasks such as crosslingual learning and the development and evaluation of multilingual syntactic parsers (Nivre, 2015). The key questions, however, are (1) to what degree English PropBank frame and role labels are appropriate for different target languages; and (2) how far this approach can handle language-specific phenomena or semantic concepts." (2016a, p.2)

Lexical resources similar to the English PB have been created for languages such as Chinese (Xue & Palmer, 2005) and Hindi (Bhatt *et al.*, 2009).

2.2.1 AnCora, a lexical resource for Spanish

One of those efforts to build a multilingual resource is AnCora (Taulé, Martí, & Recasens, 2008; Aparicio, Taulé & Martí, 2008) a multilingual corpus and lexicon for Catalan and Spanish with linguistic annotations at different, independent levels:

- Morphological level first (PoS and lemmas),
- Syntactic level next (constituents and functions), and finally,
- Semantic level, with semantic verb classes, the argument structure of verbal predicates, thematic roles associated with each argument, strong and weak named entities (NEs, following Borrega, Taulé & Martí, 2007) and WordNet

synsets for all nouns.

The annotation procedure was performed either manually, semiautomatically, or automatically, depending on the characteristics of each kind of linguistic information, sequentially from lower to upper layers. The tag sets are the same in both languages. All the layers were made independent in order to make data management easier.

2.2.1.1 AnCora Corpus

The AnCora corpus consists of 500,000 words in Catalan and the same amount in Spanish, making it the largest annotated corpus freely available for those languages. The AnCora corpora have been used to train and test several NLP systems, and it has been used in several international evaluation competitions, such as CoNLL-2006, CoNLL-2007 and SemEval-2007 for different syntactic and semantic NLP tasks; and the CoNLL-2009 Shared Task (Hajič *et al.*, 2009), dedicated to the joint parsing of syntactic and semantic dependencies in multiple languages.

AnCora was built from the previous 3LB (Civit & Martí, 2004) and CESS-ECE (Martí & Taulé, 2008) corpora, which come mostly from newspaper and newswire articles. The Spanish corpus (AnCora-Es) contains 75,000 words from Lexesp, a 6-million-word corpus by Sebastiá n-Gallés *et al.* (2000), 225,000 words from the EFE Spanish news agency, and 200,000 from the Spanish version of the *El Periódico* newspaper.

The Spanish corpus contains 17,709 sentences (with 29.84 lexical tokens on average); 54,075 predicates (3.05 per sentence on average) and 122,478 arguments (2.26 per predicate on average); 73.34% core arguments and 26.66% adjuncts.

The authors noticed a lack of such a multilevel resource for these languages. At present, it is the largest corpus annotated at different linguistic levels for Spanish and Catalan and it is freely available.[4]

2.2.1.2 AnCora Lexicon

The AnCora-Verb lexicons for English, Spanish and Catalan (AnCora-Net, Aparicio *et al.*, 2008) were obtained by deriving the syntactic schemata of each verbal

[4] Website: *http://clic.ub.edu/corpus/*

predicate (from each verbal sense) in the AnCora corpora, and include the information of the Unified Verb Index (UVI)[5], which merges different resources:

- Verbal senses, with their corresponding arguments, and thematic roles from PropBank. For the characterization of the argument structure, they followed the PropBank annotation system (Kingsbury & Palmer, 2002; Palmer *et al.*, 2005).
- Semantic classes, thematic roles, selectional restrictions on arguments, and frames from VerbNet.
- Conceptual frames from FrameNet.
- Verbal senses from WordNet 3.0 and OntoNotes.

The verbal lexicon in Spanish (AnCora-Verb-Es) contains a total of 1,965 different verbs and 3,671 senses.

2.2.1.3 Semantic annotation process

The semantic annotation of verbal predicates was done semi-automatically through a system with a rule-based projection of syntactic functions into argument positions and thematic roles, which permits to automatically annotate 60% of the corpus with a fairly low error (below 2%) as it was performed for the CESS-ECE corpus (Martí *et al.*, 2007). After that, the thematic role annotation is manually completed). The whole process is detailed by the authors: "A set of manually written rules automatically mapped part of the information declared in these verbal lexicons onto the syntactic structure, which made it possible to tag the treebanks with thematic roles and semantic classes. The output from the automatic stage was either full –both arguments and thematic roles– or partial –with either arguments or thematic roles. This level of annotation was finally revised and completed by hand." (Taulé *et al.*, 2008, p.2). They make use of 8 different types of semantic arguments tags and 20 different thematic role labels. Discourse elements and modality tags did not receive any semantic label.

[5] Website: *https://verbs.colorado.edu/verb-index/vn3.3/*

2.2.1.4 Semantic classes

Each verbal predicate in the lexicon is related to one or more semantic classes (the Lexical Semantic Structure or LSS, which distinguishes four events: states, activities, accomplishments, and achievements), depending also on the diatheses alternations in which a verb can occur. The characterization of the verbal predicates are based on the proposal of lexical decomposition by Rappaport-Hovav & Levin (1998), with the following inventory of lexical templates:

- States (descriptions): [X <STATE>],
- Activities (actions): [X ACT <MANNER>],
- Accomplishments (causatives):
  - [[X ACT <MANNER>] CAUSE [BECOME [Y <STATE>]]]
    or
  - [X CAUSE [BECOME [Y <STATE>]]],
- Achievements (transformations): [BECOME [X <STATE>]].

For the characterization of the argument structure, they followed the PropBank annotation system (Kingsbury *et al.*, 2002; Palmer *et al.*, 2005). The authors acknowledge that the identification of multiword expressions (MWEs) is particularly problematic. Finally, the verbal predicates were characterized in 13 semantic classes. (Table 2.1).

LSS	**Event classes**	**Percentage**
A	*Accomplishments*	
A1	Transitive-causative	4.38%
A2	Transitive-agentive	34.65%
A3.1	Ditransitive-agentive locative	1.90%
A3.2	Ditransitive-agentive beneficiary	6.82%
B	*Achievements*	
B1	Unaccusative-motion	6.16%
B2	Unaccusative-state	12.71%
C	*States*	
C1	Existence-state	6.65%
C2	Attributive-state	20.78%
C3*	Scalar-state	0.04%
C4	Beneficiary-state	1.76%

D	*Activities*	
D1	Agentive-inergative	3.87%
D2*	Experiencer-inergative	0.14%
D3*	Source-inergative	0.13%

Table 2.1 The 13 semantic classes that were used for the characterization of verbal predicates and their percentage of token distribution in the corpus (Taulé *et al.*, 2008).[6]

2.3 Cross-lingual annotation projection

Recently, there have been various efforts to develop language-independent NLP tools, including SRL systems, due to the need for multilingual processing in different contexts, however, there is still a large number of languages for which corpora with semantic annotations do not exist, due to the fact that manual annotation is costly and time-consuming.

As a promising alternative, previous approaches, such as Pado (2007), Pado & Lapata, (2009), and Van der Plas *et al.* (2011), have investigated the direct annotation projection of semantic labels from English to other target languages via parallel corpora, in order to create language-specific PBs and thus enabling the training of SRL systems for other languages.

These transfer methods, such as the one used by Van der Plas, Merlo & Henderson (2011), rely on the Direct Semantic Transfer, that is to say, semantic role dependencies and predicate senses are transferred (projected) to any pair of sentences E and F is there exists a word alignment between them (semantic equivalence), which was adapted from the Direct Correspondence Assumption by Hwa *et al.* (2005) for syntactic dependencies.

This is usually seen as a high-precision method, due to the fact that it is able to find more cross-lingual parallelisms at the more abstract linguistic levels of representation (Padó, 2007). For example, Wu & Fung (2009) report a role transfer correspondence

[6] They did not considered the C3, D2 and D3 classes, because they had less than 6 different lemmata per class.

of 84% between English and Chinese in PropBank-style annotations. Akbik *et al.*, (2016) conducted a preliminary comparison of an auto-generated Proposition Bank for Chinese and the manually created Chinese Proposition Bank (Xue & Palmer, 2005) with a significant overlap between the two resources.

However, non-literal translations and translation shifts usually imply annotation problems at the token-level in the target language. To address this problem, Van der Plas *et al.* (2014) built two separate models: 1) for the predicate annotations and 2) the transfer of semantic roles (in the PropBank gold annotated corpus), a method expected to be high in recall. The strength is that it corrects token-level mistakes and can also be combined with the direct transfer method. This predicate labeling method consists of a learning step (computing estimates for annotation transfer on the basis of the word alignments between English and French predicates), and a labeling step (French verbs are labeled with the English predicates, without the need for parallel data or alignments). The method is language-independent but requires part of speech (PoS) information on both sides.

To address this problem, the approach by Van der Plas et al. (2014) built two separate models: 1) for the predicate annotations and 2) the transfer of semantic roles (in the PropBank gold annotated corpus), a method expected to be high in recall. The strength is that it corrects token-level mistakes and can also be combined with the direct transfer method. This predicate labeling method consists of a learning step (computing estimates for annotation transfer on the basis of the word alignments between English and French predicates), and a labeling step (French verbs are labeled with the English predicates, without the need for parallel data or alignments). The method is language-independent but requires part of speech (PoS) information on both sides.

Akbik & Li (2016b) proposed an instance-based learning method (based on Aha *et al.*, 1991; Daelemans & Van den Bosch, 2005), to consider the similarity of test instances to training instances, with a variant of the k-nearest neighbors (kNN) classification algorithm (Cover & Hart, 1967) to overcome the challenge of low-frequency exceptions in the training data (data sparseness), which require a different treatment. They propose to identify the nearest neighbors using instances that share the most similar combination of atomic features (composite feature distance

function), something that can be achieved with a very small number of similar instances. Their system outperforms previous approaches in the in-domain and out-of-domain datasets from the CoNLL-2009 shared task (Hajic *et al.*, 2009).

As stated by Akbik, Kumar & Li (2016), the annotation projection method, based on parallel corpora, gives the possibility to create Proposition Banks for languages lacking such resources ('low-resource languages'). For example, Akbik *et al.* (2015) proposed a method for creating a SRL model in two stages for languages lacking the appropriate resources, by using monolingual (English) SRL models and multilingual parallel data. They generated Proposition Banks for 7 languages in three different language families. In the first stage, they use a filtered projection method for labels that give as a result high precision. In the second step, a bootstrap learning approach is used to retrain the SRL and improve the recall.

The underlying theory is that word-aligned sentences in parallel corpus share a degree of syntactic and semantic similarity, making the direct projection possible (Padó & Lapata, 2009). First, a syntactic parser and a semantic role labeler produce labels for the English sentences (usually the source language), which are then projected onto the aligned target language words.

Although they analysed some of the errors during the annotation projection (caused by non-literal translations) and defined some lexical and syntactic constraints, this method is only a starting point, and data correction ('curation') is still needed. In this article, they outlined the correction process (a crowd-sourced method, in order to minimize the involvement of experts for cost-effective resource generation), they found encouraging results in their initial study, and made the semi-automatically generated PBs (for Chinese, Finnish, French, German, Italian, Portuguese, and Spanish) publicly available[7]. This Projected Spanish PB is the one I propose to use for correction to test my automatic correction process.

Finally, Akbik & Li (2016a) presented a multilingual SRL system (called Polyglot) capable of parsing sentences in 9 different languages from 4 different language groups (English, Arabic, Chinese, French, German, Hindi, Japanese, Russian and Spanish). As a first step, they automatically generate labelled training data for each

[7] Website: *http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=7454*.

target language by means of annotation projection (based on Akbik *et al.*, 2015): "This approach takes as input a word-aligned parallel corpus of English sentences and their translations in a target language. A semantic role labeller then predicts labels for the English sentences. In a projection step, these labels are transferred along word alignments onto the target language sentences." (Akbik & Li, 2016a, p. 3). Finally, they use the labelled data to train for each language an SRL system.

2.3.1 Manual Correction method for Projected PBs

Akbik *et al.* (2016) presented a manual method to address lexicon-level inconsistencies in three Proposition Banks (PBs) that were automatically generated using annotation projection. These PBs (for French, German and Chinese) can be used to train statistical semantic role labeling (SRL) systems in other target languages. The manual method is performed in two steps: filtering to identify incorrect frames, and merging to reduce redundancy.

2.3.1.1 Filtering and merging

In the first step, entries (a target verb sense with the English frame description plus a set of five sample sentences) are evaluated as valid (and therefore they remain in the lexicon), or not (that is, they are removed), based on the semantic validity of the English frame.

The issue of multiple entries that evoke the same semantics is addressed with the second step. Each pair of entries is evaluated as synonymous in the usage (and therefore, the target verbs should be merged into a single entry) or not (a new frame must be created), taking into account the syntactic usages of the target verb, based on the annotation guidelines of the English PB (Palmer *et al.*, 2005) (Figure 2.2).

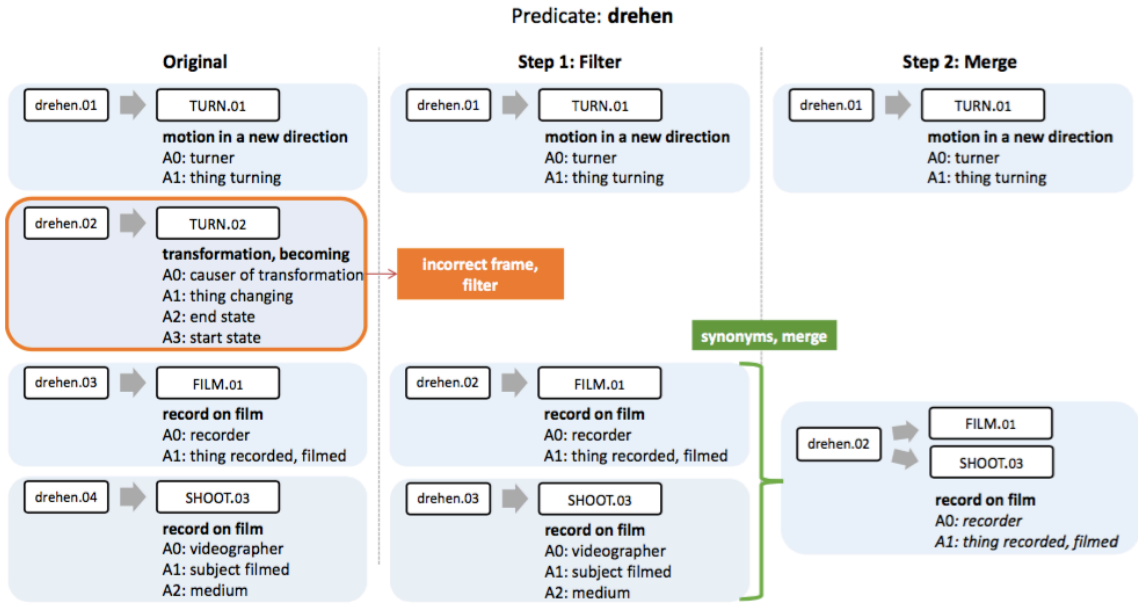


Figure 2.2 Filtering and merging steps, the second entry is removed and the last two are merged. From Akbik *et al.* (2016).

As we can see, the German verb 'drehen.02' with English source verb 'turn.02' is considered invalid (first step) and the new verb sense 'drehen.02' with English source verb 'film.01' and 'drehen.03' with English source verb 'shoot.03' are considered redundant, and therefore, are merged (and now a single verb sense has two English sources).

2.3.1.2 Evaluation

The correction process was evaluated in three different ways: 1) curator agreement scores; 2) precision, recall and F1-scores before and after the curation, and 3) a comparison against the manually created Chinese PB (Xue & Palmer, 2005).

They found that manual correction improves significantly the quality and consistency of the created PBs, although the verb coverage is also reduced, according to the linguistic distance from English (German is closer and has the largest coverage, while Chinese has the lowest). Frame redundancy is also solved through the merging procedure, as greater amounts of parallel data lead to greater redundancies (see Table). However, they recognise as a limitation that this verb-

based projection approach is unable to deal with complex predicates (target language verbs that cannot be expressed with a single verb n English).

		Proposition Bank				Evaluation		
Language	Parallel Corpus	Type	#Verbs	#Frames	#Sentences	P	R	F_1
Chinese	OpenSubtitles	projected	1,094	1,472	87,953	0.87	0.94	0.91
	(9 million sentences)	curated	942	1,003	68,829	0.93	0.96	**0.94**
French	OpenSubtitles	projected	1,323	2,249	175,636	0.82	0.94	0.87
	(15 million sentences)	curated	1,208	1,370	130,579	0.91	0.94	**0.93**
German	OpenSubtitles	projected	1,552	2,441	191,816	0.83	0.92	0.87
	(13 million sentences)	curated	1,532	1,717	150,949	0.90	0.93	**0.91**

Table 2.3 Annotation projection statistics for all three target languages. PropBanks wihtout correction are marked as ‘ Projected’ . Taken from Akbik *et al.* (2016).

2.4 Distributional Semantic Models

Distributional methods for meaning are based on the hypothesis that similar words occur in similar contexts (Harris, 1968) and thus semantic relatedness can be determined based on word occurrence in large corpora and their context patterns, and have been used for several purposes (Curran & Moens, 2002, Lin, 1998; Van der Plas & Bouma, 2005). Syntactic co-occurrences have previously been used in works on lexical acquisition (Dagan et al., 1999; Alfonseca & Manandhar, 2002). However, distributional methods for automatic acquisition of semantically related words perform less well on low-frequency words (the data sparseness problem, Van der Plas, 2008). Other authors have used this method for finding paraphrases (Ibrahim et al., 2003; Barzilay & McKeown, 2001).

Van der Plas & Tiedemann (2006) presented an alternative method using aligned multilingual data (in parallel corpora, refered to by the name of multilingual alignment-based approach), to acquire synonyms automatically as well, without the need for resources such as bilingual dictionaries, with a much higher precision and recall scores for the task of synonym extraction when compared to the monolingual approaches based on syntactic information. These methods, however, usually do not make a clear distinction between synonyms and other types of semantically related words, such as antonyms, co-hyponyms and hypernyms.

Previously, bilingual parallel corpora using distributional methods have mostly been used for tasks related to word sense disambiguation (Dagan *et al.*, 1991; Dyvik, 1998). Wu & Zhou (2003) also report and experiment for synonym extraction using bilingual as well as monolingual resources.

Other works, such as Jagfeld & Van der Plas (2015) have proposed an automatic method to improve the annotation of verbal complex predicates with already existing individual predicates by applying a multilingual distributional model.

Finally, Van der Plas (2009) presented a technique to improve the performance of distributional methods with regards of low-frequency words (phenomenon of data sparseness) by augmenting the original syntactic co-occurrences with nearest neighbours (the output of her proposed system fed into the system again in a second round). A third, high-order affinity is defined as an iterative or recursive process for calculating similarity if words share many nearest neighbours between them. The validity of the third-order affinities is dependent on the symmetric transitivity of the similarity between concepts. However, it is not always the case when dealing with separable features or ambiguous words (for example, cases of multiple inheritance, Tversky & Gati, 1978). She reports a larger percentage of synonyms found with an improvement in the performance on low and middle frequency words with respect to semantic relatedness in general, but also a gain for high-frequency words as well.

Chapter 3

Automatic Correction Method: Design & Methodology

In this chapter, I will explain the design of the automatic correction method and the methodology I followed to build all of its components. In Section 3.1, I explain the manual correction method proposed by Akbik *et al.* (2016) in two steps and the automatic method I propose for their second step. After that, I explain the motivation for creating such method, its usefulness and application (3.2).

Although I knew that the manually Curated French, German and Chinese PBs to which Akbik *et al.* (2016) had applied the manual correction method (called 'curation'), could serve for the evaluation of my automatic correction method, in Section 3.3, I explain the procedure I followed to carry out a feasability study to determine whether the semi-automatically created Spanish AnCora lexicon could be used or not for the same purpose (3.3.1) and the findings of such study (3.3.2).

My automatic correction method consisted in four steps. The first step was the collection of the different lexicons (3.4), and once I obtained those lexicons, I needed to automatically extract: 1) the already labelled instances from the three Curated PBs (3.5), and 2) the verb pair candidates in the PB to be corrected, in this case, the Projected Spanish PB (3.6).

The second step involved the creation of the main component of my automatic correction method: the distributional semantic models (DSMs). I explain its architecture, based on the collection of co-occurrence counts from bilingual aligned data (3.7.1.1), multilingual aligned data (3.7.1.2-3), and monolingual (English) syntactic information (3.7.1.4). In total, 10 DSMs with different parameters were created (3.7.1.5).

The third step is the application of the DSMs to: 1) the already labelled instances from the three Curated PBs, and 2) the verb pair candidates in the Projected Spanish PB to be corrected. I explain which programme I used (3.8), the parameter setting

(3.8.1) and how to interpret the output (3.8.2). The results were evaluated against SimVerb-3500, a gold standard resource for evaluating semantic similarity, and I explain the findings (3.9).

As a fourth and last step, I explain how I combined the output of the different DSMs to obtain better results with the use of machine learning algorithms (3.10).

3.1 Proposed automatic method

As it has been prreviously explained, the manual method proposed by Akbik *et al.* (2016) to address lexicon-level inconsistencies in Proposition Banks (PBs) that were automatically generated using annotation projection, is performed in two steps: filtering to identify incorrect frames, and merging frames to reduce redundancy.

In the first step, entries (a target verb sense with the English frame description plus a set of five sample sentences) are evaluated as valid (and therefore they remain in the lexicon), or not (that is to say, they are removed), based on the semantic validity of the English frame.

The issue of multiple entries that evoke the same semantics is addressed in the second step. Each pair of entries is evaluated as synonymous in the usage (and therefore the target verbs should be merged into a single entry) or not (a new frame must be created), taking into account the syntactic usages of the target verb, based on the annotation guidelines of the English PB (Palmer *et al.*, 2005).

In this work, I propose an automatic counterpart to the second step in the manual curation method proposed by Akbik *et al.* (2016). Although there is no way to evaluate the validity of either single instances (as shown in the sample sentences) or entire entries in an automatic manner, as it is performed in the first step of their manual correction method, there is a way to evaluate all the multiple entries by means of distributional similarity methods and decide if they evoke or not the same semantics, as it is done in the second step of the manual correction method. It is important to note that the order of the steps in the correction process does not affect

the outcome, as every entry have to be checked itself for frame validity and redundancy (or similarity) against others. [8]

The proposed automatic method is performed in four steps: 1) the collection of the different lexicons and the automatic extraction of already labelled instances from the three Curated PBs or the verb pair candidates in the Projected Spanish PB; 2) building the various DSMs; 3) the application of those DSMs to obtain different relatedness scores for each case; and 4) using a machine learning algorithm, based on those relatedness scores as features, to decide if there exists frame redundancy or not.

The purpose of this work is to evaluate the performance of my automatic method both in 1) the already corrected PBs (so called Curated PBs for French, German and Chinese) in order to make a comparison between the automatic and the manual results; and 2) in a small set of previously labelled verbs from the Projected Spanish PB, which still has to go through a correction process.

3.2 Contributions

In this work, I propose an automatic alternative to the second step in the manual frame correction method proposed by Akbik *et al.* (2016). I make use of the method multilingual distributional semantics (DS), and show it can be succesfully applied to this task, within the framework of semantic role labeling (SRL).

The main problem with the manual correction method is the number of lexical items that needs to be evaluated for each resource. Let's take the Projected Spanish PB as example. This resource has 1,584 verbs, from which 618 have multiple English verb sources (for example, the Spanish verb 'abandonar.01' has three different English

[8] It is important to note that Akbik *et al.* (2016) called this binary decision only 'merging', because in the original output all the verb senses were separate frames (and therefore the task was only to merge them). But as the reader will notice, our automatic method does it the other way around: first, it groups all the senses together and then divide them when necessary. I simply decided to call this binary decision throughout all this work 'merge' and 'divide'.

sources: ‘abandon.01’, ‘quit.01’, and ‘desert.01’), thus they all need to be evaluated for redundancy.

If any of them is found to be not redundant, that is to say, they represent a different frame from the others, another verb sense in Spanish has to be created, in this case, ‘abandonar.02’, and linked to the respective English source. On the contrary, if the frame is found to be redundant, they need to be kept together in the same verb sense (frame), in this case, ‘abandonar.01’, and the same for all the other English verb sources.

There are 1,721 different English verb sources within those 618 Spanish verbs, and every source verb needs to be compared against each other (See Sample 3.1).

Rolset id: ‘tirar.01’ **Verb sources:** 'throw.01', 'throw.07', 'toss.01', 'pull.01', 'pull.06', 'tug.03', 'fuck.01', 'shoot.03', 'lay.01']

Sample 3.1 An example of the complexity of some verbs when evaluated for redundancy.

It is evident that a manual review has to be done by a language expert, which it is costly, time-consuming, and can lead to inconsistencies if there are variations in the annotators’ criteria. I expect to alleviate this process through the automatic method. Besides, we need to take into account that, although this project focuses on correcting the Projected Spanish PB, the proposed method is language-independent, and can be used to correct all the multiple PBs in other languages generated in the same manner.

Although this method has not reached the stage where it gets to the same decisions as the language experts do and correct every case by itself, it can function in a semi-automatic manner, that is to say, the method can detect strong and weak redundant candidates, which helps to speed up and stabilize the criteria of the correction process. Finally, even when the method does not get to the same results, it reveals the thought process and the other linguistic parameters the human ‘curators’ used when performing such task.

3.3 Feasability Study with Spanish AnCora

Before I had access to the three Curated PBs, I did a small feasability study as a exploratory research comparing the Spanish AnCora lexicon and the Projected Spanish PB, in order to determine if the Spanish AnCora could serve to evaluate the results when applying the correction method.

3.3.1 Analysis procedure

Only four English verbs with all their possible translations to Spanish from the dictionary of contextual translations Reverso Context[9] were selected for the analysis. The verbs 'break', 'cut', 'hit' and 'touch' have been studied by several authors (i.e. Fillmore, 1967; Levin, 1993), because they confirm that various aspects of their syntactic behaviour are tied to the meaning. While all of them are transitive and take two arguments as subject and object, they do not have the same diathesis alternations, that is, some of these verbs can be used with different valencies (Tesnière, 1959) or subcategorization frames (Chomsky, 1965). Therefore, although these verbs share certain elements of meaning, they also fall into different verb classes for other characteristics. Levin (1993) shows that other verbs show the same patterns:

- '**break' verbs**: crack, rip, shatter, snap ...
- **'cut' verbs**: hack, saw, scratch, slash ...
- **'touch' verbs**: pat, stroke, tickle ...
- **'hit' verbs:** bash, kick, pound, tap, whack ...

As we can see, there are some ties between verb behaviour and verb meaning, as the members of each set of verbs have common syntactic as well as semantic properties. The difference of these verbs can be summarized as follows: 'touch' is a pure verb of contact, 'hit' is a verb of contact but motion is also involved, 'break' is a pure verb of change of state, and 'cut' is a verb of a caused change of state, so motion and contact of a certain instrument with the entity that changes state is also involved.

[9] Website: *http://context.reverso.net/translation/*

3.3.2 Findings

I found out that there are several issues that needed to be taken into account and kept us from considering the Spanish AnCora suitable for our evaluation in the Projected PBs.

Although it can be seen that both resources took as a starting point the verb sense distinctions made in the English PB, the verb senses are not divided with the same directives and criteria, mainly because of how both resources were created.

A first problem would be to determine how to match the senses in both resources, as it is not a simple one-to-one relation. The simplest approach would be linking the senses whose an exact match of the predicate-argument structure is found, but this would lead to a great amount of cases whose no predicate-argument structure will be found for evaluation.

However, another issue that I found out is that in many cases AnCora, and to a smaller extent the Project Spanish PB, provide for each verb case, several possible PB roleset links, which do not have always the same predicate-argument structure, and thus, but this could lead in some of the cases to an erroneous analysis and some misleading results.

It is important to remember that Spanish AnCora is a resource created semi-automatically and not through a projection method. Therefore, although both resources, Spanish AnCora and the Projected Spanish PB have links to the English PB for each of their senses, the English verbs in the case of the Spanish AnCora are not *sources*, they were just 'linked' to the English rolesets ids. As it can be seen, although the Projected Spanish PB and the Spanish AnCora lexicon are intended to represent the same information, the content is different in many respects (Samples 3.2 & 3.3).

Roleset id	Source
abandonar.01	*abandon.01, quit.01, desert.01*
abarcar.01	*encompass.01, span.01*
abastecer.01	*supply.01*
abdicar.01	*abdicate.01*
ablandar.01	*soften.01*

Sample 3.2 Projected Spanish PB with roleset ids and English sources.

Roleset id	Source
abalanzar.01	*swoop.01*
abanderar.01	*champion.01, lead.02*
abandonar.01	*abandon.01, abandon.02*
abanicar.01	*fan.01*
abaratar.01	*cheapen.01*

Sample 3.3. Spanish AnCora with roleset ids and linked English verbs.

Therefore, I decided not to include the Spanish AnCora lexicon in the evaluation but concluded both resources could be marged, in order to create an augmented resource for future work.

In any case, I decided to start a series of preliminary tests and experiments to see how the parameter setting in the distributional models affected the merging decisions. From this point, I realised that there were considerable changes in the results when I used two different singular-value decomposition (SVD) schemes for dimensionality reduction.

3.4 Collection of lexicons

As a first step, I collected the following lexicons, and a preliminary study was carried out to see their composition in detail:

- **Projected Spanish PB** (Akbik *et al.*, 2015), with Roleset id (or verb sense) in Spanish + English verb sense used as source (Sample 3.4). This lexicon has not yet been corrected and it is the purpose of this work to perform such task

in an automatic manner.

Roleset id	Source
abandonar.01	*abandon.01, quit.01, desert.01*
abarcar.01	*encompass.01, span.01*
abastecer.01	*supply.01*
abdicar.01	*abdicate.01*
ablandar.01	*soften.01*

Sample 3.4 Projected Spanish PB lexicon with roleset ids and English sources.

- **Curated French, German & Chinese PBs**[10] (Akbik *et al.*, 2016), with roleset id (verb sense) in target language + English verb sense used as source. It is important to remember that these lexicons were created with a projection method and have been already manually corrected (See Sample 3.5). We can note, for example, that three English verb sources were already merged in the verb 'abandoner.01'.

Roleset id	Source
abandonner.01	*abandon.01, forsake.01, abort.01*
abîmer.01	*ruin.01*
abolir.01	*abolish.01*
abonner.01	*subscribe.01*
aborder.01	*board.01*

Sample 3.5 Curated French PB lwith roleset ids and English sources.

- **English PB** (Kingsbury & Palmer, 2002; Palmer *et al.*, 2005), with the roleset id + roles (See Sample 3.6).

[10] Available at: *https://github.com/System-T/UniversalPropositions*

Roleset id	Roles
abandon.01	Arg0-PPT, Arg1-DIR, Arg2-PRD
abdicate.01	Arg0-PAG, Arg1-DIR
abduct.01	Arg0-PAG, Arg1-PPT
abet.01	Arg0-PAG, Arg1-GOL
abhor.01	Arg0-PPT, Arg1-PAG

Sample 3.6 English PB lwith roleset ids and roles.

This way, I obtained the lexicon in Spanish to be corrected, which now could be compared against the three manually corrected ('Curated') PBs plus the English PB, which helps as a reference and also allows me to see the information about the roles to all the previous resources, as shown in the sample. A quick comparison of the five resources allows us to see the number of verbs and verb senses in each lexicon (Table 3.7).

Resources	Verbs	Verb senses
Projected Spanish PB:	1,585	2,065
Curated French PB	1,322	1,460
Curated German PB	2,320	2,532
Curated Chinese PB	1,008	1,044
English PB:	5,649	8,632

Table 3.7 Number of verbs and verb senses in the five lexicons.

3.5 Automatic extraction of labelled instances in Curated PBs

After colleting all the resources, I extracted from the Curated PBs (French, German and Chinese) all the possible labeled instances of verb senses that were either manually merged or kept as separate senses. It is important to recall that the merging task is binary in nature: either you merged them or let them as separate senses (that is, divide them). These corrected senses served for the evaluation of my semi-automatic method, because by applying it, I could compare the outcome against what

had already been done manually. The extraction of labeled candidates was done with the following criteria:

1) If the verb sense showed several English verb senses as source, that means that each English verb was compared against each other, a decision of get them together was taken, and therefore can be used as merging instances. As shown in Sample 3.8, the verb sense 'abandonner.01' in French has three English verb senses as source ('abandon.01', 'forsake.01', and 'abort.01'). This means each pair of English verb senses was determined to be redundant with respect to the French verb, and the decision to merged was taken.

Roleset id	**Source**	**Source pairs**	**Determined outcome**
abandonner.01	[abandon.01, forsake.01, abort.01]		
		abandon.01, forsake.01	Merge
		abandon.01, abort.01	Merge
		forsake.01, abort.01	Merge

Sample 3.8 Merging decision in the verb sense ‘ abandonner.01 (Curated French PB).

2) If one verb had different verb senses, each one with a separate English source, that means that all the English sources were compared against each other, a decision of separating them was taken and therefore, they can be used as dividing instances. As shown in Sample 3.9, the French verb 'aboutir' has two senses, 'aboutir.01' and 'aboutir.02', each one with a separate English source ('lead.03' and 'result.01', respectively). This means that pair of English verb sources was determined to be different with respect to the French verb, and therefore, the decision to keep them separate (or 'divide' them) was made.

Roleset id	Source	Source pairs	Determined outcome
aboutir.01	[lead.03]		
aboutir.02	[result.01]		
		lead.03, result.01	Divide

Sample 3.9. Dividing decision step by step in the French verb ‘ aboutir’ , with two senses (Curated French PB)

3) There are cases with mixed outcomes, where merging and dividing decisions had to be made, as shown in Sample 3.10, and all serve as labelled instances. As we can see, as the number of roleset ids and sources increase, so it does the number of merging / dividing decisions. The number of individual decisions to be made manually is another argument in favor of the proposed semiautomatic method.

Roleset id	Source	Source pairs	Determined outcome
accorder.01	[grant.01, allow.01, bestow.01]		
accorder.02	[agree.01]		
		grant.01, allow.01	merge
		grant.01, bestow.01	merge
		grant.01, agree.01	divide
		allow.01, bestow.01	merge
		allow.01, agree.01	divide
		bestow.01, agree.01	divide

Sample 3.10 Mixed decisions that had to be taken manually in the French verb 'accorder’ , with two senses (Curated French PB)

4) The verbs with only one sense and one English source (i.e., the French verb 'abî mer.01', with the English source 'ruin.01') could not be used as proper instances, because no merging / dividing decision was made whatsoever.

In Table 3.11, we can see the number of merging and dividing pair instances obtained in each of the Curated PBs.

Resource	Merge pairs	Divide pairs	Total
Curated French PB	458	216	674
Curated German PB	770	295	1,065
Curated Chinese PB	238	33	271
Total	1,466	544	2,010

Table 3.11 Number of merging and dividing pair instances obtained in each PB.

Once I have these pair instances, I can measure the semantic similarity between the two elements with the help of a DSM and therefore determine wheter they should be merged or they should be divided.

3.6 Automatic extraction of pair candidates in Projected Spanish PB

A similar process was performed for the Projected Spanish PB, (in fact, all the merging and dividing candidates can be obtained in the same way for all languages), but instead of extracting separately the merging and dividing sets, what I have is all the possible unlabelled candidates together. The verbs with only one sense and one English source (i.e., the Spanish verb ' abastecer.01', with the English source 'supply.01') could not be used as proper instances, because there was no merging / dividing decision to be made. As a result, 618 Spanish verbs were extracted. They have 1,721 different English verb sources and formed a total of 1,885 pair candidates which need to be evaluated (See Sample 3.12).

Roleset id	Source	Source pairs
abandonar.01	[abandon.01, quit.01, desert.01]	
		abandon.01 quit.01
		abandon.01 desert.01
		quit.01 desert.01
abarcar.01	[encompass.01, span.01]	
		encompass.01 span.01
abordar.01	[board.01, tackle.01]	
		…

Sample 3.12 Extraction of the Spanish candidates (without labels).

3.7 Distributional semantic model architecture

In this section I explain the 10 different distributional semantic models (DSMs) that were built, and the main differences between them.

3.7.1 Collection of co-occurrence counts from corpora

3.7.1.1 Bilingual aligned data

With the verbal lexicon of the English PropBank, I could collect the data from the OPUS Search Word Alignment Database[11] (Tiedemann, 2012), and obtain all the pair-alignments for all the available English verbs in Spanish from three different corpora: EUConst, Europarl3 (Tiedemann, 2009) and OpenSubtitles2016 (Lison & Tiedemann, 2016).

The data collected contains three columns: the first column shows the headword, or the word in the source language (English in this case, as SL), the second column shows the translations found in the target language (TL, also called features or attributes) and the third one shows the frequency value.

[11] Website: *http://opus.lingfil.uu.se/lex.php*.

Co-occurrence vectors are built from the alignments found in the parallel corpus, by adding all the translations of a headword. Each aligned word type is a feature in the vector of the target word under consideration. The frequency of the headword is the sum of all the translation frequencies. This structure has a total of 14,521 alignments, with 5,264,050 co-occurrences and it served as my first, bilingual distributional semantic model (DSM). Similar bilingual models were built for French and German during the evaluation of the respective Curated French and German PBs.

3.7.1.2 Multilingual aligned data

For my second, multilingual DSM, I collected from the OPUS Search Word Alignment Database all the pair-alignments for 24 more languages[12] (Table 3.13).

Language pairs	**English verbs**	**Aligned pairs**	**Co-occurrences**
English-Bulgarian	608	2,669	60,005
English-Chinese	48	66	489
English-Croatian	101	205	1,163
English-Czech	965	4,615	105,154
English-Danish	1,664	18,325	5,299,074
English-Dutch	1,763	19,391	4,026,258
English-Estonian	378	592	8,428
English-French	2,008	15,923	5,072,451
English-Finnish	1,491	20,035	3,293,295
English-German	1,462	12,481	4,079,495
English-Greek	1,095	15,045	1,961,705
English-Hungarian	468	1,087	15,509
English-Irish	110	121	1,296
English-Italian	1,687	13,563	4,220,260
English-Latvian	213	321	2,105
English-Lithuanian	270	406	2,865
English-Norwegian	440	1,449	32,184
English-Polish	498	1,046	17,826
English-Portuguese	1,937	13,647	4,717,004
English-Romanian	841	2,804	108,210
English-Russian	137	376	2,951
English-Slovenian	944	3,527	125,688
English-Spanish	*2,070*	*14,521*	*5,264,050*

[12] Bulgarian, Chinese, Croatian, Czech, Danish, Dutch, Estonian, French, Finnish, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Russian, Slovenian, Swedish, Turkish + Spanish.

English-Swedish	1,900	13,577	5,119,923
English-Turkish	566	2,753	25,821
Total	**2,967**	**178,545**	**43,563,209**

Table 3.13. Number of English verbs collected, aligned pairs and co-occurences.

As we can see, with this information, I increased the number of English verbs to 2,967, an improvement when compared against the English-Spanish DSM and the English PB (Table 3.14).

Resources	**English Verbs**
English-Spanish Semantic Model	2,070
Multilingual Semantic Model	***2,967***
English PB	5,649

Table 3.14 Number of English verbs in the three resources.

I was from this data that I could extract the information to create two other bilingual DSMs for French and German, when evaluation the performance of the whole method in the manually Curated PBs. I must note that, due to the amount of Chinese-English pair verbs obtained from the aligned data (48), there was no point in creating a Chinese model of its own.

3.7.1.2.1 Data Pre-processing

In order to improve the quality of the alignments, all the tokens in the target languages were tagged with part of speech. For that purpose, I used the multilingual toolkit RDRPOSTagger (Nguyen *et al.*, 2014; Nguyen *et al.*, 2016), whose models were trained using the data of the Universal Dependencies (UD) v2.0[13] (Nivre *et al.*, 2016, which are also used for the gold-standard test sets in the CoNLL 2017 shared task) and report a high accuracy for all the selected languages (See Table 3.15). With this procedure, I ensure there are only verbs at the two sides of the pair-alignments.

[13] Website: *http://universaldependencies.org*

Pre-trained models	POS tagging accuracies
UD_Bulgarian	96.25%
UD_Chinese	89.12%
UD_Croatian	95.04%
UD_Czech-CAC	98.05%
UD_Danish	93.48%
UD_Dutch-LassySmall	95.77%
UD_Estonian	86.66%
UD_French-Sequoia	95.63%
UD_Finnish	92.11%
UD_German	90.24%
UD_Greek	94.24%
UD_Hungarian	87.47%
UD_Irish	82.36%
UD_Italian	96.22%
UD_Latvian	86.51%
UD_Lithuanian	*Not reported*
UD_Norwegian-Bokmaal	94.47%
UD_Polish	94.18%
UD_Portuguese-BR	95.49%
UD_Romanian	95.79%
UD_Russian-SynTagRus	96.88%
UD_Slovenian	94.62%
UD_Spanish-AnCora	*96.62%*
UD_Swedish	94.18%
UD_Turkish	92.10%

Table 3.15 POS tagging accuracies using the RDRPOSTagger, as reported by Nguyen *et al*., 2016.

Afterwards, I built a lemmatizer that uses the multilingual datasets (lists of lemma-token pairs in a machine-readable format) collected from various sources by Michal Boleslav Mě chura[14] and made available under an open-source license. With this procedure, I ensure that all the inflected forms of the verbs in the target languages are not dispersed as different co-occurrences, but they are grouped in the right lemmas.

3.7.1.2.1 Enrichment of the Multilingual DSM: a Back-Translation Assumption

A third DSM was built with the same multilingual information but including it also in the reverse order (not only source-target but also target-source alignments). Although no back-translation was carried out, in order to improve the data in the multilingual

[14] Website: *http://www.lexiconista.com/datasets/lemmatization*

DSM, I can work with the assumption that it is *possible* that the target verbs would have been translated in the same contexts as the source verbs. By using this assumption the model has 357,090 pair alignments and 87,126,418 co-occurrences. Bilingual versions with the back-translation assumption (BTA) were also created for the English - Spanish / French / German and Chinese DSMs.

3.7.1.5. Monolingual English syntactic information

In order to increase the number of alignments, a fourth monolingual DSM was built by extracting all the English verbs with nouns co-occurrences, as well as subject-verb phrases from three different corpora (Table 3.16): the Wikipedia Corpus[15] (Davies & Ferrerira, 2015), the British National Corpus[16] (2007) and the UK Web Archiving Consortium (UKWaC) Corpus[17] (Ferraresi *et al.*, 2008).[18]

Monolingual DSM	**Aligned pairs**	**Co-occurrences**
Verb-noun	7,195,188	23,484,012
Subject-verb	1,142,072	2,296,791
Total	**8,337,260**	**25,780,803**

Table 3.16 Number of alignments and co-occurences in the Monolingual English DSM.

3.8 Use of the DSMs

As it was previously explained, the use of monolingual distributional similarity is a common approach to the automatic extraction of semantically related words. The assumption in a multilingual setting, however, as proposed by Van der Plas & Tiedemann (2006), is that words that share translational contexts are semantically related.

[15] Website: *https://corpus.byu.edu/wiki*
[16] Website: *http://www.natcorp.ox.ac.uk*
[17] Website: *https://www.sketchengine.co.uk/ukwac-corpus*
[18] Available at: *http://clic.cimec.unitn.it/composes/toolkit/index.html*

For this purpose, I used the DISSECT toolkit[19] (Dinu *et al.*, 2013) to construct a programme to compose the distributional semantic representations of the verb pair-alignments extracted from various corpora, as explained in the previous section. Dissect packages are written in Python or as command-line tools. This toolkit allow us to set all the parameters of the DSMs, such as feature selection, weighting scheme and dimensionality reduction.

With this toolkit, context vectors are built from the alignments found in the parallel corpus, by adding all the translations or contexts of a given headword. Each aligned word type is a feature in the vector of the target word under consideration. The frequency of the headword is the sum of all the translation or contexts frequencies. The more similar the co-occurrence vectors of any two headwords are, the more similar they are expected to be. Finally, context vectors are compared with each other in order to calculate the distributional similarity between headwords.

The DISSECT toolkit was also used to evaluate the predicted similarity scores against SimVerb-3500 (Gerz *et al.*, 2016), a gold standard resource for semantic similarity.

3.8.1 Parameter setting

After a series of experiments with the development set of the Projected French PB, aimed to test the best parameters for the distributional models to deal with this task, I decided to use following parameters:

- I selected a high number of top features (that is to say, the 10,000 most relevant contexts) in the aligned data.

- Several weighting schemes have been proposed to calculate the distributional similarity of context vectors. The scheme with the best results for this task was obtained with the Pointwise Mutual Information (PMI or I, by Church & Hanks, 1989):

[19] Website: *http://clic.cimec.unitn.it/composes/toolkit/*

$$I(W, f) = log \frac{P(W, f)}{P(W)P(f)}$$

where the frequency of a target word (*W*) in a vector can be replaced by a weighted score in order to account for the differences in the frequency values between the several headwords and attributes (*f*).

- I used two truncated singular-value decomposition (SVD) schemes for a reduced dimension *k* (Dinu *et al.*, 2013), where, given an input matrix *X*, they compute the decomposition $X = U\Sigma V^T$, and return $U\Sigma$ truncated to dimension $min(k, rank(X))$.

3.8.2 Output scores

Finally, the DSM provides a similarity score (from 0 to 1) for each existent English verb source pair candidates within a Spanish verb sense (Sample 3.17).

Roleset ID	**English source verbs**	**Verb pair candidates**	**Similarity scores**
abandonar.01	[*abandon.01*, *desert.01*, *quit.01*]		
		abandon.01, desert.01	0.79559185038383184
		abandon.01 quit.01	0.86974114859409768
		desert.01, quit.01	0.65052407342582996
abarcar.01	[*encompass.01*, *span.01*]		
		encompass.01, span.01	0.05174248025671929
abastecer.01	[*supply.01*]		
	...		

Sample 3.17. Similarity scores for the English verb source pair candidates (Multilingual DSM).

It is important to note that, after a series of experiments measuring the accuracy with a validation set in each DSM, I manually determined that the best decision threshold for them was the following: dividing > 0.75 =< merging. That is to say, if the score was lower than 0.75, the verb pair instance was considered not redundant, and therefore a new verb sense needs to be created. On the other hand, if the score was equal to or more than 0.75, the verb pair instance is considered as redundant and they need to be grouped together, that is to say, they remain in the same verb sense.

3.9 Evaluation against SimVerb-3500

Afterwards, I evaluated the similarity scores of the different distributional models by comparing them against the SimVerb-3500 (Gerz *et al.*, 2016), a gold standard resource for evaluating the semantic similarity of verbs in English. SimVerb contains 3,500 verb pairs with ratings on a scale 0-10. The best correlation scores I obtained were for the multilingual DSM with a reduced dimensionality reduction:

- **Spearman:** 0.351961229696
- **Pearson:** 0.398037891668

These Spearman's rank correlation results are similar to the ones obtained by Vulić *et al.* (2017), when they evaluated a representative set of English vector space models with morphological constraints.

Although these results do not seem impressive, it is important to notice that my models are not aimed at giving a continuous measure of similarity (from 1 till 10), but I have designed them to make a sharp binary distinction between synonyms and non-synonyms in a multilingual context. Therefore, it was expected they would not perform well at distinguishing different levels of similarity.

Nonetheless, I confirmed it was better to use two different singular-value decomposition (SVD) schemes for dimensionality reduction, as we observed that a high (increased) dimensionality reduction (up to 30 elements) was better for the merging task, whereas a low (decreased) dimensionality reduction (up to 500 elements) was better for the dividing task.

3.10 Combined models using machine learning algorithms

After the first experiments, I found out that there was a lot of variation in the results when using any of the single DSMs, but the most concerning situation was that it was hard to select a single best model for the two tasks (dividing and merging). The best model for merging always had inferior results for dividing and vice versa. This is the reason why I started several machine learning experiments in order to see if there was a way to combine all the different models, with the idea of achieving a better optimum for the two tasks simultaneously.

The best solution was to use the similarity scores obtained in all ten DSMs as features and tested them with different algorithms for classification, mainly two logistic regression algorithms: the ***additive logistic regression***, based on the principles of additive modelling and maximum likelihood within the boosting hypothesis, by Friedman, Hastie& Tibshirani (1998), and the ***logistic model trees***, classification trees with logistic regression functions at the leaves, by Landwehr, Hall, & Frank (2005), due to the nature of the classification problem, where I estimate the probability of a binary response based on one or more predictor variables. I tested them with the set of labelled instances that were extracted from the Curated French, German and Chinese PBs. In this way, the combined model allows me to make use of the individual similarity scores that result from the different models.

Chapter 4

Results & Evaluation

In this Chapter, I will discuss the different results I obtained when evaluated the different distributional semantic models (DSMs) that were built for correction. First, the evaluation on the manually Curated PBs (4.1) using the individual semantic models (4.1.1); then, the combined semantic model, with a machine learning algorithm, and a 10-fold cross validation setting (4.1.2), or using a 20% as training set (4.1.3), and the evaluation with other languages as training models (4.1.4). Finally, I present the evaluation on the Projected Spanish PB with the other languages as training models and the role information as a constraint (4.2).

4.1 Evaluation on Curated PBs

4.1.1 Individual semantic models

Before applying the semi-automatic correction method to the Projected Spanish PB, I applied it to all the three available manually Curated PBs (French, German and Chinese). More in particular, I applied it to the labelled (merging / dividing) instances I extracted from each one of the PBs, as this data could indeed serve as a gold standard, because this is the correction process I am trying to emulate. It is important to remember that I devised 10 different DSMs, each with different features, as I was trying to compare them and improve their performance. I will shortly summarize the models for the reader's convenience.

As explained before, three main models were built with different aligned data: 1) models with bilingual information. These models could only be built for French and German, when evaluating those specific languages, due to the low amount of English - Chinese pair verbs obtained from the aligned data; 2) models with multilingual information, for 25 languages in total; and 3) models with monolingual (English) syntactic information.

Besides, two parameters showed to be relevant for the correction task, and therefore, it was necessary to test them separately, by creating 1) models with and without additional information, taking into account the back-translation assumption (BTA); and 2) models increasing and decreasing dimensionality reduction (I / D-DR). This is the list of all the models I used:

- Bilingual DSM with R-DR
- Bilingual DSM with I-DR
- Bilingual DSM with BTA and R-DR
- Bilingual DSM with BTA and I-DR
- Multilingual DSM with R-DR
- Multilingual DSM with I-DR
- Multilingual DSM with BTA and R-DR
- Multilingual DSM with BTA and I-DR
- Monolingual DSM with R-DR
- Monolingual DSM with I-DR

For these first experiments, I took 20% as training data, a 20% as development set within the training data, and the rest for evaluation. It is important to remember that after a series of experiments measuring the accuracy with the validation set, I determined that, once I obtained the similarity scores for the different DSMs, the best decision threshold for them was the following: dividing > 0.75 =< merging. That is to say, if the score was lower than 0.75, the verb pair instance was considered not redundant, and therefore a new verb sense needs to be created. On the other hand, if the score was equal to or more than 0.75 the verb pair instance is considered as redundant and they need to be grouped together.

In order to save time and space, I report only the two best models in each of the three Curated PBs (French, German & Chinese): 1) Multilingual DSM with BTA and I-DR, which obtained the best results for merging, and 2) Bilingual DSM with R-DR, which obtained the best results for filtering (Table 4.1). [20]

[20] The best model for dividing in Chinese was the Multilingual DSM with BTA and R-DR.

Resource	Model	Task	Accuracy	Error Rate	Precision	Recall	F-score
Curated French PB	Multi BTA I-DR	merge	56.3	43.6	55.6	74.8	63.8
		divide			57.9	36.7	44.9
	Bilingual R-DR	merge	49.4	50.5	70.8	02.8	05.5
		divide			48.9	98.7	65.4
Curated German PB	Multi BTA I-DR	merge	66.7	33.2	80.4	75.2	77.7
		divide			30.8	37.7	33.9
	Bilingual R-DR	merge	35.2	64.7	11.1	93.4	19.9
		divide			97.9	29.7	45.6
Curated Chinese PB	Multi BTA I-DR	merge	81.0	18.9	90.2	88.3	89.3
		divide			15.1	17.8	16.3
	Multi BTA R-DR	merge	52.0	47.9	45.7	99.0	62.6
		divide			96.9	20.0	33.1

Table 4.1. Results of the best individual models in the three Curated PBs.

As we can see, there is a lot of variation in the results, but the most concerning situation was that it was hard to select a single best model for the two tasks (merging and dividing). The best model for merging has inferior results for dividing and vice versa. For example, the best model for merging in German is the Multilingual DSM with BTA and I-DR, with a F-score of 77.7, but only a F-score of 33.9 for dividing; whereas the best models for dividing for the same language is the Bilingual DSM with R-DR, with a F-score of 45.6, but only a F-score of 19.9 for merging. This is the main reason why I attempted to create a combined DSM.

4.1.2 Combined DSM

I started several machine learning experiments in order to see if the combination of the similarity scores from different models could achieve a better optimum for the two tasks simultaneously. I finally decided to use the similarity scores obtained in each semantic model as features and tested them with different algorithms for classification with the set of labelled instances that were extracted in the three languages. This combined model allows me to make use of the separate similarity scores that resulted from the different models. The best results were obtained using

all the 10 DSMs together with an additive logistic regression algorithm (Friedman, Hastie & Tibshirani, 1998) in a 10-fold cross validation setting. (Table 4.2).

Resource	Task	Accuracy	Error Rate	Precision	Recall	F-score	Weighted Average F-score
Curated French PB	merge	76.7	23.2	82.5	83.4	83.0	76.6
	divide			64.0	62.5	63.2	
Curated German PB	merge	75.2	24.7	80.7	86.4	83.4	74.4
	divide			56.4	46.1	50.7	
Curated Chinese PB	merge	84.8	15.1	89.2	94.1	91.6	83.2
	divide			30.0	18.2	22.6	

Table 4.2. Results of the combined models in the three Curated PBs.

Before using a learning algorithm, I was only reporting the F-score for merging and dividing separately. Now, I can report a more meaningful weighted average F-score for both tasks combined. The weighted average F-score is meant to check the performance of the classifier for both tasks (the sum of all F-measures, weighted according to the number of instances in each particular class label, that is to say, merging or dividing).

As we can see, the results are much more consistent, although the numbers for the dividing task are always lower. Throughout all the results shown, in this section and in the following ones, the Curated Chinese PB is the one that shows the most drastic variations, when compared to the other two resources. I believe this is due mainly to three factors: 1) The low amount of labelled instances I obtained from the Curated Chinese PB, as it has been previously explained, 2) the skewedness of the data, and 3) the fact that I am dealing with a more distant language from English.

4.1.3 10-fold cross validation vs. 20% training data set

The results in the previous section provide an upper bound on the possibility of using semi-automatic methods, as I made use of 80% of the manually corrected (labelled) data for training. In a realistic setting, however, I would expect to be able to gather only a small portion of those manual corrections for a novel language, for example, in Spanish. Therefore, it was important to evaluate again the three language models

with only a 20% of data as training set and use the remaining 80% for evaluation (Table 4.3).

Resource	Task	Accuracy	Error Rate	Precision	Recall	F-score	Weighted Average F-score
Curated French PB	merge	71.2	28.7	73.6	88.9	80.5	68.8
	divide			61.5	35.8	45.2	
Curated German PB	merge	71.4	28.5	78.8	83.0	80.9	70.8
	divide			47.5	40.8	43.9	
Curated Chinese PB	merge	82.8	17.1	89.6	91.1	90.3	82.4
	divide			26.1	23.1	24.5	

Table 4.3. Results of the combined models with only 20% as training set.

As we can see, there is a drop of 7.7 points for French, 3.6 points for German and 0.7 points for Chinese in the weighted average F-score for the three languages. When comparing the F-scores of each task, the dividing task is the most affected, with a loss of 18 points for the Curated French PB and a loss of 6.8 points for German. We can see a gain of 1.8 points for the Curated Chinese PB, although the performance is still low in both cases. It seems that the dividing task is harder and the results are also more affected by data sparseness.

4.1.4 Evaluation with other language models as training

An alternative method when there is no large amounts of manually corrected data is to train the model for a novel language, for example, Spanish, on data which has already been manually corrected for other languages and use them cross-linguistically. It is important to remember that this semi-automatic method is expected to work in a multilingual setting.

Therefore, I took the labelled instances of each Curated PB as test set and evaluated them with a model using one or the other two remaining languages (i.e. the evaluation of the French dataset was carried out with the data on German, or Chinese or a combination of both as training set, and so on (Table 4.4).

Test Language	Train Language	Task	Accuracy	Error Rate	Precision	Recall	F-score	Weighted Average F-score
Curated German PB	**French PB**	merge	72.8	27.1	87.2	73.2	79.9	74.0
		divide			50.7	71.9	59.5	
	Chinese PB	merge	73.0	26.9	73.4	98.3	84.1	64.3
		divide			61.8	07.1	12.8	
	French + Chinese PBs	merge	75.7	24.2	83.5	82.9	83.2	75.8
		divide			56.1	57.3	56.7	
Curated French PB	**German PB**	merge	74.7	25.2	77.7	88.2	82.6	73.5
		divide			64.9	46.3	54.1	
	Chinese PB	merge	67.9	32.0	68.5	97.8	80.6	57.5
		divide			50.0	04.6	08.5	
	German + Chinese PBs	merge	74.7	25.2	76.2	91.5	83.1	72.5
		divide			73.7	74.8	72.5	
Curated Chinese PB	**French PB**	merge	79.2	20.7	92.9	82.7	87.5	81.6
		divide			30.5	54.5	39.1	
	German PB	merge	79.2	20.7	91.3	84.4	87.7	81.1
		divide			27.5	42.4	33.3	
	French + German PBs	merge	79.2	20.7	93.7	81.9	87.4	81.8
		divide			31.7	60.6	41.7	

Table 4.4. Results using other language models as training dataset.

As we can see by the weighted average F-scores, the results are consistent across all languages, except when the Curated Chinese PB is used either as the only training model (lowering the weighted average F-score up to 57.5%), or as the test set, (increasing the results up to 81.8%).

When these results are compared against the results obtained when using only a small dataset for training, as in the previous section, we can see that the use of other languages for setting the parameters is still better, although the difference is not as big as we could have expected (as there is an improvement between 0.6 – 3.7 points in the weighted average F-score among the three languages).

Finally, I also combined all the training data from the three languages and tested them. The best results in this case were obtained using a classification tree algorithm

with logistic regression functions at the leaves ('logistic model trees' by Landwehr, Hall & Frank, 2005; Sumner, Frank & Hall, 2005) in a 10-fold cross validation setting and where the test sets were also a combination of all the three languages (Table 4.5).

Resource	Task	Accuracy	Error Rate	Precision	Recall	F-score	Weighted Average F-score
French, German & Chinese PBs	merge	77.1	22.8	84.1	84.6	84.4	77.1
	divide			57.9	56.8	57.3	

Table 4.5. Results for Curated PBs using all language models as training set.

If we compare these results with all the previous experiments, we can notice the usefulness and stability of the proposed automatic method for correction across languages.

4.2 Evaluation on Projected Spanish PBs

As said previously, the Projected Spanish PB has not been curated yet and provides an ideal and realistic test bed for my method. I found 1,885 pair candidates that needed to be evaluated in 618 Spanish verbs. I manually labelled 189 pair instances (or rather 10% of all the instances), that is to say, I decided if they had to be merged or kept apart.

I decided to establish a rather simple but unbiased manual labelling criterion: if the two pair candidates appeared as synonyms in the online website Thesaurus.com,[21] they were marked as redundant (merge), otherwise, I would divide them as separate verb senses (divide). Although this simple criterion does not correspond completely to the annotator's criteria used in the manual correction process as decribed by Akbik *et al.* (2016), it serves as a first basic component, in this way I am able to provide unbiased results, and it allow us to see what other relevant linguistic factors were taken into consideration, which could be used later for further improvements towards a fully automatic correction method.

[21] Website: *http://www.thesaurus.com/*

Again, I combined the scores from the different DSMs (bilingual, multilingual and syntactic) by using them as features with different algorithms for classification, mainly logistic regression algorithms, plus the two parameters that showed to be relevant for the task, namely, the use of additional information when taking into account the BTA, and I-DR / D-DR.

I tested it with the set of instances I manually labelled from the Projected Spanish PB. The best results were obtained when using all the 10 DSMs together and the previously used logistic model tree algorithm, in order to achieve a better optimum for the two tasks (merge / divide) simultaneously (Table 4.6).

Resource	Task	Accuracy	Error Rate	Precision	Recall	F-score	Weighted Average F-score
Projected Spanish PB	merge	79.3	20.6	83.1	91.7	87.2	77.8
	divide			58.6	38.6	46.6	

Table 4.6. Results for the Projected Spanish PB using all language models as training set.

If we compare this results with those obtained when evaluating the Curated PBs using all the language models as training set, we can see that the results are consistent and make the method reliable across languages, which will help to reduce the amount of effort in the correction process.

4.2.1 Error analysis

Although the full error analysis could not be completed by the curators who had previously performed the manual correction on the three Curated PBs, I obtained some partial comments and suggestions from them, which I add to my own analysis of the Spanish dataset I labelled for the evaluation. This analysis cannot be conclusive, due to the small amount of data taken as sample in both cases.

During the analysis of some of the cases in Spanish, I realised there were some sparse phraseological expressions that could add some noise to the outcome when using a semantic model to evaluate their similarity, as the projection method only takes into account single words in the alignments, but again, I cannot determine the extent of its interference.

Another remark pointed to the fact that for the dividing task it was important to take into account if the verb pair candidates had the same group of semantic roles or not, as shown in the English PB (for example, the verb pair 'abandon.01' (with the roles Arg0-PPT, Arg1-DIR, Arg2-PRD) and 'quit.01' (with the roles Arg0-PAG, Arg1-PPT) for the Spanish verb 'abandonar.01' does not have the same group of roles, etc.) This made me add to the overall model the role comparison as a final constraint, using again all the other languages for setting the parameter of the same learning algorithm (Table 4.7).

Resource	Task	Accuracy	Error Rate	Precision	Recall	F-score	Weighted Average F-score
Spanish PB with Roles	merge	80.9	19.0	83.9	93.1	88.2	79.3
	divide			64.3	40.9	50.0	

Table 4.7. Results for the Projected Spanish PB with role comparison as feature.

As we can see, although there is an improvement in all the scores when compared with the use of the language models only as training set (without this last constraint), even in the merging task, the highest improvement is in the precision and recall of the dividing task, with a difference of 5.7 and 2.3 points respectively. It is still to be determined if a more fine grained analysis on the partial intersection of shared roles could improve the whole process.

Chapter 5

Conclusions & Future Work

In this chapter, I ouline the main conclusions derived from the many experiments I did and the results obtained (5.1) as well as some guidelines for future work on this topic (5.2).

5.1 Conclusions

In this work, I propose an automatic alternative to the second step in the manual frame correction method proposed by Akbik *et al.* (2016). I make use of a novel method in natural language understanding, multilingual distributional semantics (DS), and prove it can be adapted for this correction task, within the framework of semantic role labeling (SRL).

The main problem with the manual method proposed by the authors was the number of lexical items that needed to be evaluated for each resource. It is evident that a manual review has to be done by language experts, which it is costly, time-consuming, and can lead to inconsistencies if there are variations in the annotators' criteria.

Although this project focuses on correcting the Projected Spanish PB, it is worth noting that my method is language-independent, and can be used to correct all the multiple PBs generated with this method for other languages.

Three main distributional semantic models were built with different aligned data: 1) Models with bilingual information; 2) Models with multilingual information, with 25 languages in total, and 3) Models with monolingual (English) syntactic information. The experiments confirmed that I had to take into account two relevant parameters during the correction task: 1) the model enrichment via the back-translation assumption (BTA), and 2) the different schemes for dimensionality reduction (DR). Therefore, it was necessary to test them separately, by creating 1) Normal models and models enriched, taking into account the BTA; and 2) Models with an increased and

decreased DR (I / D-DR). At the end, I tested 10 different models, which were used as features with a learning algorithm, and included a roleset constraint as additional feature.

In general, the multilingual models performed better than the bilingual ones: the Multilingual DSM with BTA and I-DR obtained the best results for merging, and the Bilingual DSM with R-DR obtained the best results for filtering, although they were always were lower than the merging scores.

As shown in previous experiments, the models with monolingual syntactic information do not seem to improve the results for our task. This is in line with the results obtained by Van der Plas & Tiedemann (2006), when they compare and combine a multilingual alignment method and a monolingual syntax-based method to extract semantically related words.

I prove that single DSMs are not necessarily the best solution for this kind of task, and I managed to take a new approach and build a combined model, using machine learning algorithms based on the outputs (similarity scores) of the individual models as features. Finally, I discovered that role comparison is an important feature that needs to be taken into account when evaluating the precision and recall of the output.

The results for Spanish are promising and consistent with the results in other languages (French, German and Chinese), reaching a 79.3 weighted average score for the correction task. I proved it is possible to train a model with only a small amount of data, although there is a slight drop in the performance, but also using other langauges as models, which seems to be the best option.

Although my method has not reached yet the stage where it gets to the same decisions as the language experts do and correct every case by itself, it can function in a semi-automatic manner, that is to say, the method can detect strong and weak redundant candidates, which helps to speed up and stabilize the criteria of the correction process. Finally, even when the method does not get to the same results, it can shed some light on the decision process and the other linguistic parameters that the human ‘curators’ used when performing such task.

5.2 Future Work

During the research process, I attempted to add new frames to the Projected Spanish PB by two means: the integration of the Spanish AnCora and the verb alignments I extracted in Spanish, however, due to the complexity of the task, this could not be finished, but can serve for future work on this topic.

If we compare the number of verbs that could be added in an Augmented PB (Projected Spanish PB, with 1,585 verbs + Spanish AnCora, with 2,820 verbs), we can see there is a total of 3,033 different Spanish verbs. This process would basically require to re-group all the verb senses of the new augmented resource, perform my method for redundancy correction to the extracted candidates, and re-assign the new number of verb senses for each verb. I almost achieved this goal, but I decided it was more important to ensure the best possible performance of the correction method as the first objective for evaluation, before trying any other further steps. It is important to note that this enrichment method can be used for PBs in other languages as well.

I also found out that there were 12,178 verbs in the English-Spanish aligned data that did not exist in any of the previous individual resources. Just to give us an idea, the Projected Spanish PB used during the projection step 2,687 verbs from its English-Spanish aligned corpora. The reason why I found more relations is simply because I used more corpora, but nonetheless, these are verb relations that could be used to enrich and increase the coverage of the PBs. There were 3,033 different verbs by combining the Projected Spanish PB and the Spanish AnCora, and if I added the verbs in the aligned data, the resource would have 3,920 verbs, which is much closer to the 5,649 verbs available in the English PB. It is true that the aligned data does not make any verb sense distinction as such, so a different and more complex technique would be needed. But the correction process for redundancy could be applied in the same manner as in the previous step and this could be also used to enrich PBs in any other languages.

References

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1), 37-66.

Alfonseca, E., & Manandhar, S. (2002, January). An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st international conference on general WordNet*, Mysore, India (pp. 34-43).

Akbik, A., Chiticariu, L., Danilevsky, M., Li, Y., Vaithyanathan, S., & Zhu, H. (2015). Generating High Quality Proposition Banks for Multilingual Semantic Role Labeling. In *Association for Computational Linguistics* (1) (pp. 397-407).

Akbik, A., Guan, X., & Li, Y. (2016). Multilingual Aliasing for Auto-Generating Proposition Banks. In *Conference on Computational Linguistics* (pp. 3466-3474).

Akbik, A., & Li, Y. (2016a). POLYGLOT: Multilingual semantic role labeling with unified labels. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, (pp. 1-6) Association for Computational Linguistics.

Akbik, A., & Li, Y. (2016b). K-SRL: Instance-based Learning for Semantic Role Labeling. In *Proceedings of the 26th International Conference on Computational Linguistics*, (pp. 599-608). COLING.

Aziz, W., Rios, M., & Specia, L. (2011). Shallow semantic trees for SMT. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (pp. 316-322). Association for Computational Linguistics.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (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 (pp. 86-90). Association for Computational Linguistics.

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology (LiLT)*, 9.

Barzilay, R., & McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting on Association for Computational Linguistics* (pp. 50-57). Association for Computational Linguistics.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., & Xia, F. (2009,). A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop* (pp. 186-189). Association for Computational Linguistics.

Bjö rkelund, A., Hafdell, L., & Nugues, P. (2009). Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 43-48). Association for Computational Linguistics.

The British National Corpus, Version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: *http://www.natcorp.ox.ac.uk*

Borrega, O., Taulé, M., & Martı , M. A. N. (2007). What do we mean when we speak about Named Entities. In *Proceedings of Corpus Linguistics*.

Che, W., Li, Z., Li, Y., Guo, Y., Qin, B., & Liu, T. (2009). Multilingual dependency-based syntactic and semantic parsing. In *Proceedings of the thirteenth conference on computational natural language learning: shared task* (pp. 49-54). Association for Computational Linguistics.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*, 16-75.

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography, In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*.

Civit, M., & Martí, M. A. (2004). Building cast3lb: A Spanish treebank. *Research on Language and Computation*, 2(4), 549-574.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4), 589-637.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.

Curran, J. R., & Moens, M. (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition* – Vol. 9 (pp. 59-66). Association for Computational Linguistics.

Daelemans, W., & Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge University Press.

Dagan, I., Lee, L., & Pereira, F. C. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1), 43-69.

Dagan, I., Itai, A., & Schwall, U. (1991). Two languages are more informative than one. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics* (pp. 130-137). Association for Computational Linguistics.

Davies, M., & Ferreira, M. (2015). *The Wikipedia Corpus*. URL: *https://corpus.byu.edu/wiki*

Dyvik, H. (1998). A translational basis for semantics. *Language and Computers*, 24, 51-86.

Dinu, G. Pham, T.N. & Baroni, M. (2013). DISSECT-DIStributional SEmantics Composition Toolkit. *Proceedings of the System Demonstrations of ACL 2013* (51st Annual Meeting of the Association for Computational Linguistics), East Stroudsburg PA: ACL.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1535-1545). Association for Computational Linguistics.

Fillmore, C. J. (1967). *The case for case.*

Fowler, R. (1996). *Linguistic Criticism.* Oxford: Oxford University Press.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a

statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2), pp. 337-407.

Ganeri, J. (1999). *Semantic powers: Meaning and the means of knowing in classical Indian philosophy.* Oxford University Press.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3), 245-288.

Gildea, D., & Palmer, M. (2002). The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 239-246). Association for Computational Linguistics.

Gruber, J. S. (1965). *Studies in lexical relations* (Doctoral dissertation, Massachusetts Institute of Technology).

Hajič , J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Mà rquez, L., ... & Straň á k, P. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 1-18). Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146-162.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., & Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3), 311-325.

Ibrahim, A., Katz, B., & Lin, J. (2003). Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the second international workshop on Paraphrasing* - Volume 16 (pp. 57-64). Association for Computational Linguistics.

Khan, A., Salim, N., & Kumar, Y. J. (2015). A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30, 737-747.

Kingsbury, P., & Palmer, M. (2002). From TreeBank to PropBank. In *Language Resources and Evaluation* (pp. 1989-1993).

Kingsbury, P., Palmer, M., & Marcus, M. (2002). Adding semantic annotation to the Penn Treebank. In *Proceedings of the human language technology conference* (pp. 252-256). San Diego, California.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2006). Extending VerbNet with novel verb classes. In Proceedings of LREC (Vol. 2006, No. 2.2, p. 1).

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine learning*, 59(1-2), 161-205.

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

Lin, J. W. (1998). Distributivity in Chinese and its implications. *Natural Language Semantics*, 6(2), 201-243.

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Language Resources and Evaluation*.

Lo, C. K., Addanki, K., Saers, M., & Wu, D. (2013). Improving machine translation by training against an automatic semantic frame based evaluation metric. In Association for Computational Linguistics (2) (pp. 375-381).

Maqsud, U., Arnold, S., Hü lfenhaus, M., & Akbik, A. (2014). Nerdle: Topic-specific question answering using wikia seeds. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* (pp. 81-85).

Martí, M.A. & Taule, M. (eds., 2008). *CESS-ECE TreeBanks*, Barcelona. Publicacions de la Universitat de Barcelona.

Martí, M. A., Taulé, M., Má rquez, L., & Bertran, M. (2007). Anotación semiautomá tica con papeles temá ticos de los corpus CESS-ECE. *Procesamiento del Lenguaje Natural-TIMM*, 38, 67-76.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., & Grishman, R. (2004). The NomBank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation* (Vol. 24, p. 31).

Nguyen, D. Q., Dai Quoc Nguyen, D. D. P., & Pham, S. B. (2014). RDRPOSTagger: A ripple down rules-based part-of-speech tagger.

Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., & Pham, S. B. (2016). A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. *AI Communications*, 29(3), 409-422.

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D. & Tsarfaty, R. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Language Resources and Evaluation*.

Osman, A. H., Salim, N., Binwahlan, M. S., Alteeb, R., & Abuobieda, A. (2012). An improved plagiarism detection scheme based on semantic role labeling. *Applied Soft Computing*, 12(5), 1493-1502.

Padó, S. (2007). *Cross-lingual annotation projection models for role-semantic information*. Saarland University.

Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. In *Journal of Artificial Intelligence Research*, 36(1), 307-340.

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

Palmer, M., Gildea, D., & Xue, N. (2010). Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1), 1-103.

Paul, M., & Jamal, S. (2015). An improved SRL based plagiarism detection technique using sentence ranking. *Procedia Computer Science*, 46, 223-230.

Rappaport-Hovav, M., & Levin, B. (1998). Building verb meanings. The projection of arguments: Lexical and compositional factors, 97-134.

Roth, M., & Lapata, M. (2016). Neural semantic role labeling with dependency path

embeddings. *arXiv:1605.07515*.

Roth, M., & Woodsend, K. (2014). Composition of Word Representations Improves Semantic Role Labelling. In *Empirical Methods on Natural Language Processing* (pp. 407-413).

Samardž ić , T., Van Der Plas, L., Kashaeva, G., & Merlo, P. (2010). The scope and the sources of variation in verbal predicates in English and French.

Sebastiá n-Gallés, N., Martí, M. A., Carreiras, M., & Cuetos, F. (2000). *LEXESP: Una base de datos informatizada del español*. Universitat de Barcelona, Barcelona.

Shen, D., & Lapata, M. (2007). Using Semantic Roles to Improve Question Answering. In *Conference on Empirical Methods on Natural Language Processing* (pp. 12-21).

Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon.

Sumner, M., Frank, E., & Hall, M. (2005). Speeding up logistic model tree induction. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 675-683).

Taulé, M., Martí, M. A., & Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Language Resources and Evaluation*.

Tesniére, L. (1959). *Éléments de Syntaxe structurale*. Paris, I959.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Language Resources and Evaluation* (Vol. 2012, pp. 2214-2218).

Tiedemann, J. (2009). News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing* (Vol. 5, pp. 237-248).

Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and categorization*, - Vol. 1(1978), 79-98.

Van Der Plas, L. (2009). Combining syntactic co-occurrences and nearest neighbours in distributional methods to remedy data sparseness. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics* (pp. 45-53). Association for Computational Linguistics.

Van der Plas, L. & Bouma, G. (2005). Syntactic contexts for finding semantically related words. *LOT Occasional Series*, 4, 173-186.

Van der Plas, L., Merlo, P., & Henderson, J. (2011). Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* - Volume 2 (pp. 299-304). Association for Computational Linguistics.

Van der Plas, L., & Tiedemann, J. (2006). Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL* (pp. 866-873). Association for Computational Linguistics.

Vulić , I., Mrkš ić , N., Reichart, R., Séaghdha, D. Ó, Young, S., & Korhonen, A. (2017). Morph-fitting: Fine-Tuning Word Vector Spaces with Simple Language-Specific Rules. arXiv:1706.00377.

Wu, H., & Zhou, M. (2003). Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the second international workshop on Paraphrasing* - Volume 16 (pp. 72-79). Association for Computational Linguistics.

Xiong, D., Zhang, M., & Li, H. (2012). Modeling the translation of predicate-argument structure for smt. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 902-911). Association for Computational Linguistics.

Xue, N., Xia, F., Chiou, F. D., & Palmer, M. (2005). The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. In *Natural language engineering*, 11(2), 207-238.

Zhao, H., Chen, W., Kit, C., & Zhou, G. (2009). Multilingual dependency learning: A huge feature engineering method to semantic dependency parsing. In *Proceedings*

of the Thirteenth Conference on Computational Natural Language Learning: Shared Task (pp. 55-60). Association for Computational Linguistics.

Zhao, H., Chen, W., Kazama, J. I., Uchimoto, K., & Torisawa, K. (2009). Multilingual dependency learning: Exploiting rich features for tagging syntactic and semantic dependencies. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 61-66). Association for Computational Linguistics.

Ziering, P., Mü ller, S., & Van der Plas, L. (2016). Top a Splitter: Using Distributional Semantics for Improving Compound Splitting. ACL 2016, 50.