Irene Renau, Rogelio Nazar, and Daniel Mora

TOWARDS THE AUTOMATIC GENERATION OF A PATTERN-BASED DICTIONARY OF SPANISH VERBS

Abstract Corpus Pattern Analysis, CPA (Hanks, 2004a; 2004b; 2013), is a technique for identifying local semantic and syntactic information of a word and mapping it to its meanings. In verbs, it consists basically of the argument structure labelled with semantic types for each argument. CPA is used in several dictionary projects and allows systematic corpus analysis; however, it is extremely time-consuming. In this paper, we present a method for the automatic pattern identification of Spanish verbs in corpora. We used a syntactic parser for dependency analysis (Stanza), applied a named entity recognition (NER) tagger from the Flair NLP framework for NER and, for common nouns, we implemented a semantic tagger and a word sense disambiguation method, both created for the task. All resources were combined to extract CPA verb patterns. The method performs better than previous attempts and can contribute to a more efficient pattern-based lexicography.

Keywords argument structure; Corpus Pattern Analysis; pattern-based lexicography; semantic tagging; word sense disambiguation

1. Introduction

In this paper, we present a method for verb pattern recognition in corpora. For this particular experiment, we work with Spanish transitive verbs and use a series of strategies that, combined, can extract syntacto-semantic structures from discourse. A pattern can be defined as a piece of phraseology which is recurrent in discourse and can be mapped to a particular meaning of a verb (Hanks, 2004a; 2004b; 2013; Sinclair, 2004). For example, if someone says *We love Patrick,* it means having 'strong positive feelings for Patrick', while saying *We love parties* means 'liking parties very much'. The idea is that the pattern *[[Human]] love [[Human]]* is the pattern beneath the first sentence, and *[[Human]] love [[Event]]* is the one beneath the second one. Corpus Pattern Analysis, CPA (Hanks, 2004a; 2004b; 2013), offers a systematic way of analysing these patterns in corpus data and has been used in many dictionary projects of different languages. While this technique is highly sound both theoretically and methodologically, it is also very time-consuming. Our goal then is trying to find a way for the teams to extract these patterns from a corpus more efficiently.

In this article, we first present a state of the art of CPA (Section 2): its theoretical and methodological background, a summary of dictionaries and lexical databases using CPA to analyse corpus data, and a mention of some previous attempts to automate pattern recognition. We then explain the method (Section 3), dividing it into syntactic and semantic analysis, and offer an evaluation of results (Section 4). We finish with some conclusions and limitations of the experiment, which point to future work (Section 5).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

2. Corpus Pattern Analysis and Pattern-Based Lexicography: State of the Art

2.1 Theoretical and Methodological Background

CPA is a method to disambiguate words in context (Hanks, 2004a; 2004b: 2013; Hanks & Ježek, 2008; Hanks & Pustejovsky, 2005; cf. also Grefenstette & Hanks, 2023). It is described by Hanks (2004a, p. 87) as a "technique for mapping meaning onto words in text". It is based on a vast body of work considering that words do not have meaning in isolation, but most likely a meaning potential that is activated in a certain context of usage. Sinclair (1991; 2004), the closest and important antecedent of CPA, proposes that the problem of meaning must be tackled from a syntagmatic perspective, being corpus analysis of paramount importance "to observe recurrent patterns of language" (Sinclair, 2004, p. 140).

CPA develops the Sinclairian notion of pattern into a fully semi-automatic method of corpus analysis. Context of analysis is restricted to local context, which "is usually sufficient to assign a specific sense to a word and to distinguish one sense from another" (Hanks & Pustejovsky, 2005, p. 64). A pattern, thus, is broadly defined in CPA as "a semantically motivated and recurrent piece of phraseology" (Ježek & Hanks, 2010, p. 8). As we will show in sections 2.2 and 2.3, CPA is used mainly for verbs. In verbs, a brief description of a pattern is given by the acronym 'T-PAS' (Ježek et al., 2014; see Section 2.2): *Typed Predicate Argument Structure.* Indeed, a pattern consists of the argument structure of the verb and the semantic typing of each argument. Some additional syntactic information is added when required, such as pronouns for pronominal verbs (e.g., in Romance languages) or prepositions in some type of complements.

Semantic types used to label each argument are organised in an ontology suitable for everyday language, containing basic conceptual categories (Hanks & Pustejovsky, 2005; Ježek & Hanks, 2010). Semantic types in this ontology must be specific enough to distinguish one pattern from the other, especially when the syntactic structure is identical. However, categories are not fine-grained to the point that they are useless for the task, e.g., for CPA, a category such as *cooking apple* would be unnecessary.

Figure 1 shows an example of CPA for a verb extracted from the *Pattern Dictionary of English Verbs,* PDEV (Hanks & Pustejovsky, 2005).

Fig. 1: Entry for *to love* in the PDEV [last access: 20/5/2024]

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

2.2 CPA Projects

There are currently multiple projects using CPA for lexical analysis and lexicography in different languages. The pioneer project is the PDEV, developed under the direction of Patrick Hanks, mostly at the University of Wolverhampton. In its public version, it includes at the moment 1,691 verb entries, each one containing a list of the normal patterns of verbs and the description of the meaning–implicature–mapped onto each pattern. The dictionary is organised as a database in which users can also look for the semantic types of the patterns. For example, the category *[[Artefact]]* can be searched and, as a result, we obtain the list of all verbs in which this category is used at least in one of the patterns. The CPA Ontology containing the organisation of all semantic types is also provided in one of the sections of the webpage.

PDEV has been the foundational project and the inspiration for an array of other initiatives. Especially similar to PDEV are T-PAS, CROATPAS and Verbario, all lexical databases suitable to be used for downstream NLP tasks. T-PAS (Ježek et al., 2014) is a database of Italian verbs. It contains 1,000 verbs with different degrees of polysemy and can be downloaded with a Creative Commons license. CROATPAS (Marini & Ježek, 2019; Marini, 2022) is developed following T-PAS and is used to analyse the polysemy of Croatian verbs. At the moment, it contains 180 verbs and a total of 683 patterns distributed in the entries. CROATPAS shows the feasibility of using CPA for under-resourced languages. The Verbario database (Renau et al., 2019) contains around 200 verbs. Each section of the verb entry contains the pattern, the meaning (or implicature of the pattern), and a link to all concordances that were analysed to extract the patterns from corpus (around 250-1,000 concordances per verb are typically analysed in Verbario, depending on the complexity of the verb). Concordances are linked to each pattern and, in each entry, patterns are displayed with numbers. In cases of inchoative verbs, the alternance is shown by adding a letter to the number.

CPA is also used in different learners' dictionary projects. They have in common that they use CPA for corpus analysis, but they adapt the formulation of the pattern to learner's needs and to the specific project requirements, which implies different types of modifications and simplifications. It is the case of an advanced learners' dictionary of Italian currently in progress (DiMuccio-Failla & Giacomini, 2017; 2022), The *Diccionario de aprendizaje del español como lengua extranjera,* DAELE, a proposal for a learners' dictionary of Spanish (Battaner & Renau, 2011), or the DSELE, a proposal for pronominal verbs in Spanish in a learners' dictionary (Renau, 2012). Finally, the Woordcombinaties (Colman & Tiberius, 2018) is an ongoing project for a Dutch learners' dictionary. In all these cases, CPA is used for corpus analysis and, after creating the patterns, they are adapted to the dictionaries.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

2.3 Computer Methods for CPA

All dictionaries and databases described in sections 2.2 and 2.3 have in common that they use semi-automatic procedures, mainly different Sketch Engine tools (Kilgarriff et al., 2014). However, a large amount of work is still done manually and it is extremely time-consuming. There have been some proposals in this way since the beginnings of the PDEV–some of them are summarised by Giacomini and DiMuccio-Failla (2019). Pustejovsky, Hanks & Rumshisky (2004) made a preliminary attempt for automatic pattern acquisition incorporating tokenisation, POS tagging, lemmatisation and dependency analysis to corpus sentences, observing if results matched the manual analysis. In this study, semantic tagging seems to be a human operation nonetheless. Baisa et al. (2015) describe a SemEval shared task based on CPA, in which three sub-tasks were proposed, working with English data only: 1) CPA parsing (syntactic and semantic) of corpus sentences, 2) CPA clustering (of corpus sentences according to their similarities) and 3) CPA lexicography, which involved pattern building according to previous parsing and clustering. Teams report results on sub-tasks 1 and 2 but not for 3, and no team participated in more than one task.

In the case of the Verbario project, different efforts have been made to automate the procedure of pattern induction for Spanish verbs covering all these tasks, from parsing to pattern building (Renau et al., 2019). The following steps are necessary for replicating the manual procedure with computational methods: 1) Extracting corpus concordances, 2) Using a parser for dependency analysis in order to identify the main verb and its argument structure, 3) Semantic tagging of the arguments, that is, to identify the category of the head of each constituent of the sentence playing the role of argument, and 4) Pattern building. The attempts made in the Verbario project following these steps were promising but still left ample room for improvement (about 60% of the patterns were acceptable). The main problem were the difficulties of the semantic tagging and the lack of a method to deal with polysemy in the hypernymy links of the target nouns (Nazar & Renau, 2016).

3. Methods

In this section, we propose an improvement of the method used in the Verbario project (shown in Section 2.3) to automate tasks of corpus analysis and pattern building. The basic approach remains the same, as we imitate the manual process following the four steps already mentioned. However, we made adjustments in all steps to improve precision, especially concerning semantic tagging. For this experiment, we dealt only with transitive verbs and, specifically, only those with two arguments: subject and direct object. A summary of the method is shown in Figure 3.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

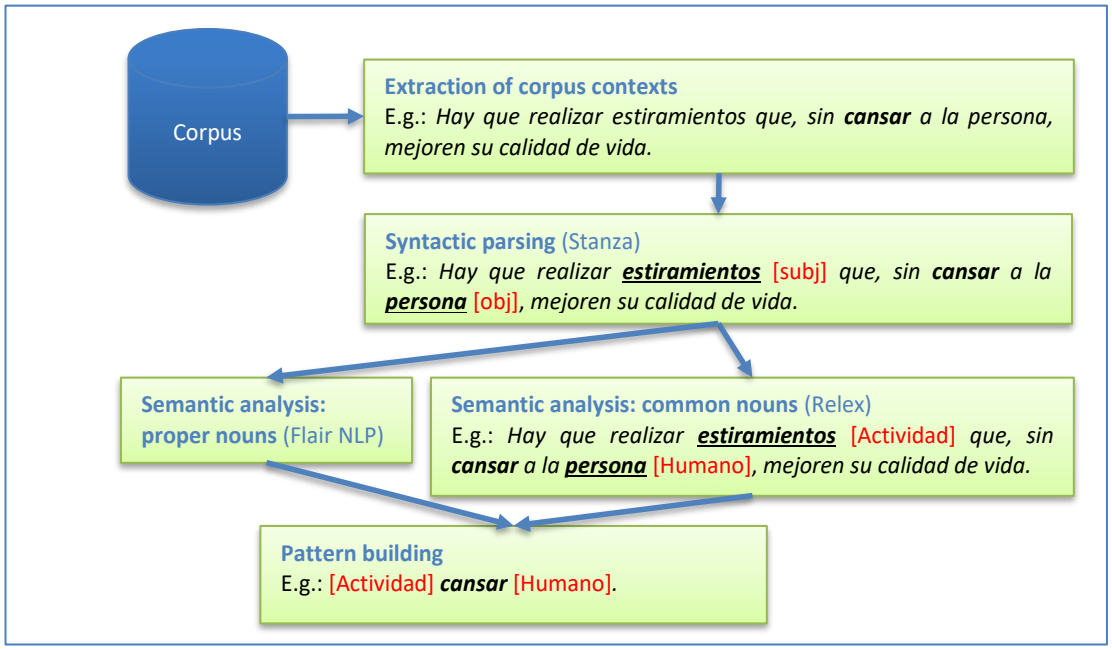

Fig. 3: Synthesis of the method

3.1 Corpus and Data Processing

We used the Spanish section of the Opus Corpus (Tiedemann, 2009), consisting of 25 billion words of texts from a variety of genres and sources. For the analysis we excluded very long sentences which may contain complex clauses. Our idea was to try to simplify the task for the syntactic parser, as previous attempts of parsing without filtering led to unsatisfactory precision (see Section 4.1). Also, we used the esTenTen corpus (Kilgarriff & Renau, 2013) to extract word association coefficients (see Section 3.3.3).

3.2 Syntactic Parsing

The syntactic parsing was carried out using the Stanza library (Qi et al., 2020), as it is slightly better compared to UDPipe 2.0 (Straka, 2018) in the Spanish-AnCora corpus. Stanza is made up of a neural network pipeline that contains several processors that take care of different text processing tasks, and annotates using universal dependencies and universal features. In this case, we used its lemmatisation module, part-of-speech / morphological tagger, and dependency parsing Spanish processors.

In order to match up both the syntactic and the semantic parsing at token level, the input was first tokenised by the semantic parsing model, and then fed to Stanza. The main focus was on extracting the basic argument structure (subject and direct object), considering transitive verbs only. For simplicity, if a syntactic argument of a verb is a noun phrase, we only consider the head noun, and if two sentences are joined by a conjunction, we retain the structure of the first sentence. The output, therefore, is each token annotated at the syntactic level, and the main tree structure (subject, root, direct object) of the sentence. For example, a sentence such as *Nepal ha presentado*

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

periódicamente informes al Comité 'Nepal has periodically submitted reports to the Committee' has *presentar(Nepal, informe)* as output.

3.3 Semantic Parsing

3.3.1 Proper Nouns

We identified proper nouns using a NER tagger from the Flair NLP framework (Akbik et al., 2019a), specifically, the 'ner-multi-fast' model (Akbik et al., 2019b), available through the HugginFace library (Wolf et al., 2020). This model was preferred over other Spanish NER taggers given its trade-off between processing speed and accuracy. The 'ner-multi-fast' model is trained on six different languages at a character level, including Spanish. In our experiment, the Flair model receives tokenised text from Stanza and, for each token, it outputs its NER tag (i.e., Person, Location, Organization, Miscellaneous) in BIO format, e.g., *El texto fue escrito por el poeta de Weimar[LOC] Salomon[B-PER] Franck[E-PER]* 'The text was written by Weimar's poet Salomon Franck' (LOC: location; PER: person).

3.3.2 Common Nouns

We populated a taxonomy of common nouns using a general Spanish dictionary and a word sense disambiguation (WSD) system. The dictionary is used to obtain the hypernym from the definition of each noun, and the WSD system is used to concatenate large hyponym-hypernym sequences in a coherent manner. Without a WSD system, the hypernym chains would quickly derail by the effect of polysemy, as already noted by Amsler (1981). For instance, from the dictionary it is possible to determine that a *ñacaratiá* is a kind of *árbol* 'tree', and then that an *árbol* is a kind of *planta* 'plant'. But then *planta* is a polysemous word, like in English, where we have for instance 'industrial installation' that is not a 'living organism'. Thus, the system must figure out what sense of the word *planta* is said to be the hypernym of *árbol*. Furthermore, a WSD is also needed later during the task of semantic tagging of the analysed text, as there will be frequent encounters with polysemous nouns in the texts to be tagged, words that have to be disambiguated before assigning them a semantic type.

The application of a WSD system for the development of a taxonomy means that one has to start from each noun in the lemma list of the dictionary and work the way up to some of the semantic types of the top-ontology. The only advantage of the application of the WSD system to taxonomy extraction over the typical case scenario is that one can use the inheritance mechanism, i.e., to use the text of the definitions retrieved from the lower links of the hypernym chain. The words in the definition of *ñacaratiá* and *árbol* are helpful to select the corresponding sense of *planta* because of the semantic similarity of the words in the definitions. In other words, we have a capital of keywords to distinguish between a 'living organism' and, say, an 'industrial installation'.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

As the dictionary, we used the Spanish Wiktionary, a useful resource seemingly underused in NLP. Inspired by an old tradition in computational linguistics (Chodorow et al., 1985; Guthrie et al., 1990, and many others), we designed a rule-based algorithm to extract the hypernyms from the definitions which is able to circumvent the different obstacles presented by the lack of uniformity and systematicity of this lexicographic resource. In most cases, hypernyms correspond to the first noun in the definition, as in *ñacaratiá,* defined as "*Árbol* frutal sudamericano…" 'South American fruit *tree*…'. However, there are many exceptions and one has to avoid metalinguistic expressions (*type of..., kind of..., name of...*, etc.) and scientific names of animals and plants, among other information.

This rule-based method of taxonomy induction from a single dictionary is a significant departure from earlier attempts in the context of the Verbario project, which were based on statistical corpus analysis and did not adequately handled the cases of polysemy. We leave for future work the possibility of combining earlier ideas (e.g., Nazar & Renau, 2016) with the method now proposed.

3.3.3 Word Sense Disambiguation Methods

Regarding the WSD system, in Spanish the subject has been thoroughly investigated (Agirre et al., 2014; Bevilacqua et al., 2021, among others). However, to the best of our knowledge, there are scarce open-source implementations.

For the development of the WSD system, we again went to the basics, following the policy of first putting in place something simple but functional, that could be progressively developed. Starting from the first ideas by Lesk (1986), we conceptualise the task of WSD as follows: there is an input sentence that contains a polysemous noun to be disambiguated, we have a list of senses of such word in a dictionary, and we have to select the most likely sense on the basis of the associations we find between the context of the noun and the text of the definitions. Lesk's (1986) first idea was to use the vocabulary intersection between both sets, but this is not very promising, as the chances of finding such an intersection are minimal.

3.3.3.1. Statistical WSD With Relex

Our first attempt was to redefine Lesk's (1986) proposal and develop a statistical WSD method (*Relex*), which instead of the intersection tries to calculate a semantic similarity between definition and target noun. Of course, if any intersection, however unlikely, happens to occur, it is also considered. Most of the computation, thus, involves words that are not the same, but that are semantically related. We do not need a precise definition of semantic relatedness, and it could be very diverse. We say, for instance, that the nouns *cat* and *dog* are semantically related. Conceptually, they are co-hyponyms, but operationally, we considered that any pair of words is semantically related if they show a strong syntagmatic association in a large Spanish reference corpus, in our case, the EsTenTen. To compute these associations, we analysed a sample of the corpus (1/1,000) and produced tables of syntagmatic associations

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

between pairs of all kinds of words, i.e., independently of their grammatical category. The values of this table are $word_1$, $word_2$, total frequency of $word_1$, total frequency of $word_2$, the frequency of co-occurrence in the same sentences and an association measure. Relex then uses this table to compare the vocabulary of the target text and each definition at a time, collecting the values of the association of each pair of words. The values are summed for each sense.

For illustration, consider the case of the frequently used polysemous Spanish word *virus.* The dictionary lists two senses: one with *organismo* 'organism' as hypernym and another with *software* as hypernym. Given a certain context of usage, there will be a strong association between collocates such as *pacientes* 'patients' or *transmisión* 'transmission', etc. and the first meaning, while words such as *antivirus* 'anti-virus', *herramienta* 'tool' or *computadora* 'computer', etc. will be strongly associated to the second meaning.

3.3.3.2. LLM-Based WSD

We also developed another method with three alternative approaches that use MarIA (Gutiérrez-Fandiño et al., 2022), the most extensive RoBERTa-based large language model (LLM) trained on Spanish corpora. These three approaches followed common and novel strategies for WSD using LLM (Huang et al., 2019; Bevilacqua et al., 2021), making special use of the fill-mask task and cosine similarity.

The first approach ('fill-cooc') extends Lesk's (1986) intuition by measuring the vocabulary intersection between a candidate's definition and the definitions of 10 words generated by the model when doing fill-mask on the target word. Fill-mask is the task in which the model predicts the masked token within a sentence. For example, the target *ratón* 'mouse' from the sentence *Se pulsa el botón del ratón* 'The mouse button is pressed' is masked and fed to the model (i.e., "The [MASK] button is pressed"). It is expected that by expanding the sets of words, intersections between definitions are more likely to occur.

The second and third approaches use cosine similarity, a metric for measuring semantic relatedness between two vectors. The value ranges from -1 (very dissimilar) to 1 (very similar), with 0 indicating no relation. Word embeddings, generated by the LLM, are used to obtain these vectors. The second approach ('fill-cosine') vectorises each definition and the concatenation of the 10 generated words after applying fill-mask. Then, cosine similarity is measured between the candidate definition and the generated words, thus circumventing the direct intersection of words.

The third approach ('target-cosine') computes cosine similarity between the vectorised input sentence and each definition of the target word without doing fill-mask. This approach exploits contextualised word embeddings generated by the LLM. Contextualised word embeddings adjust based on the specific context in which the word appears. When applying cosine similarity (approach 2 and 3), the definition with the highest score is the predicted one.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

3.4 Pattern Building

Once we have all the Opus Corpus fully tagged with the different layers we have been describing in sections 3.2 and 3.3 (i.e., POS-tags, syntactic parsing and semantic tags), we proceeded with the construction of CPA patterns. For each of a list of Spanish verbs, we extracted a maximum of 5,000 sentences that have them as the main verb of a simple transitive structure (we reserve for future work the possibility of using other structures). For each sentence, we collected the semantic type of both syntactic arguments of the verb and computed a frequency table of the types.

For illustration, consider the case of the verb *aburrir* 'to bore' used in contexts such as *Él nos aburrió* 'He bored us', *Esa gente me aburre* 'This people bore me', etc. Each of these sentences is recognised as an instance of the pattern *[[Humano]] aburrir [[Humano]]* '[[Human]] bore [[Human]]'. On the other hand, cases such as *El debate me aburre* 'The debate bores me', *Los trabajos le aburrían* 'Jobs bored him/her', etc. are instances of the pattern *[[Acción]] aburrir [[Humano]]* '[[Action]] bore [[Human]]'. In this way, as the analysis progresses and the number of sentences increases, so does the importance of each pattern in the frequency table. Once all the sentences of a given verb are analysed, we retain the *n* most frequent patterns.

The number of different patterns per verb will increase or decrease depending on the frequency of the verb but also depending on how general the semantic type is. We have retained this value as an execution parameter: if the system considers as semantic types only very abstract nouns (i.e., those placed in the higher levels of the taxonomy), then the number of different patterns decrease, as different sentences are lumped into the same pattern. On the contrary, using more specific (lower) semantic types splits the patterns in many. For instance, if instead of, say, *[[Vehicle]]*, we consider more abstract semantic types such as *[[Artefact]]*, then different artefacts other than vehicles will be placed under the same pattern. A systematic method to automatically optimise this level of abstraction is left for future work.

4. Results and Evaluation

4.1 Evaluation of the Syntactic Parser

For the evaluation of the syntactic parser (see Section 3.2), we drew two random samples from the Opus Corpus of 100 sentences each: the first one had no restrictions, i.e., all types of sentences had the same probability of being in the sample, and the second was limited to simple transitive sentences, i.e., no more than two arguments, no attachments and no subordination. Table 1 shows results of the evaluation of both samples.

Table 1: Figures of precision in percentage of syntactic parsing with Stanza

	Correct	**Incorrect**	**Unanalysable**
Unrestricted sample	50	32	18
Simplified sample	81	19	0

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

We limited our error analysis to those cases in which the system was unable to correctly identify the main verb and the arguments of the predicate structure. Unsurprisingly, these were more widespread in the case of the unrestricted sample, which contains larger and more syntactically complicated sentences. In the case of the simplified sample, there are still many errors. However, an 81% success rate is quite acceptable for our purposes, and it is only moderately lower than the results reported in the evaluation of Stanza with the Spanish-AnCora corpus. The 31% difference between both samples confirms that the decision to restrict the corpus was correct.

4.2 Evaluation of the Proper Noun Semantic Tagger

For the evaluation of the NER system (see Section 3.3.1), we proceeded with the same random sample of 100 sentences used in the previous section, specifically the unrestricted sample, and analysed in this case the identification and categorisation of the proper nouns. We manually identified a total of 112 proper nouns in the sample. All but two were correctly detected by the NER tagger (98% success). With regard to the semantic categorisation of the nouns, 94 of the 110 (85%) received the correct tag. The errors were in most cases names of persons that were tagged as organisations and organisations that received the category of 'miscellanea'.

4.3 Evaluation of the Common Noun Semantic Tagger

The evaluation of the semantic tagging of common nouns is divided in its two main components: taxonomy induction (as explained in Section 3.3.2) and WSD (Section 3.3.3). For the first component, we evaluated the results of this task by random sampling 100 nouns from the taxonomy, in order to determine in which cases there was a correct ascending path to the top-node of the taxonomy. As was to be expected, a proportion of the nouns in the sample were polysemous, which means that we had to evaluate more hypernymy chains than nouns. In total, 144 chains were evaluated. Results are shown in Table 2.

Table 2: Figures of precision of hypernymy chains

Correct		**Incorrect**		**Unfinished analyses**	
n	**%**	**n**	**%**	**n**	**%**
91	63.20	30	20.83	23	15.97
Precision: 0.75 / Recall: 0.63					

Unfinished analyses are cases in which the chain production was interrupted for not reaching a semantic type after three recursive iterations, thus aborting the attempt. Errors in this task can be related to wrong detection of the hypernym in Wiktionary, but are mainly due to wrong sense disambiguation of the hypernym, which lead to a wrong connection of the noun with the ontology. There is of course ample room for improvement here that we leave for future work. There has been research on the use of supervised models for this task (e.g., Lopes et al., 2023; Tan et al., 2020), although the problem with such method arises when dealing with words not seen during the training phase.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Regarding WSD, the task of selecting the right sense of a polysemous noun in a sentence, we evaluated these models on a random sample of 100 sentences for 5 polysemous nouns (500 sentences in total). Table 3 shows the comparative figures of the performance of the different settings, including a random selection for general reference.

Table 3: Figures of precision of Spanish WSD results

Word	relex	fill-cooc	fill-cosine	target-cosine	random
bodega 'cellar, winery'	**0.75**	0.33	0.50	0.59	0.67
ratón 'mouse'	0.80	0.55	0.66	**0.83**	0.36
operación 'operation'	**0.47**	0.23	0.19	0.25	0.21
virus 'virus'	0.66	0.71	0.70	**0.76**	0.53
medicina 'medicine, drug'	0.81	**0.86**	0.57	0.44	0.53
Mean	**0.70**	0.54	0.52	0.57	0.46

As observed in the table, Relex performs better than the other systems. Two observations can be made: 1) with values concentrated in the range .66 and .81 and with the lowest standard deviation (0.14), Relex seems relatively more stable in its results compared to the LLM-based methods. The others, with wider ranges (standard deviation greater than 0.2), seem more unpredictable; 2) showing better performance, Relex is also more parsimonious, being computationally simple and fast. With regards to error analysis, we observe that in most cases the root of the problem was the lack of sufficient data to make a correct prediction, i.e., many sentences do not contain enough lexical units to disambiguate, making it difficult even for humans to interpret the correct meaning. Consider, for instance, the case of *bodega*, which can mean 'cellar', 'warehouse' or 'winery', in a sentence like *Esta bodega contará con detectores de humo* 'This warehouse will be fitted with smoke detectors'. A possibility to solve this problem in future work should be to consider larger context windows. Other problems are related with the metaphoric use of the words or with senses that are not attested in the dictionary.

4.4 Pattern Building Analysis

As a first evaluation of the quality of the produced patterns (see Section 3.4), we took a random sample of 5 verbs from the Verbario database, which, as explained in Section 2.2, contains manual analyses of corpus concordances necessary to extract the patterns. Table 4 shows the number of manually detected patterns per verb, the number of automatically detected patterns (considering the most frequent up to 7), the number of correct automatic patterns and the match between manual and automatic patterns.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Table 4: Evaluation of the patterns in a random sample of 5 verbs

Verb	Manual patterns	Automatic patterns	Correct	Match with manual
amar	2	3	3	1
atrapar	8	5	3	3
beber	8	7	6	1
generar	3	7	5	1
iluminar	5	2	2	2

A first impression upon examining these results is that, overall, the performance is quite acceptable, considering that it is a first attempt with this method and there are still many minor details to address. Part of the mismatch between the two sets is explained by the fact that the automatic patterns are only of the transitive type, while the manual annotation of the database is unrestricted and thus contains many more types of patterns, such as intransitive, ditransitive, pronominal and so on.

5. Conclusions and Future Work

In this paper, we presented a method for extracting CPA verb patterns from corpus, using a series of strategies to label sentences with syntactic and semantic information, in order to obtain the argument structure of the verb and the semantic types of each argument. The method shows a promising approach for a more effective pattern-based lexicography, as it provides the lexicographer with a set of patterns per verb automatically extracted, with acceptable precision. Pattern induction goes a step further from collocational analysis, which does not cover all the argument structure. We believe the method can be adapted to other languages with a reasonable volume of work.

At this stage, limitations of the method lay especially in the part of pattern building. We must deal with the splitting / lumping problem, trying to observe which level of the ontology is the most common for semantic typing, among other tasks to be addressed in the near future. More extensive evaluation is also necessary and, finally, we should test the system with a team of lexicographers using these results in their job.

References

Agirre, E., de Lacalle, O. L., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics, 40*(1), 57–84.

Akbik, A., Bergmann, T., & Vollgraf, R. (2019b). Multilingual sequence labeling with one model. In *NLDL 2019*, Northern Lights Deep Learning Workshop.

Akbik, A., Bergmann, T. Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019a). FLAIR: An Easy-to-use framework for state-of-the-art NLP. In J. Burstein, Ch. Doran, & Th. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (pp. 54–59). Association for Computational Linguistics.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Amsler, R. (1981). A taxonomy for English nouns and verbs. In *Proceedings of 19th Annual Meeting on ACL (Morristown, NJ, USA),* 133–138.

Baisa, V., Bradbury, J., Cinková, S., El Maarouf, I., Kilgarriff, A., & Popescu, O. (2015). SemEval-2015 Task 15: A CPA dictionary-entry-building task. In P. Nakov, T. Zesch, D. Cer, & D. Jurgens (Eds.), *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 315–324). Association for Computational Linguistics.

Battaner, P., & Renau, I. (2011). El proyecto DAELE verbos: un diccionario de aprendizaje de español como lengua extranjera. *II Encuentros (ELE) Comillas: el profesor de ELE: metodología, técnicas y recursos para el aula.*

Bevilacqua, M., Pasini, T., Raganato, A., & Navigli, R. (2021). Recent trends in word sense disambiguation: A survey. In Z.-H. Zhou (Ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)* (pp. 4330–4338). International Joint Conferences on Artificial Intelligence.

Chodorow, M., Byrd, R., & Heidorn, G. (1985). Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on ACL*, 299–304.

Colman, L., & Tiberius, C. (2018). A good match: A Dutch collocation, idiom and pattern dictionary combined. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 233–246). Ljubljana University Press.

CROATPAS: https://croatpas.baisa.cz/ [Accessed: 25/7/24]

DiMuccio-Failla, P. V., & Giacomini, L. (2017). Designing a learner's dictionary based on Sinclair's lexical units by means of Corpus Pattern Analysis and the Sketch Engine. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 Conference* (pp. 437–457). Lexical Computing.

DiMuccio-Failla, P. V., & Giacomini, L. (2022). A proposed microstructure for a new kind of active learner's dictionary. *Lexicographica*, *38*(1), 475–499.

Giacomini, L., & DiMuccio-Failla, P. (2019). Investigating semi-automatic procedures in pattern-based lexicography. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Janssen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century: Smart lexicography. Proceedings of the eLex 2019 Conference* (pp. 490–505). Lexical Computing.

Gutiérrez Fandiño, A., Armengol Estapé, J., Pàmies Massip, M., Llop Palao, J., Silveira Ocampo, J., Carrino, C. P., Armentano Oller, C., Rodríguez Penagos, C., González Agirre, A., & Villegas, M. (2022). MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural,* (68), 39–60.

Grefenstette, G., & Hanks, P. (2023). Competing views of word meaning: word embeddings and word senses. *International Journal of Lexicography*, *36*(2), 211–219.

Guthrie, L., Slator, B., Wilks, Y., & Bruce, R. (1990). Is there content in empty heads? In H. Karlgren (Ed.), *Proceedings of the 13th International Conference on Computational Linguistics, COLING'90* (pp. 138–143). Association for Computational Linguistics.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Hanks, P. (2004a). Corpus Pattern Analysis. In G. Williams, & S. Vessier (Eds.), *Proceedings of the 11th Euralex International Congress* (Vol. 1, pp. 87–97). Université de Bretagne-Sud.

Hanks, P. (2004b). The syntagmatics of metaphor and idiom. *International Journal of Lexicography, 17*(3), 245–274.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations.* The MIT Press.

Hanks, P., & Ježek, E. (2008). Shimmering lexical sets. In E. Bernal, & J. DeCesaris (Eds.), *Proceedings of the 13th EURALEX International Congress* (pp. 391–402). Universitat Pompeu Fabra.

Hanks, P., & Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée, 10*(2), 63–82.

Huang, L., Sun, C., Qiu, X., & Huang, X. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)* (pp. 3509–3514). Association for Computational Linguistics.

Ježek, E., & Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis. Journal in English Lexicology, 4*, 7–22.

Ježek, E., Magnini, B., Feltracco, A., Bianchini, A., & Popescu, O. (2014). T-PAS; A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In N. Calzolari et al. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 890–895). ELRA.

Kilgarriff, A., & Renau, I. (2013). EsTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences, 95*, 12–19.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*, 7–36.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In V. DeBuys (Ed.), *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC'86)* (pp. 24–26). Association for Computing Machinery.

Lopes, A., Carbonera, J., Schmidt, D., Garcia, L., Rodrigues, F., & M. Abel (2023). Using terms and informal definitions to classify domain entities into top-level ontology concepts: An approach based on language models. *Knowledge-Based Systems, 265*, 110385.

Marini, C. (2022). CroaTPAS: A survey-based evaluation. In H. Bunt (Ed.), *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022* (pp. 76–80). ELRA.

Marini, C., & Ježek, E. (2019). CROATPAS: A resource of corpus-derived typed predicate argument structures for Croatian. In R. Bernardi, R. Navigli, & G. Semeraro (Eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics.*

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Nazar, R., & Renau, I. (2016). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. In N. Calzolari et al. (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1485–1492). ELRA.

PDEV: https://pdev.org.uk/ [Accessed: 25/7/2024]

Pustejovsky, J., Hanks, P., & Rumshisky, A. (2004). Automated induction of sense in context. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 924–930.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. (2020). Stanza: A Python natural language processing toolkit for many human languages. In A. Celikyilmaz, & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). ACL.

Renau, I. (2012). *Gramática y diccionario: las construcciones con se en las entradas verbales del diccionario de español como lengua extranjera* [Doctoral dissertation, Universitat Pompeu Fabra].

Renau, I., Nazar, R., Castro, A., López, B., & Obreque, J. (2019). Verbo y contexto de uso: un análisis basado en corpus con métodos cualitativos y cuantitativos. *Revista Signos*, *52*(101), 878–901.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation.* Oxford University Press.

Sinclair, J. M. (2004). *Trust the Text, Language, Corpus and Discourse.* Routledge.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In D. Zeman, & J. Hajič (Eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 197–207). ACL.

Tan, Y., Wang, X., & Jia, T. (2020). From syntactic structure to semantic relationship: hypernym extraction from definitions by recurrent neural networks using the part of speech information. In J. Z. Pan et al. (Eds.), *The Semantic Web – ISWC 2020. 19th International Semantic Web Conference* (pp. 529–546). Springer.

Tiedemann, J. (2009). News from OPUS — A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing:* Vol. V (pp. 237–248). John Benjamins.

T-PAS: https://tpas.fbk.eu/ [Accessed: 25/7/24]

Verbario: http://www.verbario.com [Accessed: 25/7/24]

Wolf, T. et al. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). ACL.

Wiktionary: https://en.wiktionary.org/wiki/Spanish [Accessed: 25/7/24]

Woordcombinaties: https://woordcombinaties.ivdnt.org/ [Accessed: 25/7/24]

Acknowledgements

We would like to express our gratitude to the evaluators for their valuable comments. This research received funds from project Fondecyt Regular nr 1231594 and ESMAS-ES+ (PID2022-137170OB-I00) funded by MCIN/AEI//FEDER "Una manera de hacer Europa". We want to thank Patrick Hanks for all these wonderful years of learning, inspiration and guidance. You will always be in our hearts!

Contact information

Irene Renau
Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso
irene.renau@pucv.cl

Rogelio Nazar
Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Daniel Mora
Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso
daniel.mm91@gmail.com

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.