Spanish HPSG Treebank based on the AnCora Corpus

Luis Chiruzzo, Dina Wonsever
Universidad de la República
Montevideo, Uruguay
{luischir, wonsever}@fing.edu.uy

Abstract

This paper describes a corpus of HPSG annotated trees for Spanish that contains morphosyntactic information, annotations for semantic roles, clitic pronouns and relative clauses. The corpus is based on the Spanish AnCora corpus, which contains trees for 17,000 sentences comprising half a million words, and it has CFG style annotations. The corpus is stored in two different formats: An XML dialect that is the direct serialization of the typed feature structure trees, and an HTML format that is suitable for visualizing the trees in a browser.

Keywords: HSPG, Spanish, AnCora, corpus

1. Introduction

We present the construction of a Spanish HPSG treebank based on the Spanish corpus AnCora (Taulé et al., 2008). This is part of an ongoing project to build a statistical Spanish HPSG parser.

An earlier version of this corpus is described in (Chiruzzo and Wonsever, 2016). In this previous work we used a series of hand-crafted rules to process all constituents in AnCora in order to find the heads and binarize all the phrases, identifying when the constituents acted as specifier, complement or modifier. In this first step, clitics and relative clauses had been identified but not properly handled.

In this work we finished the transformation of the corpus by incorporating an analysis for clitics and relative clauses into the feature structure and by annotating those constructions in the corpus, as well as producing a way of visualizing the HPSG trees.

2. Background

HPSG grammars are rich grammars (Pollard and Sag, 1994) that are able to represent both syntactic and semantic information in the same parse tree. The nodes in the trees and the rules to combine them are defined as typed feature structures (Carpenter, 1992) with a unification operation. The leaves of the parse tree are the words of a sentence, and the HPSG feature structure generally includes morphosyntactic information such as part of speech, agreement features, and syntactic valence features.

The English Resource Grammar (Flickinger, 1999) is an implementation of the HPSG principles for English, built into the LinGO Grammar Matrix (Bender et al., 2002), a framework for building unification grammars. There also exists a Spanish HPSG grammar built over the same principles: the Spanish Resource Grammar (SRG) (Marimon, 2010a). Both grammars are hand crafted paying particular attention to the linguistic details of the theory and the correctness of the modeled sentences. This means the trees obtained using these grammars are very rich, but on the other hand sentences that are not perfectly written (as is the case for general text extracted from the web) would not be even partially parsed. In this work we have the final aim of building a statistical parser from scratch based on the statistics of a properly annotated corpus, so we aim to improve the robustness of the parsing process for sentences that not necessarily are well written.

We base our approach in the work done for the parser Enju (Matsuzaki et al., 2007), which is a statistical HPSG parser for English created through the conversion of the Penn Treebank corpus (Marcus et al., 1993) into a HPSG suitable format (Miyao et al., 2005). The result is a fast high coverage parser for English that returns syntactic trees as well as argument structure information. We tried to follow a similar approach by transforming the Spanish corpus AnCora (Taulé et al., 2008) from its original context free like annotations to an HPSG compatible format.

The AnCora corpus contains about half a million words of news text in Spanish (there is also a version in Catalan of the same size), annotated in a context free grammar style enriched with attributes such as the grammatical function of constituents and the predicate-argument structure annotated in PropBank style (Kingsbury and Palmer, 2002).

There exists another HPSG corpus for Spanish based on a subset of AnCora that uses the Spanish Resource Grammar called the Tibidabo Treebank (Marimon, 2010b). In this corpus only sentences up to 40 words were considered, and the original AnCora annotations were not used. Instead, the sentences were parsed using the SRG and the appropriate analysis was selected from the set of resulting trees. A related corpus called the IULA Spanish LSP Treebank (Marimon et al., 2012) was originally annotated using a HPSG scheme, but the available version of the corpus contains dependency trees annotations instead of HPSG. This corpus contains around 40,000 sentences, generally shorter than the sentences in AnCora: 80% of the sentences have 20 words or less. In our case, we aimed to leverage the existing annotations of the AnCora corpus and try to use all its information to build our corpus, adding some missing information when necessary.

3. Description of the grammar

Our grammar is largely based on the one described in (Sag et al., 1999) for English, though some adaptations had to be done in order to port it to Spanish, and also because the information that could be extracted from AnCora sometimes was not enough to complete all aspects of the theory. We include morphological and syntactic information into our grammar, but the main simplification is our treatment of semantics: In our version of the grammar, we include features for representing the argument structure of the verbs (annotated as PropBank style semantic roles) as semantic features. The original grammar uses a more complex approach to semantics based on Minimal Recursion Semantics (Copestake et al., 2005), but our approach is easier to extract from the information that is readily available in the AnCora corpus. Also, this approach could serve as a base for transforming the semantic representation into some other formats like Abstract Meaning Representation (Banarescu et al., 2012) which, in spite of not being the semantic perspective traditionally developed in HPSG, it still offers interesting insights for some semantic tasks (for example: text entailment and paraphrasing).

Figure 1 shows the lexical entry for a generic word, indicating all the possible features. A particular word might instantiate only the features that it defines, e.g. verbs in Spanish do not have a defined value for gender, so it would not be shown in the lexical entry (see section 3.3. for an example).

$$\begin{bmatrix} \text{word} & \\ \text{SYN} & \begin{bmatrix} \text{LOCAL} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} & \text{n} \mid \text{v} \mid \text{a}\ldots \\ \text{AGR} & \begin{bmatrix} \text{PER } 1 \mid 2 \mid 3 \\ \text{NUM sg} \mid \text{pl} \\ \text{GEN f} \mid \text{m} \\ \text{MOOD}\ldots \\ \text{TENSE}\ldots \end{bmatrix} \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SPEC} & \langle\text{expr}\rangle \\ \text{COMP} & \langle\text{expr}\rangle \\ \text{CLITIC} & \langle\text{expr}\rangle \\ \text{MOD} & \langle\text{expr}\rangle \end{bmatrix} \end{bmatrix} \\ \text{NONLOCAL} & \begin{bmatrix}\text{REL} & \text{expr}\end{bmatrix} \end{bmatrix} \\ \text{SEM} & \begin{bmatrix} \text{ARGS} & \begin{bmatrix} \text{ARG0} & \langle\text{expr}\rangle \\ \text{ARG1} & \langle\text{expr}\rangle \\ \ldots & \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 1: Feature structure for a lexical entry

3.1. Grammar rules

Generally HPSG grammars have very few rules and most of the combinatorial constraints are encoded in the features associated to each word or lexical entry. That is why HPSG grammars are generally said to be highly lexicalized grammars. Every rule must clearly mark which of its daughters is the head of the phrase, as the phrase will inherit the value for many of the head features. In our case, the grammar contains only thirteen rules:

- Two rules for applying a complement to the left or to the right of a head: `head_comp` and `comp_head`
- Two rules for applying a modifier to the left or to the right of a head: `head_mod` and `mod_head`
- Two rules for applying a specifier to the left or to the right of a head: `head_spec` and `spec_head`
- One special unary rule for representing the Spanish null subjects, which could be seen as a special case of an empty lexical entry: `empty_spr`
- Two rules for binarizing chains of coordinations: `coord_left` and `coord_right`
- One rule for applying a clitic to the left of a head: `clitic_head`
- One rule for joining a noun with a relative clause that modifies it: `head_rel`
- Two rules for applying a punctuation symbol to the left or to the right of a head: `head_punct` and `punct_head`

For example, consider the schematic definition of the specifier rule `spec_head` as shown in figure 2, that applies a specifier to the left of a head. This rule is used to analyze both a determinant as specifier of a noun phrase and a noun phrase as specifier of a sentence. On the right hand side there is an expression (the specifier) followed by another expression (the head) which requires a specifier. On the left hand side, the result is a phrase whose feature `HEAD` (which carries part of speech and agreement information) is coindexed with the same feature of the head expression. The `SPEC` feature of the head is coindexed with the specifier, but the value of that feature is removed from the resulting phrase on the left hand side.

$$\begin{bmatrix} \text{phrase} & \\ \text{SYN LOCAL} & \begin{bmatrix} \text{HEAD}\ \boxed{1} \\ \text{VAL} \begin{bmatrix}\text{SPEC}\ \langle\rangle\end{bmatrix} \end{bmatrix} \end{bmatrix} \rightarrow \boxed{2}\ \text{expr} \quad \begin{bmatrix} \text{expr} & \\ \text{SYN LOCAL} & \begin{bmatrix} \text{HEAD}\ \boxed{1} \\ \text{VAL} \begin{bmatrix}\text{SPEC}\ \langle\boxed{2}\rangle\end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 2: Schematic definition of the `spec_head` rule

Notice that the agreement principle defined in HPSG (Sag et al., 1999) is expressed in this rule by coindexing the `HEAD` features of the parent and the head daughter, and consequently the `AGR` features of both structures will also be coindexed.

Generally the head defines further constraints for the specifier. These constraints are not defined in the rule, but in the lexical entry corresponding to the head (see for example figure 3).

There is also another rule in our grammar that allows to combine a specifier with a head, the `head_spec` rule that takes a specifier on the right. Although Spanish is generally regarded as a SVO language, there exist many cases where it is more common to find the subject located after the verb. One example of this is the sentence *"llegó el tren"* / "the train arrived". We decided to create two different rules for allowing the subject to be applied both to the left and to the right of the verb, as well as two rules for applying the complement to the left or to the right. The existence of these rules allows many possible analysis for the sentences, so it is important that a parser takes into account the probability of applying the rules in each context.

3.2. General transformation of the corpus

The initial transformation of the AnCora corpus into a HPSG compatible format is described in (Chiruzzo and Wonsever, 2016). As AnCora contains rather flat context free grammar structures and a lot of variability between rules, we used a top down process to break all complex structures and separate them into simple units called *elementary trees* (head surrounded by arguments and modifiers). Then a bottom up process used a series of hand written rules to decide which of the elements inside a phrase was the head, and what rules to select in order to binarize the rest of the phrase. Special care had to be taken for the transformation of verb phrases, which include auxiliary and modal verb constructions, because their analysis in AnCora was different than the rest of the phrases (see section 3.5.). The process achieved an average 95.3 % precision for head detection (98.7 % without considerind coordinations) and 92.5 % average precision for argument detection.

However, this work left out the analysis of some interesting phenomena. Particularly the presence of clitics in the corpus was marked but not properly analyzed, as well as the use of relative clauses as modifiers of noun phrases. In this work we provide an appropriate analysis for these structures.

3.3. Modeling clitics in Spanish

Clitics are pronouns that can occur in different positions in verb phrases, for instance before a verb, and they could either take the place of an argument or coexist together with the argument in the phrase. Consider the sentence *"Juan le dará un regalo a María"* / "John will give a present to Mary". The lexical entry for the verb *"dará"* / "will give" as used in this sentence is shown in figure 3. Notice that in this case the clitic (*"le"*) corresponds to the indirect object (*"a María"*) which is also present in the sentence. This phenomenon is very common in Spanish and it is known as *clitic doubling*, presenting additional modeling complexity (Pineda and Meza, 2005). In our feature structure this is represented by setting both values as ARG2 in the argument structure, which corresponds to the beneficiary semantic role. If either the clitic or the explicit argument are present, then the semantic argument will point to that expression, if both are present then the list associated to the semantic argument will contain both expressions.

3.4. Relative clauses as modifiers

In this work we focus on one kind of long distance dependency that is very common in Spanish: the use of a relative clause as a modifier of a noun, where the noun at the same time is acting as an argument of the verb in the clause. Generally the noun acts as the subject (e.g. *"el perro que me mordió"* / "the dog that bit me") or direct object (e.g. *"el libro que leí"* / "the book I read"), but it could act as any argument. Unlike English, in Spanish it is mandatory that these clauses are introduced by a subordinating relative expression that always contains a relative pronoun, such as *"que"* / "that" or *"a quien"* / "to whom".

Consider for example the sentence *"los cultivos que contienen almidón"* / "the crops that contain starch". The analysis according to our grammar is shown in figure 4. The verb of the relative clause is transitive, but its corresponding subject (*"cultivos"*) is not readily available. Instead, the relative pronoun *"que"* takes the place of the subject, but keeps a non-local feature REL that points to the noun it stands for. The rule head_rel is used to unify a non-saturated NONLOCAL.REL feature to the appropriate expression it should be bound to, the resulting phrase cancels the value of the NONLOCAL.REL feature.

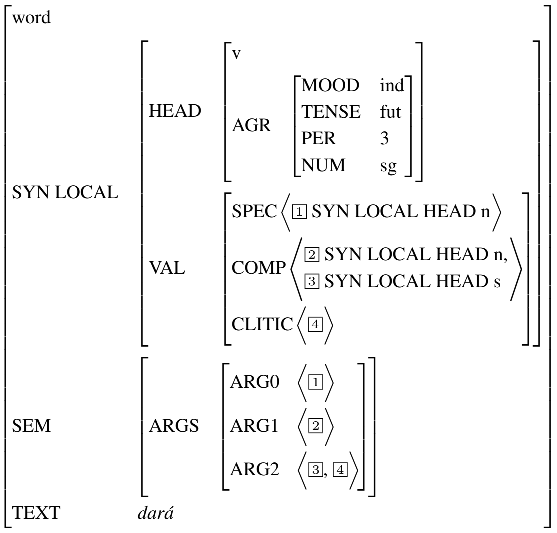

Figure 3: Feature structure for ditransitive verb *"dará"*, future indicative third person singular form of the verb *"to give"*

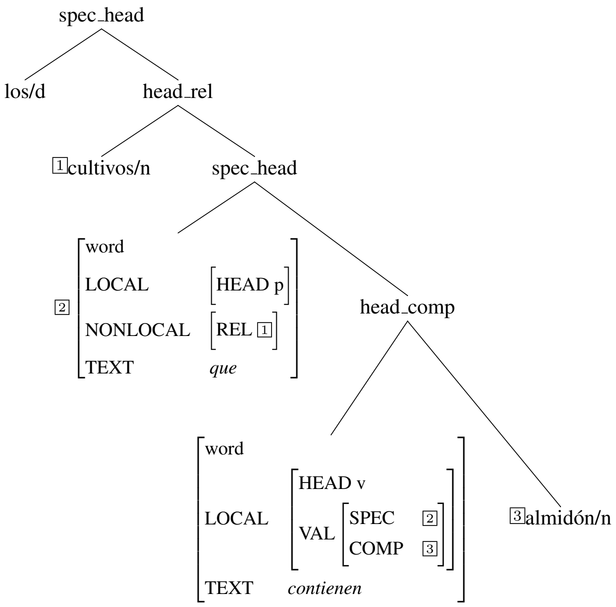

Figure 4: Simplified analysis for *"los cultivos que contienen almidón"* / "the crops that contain starch"

3.5. Verb phrases

Consider a sentence like *"ellos pueden hacer pasta"* / "they can make pasta", it contains the verbal periphrasis *"pueden hacer"* / "can make", plus a subject and a complement. A standard HPSG analysis for this phrase would first apply

the complement *"pasta"* to the verb *"hacer"*, then this verb would be applied as a complement to the verb *"pueden"* and finally the subject *"ellos"* would be applied to the resulting head. However, constructions of this type are analyzed differently in AnCora: the verbal periphrasis is considered a unit, so *"pueden hacer"* becomes a phrase that should expect a complement and a subject. We decided to keep this behavior from AnCora because it simplifies the analysis of displaced constituents.
The result of applying the complement *"hacer"* to the modal verb *"pueden"* is shown in figure 5. It is a phrase that combines the features of both daughters:

- It expects a noun phrase as specifier, which is coindexed with the specifier of both verbs.
- It expects the complement required by *hacer*.
- The agreement features are copied from the modal verb *puede*, because the subject must agree with the syntactic head.

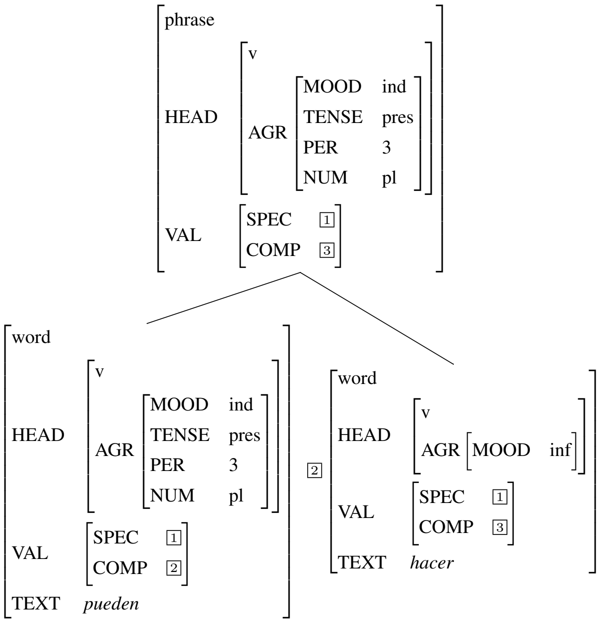

Figure 5: Simplified analysis for verb phrase *"pueden hacer"* / "can make"

In general verb phrases involving modal verbs or auxiliary verbs (e.g. *"haber visto"* / "have seen") in AnCora are analyzed like units and they are transformed in the same way in our corpus. Notice that in these cases the valence principle of HPSG is not applied in the same way, as the COMP feature of the non-head daughter (instead of the one from the head) is percolated to the mother. This is a variant of the standard HPSG complement rule that applies to a class of verbs such as modals and auxiliaries. One way of interpreting these structures is to consider that in these cases the syntactic head and the semantic head of the structure are different, and the main semantic content is provided by the semantic head. This can be extended to structures with more than two verbs, for example *"pueden querer traer"* / "might want to bring". In this structure we could consider that the main semantic content is in the verb *"traer"* / "to bring", so the arguments expected by this verb are percolated, through the successive application of the rules, to the base of the structure.

4. Composition of the corpus

The corpus has approximately half a million words in 17,000 sentences. The number of words for each part of speech is shown in table 1, while the number of times each rule is applied is shown in table 2.

POS	Instances	Unique
noun (n)	121089	22339
verb (v)	61688	11611
adjective (a)	35936	8228
adverb (r)	18951	1047
determinant (d)	76125	164
pronoun (p)	22690	183
preposition (s)	79897	342
interjection (i)	99	47
conjunction (c)	27062	125
date (w)	2731	986
number (z)	5362	3326
punctuation (f)	65538	28
Total	517168	48426

Table 1: Number of unique words and instances by part of speech

Rule	Instances
spec_head	111192
head_spec	6477
empty_spr	10776
head_comp	153473
comp_head	5777
head_mod	91913
mod_head	27873
head_punct	38709
punct_head	22848
coord_left	18226
coord_right	18226
clitic_head	8070
head_rel	7782

Table 2: Number of times each rule is applied in the corpus

The corpus is stored in two different formats[1]: XML and HTML. The XML format is a direct serialization of the typed feature structure trees into an XML dialect, we call this the TFSML format. Figure 6 shows an example of a lexical entry from the corpus in TFSML format. The HTML format is suitable for visualizing the trees in a browser. Currently we are in the process of converting the sentences into the format used by LKB (Copestake et al., 1999), a widely used grammar development environment.

5. Conclusion

We built a Spanish HPSG annotated corpus based on the AnCora Spanish corpus, containing the analysis of 17,000 sentences and half a million words. The parse trees contain syntactic and morphological information, semantic role

[1] A sample of the corpus can be found at: www.fing.edu.uy/inco/grupos/pln/recursos/spanish_hpsg/index.html

```
<tfs id="v22119" text="analizado" type="word">
  <feature name="SYN">
    <feature name="LOCAL">
      <feature name="HEAD">
        <tfs type="v">
          <feature name="AGR">
            <feature name="GEN">
              <tfs type="m"/>
            </feature>
            <feature name="NUM">
              <tfs type="s"/>
            </feature>
            <feature name="MOOD">
              <tfs type="par"/>
            </feature>
          </feature>
        </tfs>
      </feature>
      <feature name="VAL">
        <feature name="SPEC" pointer="p8085"/>
        <feature name="COMP" pointer="sn54284"/>
      </feature>
    </feature>
  </feature>
  <feature name="SEM">
    <feature name="ARG0" pointer="p8085"/>
    <feature name="ARG1" pointer="sn54284"/>
  </feature>
</tfs>
```

Figure 6: Representation of the past participle *"analizado"* / "analyzed" in TFSML format.

labels annotated in PropBank style both for explicit arguments and clitic pronouns, and the analysis of relative clauses that act as modifiers.

This is part of an ongoing project for building a statistical HPSG parser for Spanish. Some work has already been done in this direction, for example we carried some baseline parsing experiments using the data of this corpus, and we also trained a supertagger over an earlier version of the corpus that is able to classify complex verbs, nouns and adjectives using the categories of our corpus (Chiruzzo and Wonsever, 2015). The next step in this process will be to train a full deep parser for Spanish using this corpus.

6. Bibliographical References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2012). Abstract meaning representation (amr) 1.0 specification. In *Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL*, pages 1533–1544.

Bender, E. M., Flickinger, D., and Oepen, S. (2002). The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the 2002 workshop on Grammar engineering and evaluation-Volume 15*, pages 1–7. Association for Computational Linguistics.

Carpenter, B. (1992). *The logic of typed feature structures*. Cambridge University Press.

Chiruzzo, L. and Wonsever, D. (2015). Supertagging for a statistical hpsg parser for spanish. In *International Conference on Statistical Language and Speech Processing*, pages 18–26. Springer.

Chiruzzo, L. and Wonsever, D. (2016). Transforming the ancora corpus to hpsg. In *HeadLex16, Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*.

Copestake, A., Carroll, J., Malouf, R., and Oepen, S. (1999). The (new) lkb system. *Center for the Study of Language and Information, Stanford University*.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332.

Miyao, Y., Ninomiya, T., and Tsujii, J. (2005). Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the penn treebank. In *Natural Language Processing–IJCNLP 2004*, pages 684–693. Springer.

Pineda, L. and Meza, I. (2005). The spanish pronominal clitic system. *Procesamiento del lenguaje natural*, 34:67–103.

Pollard, C. and Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Sag, I. A., Wasow, T., Bender, E. M., and Sag, I. A. (1999). *Syntactic theory: A formal introduction*, volume 92. Center for the Study of Language and Information Stanford, CA.

7. Language Resource References

Flickinger, D. (1999). The english resource grammar.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *LREC*. Citeseer.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Marimon, M., Fisas, B., Bel, N., Vivaldi, J., Torner, S., Lorente, M., Vázquez, S., and Villegas, M. (2012). The iula treebank. In *Lrec*, pages 1920–1926.

Marimon, M. (2010a). The spanish resource grammar. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*, pages 17–23, Valletta, Malta.

Marimon, M. (2010b). The tibidabo treebank. *Procesamiento del lenguaje natural, 2010, vol. 45, num. 1, p. 113-119*.

Matsuzaki, T., Miyao, Y., and Tsujii, J. (2007). Efficient hpsg parsing with supertagging and cfg-filtering. In *IJCAI*, pages 1671–1676.

Taulé, M., Martí, M. A., and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC*.