Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

A Constraint Grammar Parser for Spanish

Eckhard Bick

Institute of Language and Communication, University of Southern Denmark
eckhard.bick@mail.dk

Abstract. In this paper we describe and evaluate a Constraint Grammar parser for Spanish, HISPAL. The parser adopts the modular architecture of the Portuguese PALAVRAS parser, and in a novel porting approach, the linguist-written Portuguese CG rules for morphological and syntactic disambiguation were "corrected" and appended for Spanish in a corpus-based fashion, rather than rewritten from scratch. As part of the 5 year project, a 74.000 lexeme lexicon was developed, as well as a morphological analyzer and semantic ontology for Spanish. An evaluation of the the system's tagger/parser modules indicated F-scores of 99% for part-of-speech tagging and 96% for syntactic function assignment. HISPAL has been used for the grammatical annotation of 52 million words of text, including the Europarl and Wikipedia text collections.

Keywords: Spanish parser, Constraint Grammar, NLP, Corpus annotation

1 Introduction

As results for several languages have shown (e.g. English [7], Portuguese [3], French [5], Danish [4] and Norwegian [6]), Constraint Grammar (CG) based NLP systems can achieve a high level of robustness and accuracy in the annotation of running text. However, as rule-based systems, they normally demand not only a full lexicon-based morphological analysis as input, but also a large linguist-written disambiguation grammar. In an effor to reduce development costs, various hybridization techniques have been proposed for the integration of CG and probabilistic systems, such as machine learning of rule templates and rule ordering (µTBL, [9]), correction CGs for probabilistic taggers [5] and relaxation labelling [10]. In this paper, we describe and evaluate a non-hybrid CG parser for Spanish (HISPAL), suggesting bootstrapping solutions for both the lexicon and grammar tasks as an alternvative to hybridization. Work on HISPAL was carried out over a 5-year period within the VISL framework of natural language processing (http://beta.visl.sdu.dk), where the system is used for

teaching and corpus annotation. Thus, morphological and syntactic annotation was carried out for the Spanish parts of the ECI, Europarl and Wikipedia text collections, in all 52 million words.

2 The morphological analyzer

HISPAL's morphological analyzer is a *multitagger* assigning multiple possible readings to tokenized input. The analyzer uses a full form lexicon only for about 220 closed class tokens, while everything else is treated analytically through affix classes with or without stem conditions. The ending *-aremos,* for instance, will yield the tag string *V FUT 1P IND (verb, future, first person plural, indicative)* and a lemma *'....ar',* with its stem taken directly from the token. However, a token like *compraremos,* in the example cohort below, will also allow present tense readings, for other - hypothetical - verb stems (d-e), as well as adjective and noun male plural readings (a-b).

compraremos

(a) [compraremo] ADJ M P
(b) [compraremo] N M P
(c) [comprar] V FUT 1P IND
(d) [comprarer] V PR/PS 1P IND
(e) [comprarar] V PR 1P SUBJ

Which of these alternatives are real Spanish words, will be decided by lexicon look-up. The lexicon-supported forms will then be further disambiguated by context-sensitive CG-rules. If none of the forms matches a lexicon entry, all will be submitted to contextual disambiguation, de facto turning the Constraint Grammar module into a heuristic subsidiary of the analyzer - with the added potential of being able to iteratively increase the lexicon through corpus work (jf. lexicon chapter).

The analyzer distinguishes between inflexiona affixes and derivational affixes (suffixes and prefixes). Inflexion affix classes are ordered in a way that will block more general affixes (e.g *-s* for noun plural) in the presence of more specific ones (e.g. *-anes* -> *-án* or *-enes* -> *-én).* At the same time, orthographical alternations like accents and e-insertion will be handled in order to arrive at correct (read: realistic) base forms for the lexemes involved. The largest section concerns verbal inflection, where - even with stem grouping for irregular verbs - roughly 1000 affix rules are needed to cover all cases[1]. Since all rules allow open prefixing, entries for

1 This is still, for a highly inflecting language like Spanish, much less than the 1 million entries or so that would have been necessary in the case of a full form data base, while at the same

decir will at the same time cover rarer words like *antedecir, bendecir, contradecir, desdecir, entredecir, interdecir, maldecir, predecir, redecir,* and allow for productive derivation.

Derivational affixes are checked at the lexicon look-up stage, after removal of inflexional affixes. If a derivation affix, a suffix like *-idad, -itud, -ista* or a prefix like *super-* etc. is recognized in an otherwise unknown base form, and if the remaining root (i.e. without the affix), matches a lexicon entry, the reading in question will be preferred over other, non-analytical (i.e. more heuristic) readings. Even if a root can't be matched, the word class (PoS) implied by a recognized suffix can be used to prefer the involved lexeme over readings with other word classes. For instance, in the presence of nominal suffixes, an *-s* inflexion affix will be interpreted as nominal rather than verbal.

3 The lexicon

The initial version of the HISPAL lexicon was created in 2001 with a bootstrapping method, using the following for seeding:

(1) A hand-built closed class lexicon for Spanish (pronouns, prepositions, conjunctions ...)
(2) a Spanish affix file used together with the Portuguese morpho-chunker and dummy-roots
(3) a list of safe open class word candidates, extracted from corpora using e.g. article-noun sequences and unambiguous verbal inflexions
(4) overgenerating, heuristic output from the Spanish morphological analyzer, using dummy-roots to recognize open class word candidates and their inflexion (nouns, verbs, adjectives ...)

The combined seeding system was then run on a large body of texts, covering both European and Latin-American Spanish, adding morphological and PoS tags as well as lemma cohorts for each word. Of course, for open class words from (4), not contained in (3), a lot of ambiguity would be created by hypothesizing non-existing Spanish words - for instance, a word ending in -a (say, 'xxxa'), without a recognizable affix, would trigger 3 hypothetical lemmata - 'xxxa' (female noun), 'xxxo' (female adjective) and 'xxxar' (inflected verb with several morphological readings). However, such ambiguity was then reduced by running the Portuguese CG disambiguation module to handle contextual disambiguation, removing spurious hypothetical readings. Finally, the surviving readings were used to generate new entries for the HISPAL lexicon.

---

time guaranteeing that also rare words and new productions will be covered.

After several iterations and semi-automatic consistency checking, the lexicon was large enough to support and restrict output from the morphological analyzer. Throughout the project, improvements were made with new data and better, more "Spanish" versions of the CG rule sets. A large portion of lexeme strings with more than one PoS entry in the lexicon were manually checked against published dictionaries, among them all cases of gender ambiguity for nouns (e.g. *o guarda - a guarda).* Today the lexicon contains 74.000 entries.

Table 1. Lemma distribution in the HISPAL lexicon.

	Types (m = male, f = female gender)
Nouns	m: 26.405, f: 17.280, m/f: 1.298 m, flagged for checking: 1.295
Adjectives	m: 10.068, m/f: 3637
Verbs	-ar: 8135, -er: 562, -ir: 539
Adverbs	1.302
Names	1.186

Valency. Once the parser produced more reliable output, annotated corpora were used to extract verb valency frames, such as <vt> 'transitive verb' og <vr> 'reflexive verb'. For auxiliaries and some central construction verbs *(ser, estar, dejar),* valency potential was manually added to the lexicon, departing from the Portuguese model, and for certain safe affixes, like *-izar ,* valency was added automatically. However, most entries, not least nouns and adjectives, still lack valency frames, and these should be added in the future to support valency based context restrictions in the CG rule body.

Semantic prototypes. The semantic annotation of the HISPAL lexicon, and thus, the effectiveness of semantics-based CG rules, is still largely unimplemented and highly experimental. About 150 so-called *semantic prototypes* are used for nouns (e.g. <Aorn> = 'bird', <Lh> = 'human-made place', <con> = 'container', <tool> etc.), in analogy to the PALAVRAS system, but only in a few cases (e.g. *-ista* affix for +HUM) can semantic prototypes be added automatically. Experiments are under way to extract the +TOP feature (topological) from corpora based on preposition dependency.

4 Constraint Grammar modules

In a Constraint Grammar [7] parser[2], contextual rules are used to add, remove, select or replace grammatical tags in a token-based way. Rules are usually ordered in task batches, and within a task batch, in heuristicity batches. Since the application of rules is deterministic and sequential, and removed tags can't be recovered, it makes sense to place the safest rules first. As a compiler feature, the last remaining reading of a given type can never be removed, so "late" heuristic rules can't do any damage, if safter rules already have disambiguated a token. The reductionist method implied by REMOVE and SELECT rules, in combination with the last-reading-wins convention, make Constraint Grammar a very robust formalism that will assign an analysis even to very unorthodox or even faulty language input.

(a) REMOVE VFIN IF (*-1C VFIN BARRIER CLB OR KC)
(b) MAP (@SUBJ> @<SUBJ @<SC) TARGET (PROP) (NOT -1 PRP)
(c) SELECT (@SUBJ>) IF (*-1 >>> OR KS BARRIER NON-PRE-N/ADV) (*1 VFIN BARRIER NON-ATTR)

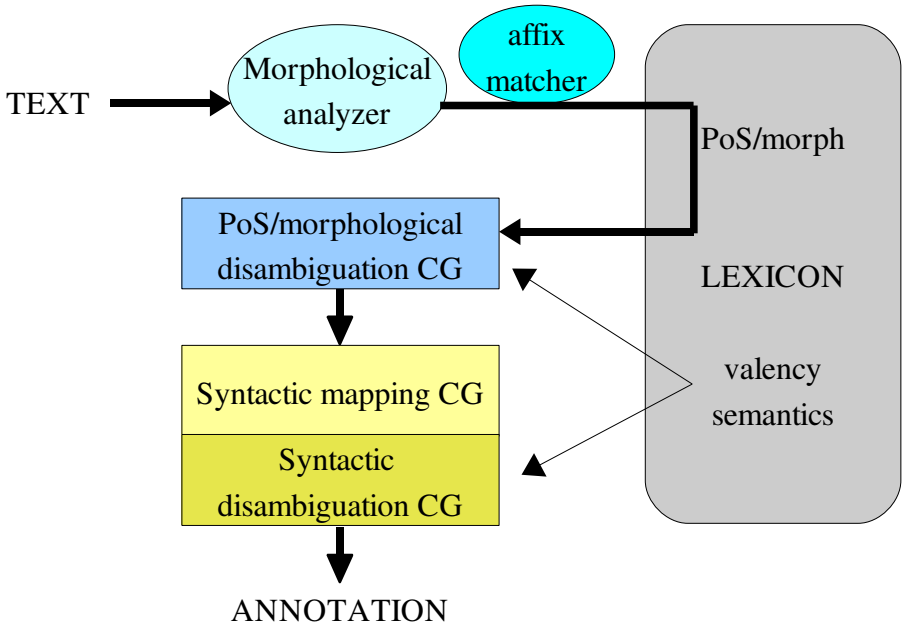
In example (a), finite verb readings (defined as a set, VFIN, in a special section of the grammar) are removed, if there is another safe, unambiguous (C) finite verb reading anywhere (*) to the left (-), counting from the first neighbouring position (1), and if there is no clause boundary (CLB) or cooordinating conjunction (KC) in between (BARRIER). The rule is a very simple example, and real rules will often incorporate more than one context, as well as negative (NOT) or conditioned (LINK) contexts. (b) is an example of a rule that *adds* information, in this case, a range of possible syntactic readings (right- and left-pointing subjects and subject complement) that are MAPped onto proper nouns (PROP) without (NOT) a left-adjacent (-1) preposition (PRP). In a later disambiguation rule section, rule (c) may SELECT the right-pointing subject reading, if there is a sentence boundary (>>>) or subordinating conjunction (KS) to the left with nothing but prenominals or adverbs in between (NON-PRE-N/ADV), i.e. if the subject-candidate is placed clause-initial, and if it is followed (*1) by a finite verb with nothing but attributes (NON-ATTR) in between.

In HISPAL, three CG modules are used, represented by the three examples, to (a) disambiguate input from the morphological analyzer, and then add (b) and disambiguate (c) syntactic readings:

---

2 The CG compiler currently used with the HISPAL rule file is the open source vislcg compiler (http://www.divvun.no/doc/tools/vislcg.html).

Illustration 1. Lemma distribution in the HISPAL lexicon.



Mature Constraint Grammars are of considerable size, with thousands of hand-written rules. Usually, a rule set will be built from scratch for each individual language. In a novel approach[3], the HISPAL project tried to cut down on grammar production time by importing a fully developed CG (PALAVRAS) from a related language, Portuguese [3]. The reason why grammar-porting is at all feasible is the robust reductionist way in which rules are applied by a CG compiler. Unlike rewriting rules in a PSG, CG rules do not strive to describe a language in a complete and positive way - rather, rules focus on what is contextually NOT possible, in effect annotating through disambiguation. This way, superfluous rules don't hurt and heuristic (Portuguese) rules can function as a harmless back-up in the presence of newer, non-heuristic Spanish rules. If a higher error rate is accepted, the Portuguese grammar will work after some minor adaptations of tag and set definitions, and can be improved incrementally, for instance by corpus based error-analysis. With every change made throughout the project, the grammar became a little more Spanish and a little less Portuguese:

1. Token- and lexeme-references in sets and rules were translated, i.e. structural words like prepositions or conjunctions: *quando -> cuando (when), e -> y (and),*

3 Another, though apparently unevaluated grammar transfer was suggested from Catalan to Spanish (http://prado.uab.es/English/corpus.html).

ou ->o (or), as well as semantically inspired lists (months, days of the week, units).
2. Specific Spanish rules were added early in the rules file to cover phenomena like the use of the preposition *a* with (especially human) direct objects.
3. Error-producing rules were traced and changed, replaced or deleted. Often, rules could be "repaired" by adding further context conditions, or by restricting the target set.

It must be kept in mind that due to the reductionist character of the grammar, many changes may appear piecemeal and unsystematic. Problem patterns, like differences in ambiguity classes between Portuguese and Spanish, such as the many-to-one relation between *muito* and *mucho/muy* (in favour of Spanish) or *lhe @DAT* and *le @DAT/ACC* (in favour of Portuguese), are difficult to exploit in a pre-emptive way, since rules interact in myriad ways and it is practically impossible to simply list all rules that have an influence on the disambiguation of a given feature. Only corpus runs guarantee that problematic rules will show up. Also, it may be difficult to tell if a given error really was caused by language differences, or if it was an inherent problem of the Portuguese rule file already. As the Portuguese and Spanish grammars grew apart, it became risky, if not impossible, to transfer changes and additions from one grammar to the other - simply because the complex reductionist interaction of Constraint Grammar rules, without extensive corpus testing, makes it difficult to predict how a given change will interact with the thousands of *other,* more and more language-specific, rules. As a result, the two grammars must now be seen as separate entities. Since its inception in 2001, at the start of the HISPAL project, the Spanish grammar has grown somewhat slower than the more actively developed Portuguese PALAVRAS[4], containing at the time of writing 1418 morphological disambiguation rules, 1249 mapping rules and 1862 syntactic disambiguation rules.

5 Evaluation and outlook

The system's morphological analyzer and lexicon were evaluated on a 43.000 word chunk representing interview-based newspaper articles and the Europarl corpus [3]. Roughly 3 % were proper nouns, of which between 5/6 (Interviews) and 2/3

4 PALAVRAS has attracted considerable attention in the Portuguese corpus linguistics community, and has been used in a large number of corpus projects, such as the ACDC and Floresta Sintá(c)tica projects (cp. http://www.linguateca.pt), motivating continuous development. Though it has been used for some annotation and teaching, HISPAL has yet to raise comparable interest, and development intensity reflects this.

(Europarl) were not covered by the lexicon. Coverage[5] for non-name words was around 99.4%. For the remaining 0.6 %, derivation-based recognition of a lexicon-root was possible for about two thirds, leaving about 0.2 % to heuristics proper. CG-disambiguation was able to assign a correct PoS class to 84-90% of lexicon failures, with derivation-supported analyses fairing a little better than purely heuristic ones.

Table 2. Coverage of the HISPAL lexicon and morphological analyzer.

	Interviews 8.058 words (9.554 tokens)	***Europarl*** 35.164 words (40.038 tokens)
Lexicon failures	0.22 %	0.13 %
-- orthographical errors	6 %	11 %
-- morphological	6 %	2 %
-- foreign spelling	39 %	27 %
-- hyphenation/MWE	6 %	11 %
-- other	50 %	49 %
-- PoS correct after CG	89 %	84 %
Derivation used	0.41 %	0.39 %
-- orthographical errors	6 %	2 %
-- PoS correct after CG	88 %	90 %
-- suffix/prefix	58% / 42%	61% / 39%
names (unknown)	3 % (83.7 %)	3 % (67.3 %)

In order to perform a detailed evaluation of HISPAL's Constraint Grammar based disambiguation system og syntactic parser, a gold standard corpus was built by manual revision of a smaller part of the interview corpus (2567 words, 3025 tokens). Against this gold corpus, the parser's F-Scores were 99% for part of speech, similar to the accuracy of CGs reported for other languages [3,5,6,7], and 95-96% for syntactic function, even when errors in dependency direction and "boundness"[6] were counted. In-clause syntactic functions fared slightly better (F=96.2) than functions assigned to entire subclauses (F=95.3).

5 Coverage was defined as cases where at least one reading lexicon- and analyzer-supported reading was found. Coverage is thus equivalent to *multitagger*-recall, *not* (disambiguated) annotation recall.

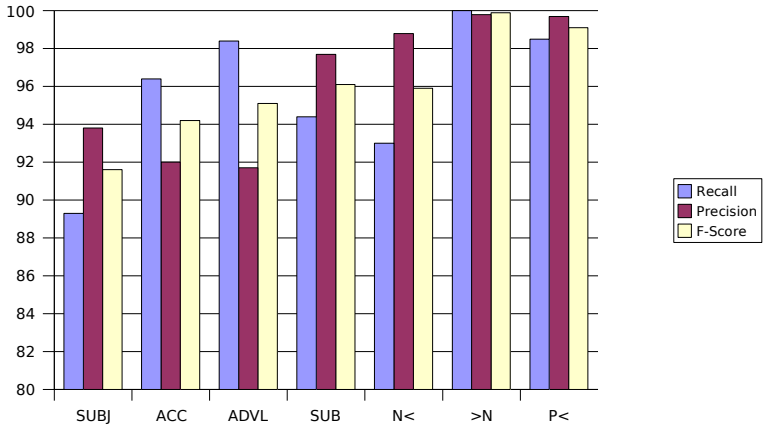
6 HISPAL's annotation system makes a distinction between arguments and adjuncts even where these share the same syntactic function, i.e. free adverbials (@ADVL) and valency bound adverbials (@ADV), as well as between subject predicatives/complements (@SC) and free subject predicatives (@PRED).

Table 3. Performance of the HISPAL parser, global values

	Recall	***Precision***	***F-score***
PoS (word class)[7]	99.03	99.03	99.03
Syntactic function, in-clause	96.42	96.44	96.43
... + dependency direction + boundness	96.22	96.24	96.23
Syntactic function, subclause	95.31	95.31	95.31

Finally, syntactic function errors were evaluated individually. As would be expected, in-phrase errors (e.g. pre- and postnominals, preposition complements) were rarer than clause level function errors (e.g. subject, object, adverbial), reflecting the closer disambiguation context for the former and long-distance dependencies of the latter.

Table 4. Performance of the HISPAL parser, specific syntactic functions



Interestingly, for subjects precision is higher than recall, while the opposite is true for objects and adverbials, suggesting a parser bias (i.e. tougher disambiguation) against the former. No evaluation figures could be found for other Spanish, CG-comparable systems, such as Connexor's Machinese (http://www.connexor.com/demo/syntax/) or Freeling (Atserias et.a. [1,2]), but at first glance, HISPAL's overall syntactic accuracy (95-96% on raw text) appears to compare favourably with the results of systems

7 Allmost no morphological errors were found for correct PoS, implying little in-class ambiguity. This may be due in part to the fairly distinctive inflexional morphology of Spanish, but can also be explained by the use of underspecified tags for systematically ambiguous morphology (e.g. gender in '-ista' nouns: M/F).

based solely on machine learning. Thus, the best scoring system in the CoNLL X shared task on dependency parsing (http://nextens.uvt.nl/ ~conll/results.html) achieved a syntactic label accuracy of only 90.4% even for "non-raw" input with manually corrected PoS and morphology, using training and test data from the Cast3LB treebank [7]. On the other hand, our evaluation was less rigid than optimally desirable, since it did not use multiple annotators and manual revision was performed on top of an automatic analysis, potentially creating a parser-friendly bias in ambiguous cases. Therefore, for a future, more detailed comparison, descriptive differences and differences in category set size would have to be addressed and reconciled, allowing the use of the same treebank test data and evaluation metrics.

References

1. Atserias J. et al.: Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text. Proceedings of LREC'98. Granada, Spain (1998)
2. Atserias, J. et al.: FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In: Proceedings of LREC'06. Genoa, Italy (2006)
3. Bick, Eckhard: The Parsing System 'Palavras' - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, Århus (2000)
4. Bick, Eckhard: A CG & PSG Hybrid Approach to Automatic Corpus Annotation. In: Kiril Simow & Petya Osenova (eds.), *Proceedings of SProLaC2003* (at Corpus Linguistics 2003, Lancaster), pp. 1-12 (2003)
5. Bick, Eckhard: Parsing and Evaluating the French Europarl Corpus. In: Paroubek, P., Robba, I. and Vilnat, A. (red.): Méthodes et outils pour l'évaluation des analyseurs syntaxiques (Journée ATALA, May 15, 2004). pp. 4-9. Paris: ATALA (2004)
6. Hagen, K., Johannessen, J. B., Nøklestad, A.: A Constraint-Based Tagger for Norwegian". In: Lindberg, C.-E. og Lund, S.N. (red.): 17th Scandinavian Conference of Linguistic, Odense. Odense Working Papers in Language and Communication, No. 19, vol I. (2000)
7. Karlsson, Fred et al.: Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin (1995)
8. Koehn, Philipp: Europarl, A Parallel Corpus for Statistical Machine Translation. In: Proceedings of MT Summit X, Phuket, Thailand (2005)
9. Lager, Torbjörn: The µ-TBL System: Logic Programming Tools for Transformation-Based Learning. In: Proceedings of CoNLL'99, Bergen, (1999)
10. Padró, L.: POS Tagging Using Relaxation Labelling. In: Proceedings of COLING '96. Copenhagen, Denmark (1996)
11. Palomar, M. et al.: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y espanñol. In: *Proceedings of SEPLN XX*, pp 81-88. Barcelona (2004)