LXGram: A Deep Linguistic Processing Grammar for Portuguese

Francisco Costa and António Branco

Universidade de Lisboa
fcosta@di.fc.ul.pt
Antonio.Branco@di.fc.ul.pt

Abstract. In this paper we present LXGram, a general purpose grammar for the deep linguistic processing of Portuguese that delivers high precision grammatical analysis and detailed meaning representations. We present the main design features and evaluation results on the grammar's coverage as well as its ability to produce correct grammatical analyses.

Key words: deep linguistic processing, unification grammars, parsing

1 Introduction

We present what is, to the best of our knowledge, the first general purpose grammar, distributed under an open-source license, for the deep linguistic processing of Portuguese, that delivers a thorough and principled linguistic analysis of sentences, including their formal semantic representation. LXGram is part of the DELPH-IN Consortium, an international group of researchers working on deep linguistic processing for a variety of languages. In Section 2 the main design features of this grammar are described. Evaluation results are presented in Section 3, based on an experiment consisting of parsing spontaneous text that was not seen during the development phase. Finally, we conclude in Section 4.

2 Scope and Design Features

LXGram is based on hand coded linguistic generalizations supplemented with a stochastic model for ambiguity resolution of parses. It follows the grammatical framework of Head-Driven Phrase Structure Grammar (HPSG [1]), one of the most prominent linguistic theories used in natural language processing.

HPSG is a linguistic framework for which there is a substantial amount of published work. This allows for the straightforward implementation of well known grammatical analyses, which are linguistically grounded and have undergone scientific scrutiny. It also has a positive impact in reusability and extendibility, because more people can understand it immediately. The HPSG literature has produced very accurate analyses of long distance dependencies, and a general strong point of computational HPSGs, among many others, is precisely the implementation of this key phenomenon of natural language syntax.

HPSGs associate grammatical representations to natural language expressions, including the formal representation of their meaning. Like several other computational HPSGs, LXGram uses Minimal Recursion Semantics (MRS [2]) for the representation of meaning. An MRS representation is a description of a set of possible logic formulas that differ only in the relative scope of the relations present in these formulas. In other words, it supports scope underspecification. Semantic representations provide an additional level of abstraction, as they completely abstract word order and language specific grammatical restrictions. Additionally, the MRS format of semantic representation that is employed is well defined in the sense that it is known how to map between MRS representations and formulas of second order logic, for which there is a set-theoretic interpretation. Because of space limitations, it is impossible to provide a detailed account of MRS representations here. [2] provides a very clear description of it.

LXGram is developed in the Linguistic Knowledge Builder (LKB) system [3], an open-source development environment for constraint-based grammars. This environment provides a GUI, debugging tools and very efficient algorithms for parsing and generation with the grammars developed there. Several broad-coverage HPSGs have been developed in the LKB; the largest ones are for English [4], German [5] and Japanese [6]. The grammars developed with the LKB are also supported by the PET parser [7], which allows for faster parsing times due to the fact that the grammars are compiled into a binary format.

LXGram is in active development, but it already supports a wide range of linguistic phenomena, such as long distance dependencies, coordination, subordination, modification and many subcategorization frames. and its lexicon contains 25000 entries. At the moment, LXGram contains 64 lexical rules, 101 syntax rules, around 850 lexical leaf types (determining syntactic and semantic properties of lexical entries), and 35K lines of code (excluding the lexicon). LXGram supports both European and Brazilian Portuguese. It contains lexical entries that are specific to either of them, and it covers both European and Brazilian syntax ([8]). A statistical disambiguation model was also trained, in order to automatically select the most likely analysis of a sentence when the grammar produces multiple solutions. This model was trained from a dataset comprising 2000 sentences of newspaper text, using a maximum entropy algorithm. The linguistic analyses that are implemented in the grammar are documented in a report that is updated and expanded with each version of the grammar. The grammar is available for download at `http://nlx.di.fc.ul.pt/lxgram`, together with its documentation.

3 Evaluation

We conducted an experiment to assess the coverage of LXGram's current version on spontaneous text. We used a subset of the Portuguese Wikipedia, as well as part of two publicly available corpora: CETEMPúblico and CETENFolha, which contain newspaper text from "O Público" and "Folha de São Paulo" respectively.

	Wikipedia	CETEMPúblico	CETENFolha	Total
Sentences	66304	30000	30000	126304
Avg. words/sentence	25	27.5	18.6	24
Avg. seconds/sentence	2.6	4.7	2	3
Parsed sentences	20995	8455	11173	40623
Parsed percent	32%	28%	37%	32%
Avg. readings/parsed sentence	67	87	75	73
Avg. words/parsed sentence	11	13	11	11

Table 1. Evaluation data and grammar coverage

The Wikipedia corpus consists of a selection of articles downloaded from the Portuguese Wikipedia, by following the links on the page "Artigos Destacados" ("Featured Articles"). 318 pages were obtained in this way and preprocessed in order to remove HTML markup.

As for the two newspaper corpora, we randomly selected 30000 sentences from each of them. We removed all XML-like tags (such as `<s>` for sentence boundaries) but kept each sentence in its own line, to be processed separately.

Before parsing these texts, we fed each sentence to a part-of-speech tagger [9] and a morphological analyzer [10, 11], in order to handle out-of-vocabulary words and to constrain the parser search space. For each sentence, we kept the 250 most likely analyses, as determined by the disambiguation model presented.

LXGram was able to successfully parse 32% of the sentences in the Wikipedia sample, 28% of the CETEMPúblico sentences and 37% of the CETENFolha sample. Table 1 summarizes our results, using a 2,5 GHz Intel processor.

The fact that the average length of parsed sentences is very similar for both CETEMPúblico and CETENFolha indicates that the large difference in coverage on these two datasets may be more related to average sentence length than to differences between European and Brazilian Portuguese.

When comparing these results to the other computational HPSGs, it should be mentioned that [12] reports values of 80.4% coverage on newspaper text for the English grammar, 42.7% for the Japanese grammar and 28.6% for the German grammar.[1] All of these grammars have been in development for over 15 years now, and they are all substantially older than LXGram, with 4 years of development. A more recent HPSG Grammar, for Spanish—a language quite similar to Portuguese—, is the Spanish Resource Grammar [13], approximately as old as LXGram. The SRG is reported in [12] to have a coverage of 7.5%.

In order to assess the accuracy of the grammar, we inspected a sample with the first 50 parsed sentenced in the CETENFolha subcorpus. 20 sentences were correctly parsed, and furthermore the preferred reading was the one chosen by the disambiguation model. Another 10 sentences also received a correct parse, although the disambiguation model did not choose the preferred reading for these sentences as the best one. From the 20 sentences that did not receive a correct parse, 12 sentences were affected by errors from the part-of-speech tagger or the

[1] However, the German grammar has close to 40% coverage on newspaper text (personal communication by Berthold Crysmann) using a more recent method to integrate information coming from preprocessing tools.

morphological analyzer, and 8 of them were due to genuine limitations in the grammar or the disambiguation model (for instance, lack of some subcategorization frames for some words in the lexicon).

4 Conclusions

We presented a resource grammar for Portuguese which is based on HPSG. To the best of our knowledge, it is the only deep linguistic parser for Portuguese that outputs fine-grained semantic representations.

The grammar keeps being developed, but it already features interesting coverage of unrestricted text, achieving over 30% coverage on newspaper text, which is usually hard to parse by symbolic systems. Additionally, a sample of those parsed sentences was manually evaluated, and it indicates that a good portion of the parsed sentences got a correct representation (60%) and are disambiguated correctly (40%), while 60% the of parse failures were due to preprocessing errors.

Our ongoing work includes grammar expansion, and also the creation of a treebank of sentences parsed with the grammar and manually disambiguated.

References

1. Pollard, C., Sag, I.: Head-Driven Phrase Structure Grammar. Chicago University Press and CSLI, Stanford (1994)
2. Copestake, A., Flickinger, D., Sag, I.A., Pollard, C.: Minimal Recursion Semantics: An introduction. Journal of Research on Language and Computation **3**(2–3) (2005)
3. Copestake, A.: Implementing Typed Feature Structure Grammars. CSLI, Stanford (2002)
4. Copestake, A., Flickinger, D.: An open-source grammar development environment and broad-coverage English grammar using HPSG. In: LREC-2000, Athens (2000)
5. Crysmann, B.: Local ambiguity packing and discontinuity in German. In: ACL Workshop on Deep Linguistic Processing, Prague (2007)
6. Siegel, M., Bender, E.M.: Efficient deep processing of Japanese. In: The 3rd Workshop on Asian Language Resources and International Standardization. Coling 2002 Post-Conference Workshop, Taipei (2002) 31–38
7. Callmeier, U.: PET — A platform for experimentation with efficient HPSG processing techniques. Natural Language Engineering **6**(1) (2000) 99–108
8. Branco, A., Costa, F.: Accommodating language variation in deep processing. In King, T.H., Bender, E.M., eds.: GEAF 2007, Stanford, CSLI (2007) 67–86
9. Branco, A., Silva, J.: Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In: LREC2004, Paris, ELRA (2004) 507–510
10. Branco, A., Silva, J.: Very high accuracy rule-based nominal lemmatization with a minimal lexicon. In: APL XXI, Lisbon (2007)
11. Branco, A., Nunes, F., Costa, F.: The processing of verbal inflection ambiguity: Characterization of the problem space. In: APL XXI, Lisbon (2007)
12. Zhang, Y., Wang, R., Oepen, S.: Hybrid multilingual parsing with HPSG for SRL. In: CoNLL 2009, Boulder, USA (2009)
13. Marimon, M., Bel, N., Espeja, S., Seghezzi, N.: The Spanish Resource Grammar: pre-processing strategy and lexical acquisition. In: ACL Workshop on Deep Linguistic Processing, Prague (2007)