# Annotating Portuguese Corpora with Word Senses Using LX-SenseAnnotator

Steven Neale and António Branco

NLX - Natural Language and Speech Group
Faculty of Sciences, Department of Informatics
University of Lisbon, Portugal
{steven.neale, antonio.branco}@di.fc.ul.pt

**Abstract.** This paper describes LX-SenseAnnotator, an accessible and easy-to-use interface tool for manual annotating text with word senses. We demonstrate how the tool was used to manually annotate the CINTIL-WordSenses corpus, outlining the loading and browsing of Portuguese texts and how word senses themselves are selected and assigned. We also describe the potential for LX-SenseAnnotator to be adapted for other languages besides Portuguese.

**Keywords:** manual annotation, word senses, corpora, interface tools

## 1 Introduction

As part of our research on word sense disambiguation (WSD) in Portuguese, we have encountered the need for a simpler, more accessible way to annotate texts with word senses in order to quickly and easily create gold standard corpora for training and evaluation. Annotated corpora are hugely important, supporting both the analysis of large quantities of text [3] and the training and evaluation of natural language processing (NLP) tools, and there is an increasing interest in producing corpora containing "high-quality linguistic annotations at the semantic level" [6].

Many of the current unsupervised approaches to WSD assign the eight-digit 'synset identifiers' from ontologies and lexicons such as the Princeton WordNet – within which nouns, verbs, adjectives and adverbs are grouped into sets of synonyms or 'synsets' [2] – to unlabelled raw text. However, despite the obvious need for corpora manually-annotated with synset identifiers against which to evaluate these approaches, manual word sense annotation tools are either difficult to come by or seem intrinsically tied to specific corpora or lexica.

In this paper, we describe LX-SenseAnnotator, a user-interface tool designed specifically to offer a more open and flexible way to annotate texts with senses pulled from WordNets in the Princeton format. While LX-SenseAnnotator has the scope to become flexible enough to be adapted to other languages in future versions, its current implementation is formatted to handle Portuguese texts specifically and was recently used to manually annotate the first version of the gold-standard CINTIL-WordSenses corpus [5].

## 2 Using LX-SenseAnnotator

In its current Portuguese implementation, LX-SenseAnnotator is designed for importing text files that have already been part-of-speech (POS) tagged and lemmatized using the LX-Suite, an existing pipeline of shallow processing tools for Portuguese [1]. The POS tags are then used to organize each word in the imported text according to whether they are or are not sense-taggable. Nouns, verbs, adjectives and adverbs whose lemma is present in the WordNet being used (potential candidates for sense tagging) are marked in red – so that they can be easily distinguished from the rest of the text, which is marked in dark blue – unless they have already been assigned a synset identifier, in which case they are marked in green (Figure 1).

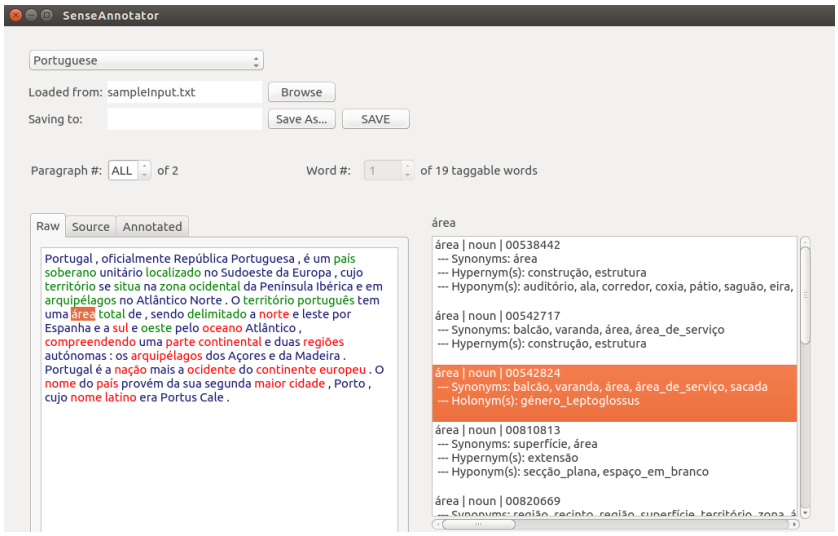


**Fig. 1.** Displaying a list of senses for the word 'área' (English 'area') using LX-SenseAnnotator.

Annotators can either click on red, sense-taggagable words or use the scroll box in the middle of the interface to browse through all currently sense-taggable words in the text, and when a sense-taggable word is selected it is highlighted and available senses from a pre-loaded WordNet are displayed in the results panel on the right-hand side of the interface. The lemmas and POS tags of specific words queried against the WordNet's index.sense file when they are highlighted by the annotator, and for every entry that the selected word has in the WordNet the appropriate eight-digit synset identifier – as well as additional contextual information from the synset such as synonyms, hypernyms, hyponyms, holonyms, antonyms and so on – is displayed in the list of available senses.

Once an annotator has decided which of the available senses is most appropriate for a given word – taking into account its discursive context – they simply double click on the sense to automatically assign it to the selected word, which now becomes green in the text display on the left-hand side of the interface. The word is removed from the list of sense-taggable words, although they can still be selected should annotators decide to remove the annotated sense or choose a more appropriate one later. This process was used to assign senses from the Portuguese MultiWordNet [4] to 45,502 words across 23,825 sentences for the first version of CINTIL-WordSenses.

## 3 Flexibility and Adaptability

For producing the first version of CINTIL-WordSenses the pre-loaded WordNet used was the Portuguese MultiWordNet, but because any WordNet in any language can be loaded into LX-SenseAnnotator providing that it adheres to the traditional Princeton format the tool is potentially extremely flexible. An important direction for future work is to take advantage of this by adapting the way texts are imported such that different pre-tagged text formats and POS tagsets are supported, which coupled with the flexibility of pre-loading WordNets would extend LX-SenseAnnotator for use with many more languages.

## Acknowledgements

This work has been undertaken and funded as part of the EU project QTLeap (EC/FP7/610516) and the Portuguese project DP4LT (PTDC/EEI-SII/1940/2012).

## References

1. Branco, A., Silva, J.R.: A Suite of Shallow Processing Tools for Portuguese: LX-suite. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations. pp. 179–182. EACL '06, Association for Computational Linguistics, Trento, Italy (2006)
2. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
3. Leech, G.: Adding Linguistic Annotations. In: Wynne, M. (ed.) Developing Linguistic Corpora: A Guide to Good Practice. AHDS Literature, Languages and Linguistics (2004)
4. MultiWordNet: The MultiWordNet project. `http://multiwordnet.fbk.eu/english/home.php` (nd), accessed: 2014-01-13
5. Neale, S., Pereira, R.V., Silva, J., Branco, A.: Lexical Semantics Annotation for Enriched Portuguese Corpora. In: Proceedings of the 12th International Conference on the Computational Processing of the Portuguese Language. PROPOR 2016, Tomar, Portugal (2016)
6. Passonneau, R.J., Baker, C., Fellbaum, C., Ide, N.: The MASC Word Sense Sentence Corpus. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. European Language Resources Association, Istanbul, Turkey (2012)