# Language Resources and Processing Tools at the University of Lisbon in the NLX Group Collection

António Branco, João Silva, Francisco Costa, João Rodrigues, Pedro Martins, Eduardo Ferreira, Filipe Nunes, Sérgio Castro, Steven Neale, Catarina Carvalheiro, Sílvia Pereira, Mariana Avelãs, Clara Pinto, Andreia Querido, Rita de Carvalho, Marisa Campos, Nuno Rendeiro, Catarina Correia, Patrícia Gomes, Diana Amaral, and Rita Valadas Pereira

University of Lisbon, Faculty of Sciences, Department of Informatics

**Abstract.** In this paper we present many of the language resources and processing tools developed and made available at the University of Lisbon by the NLX - Natural Language and Speech Group. These were developed over the years to support the development of a wide array of natural language applications, including machine translation.

**Keywords:** Portuguese, language technology, natural language processing, language resources, language processing tools.

## 1 Introduction

The development of machine translation solutions requires a number of instrumental and auxiliary language processing tools as well as appropriate companion data sets for the training and evaluation of these tools and applications. This paper aims at providing a brief introduction to the collection of processing tools and language resources for the Portuguese language developed and made available at the University of Lisbon by the NLX Group, the Natural Language and Speech Group of the Department of Informatics of the University of Lisbon.

These resources and processing tools are made available from the NLX-Group website.[1] Most of them support also free online linguistic processing services and demos that are available at the LX-Center.[2]

This paper is organized as follows: Section 2 presents the collection of tools that are instrumental for natural language processing and machine translation, and Section 3 covers the language resources. The paper closes with final remarks in Section 4.

[1] http://nlx.di.fc.ul.pt/
[2] http://lxcenter.di.fc.ul.pt/

## 2 Language processing tools in the NLX Collection

In the NLX-Group we have developed language processing tools that virtually cover the full range of tasks from shallow to deep processing. Some of tools address simple procedures, such as the LX-Tokenizer, and others tackle more sophisticated functionalities, such as the Lx-DepParser, and others cover named entity recognition, verbal and nominal inflection or word-sense disambiguation, etc. These tools are useful at different levels to support machine translation and are presented below.

- **LX-Lemmatizer** is a verbal lemmatizer that takes a Portuguese verb form as input and delivers a ranked list of the corresponding lemmata (infinitive forms) together with inflectional feature values[18]. Its performance was evaluated as delivering 96.5% accuracy.
- **LX-Inflector** is a language processing tool for nominal lemmatization and inflection [4], taking a Portuguese word form that follows the nominal inflection paradigm and an inflection feature bundle, and delivers both the corresponding lemma and the indication of its feature bundle, and the resulting form that conveys the feature bundle entered. It is based on principled linguistic generalizations captured by regular expressions and the appropriate lexica of affixes, thus handling neologisms. The lemmatization function has 97.7% f-score. Figure 1 shows LX-Inflector online service.

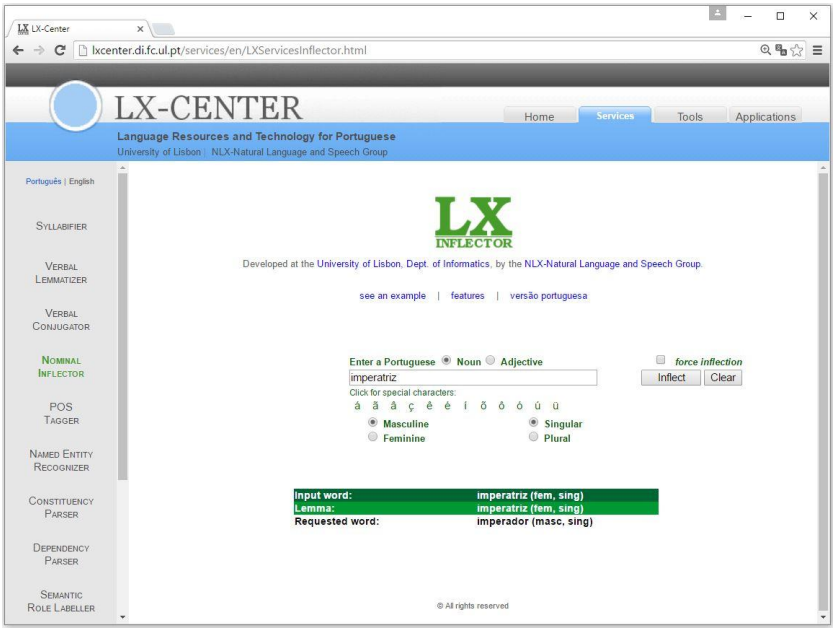


**Fig. 1.** LX-Inflector online service for the nominal lemmatization and inflection of Portuguese

- **LX-Chunker** is an identifier of paragraphs and sentences for Portuguese. It seeks to cope with the ambiguity and ambivalence of symbols that in some occurrences are indicators of separations among sentences and in other contexts are not. It is a hybrid tool, based on regular expressions and hidden Markov models, with an f-score of 99.9%.
- **LX-Conjugator** is a verbal conjugator for the Portuguese language [18]. It takes a Portuguese infinitive verb form as input and delivers the corresponding conjugated forms. It is the only available tool for fully-fledged Portuguese verb conjugation, including the full range of pronominal conjugation forms. Its capacity includes the handling of pronominal conjugation, compound tenses, double forms of past participles, past participle forms inflected for number and gender, negative imperative forms, and courtesy forms for second person. Given that it is based in principled linguistic generalizations captured by regular expressions and the appropriate lexica of affixes, it is the only available conjugator to handle neologisms. Their occasional faults have been correct along the time as it has been put to use, and at present no defect is known.
- **LX-Tokenizer** is an identifier of the boundaries of relevant word-level tokens in Portuguese text. It seeks to cope with the ambiguity of strings that in some contexts are single-word tokens and in some other contexts are contractions, i.e. double-word tokens. It achieves an f-score of 99.7%. It is incorporated in Lx-Suite[7], available at the Lx-Center.[3]
- **LX-Tagger** is a part-of-speech tagger with disambiguation and full coverage for the Portuguese language. For each word occurring in a text and from the possible different morpho-syntactic categories that word may have in the lexicon, it assigns a single tag to it that indicates the morpho-syntactic category that it bears in that occurrence in the text. It scores 96.8% accuracy.
- **LX-NER** is an identifier and classifier of named entity expressions for the Portuguese language [9]. Its number-based part evaluates to an f-score of 85.6%, and the name-based to 85.7%.
- **LX-NED** is a named entity disambiguator that annotates the occurrence of an input expression with the Wikipedia entry it refers to in its context, with an f-score of 67.0%.
- **LX-WSD** is a word sense disambiguator that annotates the occurrence of an input word with the MWN.PT wordnet concept it expresses in its context, with an fscore of 65.0%.
- **LX-Parser** is a stochastic parser that performs the syntactic analysis of Portuguese sentences in terms of their constituency structure[17][16]. It achieves an f-score of 88% under the Parseval metric.
- **LX-DepParser** is a parser of grammatical dependency relations for sentences of Portuguese that for each input sentence delivers a graph connecting its words and whose directed arcs represent grammatical dependencies and the labels at the said arcs represent the grammatical function of those dependencies. The evaluation of its performance obtained 91.2% in terms

[3] http://lxcenter.di.fc.ul.pt/

of labelled attachment score (LAS). Figure 2 shows LX-DepParser online service.

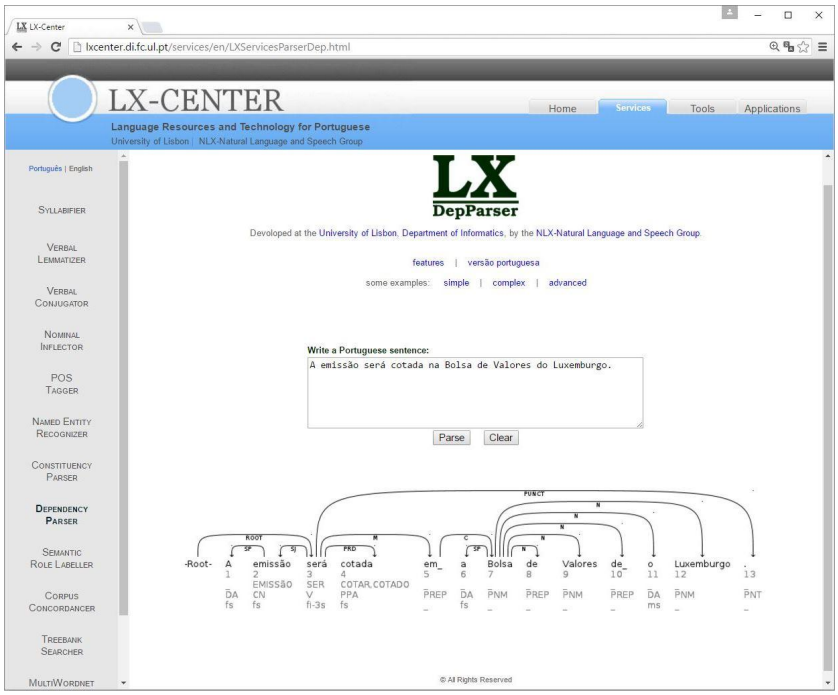


**Fig. 2.** LX-DepParser online service for the syntactic analysis of Portuguese

## 3 Language Resources in the NLX collection

In this section we briefly introduce language resources in the NLX-Group collection that are relevant for the theme of the present workshop.

- **CINTIL-International Corpus of Portuguese**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including morphological analyzers, POS taggers and named entity recognizers. This corpus contains 1 million words manually annotated by experts in natural language science and technology. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It was developed in cooperation with CLUL-Center of Linguistics of the University of Lisbon[1].
- **CINTIL-DeepBank**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including language models for deep linguistic processing grammars [15]. This corpus contains around 10 000 sentences (approximately 100000 words) manually annotated by experts

in natural language science and technology. Each sentence is associated to exhaustive characterization of its grammatical features in lexical, morphological, syntactic and semantic terms.

- **CINTIL-Treebank**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including constituency parsers[15]. This treebank contains around 10 000 sentences (100000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to linguistic information about its syntactic constituency tree tagged with phrase categories. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **CINTIL-DependencyBank**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including grammatical dependencies parsers[15]. This corpus contains around 10 000 sentences (approximately 100000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the graph that represents the grammatical functions holding between its words[15]. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **CINTIL-PropBank**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including semantic role labellers[3]. This corpus contains around 10 000 sentences (approximately 100000 words) manually annotated by experts in natural language science and technology. The syntactic constituents of sentences are associated to linguistic information about its semantic role. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **CINTIL-LogicalFormBank**: Set of text materials to support the evaluation and training of tools for the semantic processing of Portuguese[15][2]. This corpus contains around 10 000 sentences (approximately 100000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the logical form that represents its meaning in a logical language for semantic description[15].
- **CINTIL-WordSenses**: Set of text materials to support the evaluation and training of word sense disambiguators[13]. This corpus contains around 24 000 sentences with 45 000 words that are manually annotated by experts in natural language science and technology with the identifiers of concepts (synsets) that they convey in terms of the lexical semantic network MWN.PT [12]. Additionally, each word is associated to the linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **CINTIL DependencyBank PREMIUM**: Set of text materials similar in design to the previous one and differing from it in the sentences that were treebanked and in the circumstance that the support tool to draw the grammatical dependency graphs is not the LXGram but the full text coverage

LXDependencyParser [8][6]. It contains 3 000 sentences (approximately 79 000 words).

- **CINTIL-NamedEntities**: Set of text materials to support the evaluation and training of named entity disambiguators. This corpus contains around 30 000 sentences with 26 000 named entities that are manually annotated by experts in natural language science and technology with identifiers of the corresponding entities in the DBpedia ontology[13]. Additionally, each word is associated to the linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **QTLeap Multilingual Parallel Corpora**: Set of 4 000 question and answer pairs in the domain of computer and IT troubleshooting for both hardware and software[11]. This textual material was collected using a commercial support service via chat, in Portuguese, and the corpus is thus composed by naturally occurring utterances produced by users while interacting with that service. Each question answer pair is translated into seven languages, other than Portuguese, namely Czech, Basque, Bulgarian, Dutch, English, German and Spanish.
- **QTLeap WSD/NED Multilingual Corpora**: Set of text materials comprising the QTLeap Multilingual Parallel Corpora and the Europarl multilingual corpora for the Czech (9.2 Million tokens), Basque (5.2 Million), Bulgarian (4.9 M), English (53 M), Portuguese (5.7 M) and Spanish (57.1M) languages, automatically annotated at multiple semantic levels by processing tools for tokenization, lemmatization, part-of-speech tagging, named-entity recognition and classification, named-entity disambiguation, word sense disambiguation and coreference resolution[14].
- **DeepBankPT**:Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the processing of Portuguese, including language models for deep linguistic processing grammars [10]. This corpus contains around 3 500 sentences (approximately 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to exhaustive characterization of its grammatical features in lexical, morphological, syntactic and semantic terms.
- **TreebankPT**: Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the processing of Portuguese, including constituency parsers. This treebank contains around 3 500 sentences (approximately 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to linguistic information about its syntactic constituency tree. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **PropBankPT**: Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the processing of Portuguese, including semantic role labellers. This corpus contains around 3 500 sentences (approximately 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associ-

ated to its syntactic constituency tree decorated with semantic roles. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.

- **DependencyBankPT**: Set of text materials translated into Portuguese from the Penn Treebank to support the evaluation and training of tools for the processing of Portuguese, including grammatical dependencies parsers. This treebank contains around 3 500 sentences (approximately 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the graph that represents the grammatical functions holding between its words. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **LogicalFormBankPT**: Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the semantic processing of Portuguese. This corpus contains around 3 500 sentences (ca. 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the logical form that represents its meaning in a logical language for semantic description[5].

## 4 Final Remarks

The NLX Group has developed the language processing tools and resources briefly introduced above. These datasets and tools are distributed from the NLX-Group website or at the META-SHARE [4] distribution platform. They are made available with the goal of being of help for further research and progress of the computational processing of the Portuguese language.

## References


1. Barreto, F., Branco, A., Ferreira, E., Mendes, A., Nascimento, M.F., Nunes, F., Silva, J.: Open resources and tools for the shallow processing of portuguese: the tagshare project. In: Proceedings of LREC 2006. Citeseer (2006)
2. Branco, A.: Logicalformbanks, the next generation of semantically annotated corpora: key issues in construction methodology. Recent Advances in Intelligent Information Systems, Exit, Warsaw pp. 3–11 (2009)
3. Branco, A., Carvalheiro, C., Pereira, S., Silveira, S., Silva, J., Castro, S., Graça, J.: A propbank for portuguese: the cintil-propbank. In: LREC. pp. 1516–1521 (2012)
4. Branco, A., Nunes, F.: Verb analysis in a highly inflective language with an mff algorithm. In: Computational Processing of the Portuguese Language, pp. 1–11. Springer (2012)
5. Branco, A., Silva, J., Gonçalves, P., Costa, F., Silveira, S., Del Gaudio, R., Rodrigues, J., Castro, S., Rodrigues, L., Martins, P., et al.: The cintil and lx companion collections of language resources and tools for portuguese


[4] http://metashare.metanet4u.eu

6. Branco, A., Silva, J., Querido, A., Carvalho, R.: Cintil dependencybank premium handbook: Design options for the representation of grammatical dependencies (2015)
7. Branco, A., Silva, J.R.: A suite of shallow processing tools for portuguese: Lx-suite. In: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations. pp. 179–182. Association for Computational Linguistics (2006)
8. de Carvalho, R., Querido, A., Campos, M., Valadas Pereira, R., Silva, J., Branco, A., in Press: Cintil dependencybank premium: A corpus of grammatical dependencies for portuguese. In: Proceedings, LREC2016 - 10th Language Resources and Evaluation Conference (May 23-28 2016)
9. Ferreira, E., Balsa, J., Branco, A.: Combining rule-based and statistical methods for named entity recognition in portuguese. In: Actas da 5a Workshop em Tecnologias da Informaçao e da Linguagem Humana (2007)
10. Flickinger, D., Kordoni, V., Zhang, Y., Branco, A., Simov, K., Osenova, P., Carvalheiro, C., Costa, F., Castro, S.: Pardeepbank: Multiple parallel deep treebanking. Proceedings of TLT-2012, Lisbon, Portugal pp. 97–108 (2012)
11. Gaudio, R.D., Burchardt, A., Branco, A., in Press: Evaluating machine translation in a usage scenario. In: Proceedings, LREC2016 - 10th Language Resources and Evaluation Conference (May 23-28 2016)
12. MultiWordNet: The MultiWordNet project. http://multiwordnet.fbk.eu/english/home.php (nd), accessed: 2015-01-13
13. Neale, S., Valadas Pereira, R., Silva, J., Branco, A., in Press: Lexical semantics annotation for enriched portuguese corpora. In: Springer (ed.) Lecture Notes in Artificial Intelligence
14. Otegi, A., Aranberri, N., Branco, A., Hajič, J., Popel, M., Simov, K., Agirre, E., in Press: Qtleap wsd/ned corpora: Semantic annotation of parallel corpora in six languages. In: Proceedings, LREC2016 - 10th Language Resources and Evaluation Conference (May 23-28 2016)
15. Silva, J., Branco, A., Castro, S., Costa, F.: Deep, consistent and also useful: Extracting vistas from deep corpora for shallower tasks. In: Proceedings of the Workshop on Advanced Treebanking at the 8th Language Resources and Evaluation Conference (LREC). pp. 45–52 (2012)
16. Silva, J., Branco, A., Castro, S., Reis, R.: Out-of-the-box robust parsing of portuguese. In: Computational Processing of the Portuguese Language, pp. 75–85. Springer (2010)
17. Silva, J., Branco, A., Gonçalves, P.N.: Top-performing robust constituency parsing of portuguese: Freely available in as many ways as you can get it. In: LREC (2010)
18. da Silva, J.R.M.F.: Shallow processing of portuguese: From sentence chunking to nominal lemmatization (2007)