The CINTIL and LX companion collections of language resources and tools for Portuguese

António Branco, João Silva, Patricia Gonçalves, Francisco Costa, Sara Silveira, Rosa Del Gaudio, João Rodrigues, Sérgio Castro, Lino Rodrigues, Pedro Martins, Filipe Nunes, Eduardo Ferreira, João Alves, Rita Carvalho, Andreia Querido, Mariza Campos, and Nuno Rendeiro

University of Lisbon,
Department of Informatics
{nlxgroup@di.fc.ul.pt}

Abstract. This paper supports the presentation and demonstration of the LX collection of processing tools and the CINTIL companion collection of language resources.

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

1 Introduction

This paper aims at supporting the presentation and demonstration of the LX collection of processing tools and the CINTIL companion collection of language resources. These are resources and tools for the Portuguese language developed at the NLX Group[1], the Natural Language and Speech Group of the Department of Informatics of the University of Lisbon, and are made available to foster the education, research and development in natural language science and technology. This paper is organized as follows, Section 2 presents the content to be demonstrated concerning the collection of tools, and Section 3 is concerned with the language resources.

2 The LX collection of online services, tools and applications

The LX collection includes online services, tools and applications for the computational processing of Portuguese. They are listed, described and accessible from LX-Center[2].

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

[2] Available on http://lxcenter.di.fc.ul.pt

Online Services

The following online services are freely available on a best-effort basis:

- LX-Lemmatizer provides fully-fledged lemmatization of Portuguese verbs [25]. It takes a verb form and delivers all the corresponding lemmata (infinitive forms) together with the inflectional feature values. This system achieves 97.67% of f-score.
- LX-Conjugator is a fully fledged verbal conjugator, including all forms of clitic conjugation [25].
- LX-Inflector offers rule-based nominal lemmatization and inflection [7]. The service delivers (i) the input form with the corresponding values for the inflectional features of gender and number; (ii) the lemmata (singular and masculine forms when available) possibly corresponding to the input form and (iii) the inflected forms (when available) of each lemmata in accordance with the values for inflectional features requested by the user.
- LX-Suite performs the shallow processing of Portuguese tasks [12]. It is composed by a set of shallow processing tools: (i) LX-Chunker which marks sentence and paragraph boundaries; (ii) LX-Tokenizer which segments text into lexically relevant tokens, using whitespace as the separator; (iii) LX-Tagger which assigns a single morpho-syntactic tag to each token; (iv) LX-Featurizer which assigns inflection feature values to words from the nominal categories, namely, gender (masculine or feminine), number (singular or plural) and, when applicable, person (1st, 2nd and 3rd) and (v) LX-Lemmatizer which assigns a lemma to words from the nominal categories (adjectives, common Nouns and past participles).
- LX-NER undertakes the recognition of expressions for named entities [18]. It takes a segment of text and identifies, circumscribes and classifies the expressions for named entities it contains. The system handles the following types of expressions: (i) Number-based expressions; number expressions are marked as `NUMEX` (e.g. decimal, roman, cardinal, fraction, etc.); measures expressions are marked as `MEASEX` (e.g. currency and scientific units); time expressions are marked as `TIMEX` (e.g. date, month, century, etc.); addresses expressions are marked as `ADDREX` (location address, zip code, etc.); (ii) Name-based expressions are marked as `NAMEX`. Names are classified in subtypes `PER` (person, e.g. *Presidente Cavaco Silva*), `ORG` (organization, e.g. *LG Electronics*), `LOC` (locations, e.g. *Portugal*), `EVT` (events, e.g. *International Conference on Computational Processing of Portuguese*), `WRK` (movies, books, paintings, etc, e.g. *Mona Lisa*), `MISC` (entities that can't be classified according to any of the previous subtypes, e.g. *Boeing 747*). The system scores 85.19% precision and 85.91% recall for Number-based expressions and 86.53% precision and 84.94% recall for Name-based expressions.

- LX-Parser is service that provides for the constituency parsing of sentences. LX-Parser is supported by the Stanford Parser[3]. A total of 5422 sentences from CINTIL Treebank were used for training. Under the Parseval metric it achieves an f-score of 88% (value obtained through 10-fold cross-evaluation) [28] [27].
- LX-DepParser allows the parsing of sentences in terms of their grammatical functions. The system was build using the MaltParser[4] trained with Portuguese data. The training was conducted with the help of MaltOptimizer,[5] a tool for automatic tuning of parser parameters. For the training of the parser, 14,052 sentences were used from the CINTIL-Treebank. The system achieves 91.21 of LAS (labeled attachment score).
- LX-SRLabeler permits the constituency parsing and semantic role labeling of sentences. It achieves an f-score of 82% in Parseval metrics (value obtained through 10-fold cross-evaluation).
- CINTIL Concordancer is an advanced concordancer for the CINTIL corpus [1]. It allows the use of generic patterns to specify the occurrences to be retrieved. This permits to uncover linguistic structures of high complexity and use this service as a powerful research tool.
- CINTIL-Treebank Searcher is a service to search and view the parser and dependency tree of the CINTIL Treebank [21]. The searcher allows the use of generic structural patterns of the syntactic trees in order to find those trees in the treebank that conform to these patterns.
- MWN.PT Browser for MultiWordnet of Portuguese, a lexical semantic network for the Portuguese language shaped under the ontological model of wordnets. It spans over 17,200 manually validated concepts/synsets, linked under the semantic relations of hyponymy and hypernymy.
- LX-TimeAnalyzer extracts temporal information from Portuguese texts [16]. Given an input text, it finds the following elements: (i) temporal expressions, which are expressions that occur in the input text and that refer to dates and times; (ii) events terms, which are words that refer to events that happen or hold at some point in time and (iii) temporal relations between these times and events.

Processing Tools and Webservices

- LX-Tokenizer segments text into lexically relevant tokens [14].
- LX-Tagger is a part-of-speech tagger for Portuguese that assigns a single morpho-syntactic tag to every token [13]. Each individual token in multi-token expressions of closed POS classes gets the tag of that expression prefixed by "L" and followed by the number of its position within the expression:

[3] Available on http://nlp.stanford.edu/software/lex-parser.shtml - last access on September 20, 2014

[4] Available on http://www.maltparser.org/ - last access on September 21, 2014

[5] Available on http://nil.fdi.ucm.es/maltoptimizer/ - last access on September 21, 2014

de maneira a que — de/LCJ1 maneira/LCJ2 a/LCJ3 que/LCJ4 This tagger was developed with MXPOST software[6] over a 600k token, accurately hand tagged corpus. Accuracy of 96.24% was obtained with 10-fold cross evaluation.

- LX-Gram is a grammar for the computational processing of Portuguese [4]. It is being developed under the following major design features – (i)precision: it is a precision grammar delivering accurate and linguistically grounded information; (ii)deep processing: provides information on the major syntactic dimensions of grammatical constituency and dependency, it delivers a logical representation of the meaning of natural language sentences; (iii)large-scale:it is planned not to leave out any sort of regular grammatical construction or phenomena; (iv)multi-purpose:it is intended to make available as much linguistic information as it can possible be made explicit by automatic means and (v) technical features: the grammar is developed under the grammatical framework of Head-Driven Phrase Structure Grammar [23] and uses Minimal Recursion Semantics [15] for the representation of meaning.

Applications

- XisQuê is a Question Answering system for the Web of documents written in the Portuguese language [9] [10]. It handles open-domain factual questions. The system handles "Who", "When", "Where" and "Which-X" type of questions. The overall MRR value obtained for XisQuê is 0.73 when short and long answer are considered and 0.48 when only short-answers are taken into account.
- LX-Translator supports speech to speech automatic translator for Portuguese and English languages in both directions [24]. The system was implemented with three main components: automatic speech recognition, statistical machine translation and text-to-speech synthesis. The evaluation of the machine translation module obtained a 0.322 (Portuguese to English) and 0.294 (English to Portuguese) on BLEU score.
- LX-CEFR is a service that supports human experts in their task of classifying text excerpts suitable to be used in quizzes for learning materials and as items of exams that are aimed at assessing and certifying the language level of students taking courses of Portuguese as a second language [8].

3 The CINTIL collection of resources

At present the CINTIL collection includes 20 resources for Portuguese[7]. They are distributed through the international META-SHARE platform[8].

[6] Available on http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html - last access on September 22, 2014

[7] All these resources are listed at and described on http://nlx.di.fc.ul.pt/resources.html

[8] Available on http://metashare.elda.org/

- CINTIL-Corpus Internacional do Português: High quality, linguistically interpreted, accurately hand tagged 1Mtoken corpus with respect to POS, inflection and NER. Developed and maintained in cooperation with CLUL-Centro de Linguística da Universidade de Lisboa [1].
- CINTIL Annotation Manual Companion: manual of CINTIL corpus with explicit guidelines for annotation/interpretation[9].
- CINTIL TagSet: Exhaustive set of part of speech tags for Portuguese, including coverage of transcriptions of verbal productions, used in the annotation of the CINTIL corpus.
- CINTIL-DeepBank: Bank of deep grammatical representations: corpus annotated with their fully-fledged grammatical representations, along with a HPSG grammar [5]. The corpus is composed of 10,039 sentences and 110,166 tokens taken from different sources and domains: news (8,861 sentences; 101,430 tokens), and novels (399 sentences; 3,082 tokens). In addition, there are 779 sentences (5,654 tokens) used for regression testing of the computational grammar that supported the annotation of the corpus.
- CINTIL-Treebank: corpus hand annotated with trees of syntactic constituency. The corpus has the same size as CINTIL-DeepBank [26].
- CINTIL-DependencyBank: corpus of sentences annotated with graphs representing grammatical dependencies, whose arcs are decorated with grammatical functions and semantic roles [26]. The corpus has the same size as CINTIL-DeepBank.
- CINTIL-PropBank: corpus of annotated with trees representing syntactic constituency decorated with grammatical functions and semantic roles [3]. The corpus has the same size as CINTIL-DeepBank.
- CINTIL-LogicalFormBank: corpus annotated with logical forms representing their meaning [26]. The corpus has the same size as CINTIL-DeepBank.
- CINTIL-QATreeBank is a treebank composed of Portuguese sentences that can be used to support the development of Question Answering systems. This Treebank includes 111 declarative sentences from the pre-existing CINTIL-Treebank whose syntactic structure was manually transformed into their non-declarative counterpart: interrogative and imperative clauses [22].
- CINTIL-Definitions:The corpus presented here is a collection of several tutorials and scientific papers in the field of Information Technology with 603 annotated definitions from Portuguese. The texts were collected from the Web at the beginning of the 2006 and they are organized in 32 files of three different sub-domains (information society, information technology and e-Learning) with 268,064 tokens. [20].
- TimeBankPT Corpus annotated with rich temporal annotations, adopting the TimeML conventions [17]. It includes annotations not only of temporal expressions but also about events and temporal relations. This corpus is the result of translating and adapting the English corpus used in the first TempEval challenge to the Portuguese language. It contains around 70,000 words.

[9] The manual annotation can be found at http://nlxserv.di.fc.ul.pt/tagsharecorpus/guidelines.pdf

- LX-Abbreviations: collection of abbreviations of different types composed by 208 words. Each type of abbreviation is manually divided and annotated with grammatical categories, gender and number, and, finally, with the respective full expression [11].
- LX-StopWords: list of words composed by 2631 words of 51 types. The words are grouped in three big classes, arranged according to their morpho-syntactic category and inflectional feature value (closed classes, open classes, and multi-word units) [11].
- MWNPT-International WordNet of Portuguese: developed in cooperation with MultiWordnet project of ITC-Irst from Trento, Italy. It includes the subontologies under the concepts of Person, Organization, Event, Location, and Art works, which are covered by the top ontology made of the Portuguese equivalents to all concepts in the 4 top layers of the English Princeton wordnet. MWNPT spans over 17,200 manually validated concepts/synsets, linked under the semantic relations of hyponymy and hypernymy.
- Nexing Corpus: Corpus with the transcriptions of syllogistic reasoning protocols [6]. The corpus is made of 28 files, with around 15 000 tokens each.
- DeepBankPT: bank of deep grammatical representations sentence aligned with the Penn treebank of English: corpus of sentences annotated with their fully-fledged grammatical representations, along a HPSG grammar [19]. This dataset comprises over 10 000 sentences of newspaper text.
- TreebankPT: sentences aligned with the Penn treebank of English: corpus annotated with trees of syntactic constituency. Like in all corpus whose name is suffixed with ”PT” below, the raw text corpus results from the translation into Portuguese of the WSJ corpus of English. The corpus has the same size as DeepBankPT.
- PropBankPT: sentence aligned with the Penn treebank of English: corpus annotated with trees representing syntactic constituency decorated with grammatical functions and semantic roles. The corpus has the same size as DeepBankPT.
- DependencyBankPT: sentence aligned with the Penn treebank of English: corpus annotated with graphs representing grammatical dependencies, whose arcs are decorated with grammatical functions and semantic roles. The corpus has the same size as DeepBankPT.
- LogicalFormBankPT: bank of logical forms sentence aligned with the Penn treebank of English: corpus annotated with logical forms representing their meaning [2]. The corpus has the same size as DeepBankPT.

4 Final Remarks

The NLX Group has developed the tools and resources for Portuguese presented above, which form some of the more systematic and encompassing collections in this area. The LX-Center[10] web portal centralizes and showcases the tools and resources that have been created in our group, and is updated as new tools and resources are developed.

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

References

1. F. Barreto, A. Branco, E. Ferreira, A. Mendes, M. Nascimento, F. Nunes, and J. Silva. Open resources and tools for the shallow processing of portuguese. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*, 2006.
2. A. Branco. Logicalformalbanks, the next generation of semantically annotated corpora: keys issues in construction methodology. In *Recent Advances in Intelligent Information Systems*. Academic Publishing House EXIT, Warsaw, 2009.
3. A. Branco, C. Carvalheiro, S. Pereira, S. Silveira, J. Silva, S. Castro, and J. ao Graça. A propbank for portuguese: the cintil-propbank. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).
4. A. Branco and F. Costa. A deep linguistic processing grammar for Portuguese. In *Proceedings of the 9th Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR)*, number 6001 in LNAI, pages 86–89. Springer, 2010.
5. A. Branco, F. Costa, J. Silva, S. Silveira, S. Castro, M. Avelãs, C. Pinto, and J. Graça. Developing a deep linguistic databank supporting a collection of treebanks: the CINTIL DeepGramBank. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC)*, pages 1810–1815, 2010.
6. A. Branco, J. Leitão, J. Silva, and L. Gomes. Nexing corpus: a corpus of verbal protocols on syllogistic reasoning. In *LREC*. European Language Resources Association, 2002.
7. A. Branco and F. Nunes. Verb analysis in a highly inflective language with an MFF algorithm. In *Proceedings of the 10th Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR)*, number 7243 in Lecture Notes on Artificial Intelligence (LNAI), pages 1–11. Springer, 2012.
8. A. Branco, J. Rodrigues, F. Costa, J. Silva, and R. Vaz. Rolling out text categorization for language learning assessment supported by language technology. In *PROPOR'2014: Proceedings of the 14th international conference on Computational Processing of the Portuguese Language*, Berlin, Heidelberg, 2014. Springer-Verlag.
9. A. Branco, L. Rodrigues, J. Silva, and S. Silveira. Real-time open-domain qa on the portuguese web. In *IBERAMIA '08: Proceedings of the 11th Ibero-American conference on AI*, pages 322–331, Berlin, Heidelberg, 2008. Springer-Verlag.
10. A. Branco, L. Rodrigues, J. Silva, and S. Silveira. Xisquê: An online qa service for portuguese. In *PROPOR '08: Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, pages 232–235, Berlin, Heidelberg, 2008. Springer-Verlag.
11. A. Branco and J. Silva. Contractions: Breaking the tokenization-tagging circularity. In *Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language*, PROPOR'03, pages 167–170, Berlin, Heidelberg, 2003. Springer-Verlag.
12. A. Branco and J. Silva. Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, pages 507–510, 2004.
13. A. Branco and J. a. Silva. Portuguese-specific issues in the rapid development of state of the art taggers. Technical Report TR-2003-28, University of Lisbon, Faculty of Sciences, Department of Informatics, 2003.

14. A. Branco and J. a. Silva. Tokenization of portuguese: resolving the hard cases. Technical Report TR-2003-4, Departament of Informatics, University of Lisbon, 2003.
15. A. Copestake, D. Flickinger, I. A. Sag, and C. Pollard. Minimal recursion semantics: an introduction, 1999.
16. F. Costa and A. Branco. Lx-timeanalyzer: A temporal information processing system for portuguese. Technical Report TR-2012-01, University of Lisbon, Faculty of Sciences, Department of Informatics, 2012.
17. F. Costa and A. Branco. Timebankpt: A timeml annotated corpus of portuguese. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *LREC*, pages 3727–3734. European Language Resources Association (ELRA), 2012.
18. E. Ferreira, J. Balsa, and A. Branco. Combining rule-based and statistical methods for named entity recognition in portuguese. In *Workshop em tecnologia da Informação e da Linguagem Humana, Anais do XXVII Congresso da Sociedade Brasileira de Computação*, pages 1615–1624, 2007.
19. D. Flickinger, V. Kordoni, Y. Zhang, A. Branco, K. Simov, P. Osenova, C. Carvalheiro, F. Costa, and S. Castro. Pardeepbank: Multiple parallel deep treebanking. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, 2012.
20. R. D. Gaudio, G. Batista, and A. Branco. Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, page 1–33, 2013.
21. P. Gonçalves and A. Branco. Cintil-treebank searcher. In *the I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, 2009.
22. P. Gonçalves, R. Santos, and A. Branco. Treebanking by sentence and tree transformation: Building a treebank to support question answering in portuguese. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *LREC*, pages 1895–1901. European Language Resources Association (ELRA), 2012.
23. C. Pollard and I. Sag. *Head-Driven Phrase Structure Grammar*. Chicago, 1994.
24. J. Rodrigues. Speech-to-speech translation to support medical interviews. Master's thesis, University of Lisbon, Faculty of Sciences, Department of Informatics, 2013.
25. J. Silva. Shallow processing of Portuguese: From sentence chunking to nominal lemmatization. Master's thesis, University of Lisbon, 2007.
26. J. Silva, A. Branco, S. Castro, and F. Costa. 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)*, pages 45–52, 2012.
27. J. Silva, A. Branco, S. Castro, and R. Reis. Out-of-the-box robust parsing of portuguese. In T. A. S. Pardo, A. Branco, A. Klautau, R. Vieira, and V. L. S. de Lima, editors, *PROPOR*, volume 6001 of *Lecture Notes in Computer Science*, pages 75–85. Springer, 2010.
28. J. Silva, A. Branco, and P. Gonçalves. Top-performing robust constituency parsing of Portuguese: Freely available in as many ways as you can get it. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC)*, pages 1960–1963, 2010.