

Port4NooJ Resource Package

This package can be downloaded from:

<http://www.linguateca.pt/Repositorio/Port4Nooj/>

and

<http://www.nooj4nlp.net/> - Resources - pt

Resource Description

Port4NooJ is a set of linguistic resources developed on **NooJ** linguistic environment for the automated processing of the Portuguese language. They integrate a bilingual extension and can also be used in Portuguese to English machine translation.

Original source

The main dictionary is the outcome of the conversion and revision of the single word entries in the English-Portuguese dictionary of the **OpenLogos** system (<http://logos-os.dfki.de/>). The conversion involved looking at each numeric representation code of the **SAL ontology** (<http://logosystemarchives.homestead.com/>) and replacing each automatically with understandable SAL mnemonics from the **Logos Tutorial** in the web application, **LearnLogos**. These SAL mnemonics have not been implemented in the Logos system before. Some adaptations have been made. A few sets and subsets are still not converted and most supersets relate to English and not Portuguese as a source language.

New features

For the NooJ module, we created all new inflectional paradigms, cleaned up Logos dictionary obsolete entries and enhanced it with new properties such as derivation (*DRV* - e.g. noun *apresentação* derives from verb *apresentar*), and annotations for predicate nouns (*NPred*) and for nominalizations (*NPred+Nom*), among others.

Package content

The package has three directories:

1. Lexical analysis/ - contains the main dictionary and a sample of the inflectional and derivational rules
2. Syntactic analysis/ - contains a few grammars
3. Projects/ - contains two sample texts

File description

1. *PT-Dict.dic* - a general dictionary of approximately 40,000 single word entries which includes: (optional) inflectional paradigm information; (optional) derivational paradigm information; syntactic-semantic information; (optional) support verb information; (optional) argument-structure information and English translation for each entry.
2. *PT-Morph.nof* - contains a sample of the Portuguese morphological rules which cover both inflection and derivation. Inflectional rules formalize inflectional paradigms for simple words. E.g. *falamos* (we speak) is the 1st person plural of the verb *falar* (speak) in the present tense. Derivational rules formalize nominalizations. E.g. *citação* (quotation) is a nominalization derived from the verb *citar* (quote).
3. *PT-Contr.nom* - a morphological grammar to process contracted forms such as *das* = *Prep(de)* + *Det(as)* (of the) or *neste* = *Prep(em)* + *DemPron(este)*
4. *PT-Disamb-V-Pron.nog* - a syntactic grammar to annotate and disambiguate pronouns after verbs and before verbs which occur after the relative pronoun *que* (that);
5. *PT-NE-PersonType.nog* - a grammar to recognize and annotate named entities of the type PESSOA (PERSON).
6. *PT2EN-Dates.nog* - a syntactic grammar which recognizes and translates from Portuguese into English dates such as "*terça-feira, 27 de Fevereiro*" (Tuesday, February the 27th)
7. the Portuguese version of the Universal Declaration of Human Rights (*Declaração Universal dos Direitos Humanos*)
8. the XIXth century novel - *Viagens na Minha Terra* - Almeida Garret

Licenses

Files contained in the Port4NooJ Resource Package are open source. The OpenLogos GPL license is also included in the package.

Documentation

1. Port4NooJ Linguistic Resources Overview
2. Port4NooJ tutorial (in Portuguese) - Tutorial Básico do Port4NooJ
3. Barreiro, Anabela, 2008. "Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation". In *Proceedings of the 2007 International NooJ Conference* (Barcelona, Spain, June 7-9, 2007), Cambridge Scholars Publishing, 2008, pp 19-47.

How to cite this package

This package should be cited as "Port4NooJ Resource Package".

Acknowledgments

This work was developed by Anabela Barreiro within the scope of a PhD program that was supported by Fundação para a Ciência e Tecnologia, Programa POS_C through grant SFRH/BD/14076/2003. Many thanks to Max Silberztein and Slim Mesfar for their support on technical aspects of NooJ, to Andrei Telenkov for helping with OpenLogos data extraction, and to Linguatca for their support in developing, enhancing and making these resources publicly available.