Some Language Resources and Tools for Computational Processing of Portuguese at INESC

Luzia Wittmann, Ricardo Daniel Ribeiro, Tânia Pêgo, Fernando Batista

Instituto de Engenharia de Sistemas e Computadores
Rua Alves Redol, 9 - Apartado 13069 - 1000-029 LISBOA - Portugal
{luzia.wittmann, ricardo.ribeiro, tania.pego, fernando.batista}@inesc.pt

Abstract

In the last few years automatic processing tools and studies based on corpora have became of a great importance for the community. The possibility of evaluating and developing such tools and studies depends on the availability of language resources. For the Portuguese language in its several national varieties these resources are not enough to meet the community needs. In this paper some valuable resources are presented, such as a multifunctional lexicon, general-purpose lexicons for European and Brazilian Portuguese and corpus processing tools.

1. Introduction

Reusable and multifunctional language resources of Portuguese available are rather scarce when compared to resources available for languages such as English, French and German[1]. The Natural Language Group of INESC (R&D Institution) and the Centro de Linguística da Universidade de Lisboa have been dedicating part of their efforts to the development of such written language resources for Portuguese. Both institutions have been participating in European projects like PAROLE, SIMPLE and ELAN, in National funded projects and promoting internal studies on this area.

The objective of this document is to describe some language resources and tools for computational processing of (written) Portuguese that have been developed at INESC. Section 2 contemplates a description of a corpus based lexicon including frequency information. In section 3 lexicons for the European and the Brazilian national varieties of Portuguese are presented, focusing the problematic issue of the national varieties contrasts. A set of corpus processing tools - Palavroso System - is presented in section 4. This system is composed of several tools, but the most relevant ones are a morphological analyser - the core of the system -, a morphological disambiguator, a browser/retrieval and statistical system for corpora extraction, a morphological generator and a spelling checker.

2. Multifunctional Computerised Lexicon of Contemporary Portuguese

This lexicon is being developed within the "Multifunctional Computerised Lexicon of Contemporary Portuguese" project, with the participation of CLUL-Centro de Linguística da Universidade de Lisboa (Coordinator partner), INESC, Verbo Editora and ILC (Instituto de Linguistica Computazional). The project is funded by the national agency Fundação para a Ciência e Tecnologia (Foundation for Science and Technology).

[1]References to existent language resources for Portuguese can be found at www.portugues.mct.pt, an official site of Ministério da Ciência e da Tecnologia, organised by Diana Santos.

The lexicon contains about 30 000 entries (lemmas), extracted from a corpus of 15 000 000 lexical occurrences of spoken and written contemporary Portuguese. Each lexical item (lemma) is being encoded with quantitative information (frequency and repartition data, both oral and written, in percentages), and morpho-syntactic classification.

2.1. Population

The 15 million word corpus is actually a sub-corpus, named CORLEX, drew out from the Corpus de Referência do Português Contemporâneo. CORLEX is made of 1 million spoken words and 14 million written words, pulled out of didactic, literary, miscellaneous, newspaper, periodical (magazine) and techno-scientific texts. 198 066 tokens were extracted, in order to obtain a lexicon of about 30 000 words. The word forms were then automatically lemmatised and morpho-syntactically classified using the INESC's tagger Palavroso. A manual verification of the morpho-syntactic classification was done, added by corpus consultation and following a criterion previously established.

2.2. Frequency information

Frequency calculation is based on the PAROLE Portuguese Annotated Sub-corpus. This sub-corpus is made of 250 thousand running words coming from newspapers, books, periodicals and miscellaneous. It was annotated with Palavroso and manually disambiguated on the morpho-syntactic level. In order to retrieve linguistic and quantitative information from the corpora, INESC developed a tool named Encontra&Estatic. It can handle very large corpora of Portuguese and it is a basic and essential tool for linguistic and statistical information extraction. Based on a morphological analyser, it also allows queries and collocation extraction using morphological classes in non annotated corpora. (see section 4, subsection 4.3)

2.3. Format

In order to assure the reusability of the lexicon, it is stored in SGML format and European standards proposed by projects like NERC - Network of European Reference

Corpora -, EAGLES - Expert Advisory Group for Language Engineering Standards - and PAROLE are observed.

2.4. Applications

This project exploits one of the results of the Telematics Project LE-PAROLE and offers a new important product for the Portuguese language processing development. It is indeed the first fully corpora based lexicon, including frequency information for the European variant of Portuguese.

The creation of this lexicon promises great strategic significance in the development of related areas, providing for pilot applications in different domains such as: establishment of lexical entries, identification and specification of linguistic phenomena, establishment of descriptive categories, creation of corpora and lexicon checking systems and also improvement of tools for automatic analyses and disambiguation.

3. Lexicons for European and Brazilian variants of Portuguese

Lusolex and Brasilex are two multifunctional and easily reusable lexicons for, respectively, European and Brazilian Portuguese. Lusolex contains about 65 thousand entries and Brasilex contains about 68 thousand, and respective morphological inflection paradigms. The two lexicons have a common core and are stored in the same format.

The significance of this work is related to the importance of developing NLP tools, thus, language resources, covering National Varieties of the same language. If it is important to support and incentivate the development of NLP for all languages, in order to respect and guarantee the cultural diversity and specific views of each language community, it is not less important to study and assess the differences between national varieties for computational processing.

In the next subsections each lexicon will be briefly described and some of the differences will be referred to.

3.1. Lusolex

Lusolex (EP lexicon) was based mainly on the INESC's Palavroso lexicon and completed with corpora extraction. The Palavroso (see section 4) inflection rules were converted from a lexicon independent rule system into a paradigmatic one. The lexicon was completely revised and all classification and inflection paradigms were checked. About 5 thousand new words were extracted from corpora (including Parole Portuguese Corpora) and other sources (completion of word families, domains, etc.) and added to the lexicon.

3.2. Brasilex

Brasilex (BP lexicon) was created using the Palavroso lexicon and, afterwards, Lusolex (EP lexicon) as a base. Lexical items used only in EP were then excluded and new items used only in BP were added. These BP words were gathered from several different sources, but mainly from dictionaries and corpora. Most of these words are originally from Indian and African languages, integrated in the Brazilian Portuguese. For the coincident words, i.e. used both in EP and in BP, the morphological classification was checked, as well as the corresponding inflection rules, in order to identify any contrast at this level.

3.3. Format

The lexicons are stored in SGML format. The morphological encoding follows a paradigmatic model designed with the background and same philosophy of the Parole Lexicon morphological layer encoding. Each lexical entry is encoded with an ID, grammatical class and subclass information and the correspondent inflection paradigm.

All entries of the lexicon have the following format:

ID	Word	GC	GSC	InP	Stem0 Stem1 ...

The inflexion paradigms are described by

```
ID Example
        MorphoFeat StemNumber Withdraw Add
        ...
$
```

The lexicon is in ASCII format and the character set used is the standard ISO-8859-1, also known as ISO-Latin-1. Concerning the entries format:

Field	Description
ID	unique identifier for each entry
Word	spelling of the word
GC	grammatical category of the word
GSC	grammatical subcategory of the word
InP	id of the inflexion paradigm of the word
StemN	stems needed to the inflexion paradigm (it is necessary to have at least one stem, probably the word itself)

Table 1: Entries format.

3.4. Differences

As the lexicons are encoded only on the morphological layer, the differences observed are of three types only: morphological, orthographic and lexical differences. The orthographic contrasts cover different spelling in each variant (ex. projecto-projeto, pinguim-pingüim, ideia-idéia), and represent about 2.4% of the entries. The morphological differences found are mainly different derivation as in doutoramento-doutorado. Several different grammatical categorisations were also observed, but these are due to different meanings of the same entries. The so-called lexical differences correspond here only to absolute contrasts, which correspond to words used only in one of the varieties (ex. PE: autocarro - PB: bonde) and some relative contrasts, which correspond to words with frequent use in one variety and hardly any use in the other. These lexical differences represent about 6% of the lexicon.

Nevertheless, the lexical differences that appear in this analysis represent only part of the whole contrastive problematic issue. On the semantic layer, for instance, the differences are much more complex, since the absolute word contrasts (completely different words in each variant for the same sense as in bilingual dictionaries), do not cover the whole range of differences relevant for Language Engineering.

The relative contrasts assume a high complexity when considering the combination of use frequency following the different senses of a word and the exclusion of one or more senses of the word in one of the varieties. There are words with a high use frequency in one variant and a very low one in the other one. In these cases, even if the word is used with the same sense in both variants, the use frequency has to be taken into account as a contrast for LE applications. Cases where words have a slight difference in its conceptual definition should also be considered. Other problematic contrasts are related to official systems as scholar graduation, official institutions such as Government organisation etc. Words referring to specific cultural and physical phenomena of a country (religion, cookery, flora and fauna) also demand a reflection, since words as, for example, the name of some Brazilian tropical fruits are well known in Portugal, but others are not; the frontiers are fuzzy. Besides, nowadays influence of BP over EP mainly through television has to be considered.

The creation of Lusolex and Brasilex should be considered as a first step of the project. The morphologic encoding is expected to be completed by syntactical and semantic encoding in the future, creating something like a bi-variant lexicon. The complexity of the lexical differences between EP and BP was raised during an experience in constructing a contrastive dictionary of EP and BP for Machine Translation. The experience was interrupted, but about 4,000 contrastive pairs were achieved.

4. Corpus Processing Tools: the Palavroso System

Palavroso is a morphological processing system for Portuguese conceived to cover largely the written language.

The objectives that driven its development comprise a direct use, as a tool for linguistic research, and an indirect use, as an element of broader applications in the area of the Computational Processing of Portuguese.

The main component of Palavroso is a morphological analysis module, fundamental for the development of applications that require services in the area of natural language processing, such as computational grammars, machine translation, text analysis or text retrieval.

This system is composed of several tools. The most relevant ones - the morphological analyser, the part-of-speech disambiguator and the browser - are going to be presented in more detail. The others are described in the following table.

Tool	Description
Compila	Tool used for the dictionary creation.
Léxico	Lexicon collector. It uses the morphological analyser to explore corpora, registering all words that do not exist in the used dictionary.
Gerando	Morphological generator. It is able to generate all word forms of a verb, noun, adjective or adverb present in the dictionary.
Correcto	Spell checker.
Desambig	Textual interface for the disambiguation of texts in an interactive way.

Table 2: Some of Palavroso tools.

4.1. The morphological analyser

The morphological analysis module - or the morphological analyser - is the core of Palavroso. All tools are clients of the analyser. This module is a rule oriented morphological analyser to which a lexicon can be attached. This approach was motivated by the need to build a morphological analyser that could work without a lexicon (Santos et al., 1992; Medeiros et al., 1993; Medeiros, 1995). The advantage of such a system is of always getting an answer, even if the input word (i.e. its lemma) is not present in the dictionary. Another important feature of the analyser is the well defined separation between the lexicon and the analysis rules.

Palavroso has two standard lexicons, a European Portuguese lexicon with about 61 000 entries and Brazilian Portuguese lexicon with about 65 000 entries.

The morphological analyser was developed to address specific problems of the Portuguese language like composed words, enclitic pronouns and adjectives degree. It is possible to use it interactively and as a tagger, through the interface known as Morfolog. It is also possible to specify several operation modes adequate to the users needs.

4.2. The part-of-speech disambiguator

The part-of-speech disambiguator is being developed in the context of the "Multifunctional Computerised Lexicon of Contemporary Portuguese" project. It is now reaching a prototype version.

The development of the disambiguator was based on PAROLE Portuguese sub-corpus of 250 000 running words manually disambiguated. Nevertheless, the disambiguator is a hybrid tool. It uses both probabilistic and rule-oriented methods to disambiguate the input texts.

The probabilistic approach is based on HMMs (Hidden Markov Models). The language model was obtained from the above mentioned corpus. This part of the system is based on the work of Church (1988). The linguistic rule-oriented approach tries to capture significant patterns from corpora and establish rules from this patterns. It is relevant to mention the linguistic work done by NLG team. This part has as references the work done by Voutilainen (1995), Tapanainen and Voutilainen (1994) and Chanod and Tapanainen (1994).

4.3. The browser

The browser known as Encontra&Estatic is an application for extracting contextual, morphological and statistical information from corpora. It can handle very large corpora and was designed considering the specific aspects of the Portuguese language. Using the morphological analyser Palavroso in order to deal with corpora in the morphological context, it not only supplies word superficial information, but also takes into consideration characteristics such as grammatical class, gender, number, stem, etc. As it is based on a morphological analyser, it can be used over annotated and non annotated corpora, allowing queries and collocation extraction using morphological classes in non annotated corpora.

With Encontra&Estatic it is possible to extract the information needed for the probabilistic calculus used to develop disambiguation tools of probabilistic type. On the other hand, it also facilitates the establishment of patterns, which are the basis to infer linguistic rules for disambiguation tools.

Encontra&Estatic was developed in C language for extended portability and performance. This application can be split in to two different modules: Encontra, which is used for corpora searching, and Estatic, which provides different kinds of statistical information. Encontra&Estatic can be used in two different ways: as a module or as standalone application. Using it as a module the access is made through its API (Application Programming Interface), which provides methods for setting options, performs queries and gets results. As a standalone application, the user specifies command line arguments and gets results as text. The API provides extended capabilities once it can be included in and used by many other applications.

5. References

Anabela Barreiro, Luzia Wittmann, and Maria de Jesus Pereira. 1996. Lexical differences between european and brazilian portuguese. *The INESC journal of Research and Development*, 5(2):75–101.

Jean-Pierre Chanod and Pasi Tapanainen. 1994. Statistical and constraint-based taggers for french. Technical Report MLTT-016, Rank Xerox Research Center, Grenoble, Switzerland.

Kenneth W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Procedings of the Second Conference on Applied Natural Language Processing (ANLP-88), ACL*, Austin, Texas, USA.

José Carlos Medeiros, Rui Marques, and Diana Santos. 1993. Português quantitativo. In *Actas do 1º Encontro de Processamento de Língua Portuguesa EPLP'93*, pages 33–37, Lisbon.

José Carlos Medeiros. 1995. Processamento morfológico e correcção ortográfica do português. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon.

Diana Santos, Carla Fernandes, Rui Marques, and José Carlos Medeiros. 1992. Gramática sem dicionário: Relatório preliminar. Technical Report RT/15-92, INESC, Lisbon.

Pasi Tapanainen and Atro Voutilainen. 1994. Tagging accurately - don't guess if you know. In *Proceedings of the 4th Conference on Applied Natural Language Processing, ACL*, Stuttgard, Germany.

Atro Voutilainen. 1995. Morphological disambiguation. In Fred Karlsson, Atro Voutilainen, Juha Heikkila, and Arto Anttila, editors, *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin and New York.

Luzia Wittmann and Maria de Jesus Pereira. 1994. Português europeu e português brasileiro: alguns contrastes. In *Actas do X Encontro Nacional da Associação Portuguesa de Linguística*, Évora.

Luzia Wittmann and Ricardo Ribeiro. 1998. Recursos linguísticos e processamento morfológico do português: o palavroso e o projecto le-parole. In Vera Lúcia Strube de Lima, editor, *Actas do III Encontro para o Processamento Computacional de Português Escrito e Falado (PROPOR'98)*, pages 109–117, Porto Alegre, Brasil.

Luzia Wittmann, Tânia Pêgo, and Diana Santos. 1995. Português brasileiro e português de portugal: algumas observações. In *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística*, pages 465–487, Lisbon.