Portuguese Large-scale Language Resources for NLP Applications

Elisabete Ranchhod[1], Paula Carvalho[1], Cristina Mota[2], Anabela Barreiro[1]

[1]Universidade de Lisboa and LabEL (CAUTL/IST), [2]LabEL (CAUTL/IST)
Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
{elisabet, paula, cristina, anabela}@label.ist.utl.pt

Abstract

The paper describes Portuguese large-scale linguistic resources, mainly computational lexicons and grammars, developed by LabEL. These resources are formalized and applied to texts by means of finite-state techniques, more and more acknowledged in Natural Language Processing. On the one hand, it illustrates methods on lexical representation for simple words and multi-word expressions; on the other hand, it provides examples (in form of concordances) of linguistic structures recognized after the application of disambiguation and parsing grammars to texts. The paper ends with a short reference to the publicly available data highlighting its contribution towards dissemination of LabEL's knowledge on language technology.

1. Introduction

The increasing interest in NLP has pointed out the growing needs for linguistically precise broad-coverage language resources. In this context, LabEL has been developing large-scale lexicons and grammars for Portuguese. These language data are formalized and applied to texts using finite-state techniques.

Automata and finite-state transducers (FST) are particularly suitable for easy and compact representation of different types of linguistic data. They reduce space and time overhead in text processing operations.

Finite state technology is used to:

(i) Build electronic dictionaries, by formalizing and generating precise linguistic information related to simple and multi-word lexical units: e.g. *linguística*, linguistics; *linguística computacional*, computational linguistics.

(ii) Develop grammars for lexical disambiguation (POS, lemma, inflectional information, etc.): e.g. *ama*, nurse (N) and loves (V); *forte*, fort (N) and strong (A); *cobre*, cooper (N), covers (V) and collects (V).

(iii) Develop local grammars for identification and tagging of linguistic expressions with strong lexical-syntactic constraints (such as: adverbials of time and space, units of measure, etc.): e.g. *há cerca de dez anos*, about ten years ago; e*stão 37,5ºC à sombra*, it's 37,5ºC in the shade.

(iv) Develop grammars to parse different syntactic constructions, such as noun phrases, and complex predicates.

The Portuguese resources are integrated in two public FST based corpus processors, INTEX and UNITEX.

2. Electronic Dictionaries

The lexicon is the foundation of any sound NLP application. LabEL's dictionary system consists of a set of modules, each organized according to the formal complexity of the lexical units it represents.

2.1. Simple Word Dictionaries

The core module of the dictionary system contains about 120,000 simple words (lemmas), each having its own systematically encoded morphological attributes. Codes specify information about the particular entry POS and inflectional information, such as gender, number, person, case, tense, mood, diminutives, augmentatives, and superlatives, that can change according to the POS involved. Sample 1 illustrates a few entries of the simple word dictionary.

simples,A116+Det	[simple]
simples,A116.ss024.sr001+Pd	[simple]
sobreviver,V102x	[survive]
sol,N213.dh247.dt247	[sun]

Sample 1: Simple word lemmas

Syntactic and semantic information is being encoded incrementally. For instance, adjectives are being refined with information about their syntactic sub-classification. Such refinement allows the separation of homograph adjectival entries on a formal basis. For instance, *simples*, represented above, is described in two different entries, which correspond to both predicative (*Pd*) and determinative (*Det*) adjective analyses, as in examples (1) and (2) below.

(1) This is a very simple question
(2) He did not find a simple reason to come

The inflected simple forms (about 1,250,000) are system generated from the inflectional FSTs referenced in lemmas. For instance, FST *A116.ss024.sr001* allows the generation of inflected forms for the predicative adjective *simples*, as well as for other adjectives with similar morphological behavior (invariable adjectives which can inflect by means of the superlative morphemes *–íssimo* and *-érrimo*). FSTs also assign linguistic information corresponding to each generated form, as in Sample 2.

simples,simples.A+Pd:ms:fs:mp:fp
simplérrima,simples.A+Pd:Sfs
simplérrimas,simples.A+Pd:Sfp
simplérrimo,simples.A+Pd:Sms
simplérrimos,simples.A+Pd:Smp
simplicíssima,simples.A+Pd:Sfs
simplicíssimas,simples.A+Pd:Sfp
simplicíssimo,simples.A+Pd:Sms
simplicíssimos,simples.A+Pd:Smp

Sample 2: Inflection of *simples*

2.2 Multi-Word Dictionaries

It is impossible to envisage automatic text analysis without adequate identification and treatment of multi-word lexical

units. The meaning of a text is mostly supplied by frequent occurrence of multi-word units, especially compound nouns. Therefore, most NLP applications (Information Extraction and Retrieval, Machine Translation, Text Summarization, etc.) can not succeed if such expressions are not processed properly.

LabEL has given priority to the gathering and formalization of common usage compound nouns (e.g. *hora de ponta*, rush hour) and adverbs (e.g. *de cor e salteado*, by heart), but it has also been developing mechanisms to recognize and classify multi-word expressions commonly known as named entities (e.g. ACL – Association for Computational Linguistics).

2.2.1. Dictionaries of General Multi-Word Expressions

The dictionary of general multi-word expressions, still under development, contains about 79,000 entries.

Multi-word expressions are sequences of simple words, but their meaning is not always compositional. They present morphological, combinatorial, and other linguistic constraints. Some of them are completely invariable (most adverbs, conjunctions, prepositions, determiners), while others (most nouns and adjectives) generally inflect in gender and number, as in Sample 3.

caixa(N301)-forte(A301),N+NA+Conc	[safe]
lua(N301) de mel,N+NDN+Abst	[honey moon]
pele(N101) vermelha(A101),N+NA+Hum	[redskin]
novo(A001.dh012.ss001) em folha,A+APN+Pd	[brand-new]

Sample 3: Compound words

The elements that can inflect are specified with inflectional codes similar to those used to inflect simple words. However, these codes may not be always the same. This is the case of *pele vermelha* above. The code for *pele* in the simple word dictionary is N301 (feminine noun); in the multi-word dictionary, the same form is codified as N101 (both masculine and feminine noun). On the other hand, in the simple word dictionary, the feminine form *vermelha* is obtained from the adjectival lemma *vermelho* by means of the code A001; in the multi-word dictionary, *vermelha* is coded as A101, a feminine or masculine adjective, corresponding to feminine or masculine *pele*.

Thus, from the compound entries presented in Sample 3, the following inflected forms are generated as can be seen in Sample 4.

caixa-forte,caixa forte.N+NA+Conc:fs
caixas-fortes,caixa forte.N+NA+Conc:fp
lua de mel,lua de mel.N+NDN+Abst:fs
luas de mel,lua de mel.N+NDN+Abst:fp
pele vermelha,pele vermelha.N+NDN+Hum:ms:fs
peles vermelhas,pele vermelha.N+NDN+Hum:mp:fp
nova em folha,novo em folha.A+APN+Pd:fs
novas em folha,novo em folha.A+APN+Pd:fp
novo em folha,novo em folha.A+APN+Pd:ms
novos em folha,novo em folha.A+APN+Pd:mp
novinha em folha,novo em folha.A+APN+Pd:Dfs
novinhas em folha,novo em folha.A+APN+Pd:Dfp
novinho em folha,novo em folha.A+APN+Pd:Dms
novinhos em folha,novo em folha.A+APN+Pd:Dmp
novíssima em folha,novo em folha.A+APN+Pd:Sfs
novíssimas em folha,novo em folha.A+APN+Pd:Sfp
novíssimo em folha,novo em folha.A+APN+Pd:Sms
novíssimos em folha,novo em folha.A+APN+Pd:Smp

Sample 4: Inflected compound words

Semantic and syntactic attributes have been added to compound nouns and adjectives. Simple lexical units are often ambiguous, and in most cases it is only possible to determine their meaning in context. Multi-word units are much less ambiguous. Therefore, it is straightforward to assign them precise semantic information. For instance, simple words such as *caixa* (box; cashier; etc.) and *forte* (fort; strong; etc.) are ambiguous at different levels; however, when combined in the multi-word noun *caixa-forte* (safe), they lose their ambiguity, and the entry can be sub-classified as a concrete noun.

2.2.2. Specific Lexicons – Named Entity Expressions

Named-entity is a term usually used to name specific entities, such as persons, organizations, and locations. Most times, they correspond to capitalized multi-word expressions. Such expressions are classified and extracted semi-automatically from corpora by means of finite-state transducers, and are manually verified by linguists. According to their semantic nature, named entities are organized in different specific lexicons, which need to be expanded. Such lexicons are briefly presented below.

- Collective proper nouns

The collective proper nouns module presently contains about 7,500 multi-word lexical units, which handles educational institutions, non-profit organizations and corporate entities (e.g. *Instituto Superior Técnico*, *Fundo Monetário Internacional*).

- Toponyms

The toponym module includes around 3,300 entries, which correspond to Portuguese or Brazilian places (e.g. *Castelo Branco, Rio de Janeiro*) and other foreign country places, spelled either in Portuguese (e.g. *Grã-Bretanha*, Great Britain; *Nova Zelândia*, New Zealand) or in their original language, whenever there is no equivalent term in Portuguese (e.g. *Buenos Aires, Manhattan*).

- Acronyms

Named entities often appear in texts either as acronyms or next to them. Acronyms create problems concerning formalization because of their linguistic specificity (sequences of letters corresponding to sequences of words). Acronyms, like *UE* (EU), are formally simple words. However, they stand for multi-word expressions: *União Europeia* (European Union). Different interconnected modules, formally similar to simple and multi-word dictionaries, were created in order to recognize and relate both types of linguistic objects. At present, 5,100 pairs of acronyms and their multi-word correspondents have been collected and formalized. Samples 5 and 6 show both types of representation.

BA,ba.N+SigE+HumCol:fs
EUA,eua.N+Sig+Top:mp
CML,cml.N+Sig+HumCol:fs

Sample 5: Acronyms

British Airways,ba.N+DSigE+HumCol:fs
Estados Unidos da América,eua.N+DSig+Top:mp
Câmara Municipal de Lisboa,cml.N+DSig+HumCol:fs

Sample 6: Acronym multi-word correspondents

Both dictionary modules are simultaneously applied to corpora by a specific FST, so that both acronyms and their full word correspondents are extracted from corpora whenever they co-ocurr, as illustrated in Concordance 1.

s» que o Banco Comercial Português (BCP) encont
mbros do Comité Olímpico Internacional (COI) te
do com o Fundo Monetário Internacional (FMI). P
pação do Grupo Espírito Santo (GES) na privatiz
lanta do PRS (Partido da Renovação Social) desp
própria União Europeia (UE) embaraçaram o Pres

Concordance 1: Acronyms and their multi-word correspondents

2.3. Numerical Expressions

Numerical expressions such as cardinals, ordinals and Roman numerals are represented not in dictionaries but in lexical FSTs. Since elements of each class can combine dynamically with one another to form new lexical units, it would not be adequate to list them exhaustively. Transducers allow, on the one hand, the easy and compact encoding of the different combinatorial sequences, and, on the other hand, the mapping from (lexical) numerals to their corresponding numbers and vice-versa.

3. Grammars

Parallel to the development and improvement of electronic dictionaries, LabEL has been building grammars for sentence identification, lexical disambiguation, recognition and tagging of different linguistic constructions.
Grammars are applied to texts in combination with dictionaries, and they are corpora independent.

3.1. Disambiguation Grammars

In Portuguese, noun and adjective homographs are a significant source of POS ambiguity. In order to decrease or eliminate this kind of ambiguity, a set of about 100 disambiguation grammars have already been built. Disambiguation grammars are represented by FSTs, and they describe linguistic constraints (morphological, syntactic, distributional) that should be observed between noun phrase constituents, namely between nouns and adjectives. These grammars make use of the linguistic information formalized in the dictionaries, in particular of adjectival sub-classification attributes.
Some constructions with predicative adjectives are expressed in the FST presented in Figure 1.

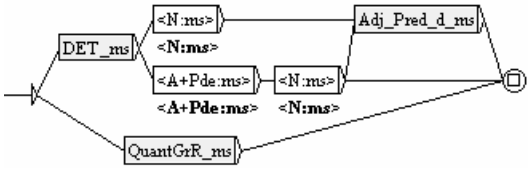

Figure 1: Disambiguation grammar

The FST analyzes ambiguous nominal constructions containing: a determiner or a combination of determiners, formalized in the sub-graph *DET_ms*, a noun, which can be preceded by a sub-type of predicative adjective (*Pde*), and a predicative adjectival phrase, represented in the sub-graph *Adj_Pred_d_ms*. There is morphological agreement among all categories of the noun phrase (masculine singular).
Concordance 2 shows some nominal constructions resulting from the application of the disambiguation grammar presented before to a journalistic corpus, which will be referred further ahead.

saber se este novo meio de transporte permite d
. E como o contribuinte português já se habituo
amilton, o mais famoso restaurante da capital.
eçado. O mais importante, agora, era traçar u
ível ter o primeiro-ministro mais simpático da
erto que o sorriso insípido, incolor e inodoro
unais. O único elemento externo que é novo e
dado por um discurso pausado mas inflexível, Va
promete um futuro tão excitante e radioso, o l
apricho, um jovem muito pouco artístico». Curio
ambém de um poder político federal forte, como

Concordance 2: NP identification

In order to estimate the disambiguation grammars performance, they were applied to a test corpus, a journalistic non-annotated text with about 163,000 words. As the result of the application, a set of 3,657 syntactic structures were matched, and 94,1% of the words analyzed by the grammars were disambiguated and tagged. The remaining 5,9% were not tagged because the ambiguity could not be completely eliminated. Only 0,7% of the tagged words were incorrect, which means that the precision rate is 99.3%. The recall was not calculated.

3.2. Local Grammars

Local grammars are used to represent constructions exhibiting strong lexical-syntactic constraints. Dates (e.g. *Sexta-feira, 31 de Outubro de 2003*, Friday, October 31st 2003), temporal expressions (e.g. *há duas semanas atrás*, two weeks ago) and percentages (e.g. *entre os 3 e os 6%*, from 3 to 6 %), among other constructions belonging to a particular semantic domain, are examples of linguistic data formalized in local grammars (FSTs).

3.3. Grammars to Parse Complex Predicates

Grammars to parse complex predicates were also built. A complex predicate corresponds to a constrained sequence of auxiliary verbs preceding a main verb or a predicative adjective, which are the syntactic and semantic nucleus of a sentence structure. In Portuguese, these strings may contain adverbs, clitics and other inserts that do not belong to the predicate. Enhanced FSTs (i.e. transducers that allow the reordering of the FSTs outputs) were used to formalize such combinations, and to move the inserts from inside the predicate to their canonical positions.
The incomplete FST represented in Figure 2 corresponds to a grammar fragment describing possible combinations of temporal and aspectual auxiliaries, preceding a main verb in the infinitive.
Aspectual auxiliaries are lexically specified in the main graph; temporal auxiliaries are represented in the embedded FST *TempK*. Sub-graphs *ClitH* and *Adv* describe some of the constituents that can be inserted either between auxiliaries or between auxiliaries and main verb. Concordance 3 illustrates complex predicates recognized by such grammar.

como tem estado a decorrer o ciclo de debates
alho tem vindo assim a tornar-se cada vez mais
lvez tenha começado a perder a conta em 1959, q
essa ter estado quase a desistir durante a leit
ora. Têm-se vindo a acumular as críticas estuda
que tinha ficado a gerir o Teatro Camões/Sala
se tivesse continuado a vender ao preço anter
rixa tivesse estado já para acontecer. «Na terç

Concordance 3: Identification of complex predicates

The tagging of auxiliaries (V+Aux) and the constituent reordering is illustrated in Concordance 4.

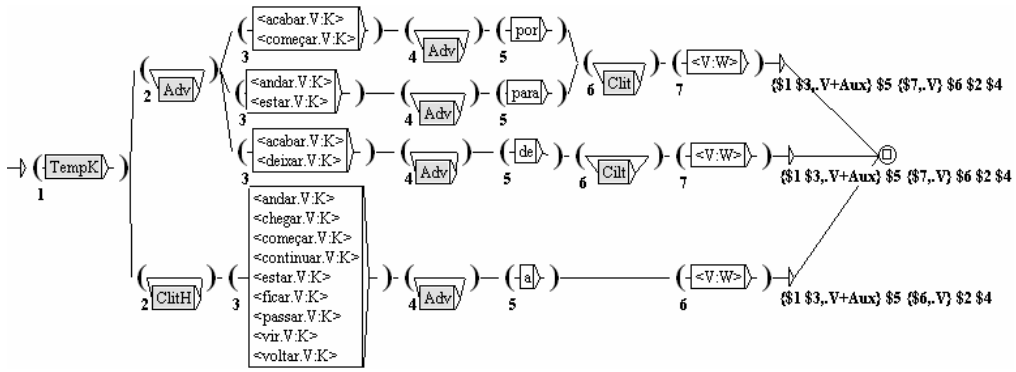

Figure 2: Grammar to parse complex predicates (fragment)

como {tem estado,.V+Aux} a decorrer o ciclo de
alho {tem vindo,.V+Aux} a tornar assim -se cada
lvez {tenha começado,.V+Aux} a perder a conta e
essa {ter estado,.V+Aux} a desistir quase duran
ora. {Têm vindo,.V+Aux} a acumular -se as críti
que {tinha ficado,.V+Aux} a gerir o Teatro Cam
se {tivesse continuado,.V+Aux} a vender ao pr
rixa {tivesse estado,.V+Aux} para acontecer já.

Concordance 4: Auxiliary tagging and constituent reordering

4. Knowledge and Technology Transfer

An important part of the lexical resources presented here is freely distributed for research purposes from the team website (http://label.ist.utl.pt). New linguistic data will be made available periodically.

A free system for tagging corpora via web, AnELL (http://www.linguateca.pt/AnELL/), has also been developed by LabEL, in cooperation with Linguateca. LabEL's linguistic resources are used by the INTEX corpus processor to produce the linguistic annotation of the corpora. Texts to be annotated may be directly typed in by the users, being the automatically annotated texts presented instantly. Alternatively, the user may upload the corpus and the result, after email notification, can be downloaded later from the website.

5. Final Remarks

In last years, some believed that statistical NLP had rendered linguistic analysis unnecessary. In fact, this is not the case. It is now widely recognized that linguistically precise broad-coverage language resources are the basis for ongoing product development in a number of application areas.

Within the scope of this paper, the main characteristics of a set of Portuguese large-scale language resources were summarized. We believe that (i) these formalized language data are a requirement for language engineering but also for theoretical linguistics and education; (ii) the availability of such broad-coverage linguistic-based resources should offer valuable opportunities to train statistical models on the output of linguistic-based processed corpora.

Acknowledgments

This work was partially supported by Fundação para a Ciência e Tecnologia, grant POSI/39806/ PLP/2001.

References

Baptista, Jorge (1995). *Estabelecimento e formalização de classes de nomes compostos*, Tese de Mestrado, Faculdade de Letras da Universidade de Lisboa.

Carvalho, P. (2001). *Gramáticas de Resolução de Ambiguidades Resultantes da Homografia de Nomes e Adjectivos*. Tese de Mestrado, Faculdade de Letras da Universidade de Lisboa.

Carvalho, P.; C. Mota; E. Ranchhod (2002). Complex Lexical Units and Automata. In Ranchhod, E. M. and Nuno J. Mamede (eds.) *Advances in Natural Language Processing*, Proceedings of PorTAL 2002, LNAI 2389 (pp. 229-238), Heidelberg: Springer.

Carvalho, P. and E. Ranchhod (2003). Analysis and Disambiguation of Nouns and Adjectives in Portuguese by FST. In *Proceedings of the Workshop on Finite State Methods in Natural Language Processing* (pp. 105-112), EACL'03.

Gross, M. (1988). Methods and Tactics in the Construction of a Lexicon-Grammar. In *Linguistics in the Morning Calm*, selected papers from SICOL-86. Seoul: The Linguistic Society of Korea.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to natural Language Processing, Computational Linguistics, and Speech Recognition*, New Jersey: Prentice Hall.

Mohri, M. (1997). Finite-State Transducers in Language and Speech Processing. *Computational Linguistics* 23:2 (pp. 269-312).

Mota, C.; Moura, P., (2003). ANELL: A Web System for Portuguese Corpora Annotation. In Mamede, N. J.; Baptista, J.; Trancoso, I.; Volpes Nunes, M. G. (eds.), Computational Processing of the Portuguese Language, LNAI 2721 (pp. 184-188), Berlin: Springer.

Paumier, S. (2002). *Unitex - manuel d'utilisation*. Rapport de recherche, IGM, Université de Marne-la-Vallée.

Ranchhod, E. M. 2001. O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais. In Ranchhod, E. M. (org.), *Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações* (pp. 13-47), Lisboa: Caminho.

Roche, E. and Schabes, Y. (eds.) (1997). *Finite-State Language Processing*. MIT Press, Cambridge, Massachusetts.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse lexicale du français. Le système INTEX*, Paris: Masson.