A Computational Lexicon of Portuguese for Automatic Text Parsing

Elisabete RANCHHOD
FLUL/CAUTL-IST
Av. Rovisco Pais, 1
1049-001 Lisboa, Portugal
elisabet@label.ist.utl.pt

Cristina MOTA
CAUTL-IST
Av. Rovisco Pais, 1
1049-001 Lisboa, Portugal
cristina@label2.ist.utl.pt

Jorge BAPTISTA
UALG/CAUTL-IST
Av. Rovisco Pais, 1
1049-001 Lisboa, Portugal
jbaptis@ualg.pt

Abstract

Using standard methods and formats established at LADL, and adopted by several European research teams to construct large-coverage electronic dictionaries and grammars, we elaborated for Portuguese a set of lexical resources, that were implemented in INTEX. We describe the main features of such linguistic data, refer to their maintenance and extension, and give different examples of automatic text parsing based on those dictionaries and grammars.

Keywords: Text parsing; large-coverage dictionaries; computational lexicons; word tagging; information retrieval.

1 Introduction and Background

The French DELA system was conceived and developed at LADL (Laboratoire d'Automatique Documentaire et Linguistique). It includes monolingual linguistic resources (mainly for French and English) specifically elaborated to be integrated into NLP systems. Standard methods and formats have been defined and are now used by other national teams working on their own languages: German, Greek, Italian, Portuguese and Spanish. Within that common framework, important fragments of the description of the languages involved have been worked out: the syntactic and semantic properties of free and frozen sentences are described and formalized. As for the lexicon, a major component of NLP, large coverage electronic dictionaries have been built. Simple and compound words have been described, and their linguistic characteristics have been hand-coded by computational lexicographers using a common method.

Most of these lexical resources can now be imported into the Intex NLP system[1], and then automatically applied to large texts. Within the scope of this article, we describe the set of lexical resources built so far for Portuguese, and we give different examples of automatic Portuguese text parsing.

2 Portuguese Electronic Dictionaries

By *electronic dictionary*, we mean a computerized lexicon specifically elaborated to be used in automatic text parsing operations (indexing, recognition of complex words, technical and common, etc.). Thus, large coverage electronic dictionaries were built for Portuguese for that purpose.

The set of lexical data is organized according to the formal complexity of the lexical units. The Portuguese DELAS is the central element of the dictionary system: it contains more than 110,000 *simple words*, whose grammatical attributes are systematically described and encoded. The set of *compound words* is structured in the Portuguese DELAC. At the moment, it is constituted by a lexicon of 22,000 compound nouns and 3,000 frozen adverbs, so it is still far from adequate completion[2].

2.1 The DELAS and DELAF Dictionaries

As said before, DELAS is the dictionary of simple words. We understand by simple words the lexical units that correspond to a continuous string of

---

[1] See http://www.ladl.jussieu.fr/INTEX/index.html

[2] The French DELAC contains (Silberztein (1997: 189) about 130,000 entries.

letters. The lexical entries of DELAS have the following general structure:

<*word*>, <*formal description*>

where *word* represents the *canonical form* (the lemma) of a simple lexical unit (in general the masculine singular for the nouns and adjectives, the infinitive for the verbs), and *formal description* corresponds to an alphanumeric code containing information on the grammatical attributes of the entries: their grammatical class (eventually, sub-class), and their morphological behavior.

The inflected forms are automatically generated from the association of a lemma to an inflectional code: the list of all inflected words constitutes the Portuguese DELAF (1,250,000 word forms).

In Portuguese, the major grammatical classes: *nouns*, *adjectives* and *verbs* have inflected forms:

- nouns and adjectives can appear in the feminine and/or in the plural; they can receive diminutive and augmentative suffixes; the superlative degree of the adjectives can be expressed by morphological means (suffixes);
- verbs are conjugated (mood, tense, person, number); furthermore, some verbal forms can undergo formal modifications induced by the presence of a clitic pronoun.

Thus, the DELAS entries:

gato, N01D1
gordo, A01D1S1

(where *N* and *A* indicate that *gato* (cat) is a noun and *gordo* (fat) is an adjective; *01* corresponds to the inflection rule for masculine, feminine, singular and plural; *D1* and *S1* explicit the type of diminutive and superlative suffixes that can be accepted by these entries) produce the following infected forms (DELAF entries):

gato, gato.N: ms (cat)
gata, gato.N: fs
gatos, gato.N: mp
gatas, gato.N: fp
gatinho, gato.N: Dms (little cat)
gatinha, gato.N: Dfs
gatinhos, gato.N: Dmp
gatinhas, gato.N: Dfp

gordo, gordo.A:ms (fat)
gorda, gordo.A: fs
gordos, gordo.A: mp
gordas, gordo.A: fp
gordinho, gordo.A: Dms (rather fat)
gordinha, gordo.A: Dfs
gordinhos, gordo.A: Dmp
gordinhas, gordo.A: Dfp
g*ordíssimo, gordo.A: Sms* (very fat)
gordíssima, gordo.A: Sfs
gordíssimos, gordo.A: Smp
gordíssimas, gordo.A: Sfp

As for the verbs, for instance, *dar* (to give):

dar, V02t

gives rise to a list of 73 inflected forms that correspond to the normal conjugation of a non-defective verb; in addition, *dar* can be constructed with clitic pronouns (*t*), in the position of accusative and dative complements. So, in:

(1) *Nós demos o livro à Maria*
(Lit.: We gave the book to Maria)

the verb form *demos* expresses: indicative mood, past tense, and first person plural.

From a syntactic point of view, *dar* is constructed with three arguments, subject: *Nós* (we) and two complements: *o livro* (the book), *à Maria* (to Maria). The complement syntactic positions can be fulfilled by clitic pronouns, respectively, *o* (it), accusative, and *lhe* (her), dative, as in:

(2) *Nós demo-lo à Maria*
(Lit.: We gave it to Maria)
(3) *Nós demos-lhe o livro*
(Lit.: We gave her the book)
(4) *Nós demos-lho*
(Lit.: We gave her_it)

In (2), the direct object has been cliticized, and, due to historical phonetic reasons, both the accusative pronoun and the verb have undergone formal modifications: *o*>*lo*; *demos*>*demo*. In (4), both pronouns (dative and accusative) are obligatorily agglutinated, forming the contraction: *lho* (<*lhe* + *o*).

So, even though the analysis of the combinations *verb-clitic* is a syntactic matter, given the morphological changes induced by such combinations in Portuguese, a first description had to be made at the morphological level.

On the other hand, the example in (4) illustrates a case where the formal notion of simple word does not correspond to an adequate linguistic analysis. Indeed, the form *lho* results from the contraction of two independent pronouns: *lhe* + *o*. In Portuguese, contracted forms issued from the

agglutination of two different words (and two different grammatical categories) are commonly observed. We give some simple examples of contractions resulting from the merging of prepositions with determiners, pronouns and adverbs:

pel(o,a,os,as) < por + (o,a,os,as)
(by the)
del(e,a,es,as) < de + (ele, ela, eles, elas)
(of (him, her, them))
daqui < de + aqui
(from here)

The relationship between contractions and their base constituent categories are established by finite-state transducers (see below).

2.2 Dictionaries for Compounds

Compound words, i.e, lexical units that are constituted by a fixed combination of simple words, represent a large amount of the lexicon of any language. One has only to underline in a text the sequences of words that are frozen together to some extent to realize that compounds constitute an important percentage of the text[3]. It is therefore illusory to envisage any sort of automatic processing before a significant lexical coverage is achieved. The issue is even more acute if one considers the description of scientific or technical texts or any specialized lexicon, where the number of compounds can rise up to appalling figures.

As said in 2. compounds are structured in the Portuguese DELAC. Priority was given to the listing and formalization of *compound nouns*, that can inflect: *lua de mel - luas de mel* (honeymoon), and to *compound adverbs*, that are invariable: *de repente* (suddenly). From the point of view of the lexicon, the main focus, especially as far as compound nouns are concerned, has been the every-day, not too technical, lexicon.

In order to identify compound words, and distinguish them from formally identical word free combinations, a set of morpho-syntactic criteria was adopted (Ranchhod (1991), Baptista (1995)). In short, compounds are the sequences of words that present restrictions to the combinatorial properties that they were supposed to have.

The formalization of compound dictionary entries is similar to that of simple words. Since compound adverbs, prepositions and conjunctions do not inflect, their formats are rather simple:

de repente, ADV+PC
(suddenly)
para com, PREP
(towards)
a fim de, CONJ
(in order to)

Compound nouns, however, have generally inflected forms. The rules for the inflection of compound nouns presented by grammarians do apply to some cases, but most compound nouns exhibit inflectional restrictions on gender or number that cannot be accounted by the morphological properties of their constituents. In the DELA format, the inflectional properties of compound nouns are specified according to the same criteria as in the dictionary of simple words. Thus, given the following nominal entries of the DELAC:

ser(21)humano(01), N + NA: ms - +
(human being)
guerra fria, N + NA: fs - -
(cold war)
visita(30) de estudo, N + NDN: fs - +
(field trip)

The first two compound nouns, *ser humano* and *guerra fria* have an internal structure *Noun Adjective* (NA), the most productive class in Portuguese, *visita de estudo* is a compound of structure *Noun de Noun*, also a very productive one. Each entry is characterized by the possibility (+) or impossibility (-) of gender and number inflection, respectively; the elements of the compound that can be inflected receive the inflectional code that they have in the DELAS: both constituents of *ser humano* inflect (in number) according to, respectively, the rules 21 and 01: *ser humano - seres humanos*; *guerra fria* is invariable, and the noun *visita de estudo* only allows the inflection of *visita*: *visita de estudo - visitas de estudo*.

As well as for other languages (e.g. French), additional information is being added, namely semantic.

[3] See 5. «Parsing Texts Using INTEX Tools»

2.3 Local Grammars

Most of the local linguistic phenomena, as well as many complex sentences, are represented in a natural way by the formalism of finite-state automata (FSA). For instance, frozen or semi-frozen structures are very naturally described by graphs, that represent FSAs (Silberztein (1997)). We illustrate the use of graphs with an elementary example, selected from the libraries of Portuguese local grammars. This grammar describes a family of adverbial expressions (dates), which refer to a period of time around the middle of the months (or, by extension, of some years) as in the underlined expression:

Isso aconteceu nos idos de Março

(That happened on the ides of March)

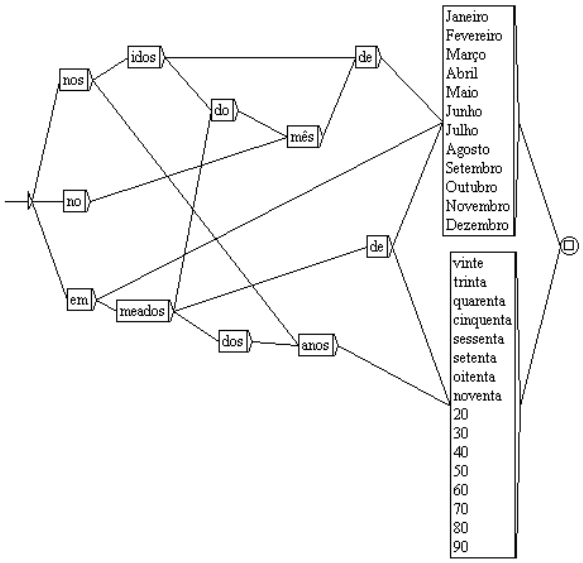

Fig. 1-AdvIdos.grf

This set of adverbial phrases corresponds to a linguistic object of clearly finite-state nature, but linguistic phenomena of a more complex nature can be efficiently described by such formalisms (Gross (1997, 1995)).

From the graphs of the local grammars, parsers (FSTs) can be automatically constructed, that applied to texts in combination with the dictionaries, allow the detection of a large variety of linguistic patterns (see below).

3 Transducers

Finite-state automata and transducers can be efficiently applied at various levels of linguistic analysis.

The following examples show how transducers are used to analyze contractions, ambiguities and compound numerical determiners.

3.1 Analysis of Contracted Words

As stated above (2.1.), contracted forms resulting from the agglutination of two independent words are commonly observed in Portuguese. To properly analyze these entities we built finite state transducers (FST) that, given a contracted form, produce an output corresponding to the decomposition of the contraction into its base constituents. For instance, the FST:

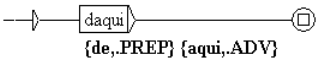

Fig. 2 – Analysis of the contracted form *daqui*

decomposes *daqui* (a contraction of the preposition *de* (from) and the adverb *aqui* (here)) in its base constituents and, simultaneously, associate to them the grammatical information of the dictionary.

3.2 Disambiguation

Disambiguation can be done at different moments of parsing.

a) *Disambiguation during normalization*

The normalization of texts for linguistic analysis uses FST to identify sentences and unambiguous compounds, to solve contractions and elisions. As an example of disambiguation at this level, we still use the case of contractions.

The form *dele* results from the contraction of *de* (of) with the ambiguous personal pronoun *ele* (he, him), which can be either a subjective (coded N) or a genitive form (coded O): *ele,eu.PRO+Pes:**N**3ms :**O**3ms*. However, only genitive forms can occur in the contraction *dele* (de + ele). So, the FST:

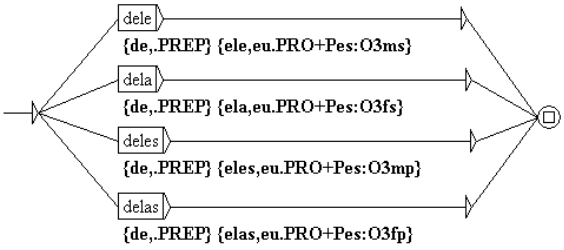

Fig. 3 – Analysis of the contracted forms *dele, dela, deles, delas*

is used not only to decompose the contracted form *dele* in its base constituents but to disambiguate the pronouns *ele, ela, eles, elas*. Identical FSTs can be used to analyze more complex situations where both constituents of a contraction can inflect independently.[4]

b) *Disambiguation for tagging*

In Portuguese, a word such as *compra* can be either a noun or a verb; the form *o* can be a determiner, a demonstrative pronoun and a personal pronoun. So, the linear combination of these elements allows six different analyses. However, in sentences like:

Ela compra-o hoje (She buys it today)

compra is only a verb, and *o* is only a personal pronoun, bound to the verb by an hyphen.

The following FST:

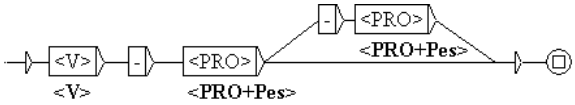

Fig. 4 – FST for the disambiguation of verbs and clitics

was built to solve these ambiguities: the five erroneous analyses are not taken into account; *compra* and *o* receive the correct tags.

3.3 Numerical Determiners

The Portuguese numerical determiners from *dois* (2) to *novecentos e noventa e nove mil novecentos e noventa e nove* (999,999) are plural forms. However, some of them can inflect in gender:

dois <*livros*>
(two <books>)
duas <*cadeiras*>
(two <chairs>)
trezentos e vinte e dois <*livros*>
(three hundred and twenty-two <books>)
trezentas e vinte e duas <*cadeiras*>
(three hundred and twenty-two <chairs>)

Others are invariant in respect to gender:

vinte <*livros*>
(twenty <books>)
vinte <*cadeiras*>
(twenty <chairs>)

mil e sete <*livros*>
(one hundred and seven <books>)
mil e sete <*cadeiras*>
(one hundred and seven <chairs>)

Numerical determiners such as *dois*, *duas* and *vinte* are simple words and therefore they are formalized in the DELAF dictionary; numerical determiners such as *trezentos e vinte e dois*, *trezentas e vinte e duas* and *mil e sete* can be seen as special compound words that are more adequately described by FST.

The first FST in figure:

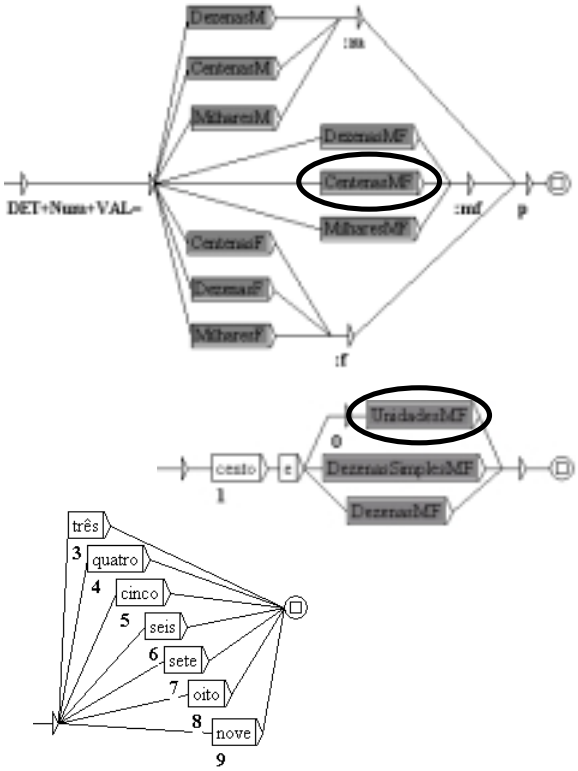

Fig. 5 – FST for the identification of Numerical Determiners

describes all the compound numerical determiners from *vinte e um* (21) to *novecentos e noventa e nove mil novecentos e noventa e nove* (999,999), including feminine and invariant forms, associating to each of them the grammatical category and the corresponding numerical value, as in the examples:

[4] That is the case of *aqueloutra* which is the contraction of the demonstrative pronouns *aquela* + *outra* (*that(fs) + other(fs)*). In Portuguese, even though contracted words are numerous, the list of contractions is still a closed set. So its description with FSTs is possible. However, this solution would not be adequate to describe productive phenomena involving agglutination, as it is probably the case of most compound nouns in German, for instance.

trezentos e vinte e dois,trezentos e vinte e dois.
DET+Num+Val=322:mp
trezentas e vinte e duas,trezentas e vinte e duas.
DET+Num+Val=322:fp
*mil e sete,mil e sete.*DET+Num+Val=1007:mfp

The FST shaded nodes refer to embedded FST; for instance, *CentenasMF* refers to the sub-graph that represents all invariant compound determiners from *cento e três* (103) to *cento e noventa e nove* (199) and *UnidadesMF* represents all invariant units from *três* (3) to *nove* (9).

4 Parsing Texts Using INTEX Tools

The linguistic resources that we briefly described have been imported into INTEX, that apply them to large texts. We give here some examples of text processing, using a small text.

a) *Recognition of all compound words of the text*

> À semelhança de um código de barras que permite identificar uma infinidade de produtos, dependendo da sequência de números, o genoma humano também encerra quase todos os nossos segredos e, grosso modo, basta uma ligeira mutação num gene para que se manifeste uma doença ou, pelo contrário, uma resistência à mesma. A toda a hora novos genes são identificados: um dia é um gene associado à repulsa do tabaco, noutro um que traduz uma maior susceptibilidade de se ficar infectado por determinado vírus.
>
> Há um código para tudo. Mas todos estes dados constituem apenas 10 por cento do património genético humano conhecido. Um facto que deverá ser alterado em Fevereiro do próximo ano, se se puderem cumprir as previsões dos responsáveis pelo ambicioso Projecto do Genoma Humano.

In the example, the compound words have been underlined.

b) *Indexing all utterances of a given word*

All the forms associated to the infinitive of the verb *ser* (to be):

> ão identificados: um dia é um gene associado à rep
> toda a hora novos genes são identificados: um dia
> o. Um facto que deverá ser alterado em Fevereiro

were identified and extracted into a concordance.

c) *Indexing a morphological pattern*

The rational expression:

<DET+Art+Ind:fs> (<E>+<A:fs>) <N:fs> ...
... (<E>+<PREP><N>)

or the equivalent FST:

identify feminine singular (*:fs*) noun phrases, that are specified by a determiner (*DET*) belonging to the class of indefinite articles (*Art + Ind*); the head of the noun phrase is a feminine singular noun (*N:fs*), optionally (*E*) modified by an adjective in pre-nominal position or a prepositional phrase (*PREP N*). In the first paragraph of the sample text, the NPs corresponding to those structures are (underlined):

> À semelhança de um código de barras que permite identificar uma infinidade de produtos, dependendo da sequência de números, o genoma humano também encerra quase todos os nossos segredos e, grosso modo, basta uma ligeira mutação num gene para que se manifeste uma doença ou, pelo contrário, uma resistência à mesma. A toda a hora novos genes são identificados: um dia é um gene associado à repulsa do tabaco, noutro um que traduz uma maior susceptibilidade de se ficar infectado por determinado vírus.

d) *Locating lexico-syntactic patterns*

A regular expression (or a local grammar) of the form:

(<dever>+<poder>) (<ADV>+<E>) <V:W>

corresponds to syntactic constructions with modal verbs: *dever*, *poder* (must, can). The main verbs are in the infinitive form: <V:W>; an insert or an adverb (simple or compound) can occur between the two verbs. In the text, there are two constructions of such type:

> conhecido. Um facto que deverá ser alterado em Fe
> ro do próximo ano, se se puderem cumprir as previs

5 Maintaining and Increasing Dictionaries, using INTEX features

5.1 Simple words

To evaluate the coverage of the existing dictionary we apply it to varied *corpora*: the non-recognition of a word form indicates in general that: (i) it is not in the dictionary; (ii) it was incorrectly formalized (iii) it is a proper name; (iv) it is an acronym; (v) it is misspelled.

Each of these failures require different solutions: (i) all the new words (with good prospects to

remain in the lexicon of the language) are formalized and added to the dictionary; (ii) the erroneous entries must be corrected; (iii) proper names must be listed in special dictionaries, built from the exploration of existing catalogs. However a lot of proper nouns are homographs with common ones, that in some contexts are written in capitals (*Bush* and *Rose* can be either a proper noun or a common one); (iv) acronyms (if they have good prospects to survive) must be listed and associated with the words that they represent. In general, acronyms are formally simple words, but they represent compounds. Our experiment of building such dictionaries indicates that the association of both types of lexical units it is not a trivial task.

5.2 Compound nouns

The dictionaries of compound nouns are being enlarged in a semi-automatic way. We write regular expressions that correspond to typical patterns of compound nouns (e.g. <N: ms> <A: ms>), and then we ask INTEX to extract from texts (to which dictionaries have been applied previously) all patterns that match that structure.

The resulting lists, integrated into a concordance, contain not only the combinations of a noun and an adjective but also compound nouns of that form that are followed by an adjective. Linguists interactively validate the lists of candidates to binary or ternary compounds.

References

Baptista J. (1995), *Estabelecimento e formalização de classes de nomes compostos*, M.A. Thesis, Universidade de Lisboa.

Courtois B. (1990), Un système de dictionnaires électroniques pour les mots simples du français, *Langue Française*, 87, «Dictionnaires électroniques du français», Paris: Larousse, pp. 11-22.

Eleutério S.; Ranchhod E.; Freire H.; Baptista J. (1995), A System of Electronic Dictionaries of Portuguese. *Lingvisticae Investigationes*, XIX:1, Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 57-82.

Gross M. (1995), Representation of Finite Utterances and the Automatic Parsing of Texts, *Language Research*, Vol. 31, No 2, Seoul: Language Research Institute, pp. 291-307.

Gross M. (1997), The Construction of Local Grammars, *Finite-State Language Processing*, Cambridge, Mass./London: MIT Press, pp.329-354.

Laporte E. (1997), Les mots. Un demi-siècle de traitements, *t.a.l.*, vol. 38, nº 2, Paris: Association pour le Traitement Automatique des Langues, pp. 47-68.

Ranchhod E. (1991), *Frozen Adverbs. Comparative Forms* como C *in Portuguese*, Lingvisticae Investigationes, XV: 1, Amsterdam/Philadelphia: John Benjamins, pp. 141-170.

Ranchhod E. (1998a), Dicionários Electrónicos e Análise Lexical Automática, In *Actas do Workshop sobre Linguística Computacional da APL*.

Ranchhod, E. and Mota C. (1998b), Elaboração de dicionários termiológicos. Seguros. In *Actas do Workshop sobre Linguística Computacional da APL*.

Silberztein M. (1993), Dictionnaires électroniques et analyse automatique de textes: le système INTEX, Paris: Masson, 233 p.

Silberztein M. (1997), The Lexical Analysis of Natural Language, *Finite-State Language Processing*, Cambridge, Mass./London: MIT Press, pp.175-203.

Acknowledgements

This research was partly supported by the FCT (Programme PRAXIS XXI, 2/2.1/CSH/775/95).