Multiword Lexical Acquisition and Dictionary Formalization

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

Paula CARVALHO
University of Lisbon and
LabEL, CAUTL, IST
Av. Rovisco Pais
Lisboa, Portugal, 1049-001
paula@label.ist.utl.pt

Elisabete RANCHHOD
University of Lisbon and
LabEL, CAUTL, IST
Av. Rovisco Pais
Lisboa, Portugal, 1049-001
elisabet@label.ist.utl.pt

Abstract

In this paper, we present the current state of development of a large-scale lexicon built at LabEL[1] for Portuguese. We will concentrate on multiword expressions (MWE), particularly on multiword nouns, (i) illustrating their most relevant morphological features, and (ii) pointing out the methods and techniques adopted to generate the inflected forms from lemmas. Moreover, we describe a corpus-based aproach for the acquisition of new multiword nouns, which led to a significant enlargement of the existing lexicon. Evaluation results concerning lexical coverage in the corpus are also discussed.

1 Introduction

MWEs have been viewed, for long time, as marginal idiosyncratic combinations of words. In recent years, however, there has been a growing awareness in the NLP community of the problems that MWEs pose and the need for their robust handling. Several major conferences and satellite workshops have been dedicated to the subject (ACL, EACL, LREC, for instance); major publications devote thematic issues to MWEs.

Anticipating that growing interest, over the last years, a significant part of LabEL's research has been devoted to the development of large-scale, linguistically precise language resources, namely to the construction of computational lexicons for simple and multiword units (Eleutério et al., 1995; Ranchhod et al., 1999; Ranchhod et al., 2004).

In fact, we have observed that MWEs are used frequently in both everyday language and technical and scientific texts to express ideas and concepts that in general cannot be stated by "free" linguistic structures. They include a large range of different linguistic phenomena, such as: (i) lexical compounds (nouns: *cellular phone*, *rush hour, New Jersey*; adjectives: *well-known*; adverbs: *for the time being*, *in short*; prepositions and conjunctions: *in spite of, in order to*) (ii) phrasal verbs (*give up*); (iii) light verbs (*give a lecture*); (iv) fixed (proverbs and maxims) and semi-fixed sentences (*to see the light at the end of the tunnel*; *to take the Lord' name in vain*). From a linguistic point of view, all these expressions exhibit distributional and selectional constraints, i.e. they lack compositionality, and frequently have idiomatic interpretations.

In this paper, we focus on multiword nouns. Special attention will be given to their formalization and generation, using INTEX, a public FST (Finite-State Transducer) based NLP system [Silberztein, 1993]. In this context, we present the main characteristics of a new inflectional module, conceived at LabEL, fully compatible with this system. Next, we describe the acquisition methodology used to gather new dictionary entries in a fragment (extracts 1,520,001 to 1,567,625) of the non-annotated version of a public Portuguese corpus, CETEMPublico[2]. Finally, based on this experiment, we assess, on the one hand, the dictionary increase, and, on the other hand, the lexical coverage improvement in the referred corpus.

2 Characterization of Multiword Nouns

Multiword (or compound) nouns are composed of non-capitalized simple words. Superficially, they seem to result from general rules of word combinations but they present constraints (morphological, combinatorial, etc.) concerning the properties they were supposed to have. Regarding inflection, general rules presented by grammarians do apply to some cases, but most compounds exhibit inflectional restrictions on gender or number that cannot be described by the morphological properties of their constituents.

Table 1 presents a few examples of the most representative classes of compound nouns in Portuguese.

[1] LabEL (Laboratório de Engenharia da Linguagem) http://label.ist.utl.pt

[2] CETEMPublico is a journalistic corpus containing about 180 million words (see Santos and Rocha, 2001 for techical information).

Class	Structure	Example
NA	Noun Adj	bomba atómica [atomic bomb]
NDN	Noun *de* Noun	consellho de guerra [council of war]
AN	Adj Noun	mau pressentimento [bad feeling]
NPN	Noun Prep Noun	barco a remos [rowing boat]
NPV	Noun Prep Verb	máquina de lavar [washing machine]
VN	Verb Noun	cessar-fogo [cease(-)fire]
NN	Noun Noun	bomba-relógio [time bomb]
NCN	Noun Conj Noun	prós e contras [pros and cons]

Table 1: Some binary compound noun classes

These classes represent binary compounds, comprised of two content words (where one of them is a noun), eventually connected by a grammatical word[3].

The classification criteria are based on the noun's internal structures, which are generally associated with a characteristic inflectional pattern. For instance, compound nouns belonging to the NA class usually allow the inflection in gender and/or number of both constituents (e.g. *bomba atómica*, *bomba**s** atómica**s***); on the contrary, in the majority of NDN compound nouns, only the first noun can inflect (*conselho de guerra*, *conselho**s** de guerra*).

In the following sections, further relevant information on inflection, formalization and generation of inflected forms will be given.

3 Formalization of NA and NDN Nouns

Following methods and formalisms introduced at LADL [Gross, 1988; Courtois and Silberztein, 1990], linguistic attributes of simple and multiword units are systematically encoded in dictionaries compatible with INTEX.

In this system, compound word entries are handled depending on their internal structure. They are formalized and processed separately by different programs.

In order to simplify the formalization of linguistic attributes, and make the generation process easier, we implemented a new inflectional module compatible with INTEX system. The main strength of this tool is allowing the simultaneous generation of all compounds, regardless of their internal structure, reusing the inflectional graphs already built for simple words [Mota, forthcoming]. The morphological constraints are specified manually, assigning to each constituent the inflectional code that corresponds to its inflectional behavior within the compound, as illustrated by the following dictionary entries:

```
actor(N040) secundário(A001),N+NA+Hum
ser(N205) humano(A201),N+NA+Hum
ponto(N201) de vista,N+NDN+Pred
direitos(N292) de autor,N+NDN
```

In the first compound, both constituents keep their simple word dictionary inflectional code; they inflect in gender and number, according to the compound inflectional behavior.

On the other hand, the compound noun *ser humano* inflects only in number, which means that the masculine nominal constituent *ser* preserves its inflectional code, but the adjectival constituent *humano* (which also inflects in gender, as simple word) receives a new code that just allows its inflection in number within the compound.

In the case of the NDN compounds, as previously mentioned, only the head can inflect: *ponto de vista* inflects in number, so the head receives a code allowing its inflection; *direitos de autor* does not inflect (it is an exclusive masculine plural noun), hence the head is assigned an inflectional code that simply transmits these gender and number features to the compound.

4 Generation of Inflected Forms

The following example illustrates how the new inflection module uses the dictionary information, briefly presented in the previous section:

```
actor(N040) secundário(A001),N+NA
```

Inititally, the inflectional module generates the inflected forms of the noun *actor*, based on the inflectional paradigm described in the graph *N040*. Then, it combines the resulting inflected nouns with all inflected forms of the adjective *secundário*, generated given graph *A001*. Subsequentely, the combinations that do not verify the gender and number agreement constraints are eliminated. Additionally, the constituent inflectional attributes are inherited by the compound.

As a result, the following entries are obtained:

```
actor secundário,actor secundário.N+NA:ms
actriz secundária,actor secundário.N+NA:fs
actores secundários,actor secundário.N+NA:mp
actrizes secundárias,actor secundário.N+NA:fp
```

This example illustrates the case where both words that constitute the compound have similar inflectional features. These attributes are directly transferred to the inflected forms of the compound.

When one of the compound constituents does not have either gender or number explicit morphemes, as *artista* (which can be either a masculine or a feminine singular form) in the following entry:

```
artista(N101) plástico(A001),N+NA
```

[3] Even though less productive than the previous structures, there are longer multi-word combinations that may involve more than one compound form (e.g. *cabo de alta tensão*, high-tension electricity cable).

the inflectional module assigns to the compound the morphological attributes of the constituent that has explicit gender and/or number morphems (in this case, *plástico*).

The compounds just illustrated belong to the NA class[4]. The inflection of NDN forms simply corresponds to the inflection of the head noun and assigment of its attributes to the compoud.

5 Acquisition of New Entries

With the purpose of increasing the number of the most representative nominal entries (NA and NDN) in the common multiword dictionaries, we used a corpus-based approach.

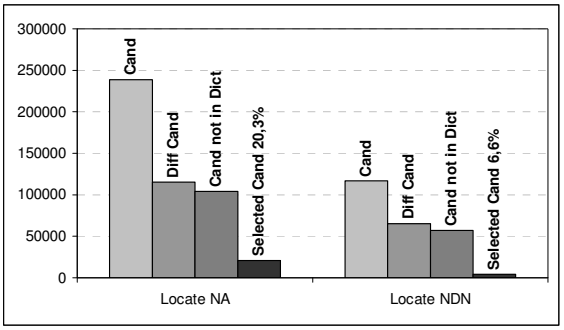
In the first stage, candidates were automatically extracted from a fragment of CETEMPublico corpus (from now on, acquisition corpus), using INTEX. After tokenized by INTEX, 6,385,531 tokens (corresponding to 138,230 different tokens) were identified in the acquisition corpus. From those, 5,162,111 (138,174 different forms) are alphabetic words.

LabEL's simple and multiword electronic dictionaries were then applied to the acquisition corpus. These dictionaries contained 171,159 nominal entries, from which 82% were simple words and 18% compounds. From the 22,581 compounds, 61,7% are NA, 33,6% are NDN and the remaining 4,7% belong to other strucutres.

Candidate identification was performed using the elementary regular expressions *<N>de<N>* and *(<N:ms><A:ms> + <N:fs><A:fs> + <N:mp><A:mp> + <N:fp><A:fp>)*[5]. They match nominal compounds presenting, respectively, an NDN and an NA structure. With respect to the latter structure, the expression also guarantees morphological agreement between nouns and adjectives.

Such expressions recognized in the acquisition corpus 242,527 (187,146 different forms) candidates, from which 117,616 (69,066 different forms) are NDN structures and 230,761 (118,080 different forms) are NA structures.

Each class's candidates were integrated into a concordance, to which were applied the existing compound dictionaries. The resulting list of non-recognized candidates was then manually reviewed by linguists, aiming the selection and linguistic formalization of valid compounds. Graphic 1 reflects the effort involved in the selection procedure.



Graphic 1: NA and NDN candidates

One clear observation is that there is a great discrepancy between the initial candidate lists (NA: 238,313; NDN: 116,246) and the final selected compound forms (NA: 21,289; NDN: 3750)[6]. The percentage in the graphic was calculated based on the number of non-recognized different candidates (NA: 104,715; NDN: 56,741).

Another interesting observation regarding Graphic 1 is that the size of the NA candidate list is slightly more than a double of the NDN candidate list. Nevertheless, the NA candidate list includes proportionally more valid compound forms (NA: 20%; NDN: 7%). In addition, the final selected NA compound list contains about five times more entries than the corresponding NDN list.

The selected compound forms resulted in a total of 19,825 NA and 3,769 NDN canonical entries, which correspond respectively to 41,267 NA and 7,722 NDN inflected forms.

6 Evaluation

Before lexical acquisition, the application of LabEL's lexical resources to the acquistion corpus allowed assigning 94,229 different simple word tags (34,892 nouns) and 15,120 different compound tags (13,594 nouns). Hence, the compound noun percentage is low (28%), compared to the number of different nominal tags assigned to simple forms. This discrepancy is not unexpected, regarding the low number of nominal compound entries in dictionaries.

As previously mentioned, the gathering of NA and NDN compounds in the acquisition corpus led to the formalization of 23,594 canonical entries. Accordingly, the inflected form dictionary increased approximately 3 to 4 times (more 53,815 entries, in a total of 76,396 entries).

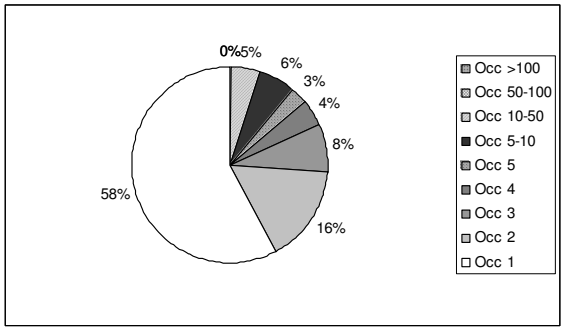
[4] This procedure also applies to other compounds composed of two or more elements that inflect (e.g. *acidente(N201) vascular(A205) cerebral(A211)*, stroke).

[5] Regular expressions are written according to the INTEX format.

[6] In this study, we did not assess the list of candidates recognized by dictionaries, which means that we did not count the hypothetical cases of embedded compound forms (e.g. *cabo de alta-tensão*, high-tension electricity cable).

When we apply the enlarged dictionary to the same corpus, we observe that the percentage of compound nouns with respect to the total of nominal tags incresead significantly. Now, 40,902 different compound forms were identified, which means that more than a half of the nouns in the corpus correspond to multiwords.

Considering the compound occurrences in the acquisition corpus, Graphic 2 illustrates their frequency distribution.



Graphic 2: Compound noun frequency

It is important to draw attention to the fact that 89% of compound forms occur less than five times; in particular, 58% occur just once.

These figures demonstrate that, contrary to what is observed with simple nouns, which are very recurrent in texts, the average number of compound occurrences is, in general, extremely low. This evidence raises the question whether statistical methods, based on frequencies, can adequately handle the majority of compound forms.

Regardless of wether the compound acquisition has been done exclusively in a fragment of CETEMPúblico, the application of the new dictionary to the remaining fragments of this corpus also increased the number of tags assigned to compound words.

On average, before lexical acquisition about 13,000 different compound nouns were recognized; this number more than doubled (approx. 33,000) when we applied the enlarged dictionaries to the other fragments. As mentioned before, in the acquisition corpus, the number of compound tags exceeded the number of simple noun tags. So, we may infer that a similar behavior would be expected in the remaining fragments, if they were also considered.

7 Final Remarks

In this paper, we described a new FST-based compound inflectional tool. The main advantage of this inflectional module is reusing INTEX simple word inflectional graphs in the simultaneous generation of all compound, regardless of their internal structure. Even though it has only been tested in the formalization and generation of Portuguese compound nouns, we believe it can be easily adapted to handle other languages having similar compound inflectional behavior.

A corpus-based approach to multiword acquisition was also presented. We showed that, in spite of involving human effort, the results obtained effectively improved dictionary coverage. Moreover, the results concerning compound noun frequency raised the question whether statistical approaches, based on word frequencies, are (un)adequate to extract multiword nouns from texts.

8 Aknowlegments

This work was partially supported by FCT (POSI/39806/PLP/2001; POSI/39806/PLP/2001).

References

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

Courtois, B. ; M. Silberztein (eds). 1990. *Langue Française*, 87, «Dictionnaires électroniques du français», Paris: Larousse.

Eleutério S.; E. Ranchhod; H. Freire; J. Baptista. 1995. A system of electronic dictionaries of Portuguese. In *Lingvisticae Investigationes*, XIX:1, pp. 57-82, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gross, M. 1997. The Construction of Local Grammars. In E. Roche; Y. Schabes (eds.), *Finite State Language Processing,* pp. 329-352, Cambridge, Mass./London: MIT Press.

Mota, C. (forthcoming). Inflection of the Portuguese DELAS using FST. In *Proceedings of the 4th and 5th INTEX Workshops*, Presses Universitaires de Franche-Comté.

Ranchhod, E; C. Mota; J. Baptista. 1999. A Computational Lexicon of Portuguese for Automatic Text Parsing. In *Proceedings of SIGLEX99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL*, pp. 74-80, Maryland, USA..

Ranchhod, E.; P. Carvalho; C. Mota; A. Barreiro. 2004. Portuguese Large-scale Language Resources for NLP Applications. In *Proceedings of 4th LREC*, pp. 1755-1758, Lisbon.

Santos, D.; P. Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 442-449, Toulouse.

Silberztein, M. 1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, Paris: Masson Ed.