Using Corpora to Increase Portuguese MWU Dictionaries.
Tagging MWU in a Portuguese Corpus.

Elisabete Marques Ranchhod
Department of Linguistics & LabEL / Onset
University of Lisboa
elisabeto@label.ist.utl.pt

1. Introduction

It is impossible to envisage automatic corpus analysis without adequate identification and treatment of multiword expressions (MWE). The meaning of a text is mostly supplied by frequent occurrence of multiword units (MWU), especially compound nouns.

MWE are usually built from the vocabulary of simple words, but their meaning is not always compositional. They include a large range of different linguistic objects, such as: (i) lexical compounds (nouns: balance of trade, bull's-eye, magnetic field; adjectives: blow by blow, high-flying, well-known; adverbs: above all, in crude terms, time and again; prepositions and conjunctions: as far as, in spite of, in order to); (ii) phrasal verbs (carry out, give up); (iii) light verbs (give a lecture, make a speech); (iv) fixed and semi-fixed sentences (burn the candle at both ends, take the bull by the horns).

MWE 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 they pose and the need for their robust handling. In fact, it is now clear that MWE play an important role in real-world applications, particularly in those that require some degree of semantic processing (e.g. machine translation, question-answering, summarisation, information retrieval and extraction, etc.).

Anticipating that growing interest, M. Gross (1986) identified different forms of multiword expressions (compound verbs, nouns, adverbs, frozen sentences), and discussed the problems that they pose to natural language processing. Later, he demonstrated (1989) that finite-state methods are adequate and suitable to represent the different types of linguistic structures, including lexical compounds.

Following M. Gross’ approach to the lexicon (and grammar) of natural languages, a significant part of LabEL’s research has been devoted to the development of Portuguese large-scale language resources, in particular to the construction of computational lexicons for simple and multiword lexical units. Linguistic data are formalized using finite-state techniques, and are applied to corpus processing by the same mechanisms.

In the scope of this paper, after a short description of Portuguese lexicons (2.1), special attention will be given to multiword lexical units: criteria for MWUs identification, linguistic attributes and formalization (2.2); dictionary enlargement by corpus exploitation (2.3). UNITEX, a corpus processing system, based on automata technology (www-igm.univ-mlv.fr/~unitex/) will be used to apply the MWU dictionary to a Portuguese raw
corpus. Different examples of corpus annotation and linguistic knowledge extraction will be provided in form of concordances (section 3.).

2. LABEL-LEX - Portuguese Large-Scale Language Resources

The Portuguese language resources developed at LabEL – LABEL-LEX – have been described in a number of publications (Ranchohd and al., 1999, 2002, 2003, 2004). They consist of lexicons and grammars formalized in finite-state transducers (FST). The former contain both simple and multiword lexical units; the later specify lexical and syntactic restrictions on word combinations.

The lexical data are organized in several modules, according to the formal and linguistic characteristics of the lexical entries. The two main dictionary modules are:

- **LABEL-LEX-sw**, which comprises more than 1,500,000 inflected word forms, automatically generate from a lexicon of about 120,000 lemmas;
- **LABEL-LEX-mw**, which contains about 80,000 multiword lexical units, mainly compound nouns (76,400), adverbs and adjectives (3,500).

[Notice that these figures do not include proper nouns or ‘named entities’, which are not taken into account in the scope of this paper].

Each dictionary entry is described in terms of part-of-speech (PoS) and morphological attributes. Syntactic and semantic information is being introduced progressively.

2.1 LABEL-LEX-sw

As said above, inflected word forms are system generated from their corresponding lemmas. The lexical structure and description of lemmas is illustrated in the Sample 1 [next to Portuguese examples, an approximate translation to English is enclosed in square brackets].

```
alto.,INTERJ     [halt]
alto,A001_sst001+Pd [high, tall, loud]
alto,N201_fh201   [hillock; bump]
antes.,ADV       [before]
atacar,V102t     [to attack]
atômico,A001+Rel [atomic]
```

Sample 1 – Dictionary of lemmas

The «» mark separates different syntactic and/or semantic attributes; alphanumeric codes that follow each lemma represent an inflectional FST. The codified information concerns part-of-speech (Verb, INTERJection, Adjective, Noun, ADVerb); the numerical codes represent inflectional rules (number and gender for nouns and adjectives, number, person, tense and mood for verbs); dh, and ss correspond to diminutive and superlative suffix. Syntactic information is also specified: Pd and Rel refer to predicative and relational adjectives, respectively; t, in the verb code, means that atacar can be followed by an enclitic pronoun. The six lemmas above generate 77 inflected word forms; examples are given in Sample 2.
All the inflected word forms are associated to their lemma. Some lemmas belong to more than one grammatical category (alto, Sample 1). That homography, very frequent in Portuguese, particularly between nouns and adjectives, affects all the grammatical categories, and is an important source of ambiguity. The homography of simple words increases considerably when the lemmas are inflected. In Sample 2, the verb forms ataque (subjunctive present, first, second (formal) and third person singular: S1:s:S2’s:S3s) and ataques (subjunctive present, and negative imperative, second person singular: S2:Y2s) are homographs of a masculine noun that inflects in number, ataque, ataques [attack, attacks]; the past participle atacado (noted K) is homograph of an adjective that inflects in gender and number (atacado, atacada, atacados, atacadas). The feminine forms of the adjective alto (alta, altas) are homographs of a feminine noun alta, altas [discharge, discharges]. To resolve word ambiguity, specific grammars for word sense disambiguation are being developed (Carvalho, 2001).

2.2 LABEL-LEX-mw

Although multiword lexical units are not immune to ambiguity, they are less ambiguous than simple words. In fact, MWUs are constrained combinations of simple words, and can be considered as the first level of frozen expressions. Most of the words included in Samples 1 and 2 occur in a number of multiword nouns, adjectives and adverbs, as illustrated by a few examples given in Sample 3.

```plaintext
alto(A001) comissário(N001),N+AN+Hum+Cargo  [high commissioner]
alto(A216)mar,N+AN+Geo  [open sea]
altos(N292) e baixos.,N+NConj+N+Abst  [ups and downs]
ângulo(N201) de ataque,N+NDN+AerDin  [angle of attack]
antes de mais.,ADV+PDC  [first of all]
antes que.,CONJ  [before]
ataque(N201) de asma,N+NDN+Med  [attack of asthma]
ataque(N201) pessoal(A211),NA+Abst  [personal attack]
chapéu(N201) alto(A201),N+NA+Vest  [top hat]
de alto a baixo.,ADV+PCPC  [from top to bottom]
dera(N392) atómica.,N+NA+Temp  [atomic era]
saxofone(N201) alto(A201),N+NA+Mus  [alto saxophone]
segredo(N201) atómico (A201),NA+Abst  [atomic secret]
```
In such combinations, the ambiguous simple words of Samples 1 and 2 are no more ambiguous. Actually, the constituent words of a given MWU loose, in part, their lexical independence, they are an element of a new lexical entity. Differently to simple words, the bulk of MWUs can be assigned a single PoS tag (N, A, ADV, ...) as well as specific semantic attributtes. So the adequate identification of MWUs constitutes a necessary activity in its own, but, in addition, it contributes significantly to reduce ambiguity. For instance, the compound noun ataque pessoal [personal attack] is internally constituted by a noun, ataque, and an adjective, pessoal (N+NA), both exhibiting PoS ambiguity: a noun and a verb the former, a noun and an adjective the latter. In the compound, that PoS ambiguity disappears. From a computational point of view, the adequate identification of lexical compounds avoids a number of erroneous analyses resulting from the different values of their constituent simple words.

Although some compounds are semantically ambiguous, the majority of them have a unique interpretation. The main types of ambiguity are: (i) the same sequences of words correspond to more than one compound; bomba relógio [time bomb] can refer to a type of bomb or to a tense situation; this ambiguity requires, like in the case of simple words, the creation of more than one dictionary entry; (ii) a given sequence of words can be analysed as a compound or as a free combination of words: mesa redonda [round table] can designate a meeting, in that case mesa redonda is a compound noun, or a round object. Both types of ambiguity can be solved by grammars (Carvalho and al. 2003).

While it is difficult to associate semantic values to simple words on a non-intuitive basis, concerning MWUs this is a realistic goal. So, semantic attributes are being introduced into MWUs dictionary (even if this task is far to be complete). Semantic values are established on a syntactic basis. For instance, the notion «human», coded Hum, is associated to nouns that can be found in subject position of verbs such as dizer [to tell] or pensar [to think]. Such syntactic positions can be filled by individual or collective human nouns. The later are integrated in the subtype Hum+Col (which comprises different subclasses: institutions, organisations. etc.). Concerning individual human nouns, a hierarchical structure is also being organised; typically human occupations are coded Hum+Cargo (alto comissário, Sample 3).

The notions «abstract» and «concret» are too vague, and poorly operative. The attribute Abst is being used (for abstract) while a better solution is not found (for instance, ataque pessoal, in sample 3.). Concret nouns are being structured in a number of types. For example, the nouns occurring in complement position of verbs such as usar [to wear] are classified as Vest [clothing], in Sample 3, chapéu alto.

2.2.1 Inflectional and Morphological Constraints

In Portuguese, a large number of MWUs can inflect, particularly nouns and adjectives, others are completely invariable (adverbs, conjunctions, prepositions, some nouns and adjectives). The examples in Sample 3 illustrate uninflected MWUs. The corresponding inflected forms are system generated by the same inflectional FSTs (represented by the numerical codes in brackets) that inflect simple words. Now, compound words exhibit inflectional restrictions relative to the inflectional behaviour of their constituents. For example, the rule to generate the inflected forms (gender and number) of the adjective alto is formalized in the FST A001 (Sample 1). In compound nouns (Sample 3), alto (i) inflects
in gender and number (A001): *alto comissário*; (ii) inflects in number but not in gender (A201): *chapéu alto; saxofone alto*; (iii) is exclusively masculine singular (A216), and transmits these attributes to the compound, in *alto mar*; (iv) is exclusively masculine plural (N292) in *altos e baixos*. Identical observations apply to the adjective *atómico*. The noun *mar* [sea] can be either singular *o mar* [the (ms) sea] or plural *os mares* [the (mp) seas], but *mar alto* is an invariable masculine singular noun.

The 22 inflected MWU forms, generated from those represented in Sample 3, are presented in Sample 4.

| alta comissária, alto comissário. N+AN+Hum+Cargo:fs | [high commissioner] |
| altas comissárias, alto comissário. N+AN+Hum+Cargo:fp | |
| alto comissário, alto comissário. N+AN+Hum+Cargo:ms | |
| alto mar, alto mar. N+AN+Geo:ms | [open sea] |
| altos comissários, alto comissário. N+AN+Hum+Cargo:mp | |
| altos e baixos, altos e baixos. N+NCNJ+N+Abst:mp | |
| ângulo de ataque, ângulo de ataque. N+NDN+AerDim:ms | [ups and downs] |
| ângulos de ataque, ângulo de ataque. N+NDN+AerDim:mp | [angle of attack] |
| antes de mais, antes de mais. ADV+PDC | [first of all] |
| antes que, antes que. CONJ | [before] |
| ataque de asma, ataque de asma. N+NDN+Med:ms | [attack of asthma] |
| ataque pessoal, ataque pessoal. N+NA+Abst:ms | [personal attack] |
| ataque de asma, ataque de asma. N+NDN+Med:mp | |
| ataque pessoal, ataque pessoal. N+NA+Abst:mp | |
| chapéu alto, chapéu alto. N+NA+Vest:ms | [top hat] |
| chapéus altos, chapéu alto. N+NA+Vest:mp | |
| de alto a baixo, de alto a baixo. ADV+PCPC | [from top to bottom] |
| era atómica, era atómica. N+NA+Temp:fs | [atomic era] |
| saxofone alto, saxofone alto. N+NA+Mus:ms | [alto saxophone] |
| saxofones altos, saxofone alto. N+NA+Mus:mp | |
| segredo atómico, segredo atómico. N+NA+Abst:ms | [atomic secret] |
| segredos atómicos, segredo atómico. N+NA+Abst:mp | |

Sample 4 – Inflected MWU

The PoS information (N, ADV, etc.) is followed by the specification of the lexical structure of MWUs. Regarding nouns, NA, the most productive structure, means that the compound is constituted by a noun and an adjective, for instance *ataque pessoal* [personal attack] *chapéu alto* [top hat], *era atómica* [atomic era]; NDN represents the structure ‘noun of noun’, *ângulo de ataque* [angle of attack] *ataque de asma* [attack of asthma], a very productive one as well. The codification of this information is useful for various reasons, for instance, the nouns can be searched for by their PoS or by their lexical structure (see 3.1 below).

Concerning compound adverbs, their lexical structure is more irregular; some are constituted by peculiar sequences of grammatical categories: preposition, noun, preposition, adverb: *de vez em quando* [from time to time], preposition, noun, conjunction noun: *a par e passo* [continuously]; others contain words that only exist in the adverb: *a contragosto* [in an unwilling way] *a trouxe-mouche* [higgledy-piggledy]. As a matter of fact, inside the adverb, the notion of grammatical category loses its relevance. The codes that follow the PoS information reflect vaguely the internal structure of compound adverbs: the capital letter P stands for a preposition (that sometimes does not exist: *vezes sem conta* [time and time again]), lexical words are represented by the capital letter C: *de alto a baixo* (PCPC).
2.2.2 Lexical and Syntactic Constraints

As already mentioned, MWUs are constrained combinations of simple words. At the lexical level, such constraints are observable by paradigmatic ruptures. For instance, in *chapéu alto* [top hat] and *alto mar* [open sea], two NA (noun + adjective) multiword nouns, neither the nouns *chapéu* and *mar* nor the adjective *alto* commute, respectively, with other nouns and adjectives of the same lexical family (synonyms, antonyms). The commutation produces unacceptable word sequences, as the following examples show:

1. *Ele usava um chapéu alto*
   [He wear a top hat]

2. *Ele usava um boné alto*
   [He wear a top cap]

3. *Ele usava um chapéu baixo*
   [He wear a bottom hat]

4. *Ele navegava no alto mar*
   [He sailed on the open sea]

5. *Ele navegava no alto oceano*
   [He sailed on the open ocean]

6. *Ele navegava no baixo mar*
   [He sailed on the close sea]

At the syntactic level, the predicative adjective *alto*, occurring in free combinations with nouns, is gradable and can be quantified by adverbs:

7. *Construíram um edifício alto*
   [They built a high building]

8. *Construíram um edifício muito alto*
   [They built a very high building]

9. *Construíram um edifício altíssimo*
   [They built a highest building]

Such possibilities do not exist when *alto* is part of a multiword noun:

10. *Eles usavam chapéus altíssimos*
    [They wear very top hats]

11. *Ele navegava no altíssimo mar*
    [He sailed on the very open sea]

As to compound adverbs, they commute with and have the same syntactic value as simple adverbs (Ranchhod, 1991):

12. *Ao mesmo tempo = simultaneamente* [at the same time = simultaneously]

13. *De vez em quando = ocasionalmente* [from time to time = occasionally]
2.2.3 Lack of Semantic Compositionality

Such a linguistic behaviour indicates clearly that, even if familiar compounds seem to be semantically compositional, it is not so. Indeed, concerning semantic transparency/opacity there is a continuum from compounds that are totally idiomatic: cara-metede [better half], to compounds whose interpretation is close to compositionality: ataque aéreo [air attack/raid]. But even in the later situation the semantic value of the compounds does not correspond to the sum of the individual meanings of their constituent words. It is not clear that the meaning of ataque [attack] is the same in ataque aéreo [air attack], ataque pessoal [personal attack], ataque de asma [attack of asthma] and ângulo de ataque [angle of attack]. In the case of adverbs the lack of compositionality is still more obvious. For instance, the meaning of the adverbs de cor [by heart] and por alto (‘by high’, cursorily):

(14) Ele conhece todas as definições de cor
[He knows all the definitions off by heart]

(15) Ele leu esse artigo por alto
[He read that paper cursorily]

Can’t be associated neither to the prepositions de and por nor to the words cor and alto. As a matter of fact, in contemporary Portuguese cor, as a simple word, does not have any linguistic value, it only exists associated to the preposition de, forming together the adverb de cor.

2.3 Using Corpora to Gather Multiword Lexical Units

MWUs have been considered as marginal, idiosyncratic linguistic objects, and, for that reason, largely ignored by theoretical linguistics. References to MWUs made by grammarians are trivial and inconsequential (Cunha & Cintra, 1984). Lexicographers only introduce in dictionaries a few MWU entries (mostly nouns) written with an orthographic hyphen (bem-estar [wellbeing], guarda-costas [bodyguard], livre-arbitrio [free will]); inside the dictionary entries, some examples of MWUs can also be found. The first version of LABEL-LEX-fw contained about 22,000 compounds, collected in Portuguese grammars and dictionaries. The formalization and classification of those compounds made clear that, while the lexical structure of adverbs is unpredictable, the majority of common compound nouns correspond to characteristic combinations of words. Table 5 (adapted from Mota and al., 2004) shows the most productive classes of Portuguese multiword nouns.

<table>
<thead>
<tr>
<th>Class</th>
<th>Structure</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>Noun Adjective</td>
<td>mercado negro [black market]</td>
</tr>
<tr>
<td>NDN</td>
<td>Noun de Noun</td>
<td>estado de coisas [state of affairs]</td>
</tr>
<tr>
<td>AN</td>
<td>Adjective Noun</td>
<td>mau pressentimento [bad feeling]</td>
</tr>
<tr>
<td>NPN</td>
<td>Noun Prep Noun</td>
<td>barco a remos [rowing boat]</td>
</tr>
<tr>
<td>NPV</td>
<td>Noun Prep Verb</td>
<td>canção de embalar [lullaby]</td>
</tr>
<tr>
<td>VN</td>
<td>Verb Noun</td>
<td>cessar-fogo [cease(-)fire]</td>
</tr>
<tr>
<td>NN</td>
<td>Noun Noun</td>
<td>bomba-relógio [time bomb]</td>
</tr>
</tbody>
</table>

Table 5 – Productive Classes of MWU nouns

In order to enlarge the dictionaries, corpora were used to extract new multiword nouns. The CETEMPúblico, a Portuguese non-annotated public corpus was the major source of data. This corpus is constituted by fragments of the Portuguese daily newspaper Público,
and contains about 180 million words (see Santos and Rocha, 2001 for technical information).

Using the NLP system INTEX (Silberztein, 1993), and the LABEL-LEX-sw dictionaries, regular expressions, representing the two more productive lexical patterns of compound nouns (NA and NDN), were applied to a sample of the entire corpus. That sample (sample-corpus from now on) comprises extracts 1,520,001 to 1,567,625 (about 5,162,111 word forms, 138,230 different forms). The existing multiword dictionaries were applied to the sequences of words matching the regular expressions. The sequences that did not exist in the dictionaries were then analysed by linguists. All the word combinations that corresponded to compound nouns (23,594 canonical forms) were formalized. After generation of inflected forms, about 53,800 new compound nouns were introduced into dictionaries (see Mota and al., 2004, for more details and other numeric values).

3. Identifying and Tagging MWU in a Portuguese Corpus

The sample-corpus mentioned before will be used now to illustrate how the linguistic information associated to MWUs can be extracted from de corpus (3.1), and how that information can be merged in the corpus for linguistic tagging (3.2).

UNITEX, a corpus processing system, based on automata-oriented technology (http://infolingu.univ-mlv.fr/), will be used to apply the LABEL-LEX-mw to the sample-corpus. After the application of the dictionaries to the corpus, UNITEX recognized 210,315 compounds (all grammatical classes included: nouns, adjectives, adverbs, prepositions and conjunctions), corresponding to 6,259 % of the sample-corpus. These values include all the occurrences; the number of compound lexical entries is: 39,021 (see Figure 6).
Examples of the totality of recognized compounds in the sample-corpus are given in Concordance 7.

Concordance 7 – Sample of recognized compounds

The concordance contains a diversity of MWUs: compound nouns (auto-estrada [highway], bairro degradado [slum area], cabeça-de-lista [list leader], cessar-fogo [cease-fire], feira do livro [book fair], fim de semana [week-end], etc.); compound adverbs (a curto prazo [shortly], à tarde [in the afternoon], ao mesmo tempo [at the same time], etc.); compound adjectives (bem sucedida [successful]); prepositions (dentro de [within]) and conjunctions (a fim de [in order to], antes de [before]).

3.1 Corpus Exploitation Using Encoded Linguistic Information

The linguistic information (morphological, lexical, semantic) associated to MWUs can now be used in corpus exploitation. A few examples will illustrate a variety of possibilities.

Information about the lexical structure may be used to look at the occurrence of compounds. The regular expression (Unitex format): <N+NA> extracts from de sample-corpus all the compounds constituted by a noun and an adjective. A small sample of the 97,119 sequences matching that pattern is given in the Concordance 8.

ou desde 1984 de penalizar o em introduzir na cena. (S)A o -- 9,1% do PIB (S)Saldo da ine a missa na Sé Velha, com nada mudou, com excepção das il, como quem está a receber speculativos se convertem no ou de ser confrontado com um orcionavam uma vasta gama de alar para o trágico e para o do com tal fascínio no nosso usando se espera pelo dia do ampeão argentino, verdadeira o operar, basta escolher no S)Esta adaptação francesa da aborto terapêutico, por motivos de saúde, ma bailarina principal teria de usar um cacheco balança comercial -- .+ de 6,3 mil milhões canto gregoriano, por volta das 10h00, começ caras novas na direcção do PC. (S)Testemunho dinheiro vivo. (S)Segundo o mesmo documento, elxo central da economia do país. (S)O que a facto consumado, sem qualquer esclarecimento gêneros agrícolas, que nasceu Huambo, no cen humor negro, passando pelo sórdido. (S)A re imaginário colectivo, a sua silhueta aumento juizo final» da TVI, tema da ordem do dia sã lenda viva do automobilismo de competição, a mapa celeste que surge no ecrá o astro que t obra-prima de Victor Hugo ao cinema tem, na
campo. (S)Alguém prepara um ovo estrelado. (S)Ouve-se um atendedor autom depois, o governo anunciou a queda livre das suas reservas em divisas (me s. «Fiz Álgebra II e demos a raiz quadrada -- uma coisa que nós aprendemo lém, na gestão das verbas do saco azul do Instituto Português de Oncologi

Concordance 8 – Nouns: NA

The extraction of compound nouns using their morphological attributes is also a possibility. Concordance 9 contains a sample of the 15,863 NDN masculine singular nouns (the totality of NDN nouns in the corpus is 35,588), matching the regular expression: <N+NDN-ns>.

Concerning semantic information, in spite of the dictionary incompleteness, the extraction of nouns by semantic attributes is already a possible option. Regular expressions such as: (<N+Hum> + <N+Vest> + <N+Econ>) match all the nouns having the attributes: «human», «clothing» and «economy». We illustrate this possibility with «clothing» nouns of the form NA and NDN, using the regular expression: (<N+NA+Vest> + <N+NDN+Vest>). Examples of matching patterns are provided in Concordance 10.

Concordance 9 – Masculine singular nouns

Concordance 10 – «Clothing» Nouns

3.2 Linguistic Corpus Annotation

Using regular expressions and automata, the linguistic information associated to compounds can be merged in the corpus to tag MWUs. Different examples of corpus tagging will be provided in the following sections. A simple format of annotation will be used: square brackets demarcate compounds from the surrounding text, angled brackets delimit linguistic tags.
Part-of-speech tagging will be illustrated with the annotation of compound adverbs in the sample-corpus. The lexicon contains about 2,400 entries, a small number when compared with the totality of nominal entries. However, adverbs have a high rate of recurrence in texts. In the sample-corpus, 45,259 multiword adverbs (representing 1,516% of the text) were recognized (even if, due to the ambiguity of adverbs, some of the recognized sequences are incorrect). Concordance 11 contains a sample of those tagged adverbs.

Due to the idiomatic flavour of most adverbs and lack of corresponding literal expressions in other languages, their identification and tagging have an obvious interest to different domains, particularly to translation (both human and automatic).

Semantic tagging

The semantic attributes worked up so far can be introduced into texts. Almost all the nouns designating individual and collective «humans», as all as some subclasses, such as «human occupations», have been semantically encoded. Concordances 12 and 13 contain, respectively, a few examples of the 7,415 Hum and 7,473 Hum+Cargo nouns annotated in the corpus.
Concordance 12 – Semantic Tagging: human nouns

Notice that a number of compound nouns classified as **humans** are constituted by words that, isolated, do not have that semantic value: **bombo da festa**, **cabeças coroadas**, **fontes militares**, **lendas vivas**, etc. The nouns included in the subclass «human occupation» seem to be more compositional (Concordance 13).

Concordance 13 – Semantic Tagging: human occupations

The attribute **Cul** was associated to «edible» nouns that can appear in direct object position of verbs such as **comer** [to eat]. A sample of those nouns is presented in Concordance 14.

Concordance 14 – Semantic tagging: edible nouns

3.2.3 Syntactic Tagging

Part-of-speech tagging can be considered as syntactic information. But a more clearly syntactic categorization involves compound adjectives. Like simple adjectives, compound adjectives were included in different subclasses, according to their syntactic properties (Carvalho, 2001). Concordance 15 is a sample of predicative adjectives, denoting colour, that have been tagged in the corpus.
In Portuguese, as well as in other languages, syntactic and semantic properties of adjectives are correlated. In Portuguese, adjectives denoting colour are a subclass of predicative adjectives, coded Pco.

4. Concluding Remarks

Following M. Gross linguistic approach to the lexicon, large-coverage computational lexicons, for simple and multiword lexical units were built for Portuguese. Finite-state methods were adopted: (i) to formalize linguistic information, (ii) to associate simple and compound lexical entries, (iii) to generate simple and compound inflected forms.

A raw Portuguese corpus was exploited to collect new multiword compound nouns. That lexical acquisition led to a significant increase of the dictionary coverage (from about 22,000 inflected compound nouns to about 76,000). Dictionary enlargement will pursue.

In this paper it has been shown that:

- Due to the linguistic characteristics of lexical compounds (constrained sequences of words), semantic attributes can be associated to dictionary entries on a non-intuitive basis. The addition of semantic features to the dictionary entries has a crucial importance to most applications that use dictionaries.

- From a computational (and linguistic) point of view, the adequate identification of lexical compounds avoids the generation of a number of erroneous analyses resulting from the different values (homography, ambiguity) of their constituent simple words. The correct identification of lexical compounds avoids their incorrect analysis as free syntactic structures;

- The application of the enhanced dictionaries to a raw Portuguese corpus provided quantitative evidence to the idea that lexical compounds are very frequent in texts.

- The linguistic information associated to compounds can be annotated in corpus;

- Morphological, syntactic and semantic tags can then be used to extract linguistic knowledge from corpus.
5. References


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


UNITEX, http://www-igm.univ-mlv.fr/~unitex/