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

Elisabete Marques Ranchhod
Department of Linguistics & LabEL / Onset
University of Lisboa
elisabet@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 demostrated (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, 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 (Ranchhod 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_ss001+Pd	[high, tall, loud]
alto, N201_dh201	[hillock; bump]
antes,. ADV	[before]
atacar, V102t	[to attack]
atómico, 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 (**V**erb, **INTERJ**ection, **A**djective, **N**oun, **ADV**erb); 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.

atacado, atacar.V:K	alta, alto.A+Pd:fs
atacam, atacar.V:P2'p:P3p	altas, alto.A+Pd:fp
atacámos, atacar.V:J1p	altinho, alto.N:Dms
atacando, atacar.V:G	altinhos, alto.N:Dmp
atacar,. V:U1s:U2's:U3s:W:V1s:V2's:V3s	altíssima, alto.A+Pd:Sfs
atacará, atacar.V:F2's:F3s	altíssimas, alto.A+Pd:Sfp
atacáramos, atacar.V:M1p	altíssimo, alto.A+Pd:Sms
atacares, atacar.V:U2s:V2s	altíssimos, alto.A+Pd:Smp
atacaria, atacar.V:C1s:C2's:C3s	alto, alto.A+Pd:ms
atacas, atacar.V:P2s	alto, alto.INTERJ
atacasses, atacar.V:T2s	alto, alto.N:ms
atacava, atacar.V:I1s:I2's:I3s	altos, alto.A+Pd:mp
ataque, atacar.V:S1s:S2's:S3s	altos, alto.N:mp
ataquei, atacar.V:J1s	antes, antes.ADV
ataqueis, atacar.V:S2p:Y2p	atómica, atómico.A+Rel:fs
ataquemos, atacar.V:S1p:Y1p	atómicas, atómico.A+Rel:fp
ataques, atacar.V:S2s:Y2s	atómico, atómico.A+Rel:ms
[...]	atómicos, atómico.A+Rel:mp

Sample 2 - Dictionary of inflected forms

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: S1s: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.

alto(A001) comissário(N001),N+AN+Hum+Cargo	[high commissioner]
alto(A216) mar,.N+AN+Geo	[open sea]
altos(N292) e baixos,.N+NConjN+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]
era(N392) atómica,.N+NA+Temp	[atomic era]
saxofone(N201) alto(A201),N+NA+Mus	[alto saxophone]
segredo(N201) atómico (A201),NA+Abst	[atomic secret]

Sample 3 – Uninflected MWU

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 contibutes 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 occuring 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), that inflection is constrained. The same element (i) inflects in gender and number (A001): *alto*

comissário; (ii) inflects in number but not in gender (A201): *chapéu alto*; *saxophone 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+NConjN+Abst:mp	[ups and downs]
ângulo de ataque, ângulo de ataque.N+NDN+AerDin:ms	[angle of attack]
ângulos de ataque, ângulo de ataque.N+NDN+AerDin:mp	
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]
ataques de asma, ataque de asma.N+NDN+Med:mp	
ataques pessoais, 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 bellow).

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-mouxe* [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], 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 an can be quantifyed 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-metade* [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-arbítrio* [free will]); inside the dictionary entries, some examples of MWUs can also be found. The first version of LABEL-LEX-mw 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.

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

Table 5 – Productive Classes of MWU nouns

In order to enlarge the dictionaries, corpora were used to extract new multiword nouns. The CETEMPublico, 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 the 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).

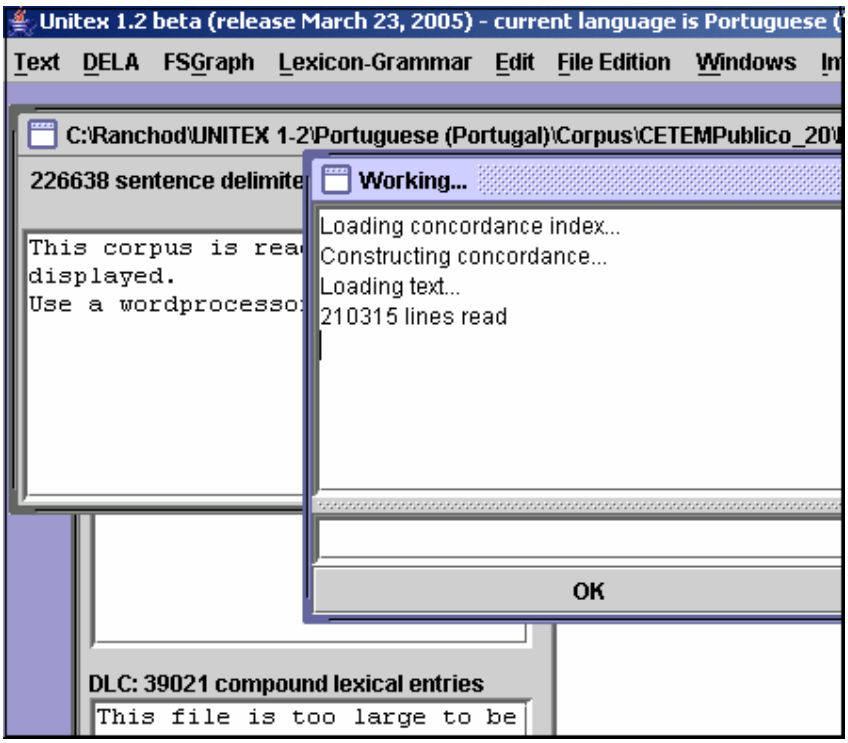

Figure 6 – Recognized compounds in the sample corpus

Examples of the totality of recognized compounds in the sample-corpus are given in Concordance 7.

al social vai ser analisada a curto prazo . Para Álvaro de Carvalho, «e
pelas 20h30, uma assembleia a fim de eleger os corpos gerentes do Salg
s estava marcado para ontem à tarde , fosse qual fosse o resultado da vo
. A Vitrina -- para quem «o acto público do concurso está inquinado de i
s milhões de contos anuais. Antes de ser comprada pelo futuro bilionári
melhor ao poder americano», ao mesmo tempo tem meios para ajudar as camp
a ligação Marateca-Caia e a auto-estrada Marateca-Grândola-Vila Real de
a, através de um túnel e do bairro degradado em que residem diversas fam
e de passagem, foi bastante bem sucedida . A participação de Phil Colli
nde confusão». Cravinho foi cabeça-de-lista nas eleições europeias de 1
azo para a data do eventual cessar-fogo . A delegação da Renamo em Portu
r regime, relativamente aos cheques pré-datados, a que o novo diploma re
a décima Supertaça FC Porto com o credo na boca Manuel Mendes e Luís Oc
primeira vez apresentado ao conselho de ministros dos Transportes que s
ar-se com os seus inimigos. De momento , bastam-lhe os amigos. James Tob
Marques Mendes a caminho do debate parlamentar do Orçamento de Estado, c
-se que as obras se iniciem dentro de alguns meses. No âmbito desta emp
r em Portugal a produção de detergente líquido para o mercado ibérico, a
divisas, operação na qual, diga-se de passagem, foi bastante bem sucedi
os próximos tempos, a maior feira do livro americana acabou na segunda-f
Portugal admitiu durante o fim de semana que os trabalhos na capital i

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 aborto terapêutico, por motivos de saúde, ma
em introduzir na cena. {S}A bailarina principal teria de usar um cacheco
o -- 9,1% do PIB {S}Saldo da balança comercial -- .++ de 6,3 mil milhões
ine a missa na Sé Velha, com canto gregoriano, por volta das 10h00, começ
nada mudou, com excepção das caras novas na direcção do PC. {S}Testemunho
il, como quem está a receber dinheiro vivo. {S}Segundo o mesmo documento,
speculativos se convertem no eixo central da economia do país. {S}O que a
ou de ser confrontado com um facto consumado, sem qualquer esclarecimento
orcionavam uma vasta gama de géneros agrícolas, que nasceu Huambo, no cen
alar para o trágico e para o humor negro, passando pelo sórdido. {S}A re
do com tal fascínio no nosso imaginário colectivo, a sua silhueta aumento
uando se espera pelo dia do «juízo final» da TVI, tema da ordem do dia sã
ampeão argentino, verdadeira lenda viva do automobilismo de competição, a
o operar, basta escolher no mapa celeste que surge no ecrã o astro que t
S}Esta adaptação francesa da 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:ms>.

para 20 contos. {S}Quanto ao abono de família, regime geral, a sua actua
{S}Esta crise é, em parte, um braço de ferro entre a classe política e o
sa autonomia é capaz de ser o calcanhar de Aquiles de toda esta temática.
tinada a retirar o excesso de dióxido de carbono -- e introduzem uma palh
es de gases responsáveis pelo efeito de estufa antes do ano 2010, apesar
riado mais celebrado nos EUA, fim-de-semana prolongado com paradas milita
mas a que falta ainda aquele golpe de asa que transforma as coisas em ar
com a pasta da Defesa, mas um homem de confiança de Eyadéma, Aboudou Asso
m défice grotesco é apenas um incidente de percurso.» idem «Se Monteiro d
onde o amor se transforma em jogo de poder («Deixas-me aqui deitado/ o t
es mentais. {S}Na posse de um livro de cheques do Banco de Galicia, passo
o em que, disse, foi mais «um moço de recados do que director». {S}Coutin
falar foi o «coronel Parra», nome de guerra com que nos grupos clandesti
será exagero falar de um novo ovo de Colombo. {S}Para que os utilizadores

Concordance 9 – Masculine singular nouns

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.

onde têm que deixar os seus bonés de pala. «Quando me disseram que tinha
S}Para este criador de moda, calças de ganga e elegância são «conceitos op
isão, mas nunca foi visto em calções-de-banho na praia e muito menos deixo
passagem e os cavalheiros de chapéu alto fizeram acenos à multidão. 1910,
colete, muitas vezes usando chapéu de coco, num pequeno país do interior
úblicas? Já constatou que do fato de banho transbordam pedaços do que em t
ltura da estreia, trazia uma saia rodada, até aos pés, feita de retalhos d
s 2h00, apenas vestida com a roupa interior, uma doméstica, de 44 anos, re
lo azul escuro e, a rematar, sapatos de ténis. {S}Sempre que perdia a bola
{S}Havia beldades com longos vestidos de baile e ombros muito marcados - a

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.

3.2.1 Part-of-speech Tagging

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.

```
uito trabalho e a elaboração, [a curto prazo <Adv>], de um plano estratégi
heiro. {S}O negócio progredia [a olhos vistos <Adv>]. {S}Primeiro abriu co
s dos estilistas portugueses, [afinal de contas <Adv>] a verdadeira razão
futuro, com Nogueira a dizer, [alto e bom som <Adv>], que a «análise e cor
e vai fazer para o Brasil? {S}[Ao fim e ao cabo <Adv>], a França de 1993 n
enager Of The Year» é a capa, [ao mesmo tempo <Adv>] grotesca e divertida,
 o empresário José Varandas: «[Até agora <Adv>], tenho preferido preparar
elo IRC: «Gosto mais de falar [cara a cara <Adv>]». {S}O Alexandre tem uma
epública como a lei exige. {S}[Com efeito <Adv>], a homologação só ocorreu
s Clérigos foi então renovada [de alto a baixo <Adv>] (trabalho que ficou
ecitador inimitável. {S}Sabia [de cor <Adv>] todas as quadras do António A
terres». «Isso representaria, [de facto <Adv>], uma forma de fomentar arti
s de Almeida {S}Está frio. {S}[De repente <Adv>] faz-se silêncio. {S}E as
da guerra, em que podia andar [dias a fio <Adv>] pelo mato, a atravessar r
 centristas, o que deverá ser [em breve <Adv>] divulgado publicamente pelo
 e Desportos por não cumprir, [em devido tempo <Adv>], os compromissos fin
que vem a público, baseia-se, [grosso modo <Adv>], em interesses: de rigor
danças e cânticos rituais. {S}[Hoje em dia <Adv>], porém, são poucos os qu
supermilitarizadas, as quais, [mais cedo ou mais tarde <Adv>], serão afect
io apenas deverá ser lançado, [na melhor das hipóteses <Adv>], em 1999, qu
a Tunísia em Maio de 1915. {S}[No entanto <Adv>], há missões que compensam
ugal, onde se preferiu seguir [passo a passo <Adv>] todas as fases da dist
chegaram e, fazendo as contas [por alto <Adv>], estimo que os estragos dev
snios. {S}Segundo: dizer-lhe, [preto no branco <Adv>], que se a ONU não de
itais, as vacas loucas, foram [pura e simplesmente <Adv>] ignorados. {S}Co
a recente que vai transformar [sem sombra de dúvida <Adv>] o quotidiano po
ção processual que se destina [tão-somente <Adv>] a obter efeitos fiscais,
enças em cujo texto se repete [tintim por tintim <Adv>] não a lei e o dire
 gosta disso é a polícia que, [volta e meia <Adv>], apreende os artigos ao
```

Concordance 11 – PoS Tagging: 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).

3.2.2 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.

```
i emprestado por um [amigo de longa data <N+Hum:ms>] e pelos seus irmãos.
oite de Natal com a [avó materna <N+Hum:fs>], mas foram depois entregues
e «Juans Guerra», o [bombo da festa <N+Hum:ms>] do carnaval de 1990. {S}T
. {S}O encontro das [cabeças coroadas <N+Hum:fp>] legítimas em Kétu -- o
r a enorme falta de [dadores humanos <N+Hum:mp>]. «O problema deste tipo
ham perdido os seus [entes queridos <N+Hum:mp>]», disse. {S}Este acidente
», defendeu uma das [fontes militares <N+Hum:fp>] ouvidas pelo PÚBLICO. {
os, que protegem as [grã-duquesas <N+Hum:fp>] da cólera popular e que se
```

está situado. {S}Um [homem de idade <N+Hum:ms>], pequeno e magro, com mad
rceiram com algumas [lendas vivas <N+Hum:fp>] do fado. {S}Nostalgia. {S}I
ulher de sociedade, [musa inspiradora <N+Hum:fs>] de dois músicos, que ac
o não», explica uma [recém-licenciada <N+Hum:fs>] em Direito. {S}Quanto a

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).

nde trabalhava como [adida cultural <N+Hum+Cargo:fs>] e, em part-time, co
e com a nomeação do [alto-comissário <N+Hum+Cargo:ms>]. {S}Feytor Pinto,
0 anos. {S}Passou a [bispo titular <N+Hum+Cargo:ms>] da diocese em 8 de S
icida», entendem os [capitães-de-fragata <N+Hum+Cargo:mp>] Brites Nunes e
com declarações da [directora-executiva <N+Hum+Cargo:fs>] da Apifarma, I
lica Luís Pinheiro, [engenheiro-chefe <N+Hum+Cargo:ms>] da circunscrição
com as palavras dos [guardas-florestais <N+Hum+Cargo:mp>], que darão a co
dos seus colegas, o [intendente-mor <N+Hum+Cargo:ms>], o superpolícia do
polícia, e de duas [juízas de instrução <N+Hum+Cargo:fp>], em casa de Du
cípio de Janeiro, o [ministro-sombra <N+Hum+Cargo:ms>] dos Negócios Estra
}Tony Galsworthy, o [negociador-chefe <N+Hum+Cargo:ms>] da Grã-Bretanha,
padas». {S}A antiga [primeira-ministra <N+Hum+Cargo:fs>] disse ter «perdo

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.

xes e afins, luziam [arroz de marisco <N+Cul:ms>] (mínimo duas doses, sen
ação», concluiu. {S}[Arroz doce <N+Cul:ms>] no forno 6 pessoas fácil/rápi
s -- papos-de-anjo, [barrigas-de-freira <N+Cul:fp>], conventuais ... --,
s cereais, galinha, [batatas fritas <N+Cul:fp>], iogurte e doces. {S}O co
ia;{S} jantei cedo, [bolachas de água e sal <N+Cul:fp>] com queijo-creme
rissóis, croquetes, [bolinhos de bacalhau <N+Cul:mp>], folhadinhos, bacon
o público fatias de [bolo inglês <N+Cul:ms>], os actores dão corpo a pers
ia às rabanadas, ao [bolo-rei <N+Cul:ms>] e ao vinho fino, no alegre rema
duas sorvedelas de [caldo verde <N+Cul:ms>] e três pieguices de saudade
garmente chamado `o [fiel amigo <N+Cul:ms>]'», lê-se na Grande Enciclopéd
os «comer» frutos e [frango assado <N+Cul:ms>] para restaurar as energias
s em calda (37500), [leite-creme <N+Cul:ms>] queimado (20000), pudim de l
guias fritas ou com [migas de bacalhau <N+Cul:fp>], sável frito, grelhado

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.

orta. Veste um pólo [azul escuro <A+AA+Pco>], um lenço grená e umas calças
cobrem-se com panos [branco cru <A+AA+Pco>] suspensos por cordéis. Logo po
cor globalizadora ([brancos sujos <A+AA+Pco>] e texturados) e com as figu
ama (LV), as cores ([castanho escuro <A+AA+Pco>] e tonalidades dourado-mat
ce um cenário muito [cor-de-rosa <A+NDN+Pco>]? «As pessoas habituam-se», c

a um tubinho de luz [verde fluorescente <A+AA+Pco>], «tees» e um pequeno
rejeitar a «boina – [verde seco <A+AA+Pco>] – o punhal e a insígnia» dos
azinha do tejadilho [verde-alface <A+AN+Pco>]. «E há outra situação que i
haga com o uniforme [verde-azeitona <A+AN+Pco>] de sempre mas apareceu no

Concordance 15 – Syntactic tagging: predicative adjectives denoting colour

In Portuguese, as well as in other languages, syntactic and semantic properties of adjectives are correlated. In Portuguese, adjectives denoting colour, coded **Pco**, are a subclass of predicative adjectives.

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 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. On the other hand, the correct identification of lexical compounds prevents 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. (2001) *Gramáticas de Resolução de Ambiguidades Resultantes da Homografia de Nomes e Adjectivos*. Tese de Mestrado, Universidade de Lisboa.

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.

Carvalho, P. and Ranchhod, E. (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*, EACL'03, 105-112 .

Gross, M. (1986) Lexicon-grammar. The representation of compound words. *Proceedings of COLING '86* (Bonn: University of Bonn).

Gross, M. (1989) The Use of Finite Automata in the Lexical Representation of Natural Language, in *Electronic Dictionaries and Automata in Computational Linguistics, Lecture Notes in Computer Science 37* (Berlin/New York: Springer), 34-50.

Mota, C.; Carvalho, P.; Ranchhod, E. (2004) Multiword Lexical Acquisition and Dictionary Formalization, in *Proceedings of the Workshop Enhancing and Using Electronic Dictionaries*, *Coling 2004*, Geneva, Switzerland, 73-77.

Paumier, S. (2004) *Unitex 1.2, Manuel d'utilisation*, Université de Marne-la-Vallée (http://www-igm.univ-mlv.fr/~unitex/)

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

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

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

Ranchhod, E.; Carvalho, P.; Mota, C.; Barreiro, A. (2004) Portuguese Large-scale Language Resources for NLP Applications, in *Proceedings of the IV Conference on Language Resources and Evaluation, LREC* , Lisboa, 1755-1759.

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

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

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