Second ACL Workshop on Multiword Expressions: Integrating Processing, July 2004, pp. 72-79

Frozen Sentences of Portuguese: Formal Descriptions for NLP

Jorge Baptista
Universidade do Algarve
Campus de Gambelas
P-8005-139 FARO
jbaptis@ualg.pt

Anabela Correia
Universidade do Algarve
Campus de Gambelas
P-8005-139 FARO

Graça Fernandes
Universidade do Algarve
Campus de Gambelas
P-8005-139 FARO

Abstract

This paper presents on-going research on the building of an electronic dictionary of frozen sentences of European Portuguese. It will focus on the problems arising from the description of their formal variation in view of natural language processing

1 Introduction

Defining frozen sentences is not easy, and many conceptual and terminological disputes on concurrent terms ('idioms', 'collocations', 'phrasèmes', etc.) can be found in the literature (M. Gross 1988; G.Gross 1996; Mejri 1997; Mel'cuk 1993; Mogorrón-Huerta 2002; Gaatone 2000; Jurafsky & Martin 2000: 571-573; McKeown & Rodev 2000; Mutsimoto 2003: 395). As a first approach to a (consensual?) definition, frozen sentences are elementary sentences where the main verb and at least one of its argument noun-phrases are distributionally constraint, and usually the global meaning of the expression cannot be calculated from the individual meaning of its component elements when they are used independently (M. Gross 1982, 1989, 1996; G. Gross 1996; Ranchhod 2003). For that reason, the whole expression must be taken as a complex, multiword lexical unit. For example, in: (1) *O João matou dois coelhos de uma cajadada* (lit: 'John killed two rabbits with one blow', John killed two birds with one stone) the verb-object combination (*matar-coelhos*) is frozen. One cannot replace *coelhos* (rabbits) for another animal and the numeral determinant is necessarily *dois* (two). Also, it is not possible to modify *coelhos* with any free adjective (*dois coelhos *gordos*/ **grandes*, two fat/big rabbits). In addition, the adverbial *de uma cajadada* (with one blow) can hardly be zeroed, or the meaning of the sentence becomes literal. On the other hand, frozen sentences usually present some, often highly constraint, formal variation. For the most part, this variation is strictly lexical. In this case, in the adverbial, the noun *cajadada* (lit: a blow with a stick) can be replaced by *assentada* and *vez* (turn), but the meaning of the expression remains unchanged. This variation does not happens elsewhere. Furthermore, if we disregard proverbs[1], completely frozen sentences are rare. Usually, one or more of its argument noun phrases are distributionally free. In this case, any human noun can adequately occupy the structural position of subject. The frozen verb-noun combination is responsible for this distributional constraint, which can be considerably different from the constraints imposed by the verb when functioning as an independent lexical unit. For example, the verb *matar* (to kill) admits both human and non-human (animate and abstract) nouns for its subject when its object is *coelhos* (rabbits): (2a) *O João/a raposa/o tiro/a praga matou dois coelhos* (John/the fox/ shooting/ plague killed two rabbits). Another way frozen sentences often differ from free sentences is the fact that they block transformations that should otherwise be possible, given the syntactic properties of the main verb and its arguments. Hence, while it is possible to form from (2a) the passive sentence (2b): (2b) *Dois coelhos foram mortos pelo João/ a raposa/ o tiro/ a praga* (Two rabbits were killed by John/ the fox/ shoot-

[1] Proverbs differ from frozen sentences (a) from not having any free elements, (b) in the way they function in discourse, which is similar to quotations and (c) for their general value as advices or as atemporal truths about human life. However, partly because of their being an oral tradition, even proverbs can show some formal variation (Chacoto 1994).

ing/ plague), the same does not happen with (1): (1a) º*Dois coelhos foram mortos pelo João de uma cajadada* (Two rabbits were killed by John with one blow), since the meaning of the sentence becomes literal (this is shown by symbol 'º'). Finally, frozen sentences constitute a non-trivial problem for many NLP applications. Since they are multiword expressions functioning as meaning units, they have to be identified as a block. However, their elements can appear discontinuously and they can also present some formal variation. They are often ambiguous, the same sequence having both a literal and a figurative meaning – and in this case, only an extended context can disambiguate them. They constitute an important part of the lexicon, comparable to (but probably much larger than) simple verbs.

2 Classification of Frozen Sentences

Many frozen sentences, especially those that are most usual or most obviously idiomatic, have already been collected both in general and in specialized dictionaries of 'idioms'. In these dictionaries, frozen sentences are usually undistinguished from other types of multiword expressions, such as compound nouns, adverbs, prepositions, conjunctions, proverbs, and so on. In order to build an electronic dictionary of frozen sentences of European Portuguese, several sources where used, including specialized dictionaries[2], and completed with newspapers, magazines, etc. and our knowledge as native speakers of Portuguese. The formal framework of M. Gross (1982, 1989, 1996; see Leclère 2002, for an updated overview) was adopted to classify frozen sentences. The classification is based on the sentence structure, the number and type of noun phrases attached to the main verb, their frozen or free nature, as well as the syntactic properties of the construction. Table 1 shows some formal classes[3], their internal structure, an illustrative example, and the approximate number of sentences collected so far. Compared with figures available for other languages – French (+20,000; M.Gross 1996), Spanish (3,500; Mogorrón-Huerta 2002), Greek (4,500; Fotopoulou 1993) and Brazilian Portuguese (3,500; Vale 2001), it is clear that these lists are still far from complete and should, in fact, be completed, probably using other corpus-based methods for lexical acquisition (McKeown & Rodev 2000, Mutsimoto 2003).

Table 1- Classification of frozen sentences (extract)

Class	Structure/Example	Size
C1	N_0 **V** C_1	
	O Pedro matou a galinha dos ovos de ouro	800
CAN	N_0 **V** (**C de N**)$_1$ = C_1 **a** N_2	
	O Pedro arrefecer os ânimos de=à Ana	200
CDN	N_0 **V** (**C de N**)$_1$	
	O Pedro queria a cabeça da Ana	100
CP1	N_0 **V Prep** C_1	
	O Pedro bateu com a porta	900
CPN	N_0 **V Prep** (**C de N**)$_1$	
	O Pedro foi aos cornos do João	100
C1PN	N_0 **V** C_1 **Prep** N_2	
	O Pedro arrastou a asa à Ana	400
CNP2	N_0 **V** N_1 **Prep** C_2	
	O Pedro tirou o relógio do prego	350
C1P2	N_0 **V** C_1 **Prep** C_2	
	O Pedro deitou mãos à obra	400
CPP	N_0 **V Prep** C_1 **Prep** C_2	
	O Pedro foi de cavalo para burro	200
CPPN	N_0 **V** C_1 **Prep** C_2 **Prep** C_3	
	O Pedro deitou o bebé fora com a água do banho	50
Total		3,500

3 Format of Dictionary

The electronic dictionary is composed of several matrices, one per formal class. In these matrices, each line is a frozen sentence and the columns contain the lexical elements of the sentence and their syntactic (distributional and transformational) properties. The set of matrices constitute the lexicon-grammar of frozen sentences. Table 2 is a sample of class CPN. This class is defined by having a prepositional phrase where only the head-noun *C* is frozen with the verb, while its determinative complement *de N* (of N) is free: N_0 *V Prep* (*C de N*)$_1$.

[2] Basically, Mello 1986, Moreira 1996, Neves 2000, Santos 1990 and Simões 1993. The electronic dictionary of frozen sentences of Brazilian Portuguese (Vale 2001) was also consulted, but many of those sentences either do not exist if European Portuguese or else present substantial syntactical and lexical differences, so that a detailed comparative study is in order.

[3] Frozen sentences with sentential subjects or objects, or with frozen subject noun phrases were not considered in this paper. *N* and *C* stand for noun phrases; *N* is a free and C is frozen noun phrase; N_0 is the subject, N_1 and N_2 the first and second complement; *V* is the verb and *Prep* a preposition.

Table 2 – Class *CPN* (extract)

A	B	C	D	E	F	G	H	I	J	K	L	M
N0=:Nhum	*N0=:N-hum*	V	*Vse*	*NegObrig*	*Prep*	*Det*	C	*N1=:Nhum*	*N1=:N-hum*	*de N=a N*	*de N=Poss*	Example
+	-	<acabar>	-	-	com	a	raça	+	+	+	+	*O Pedro acabou com a raça da Ana*
+	-	<atirar>	+	-	a	os	pés	+	-	+	+	*O Pedro atirou-se aos pés da Ana*
+	-	<chegar>	-	+	a	os	calcanhares	+	-	+	+	*O Pedro não chega aos calcanhares da Ana*
+	-	<cortar>	-	-	em	a	casaca	+	-	+	-	*O Predro cortava na casaca da Ana*
+	-	<ir>	-	-	a	as	trombas	+	-	+	-	*O Pedro foi às trombas do João*
+	-	<ir>	-	-	em	a	cantiga	+	-	-	+	*O Pedro foi na cantiga da Ana*
+	-	<ir>	-	-	a	a	cara	+	-	+	-	*A Ana foi à cara do Pedro*
+	-	<pegar>	-	-	em	a	deixa	+	-	+	+	*O Pedro pegou na deixa da Ana*
+	-	<rir>	-	-	em	a	cara	+	-	+	+	*O Pedro riu na cara da Ana*
+	-	<rir>	+	-	em	a	cara	+	-	+	+	*O Pedro riu-se na cara da Ana*
-	+	<sair>	-	-	de	o	pelo	+	-	+	-	*O salário sai-lhe do pelo*
-	+	<subir>	-	-	a	a	cabeça	+	-	+	+	*A fama subiu à cabeça do Pedro*
+	-	<viver>	-	-	em	a	sombra	+	+	-	+	*O Pedro vive na sombra da Ana*

4 Syntactic Properties

For lack of space, only some of the most prominent properties will be considered here. For each sentence, the distributional constraints (human/ non-human noun) on the free noun phrases are indicated: '+' if the sentence admits it, '-' if it does not. In this class, these are the subject and the determinative complement of *C*. Usually, the verb can inflect freely, thus its lemma is shown in brackets < >. In some sentences, *V* presents an intrinsically reflexive construction (noted *Vse*): (3) *O Pedro atirou-se aos pés da Ana <e pediu-lhe para ficar>* (Peter threw himself to the feet of Ana <and begged her to stay>. The reflex pronoun cannot be replaced by a noun phrase of the same distributional nature but not coreferent to the subject: *O Pedro atirou *o João* / **o livro aos pés da Ana* (Peter threw John/the book to the feet of Ana). Usually, this reflex pronoun cannot be zeroed. However, some verbs allow this zeroing of the pronoun: (4) *O Pedro riu*(*-se*) *na cara da Ana* (Peter laughs (himself) at the face of Ana). In this case, the simplest way is to double the entry. Also in some cases, there is an obligatory negation (`NegObrig`): (5) *O Pedro não /nunca /nem chega aos calcanhares da Ana* (lit: Peter does not/never/not even gets to the heels of Ana, Peter is not a match for Ana). The half-frozen noun phrase can often undergo an operation called dative restructuring (Leclère 1995), that splits the noun phrase in two and where the determinative complement becomes a dative complement *a N* (to N) of the verb: (6) *O Pedro foi às trombas* ***do João*** = ***ao João*** (lit: Peter went to the snouts of/to John, Peter hit John). This operation depends on the metonymical relation between *C* and the noun of its free determinative complement. The new dative complement can be reduced to a dative clitic pronoun (*-lhe*): (6a) *Pedro foi-lhe* (= *ao João*) *às trombas do João*. However, dative restructuring must be systematically checked for each sentence, since some sentences do not admit it: (7) *O Pedro foi na cantiga do João*/ **ao João*/ **-lhe* (lit: Peter went in the song of John, Peter was persuaded by John's ill-intended words). In some cases, however, the restructured noun phrase: (8a) *O salário do Pedro sai-lhe do pelo* (lit: Peter's salary gets him out from the fur, To earn his salary, Peter has to work very hard) is much more acceptable than its basic form: *?*O salário sai do pelo do Pedro* (The salary gets out from Peter's fur) [4]. The determinative complement can also be reduced to an oblique or to a possessive pronoun: (7a) *O Pedro foi na cantiga* ***do João*** = *na cantiga* ***dele*** = *na* ***sua*** *cantiga*, but in some cases, the reduction to a possessive is blocked: (6b) *O Pedro foi às trombas* ***do João*** = *?às trombas* ***dele*** = **?às* ***suas*** *trombas*. Finally, there can be some facultative, free insertions be-

[4] In this case, the pronouncing may blocked by the (double) metonymical relation of Pedro with both the head noun of both the subject and object noun phrase.

tween the verb and the prepositional phrase: (9) *A fama subiu* (*rapidamente*/*logo*) *à cabeça do Pedro* (Fame went up (quickly/soon) to Peter's head). As one can see, these properties may be independent from each other and some of them can appear in combination in the same sentence. For example, obligatory negation can combine with the reduction do possessive pronoun: (5a) *O Pedro* ***não*** *chega aos* ***seus*** *calcanhares* (lit: Peter does not get to the his heels, Peter is not a match for him) or with the dative pronoun resulting from the noun phrase restructuring: (5b) *O Pedro* ***não te*** *chega aos calcanhares* (lit: Peter does not get you to the heels, Peter is not a match for you). In both cases, the pronouns appear inserted between the characteristic (and fixed) elements of this frozen sentence.

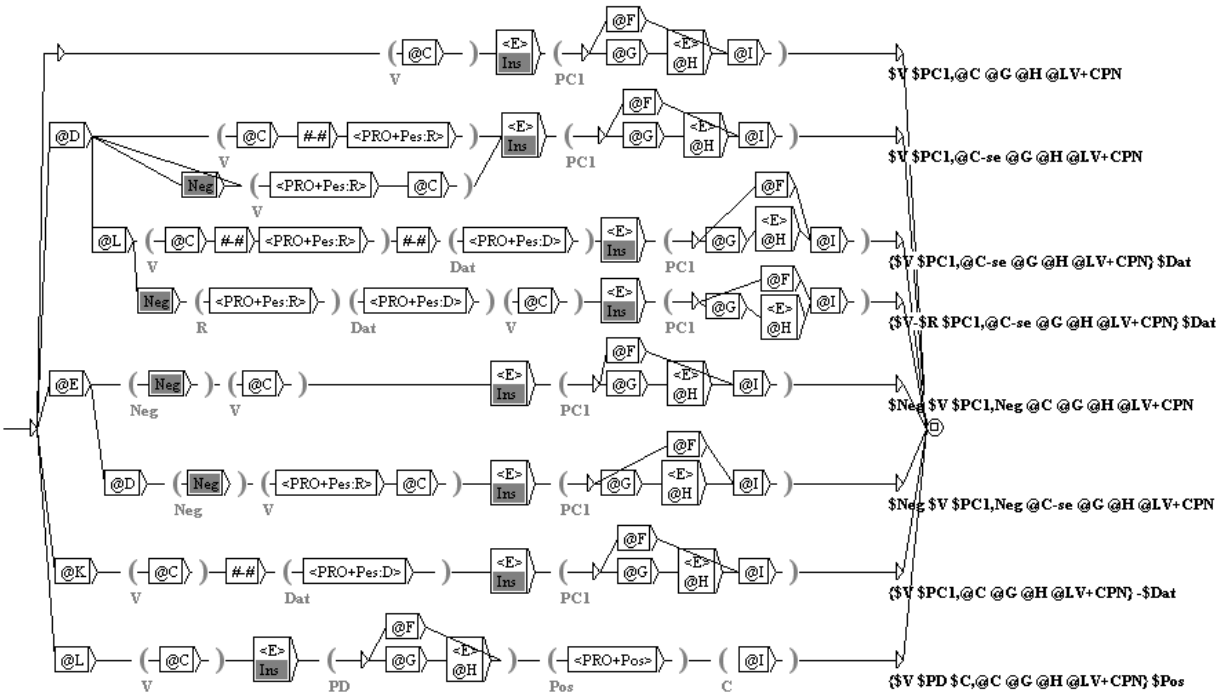

Figure 1. Reference graph for class CPN

5 Lexical Finite-State Transducers: building a reference graph

The lexicon-grammar of frozen sentences (i.e. the set of matrices) cannot be directly applied to recognize these expressions in texts. Using INTEX linguistic development platform (M. Silberztein 1993, 2004)[5] it is now possible to build lexical finite state transducer that can identify and tag frozen sentences in the texts where they occur. This is done by building a reference graph for each matrix. Fig. 1 (above) shows a simplified version of the reference graph for class CPN.

This graph describes the formal sequences of the components of the frozen sentences. In this graph, variables such as `@X` refer to the content of the matrix (at column `X`). Furthermore, this graph is an enhanced transducer, where it is possible to define variables containing strings of elements and to reuse these variables in the transduction.

For example, in the top line of the graph, variable *V* (inside brackets) stores the verb (represented by `@C`). This is followed by a facultative subgraph, for any eventual insertions (in the shadowed box), and it is then followed by variable *CP1*, containing the frozen prepositional phrase. The two variables are then reused in the transduction (the output in bold, on the right), providing the multiword lexical entry and its adequate tags: `$V $PC1,@C @F @G @H @I.V+CPN`, so that for the frozen sentence: (7c) *O Pedro* ***foi na cantiga*** *do João* the system would produce the tag: *O Pedro* `{foi na cantiga,ir em a cantiga.V+CPN}` *do João*.

[5] http://www.nyu.edu/pages/linguistics/intex/.

Notice that the inflected form of the verb appearing in the text is lemmatized (after the coma) using the `@X` variables to retrieve the lexical elements in the matrix: in this case, the infinitive *ir* (to go); also in the lemma, the contraction (*na*) of preposition *em* and the definite article *a* is split in its component elements. Finally, variables @D, @E, @K and @L function as switches, reading the '+' or '-' of the corresponding columns in the matrix so that the remaining transitions are either activated or collapsed. These variables correspond to the syntactic properties of the entries. In this way, it is possible to compile a detailed FST that complies with the syntactic restrictions described in the matrix.

After building the reference graph, the system is then able to compile the lexical transducer for this class, exploring, for each line of the matrix, all the paths of the graph, and then determinizing and minimizing the resulting transducer. With this methodology, the linguistic information stored in the lexicon grammar is represented independently from the lexical transducers and can be regularly revised or updated. The reference graph can also be seen as describing linguistic information regarding the surface disposition of the lexical elements entering the frozen sentences of the matrix. Of course, strict formal coherence in needed between the reference graph and the matrix.

6 Application to texts: some experiences

In order to evaluate the performance of the electronic dictionary on real texts, experiences were made using INTEX[6] on three different texts. Two smaller texts, one obtained from the on-line edition of the *Expresso* weekly newspaper [7] and the other a composite text [8] obtained from several sources and used on the *MorphOlimpics* evaluation campaign for Portuguese [9]. We also used a larger, publicly available, journalistic corpus (*CETEM-Público*)[10]. Tables 3 to 5 show results from the application of the modules of the four largest classes to these texts:

Table 3 – Application of dictionary of frozen sentences to text from Portuguese *MorphOlimpics*

Class	DLE	≠L	M	Precision
C1	6	6	6	100 %
C1P2	1	1	1	100 %
C1PN	5	5	6	100 %
CP1	5	5	5	100 %
Totals	17	17	18	100 %

Table 4 – Application of dictionary of frozen sentences to text from *Expresso* newspaper

Class	DLE	≠L	M	Precision
C1	15	13	16	75 %
C1P2	2	2	2	100 %
C1PN	24	18	33	94 %
CP1	37	32	39	100 %
Totals	78	65	90	93 %

Table 5 – Application of dictionary of frozen sentences to text from *CETEMPúblico*

Class	DLE	≠L	M	Precision
C1	541	217	939	78.4 %
C1P2	104	61	157	99.0 %
C1PN	434	96	1,357	96.5 %
CP1	963	309	1,270	88.2 %
Totals	2,042	638	3,723	89.6 %

DLE – number of inflected lexical entries obtained from the text; **≠L** – number of different lemmas; **M** – number of matched strings in text.

These preliminary results must take into consideration the different size of the texts (number of words) and the number of matched sequences. Several comparisons are thus made in Table 6 (below). The size (in number of simple words - *W*) of *Expresso* is approximately 2.4 times larger than *MorphOlimpics*, and the size of *CETEMPúblico* is 116 times larger than *Expresso*:

W(*Exp*)=2.4xW(*MO*);W(*Pub*)=116xW(*Exp*).

Naturally, the number of matches (*M*) does not increase in the same proportion:

[6] We also used an electronic dictionary of simple words of Portuguese (Ranchhod *et al.* 1999), from the public linguistic resources built by LabEL: http://label.ist.utl.pt.

[7] http://www.expresso.pt/. This is a 976 Kb text, with 83,269 (5,764 different) words.

[8] http://acdc.linguateca.pt/aval_conjunta/morfolimpiadas/ts_ml.txt [29-03-2004] This is a 215 Kb text, with 35,053 (10,070 different) words.

[9] http://acdc.linguateca.pt/aval_conjunta/morfolimpiadas/index.html [29-03-2004].

[10] http://www.linguateca.pt/CETEMPublico/. Only the first fragment of this corpus was used. This is a text file of approximately 60 Mb, with 9.6 million (177,368 different) words.

M(*Exp*)=5xM(*MO*);M(*Pub*)=41.38xM(*Exp*).
In fact, in the smaller texts, precision scores are not very significant in view of the reduced number of matches. However, precision does not drop abruptly in the larger corpus (*CETEM-Público*), even if the size of the text and the number of matches increase significantly.

Table 6 – Size of texts and scores of dictionary of frozen sentences (Classes C1, C1P2, C1PN, CP1)

Text	W	≠W	{S}
MO	35,053	10,070	527
Exp	83,269	5,764	8060
Pub	9,632,623	177,368	447,125

Text	DLE	≠L	M	CM	P(%)
MO	17	17	18	18	100.0
Exp	78	65	90	84	93.0
Pub	2,042	638	3,725	3,336	89.6

Text	LexDiv	LexDiv/P	CMDiv	CMDiv/P	FS/S (‰)
MO	1.000	1.000	1.059	1.059	3.416
Exp	1.200	1.286	1.292	1.385	1.042
Pub	3.201	3.574	5.229	5.839	0.746

MO – *MorphOlimpics*; **Exp** – *Expresso*; **Pub** – *CETEM-Público*; **W** – number of simple words; **≠W** – different simple words; **{S}** – sentences; **DLE** – lexical entries of frozen sentences; **≠L** –different lemmas of lexical entries; **M** – matches; **CM** – correct matches; **P** – Precision (%); **LexDiv** – lexical diversity of DLE (DLE/≠L); **CMDiv** – lexical diversity of correct matches (CM/≠L); **FS/S** – average number of frozen sentences per thousand sentences (‰).

Another measure is the ratio between the lexical diversity (LexDiv) of DLE (number of entries/different lemmas) and precision (P). The higher the diversity of the DLE, the lower should be the precision. This, however, does not happen: even if the larger corpus shows lower DLE diversity, the ratio LexDiv/P is higher than in the two smaller texts. Therefore, lower precision in *CETEMPúblico* is not significant. Similarly, one could compare ratio between the lexical diversity of correct matches (correct matches/different lemmas) and precision. In theory, this ratio should be smaller if the lexical diversity of correct matches were higher. Instead, *CETEMPúblico* shows a higher CMDiv/P ratio than the two smaller texts, therefore confirming the significance of the precision rate obtained with the frozen sentences' dictionary. Finally, we present, as an indication of frequency, the average number of correctly matched frozen sentences per thousand sentences in each text. This varies from 1.042 ‰ (*Expresso*) to 0.746 ‰ (*CETEMPúblico*).

7 Discussion

As it was said above, precision of the results is relatively high. In fact, most word combinations in the dictionary of frozen sentences are unique, therefore, unambiguous combinations. The finite-state approach adopted here is based on matching strings that could correspond to the characteristic word sequence of each frozen sentence. However, going through the concordances manually, some incorrect matches were detected and it would be impossible to comment on all those cases here, so only a few cases (all from *CETEMPúblico*), will be discussed.

Mostly, mismatches were due to morphologically (orthographically) ambiguous words. In the sentence: *O general António Ramalho Eanes está de **volta à cena** política* (General António Ramalho Eanes is back to the political scene) *volta* is not verb, but a noun. Still, this noun is a nominalization of *voltar* (with a resultative aspectual value): *voltar à cena = estar de volta a cena* , so that the overall meaning is stil the same. Other nominalizations of the same kind were also founded, e.g. *usar da palavra = fazer uso da palavra* (to speak, in a formal situation). Therefore, the study of frozen sentences with verbal predicates should be extended to their corresponding nominalizations (M. Gross 1986). Another interesting aspect of this example is the fact that the noun *cena* appears with the adjective *política* (political, referring to politics). In the dictionary, the entry of the frozen sentence was just *voltar à cena*. As one can see, it will still be necessary to complete the dictionary allowing *C* to be modified by this kind of adjectives, with which they form compound nouns (G. Gross 1988).

Sometimes, the matched string is formally ambiguous with free constructions: *Como resultado, a Comunidade **dá de si** uma imagem de paralisia* (Has a result, the Community gives of itself an image of paralysis). The expression *dar de si* (class CP1) usually has a non-human subject meaning 'to break','to fall apart'. In this case, however, we have a construction of the noun *imagem*, with two distributionally free complements, such as we find in the next example: *Isso*

deu uma imagem negativa da comunidade (That gave a negative image of the community). If it were possible to identify the noun phrase *a Comunidade* and its head as a human noun, and to associate a construction to the noun *imagem*, the ambiguity would not arise, since the distributional properties stated in the CP1 matrix for *dar de si* would prevent it from being tagged. However, this goes well beyond the mere task of lexical analysis and it would require some parsing procedure to avoid the incorrect lexical tag. Ambiguity also arises from the syntactic operations underwent by a free sentence. These transformations may produce strings that are superficially identical to frozen sentences. In the following example: […] *e em vez de se dirigir às máquinas,* ***foi para a rua*** *que Júlia Pinho teve de caminhar* […] (instead of approaching the machines, it was to the street that JP had to walk), we find a cleft sentence with the form *foi* of verb *ser* (to be). This sentence can be obtained from: *JP teve de caminhar para a rua* (JP had to walk to the street). Now, *foi* of verb *ser* (to be) is ambiguous with *foi* of verb *ir* (to go/walk) appearing in the frozen sentence *ir para a rua* (lit: to go to the street, 'to be fired'). Again, ambiguous strings will not be resolved unless some syntactic parsing is done in order to recognize transformations (in this case, clefting).

The components of certain frozen sentences may, also, be particularly apt to become ambiguous strings. For example, in the expression *fazer das suas* (lit: to do of his-fem.pl., to do mischief) the possessive pronoun is a lexical constant, invariable in both gender, and number, while agreeing in person with the subjet: *Fergie tornou* ***fazer das suas*** (Fergie has done mischief again). However, the possessive often appears as a mere determinant on a free construction, e.g. *João Honrado nunca fez das suas certezas uma muralha de arrogância* (JH has never done of his certainties a wall of arrogance). One could think that people tend avoid this kind of ambiguity, but it is not always so: 6 out of 14 matched sequences *fazer das suas* do not correspond to the frozen sentence. Likewise, with the expression *partir do zero/nada* (to start from scratch) 7 out of 14 matches correspond to the compound adverb *a partir do zero/nada* (from scratch). Notice also that this frozen adverb, undoubtedly related with the frozen sentence but appearing with various verbs, also composed of a compound preposition *a partir de* (from), so the expression is three times ambiguous grammatically. A similar ambiguity between occurs between frozen sentence N_0 *dizer respeito a* N_2 (N_0 *concerns to* N_1; 299 matches) and the related compound adverb *no que diz respeito a N* (*concerning N*; 215 matches).

Finally, while most frozen sentences constitute unique word combinations and present a clear-cut, single meaning (however difficult it may be to define it precisely), more rarely, some of them show multiple meanings. This is the case of *entrar em campo* (to enter the field) or *entrar em cena* (enter the scene) used in some sports or in theater jargon and as a general metaphor to 'begin some activity'. These multi-meaning frozen sentences can be dealt with in the same way as polisemic simple words, by multiplying the number of entries in the lexicon-grammar.

8 Final words

Being an on-going research, it is still too early to write 'conclusions'. We believe that with the continuation of current work the size of the electronic dictionary of frozen sentences of European Portuguese will still increase in a significant way. Present experiments on large corpora will undoubtedly contribute to this goal. Statistical methods for collecting frozen sentences should be combined with our more traditional method of perusing dictionaries.

Finite-state techniques prove to be adequately for the lexical analysis of frozen sentences. Frozen sentences constitute an important part of multiword lexical units of any language. Their identification is an essential part of lexical analysis of texts in view of many NLP applications. They present several, non-trivial difficulties to this task. They may be formed of non-adjacent words or allow some lexically constraint variation of some of their elements. They allow several syntactic transformations, but those operations are lexically determined. They may be ambiguous with free sentences, depending on the linear arrangement of their components. Their accurate recognition may often depend on the previous syntactic parsing of the sentence, which may include checking the semantic attributes of nearby noun phrases. On the other hand, parsing depends

on the availability of information regarding multiword lexical units.

Acknowledgement

Research for this paper was partially funded by FCT-MCES (project grant POSI/PLP/34729/99).

References

Araújo-Vale, Oto, 2001. *Expressões Cristalizadas do Português do Brasil: Uma Proposta de Tipologia* (Ph.D. Thesis). Araquara (Brazil): UNESP.

Chacoto, Lucília, 1994. *Estudo e Formalização das Propriedades Léxico-Sintácticas das Expressões Fixas Proverbiais*. (M.A. Thesis). Lisbon: FLUL.

Fotopoulou, Aggeliki, 1993. *Une classification des phrases à compléments figés en grec moderne*. (PhD Thesis). Paris : Univ. Paris 8.

Gaatone, David, 2000. A quoi sert la notion d' «expression figée» ?, in Buvet, P.-A., D. le Pesant, M. Mathieu-Colas (eds.), *Lexique, Syntaxe et Sémantique*, *BULAG* (hors série), Besançon : Centre Lucien Tesnière/PUFC, pp. 295-308.

Gross, Gaston, 1988. Degrée de figement des noms composés. *Langages* 90. Paris : Larousse, pp.57-72.

Gross, Gaston, 1996. *Les Expressions Figées en Français*. Paris: Ophrys.

Gross, Maurice 1982. Une classification des phrases 'figées' du français. *Revue Québécoise de Linguistique* 11-2. Montréal : UQAM, p. 151-185.

Gross, Maurice 1986. Les nominalisations d'expressions figées. *Langue Française* 69, Paris: Larousse, pp. 64-84.

Gross, Maurice 1988. Les limites de la phrase figée. *Langages* 90. Paris: Larousse, pp. 7-22.

Gross, Maurice 1989. *Les expressions figées : une description des expressions françaises et ses conséquences théoriques*. Rapport Téchnique 8. Paris : LADL-Univ. Paris 7 / CERIL.

Gross, Maurice 1996. Lexicon-Grammar. in K. Brown and J. Miller (eds.). *Concise Encyclopedia of Syntactic Theories*. Cambridge: Pergamon, pp. 244-259.

Jurafsky, Daniel and James H. Martin, 2000, *Speech and Language Processing*. New Jersey: Prentice Hall.

Leclère, Christian, 1995. Sur une restructuration dative. *Language Research* 31-1. Seoul: LRI- Seoul National Univ, pp. 179-198.

Leclère, Christian, 2002. Organization of the Lexicon-Grammar of French Verbs, *Linguisticae Investigationes* 25-1, Amesterdam: John Benjamins Pub. Co., pp. 29-48.

McKeown, Kathleen R. and Dragomir Rodev, 2000, Collocations, in Dale, R., H. Moisl and H. Sommers (eds.) *Handbook of Natural Language Processing*. New York: Marcel Dekker Inc., pp. 507-523.

Mejri, Salah, 1997. *Le figment lexical. Description linguistique et structuration sémantique*. La Manouba (Tunis) : Pub. Fac. Lettres.

Mel'cuk, I, 1993. La phraseologie et son rôle dans l'enseignment / apprentissage d'une langue étrangère. *ELA*, Didier Érudition, pp. 82-113.

Mello, Fernando R., 1986. *Nova Recolha de Provérbios Portugueses e Outros Lugares-Comuns* (2nd. ed.). Lisbon: Ed. Afrodite.

Mogorrón-Huerta, Pedro, 2002. *La expressividad en las locuciones verbales españolas y francesas*. Alicante: Pub. Univ. Alicante.

Moreira, António, 1996. *Provérbios Portugueses*. Lisbon : Ed. Notícias.

Mutsimoto, Yuji, 2003. Lexical Knowledge Acquisition, in Miktov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford: OUP, pp. 395-413.

Neves, Orlando, 2000. *Dicionário de Expressões Correntes* (2nd. ed.) Lisbon: Ed. Notícias.

Ranchhod, Elisabete, Cristina Mota, Jorge Baptista, 1999. A Computational Lexicon for Automatic Text Parsing, *Proceedings of SIGLEX'99*: ACL/NScF, pp. 74-80.

Ranchhod, Elisabete M., 2003. O lugar das expressões 'fixas' na gramática do Português. in Castro, I. and I. Duarte (eds.), *Razão e Emoção*, vol. II, Lisbon: INCM, pp. 239-254.

Santos, António, 1990. *Novos Dicionários de Expressões Idiomáticas*. Lisbon: João Sá da Costa.

Silberztein, Max, 1993. *Dictionnaires électroniques et analyse automatiques de textes : le système INTEX*. Paris : Masson.

Silberztein, Max, 2004. *Intex Manual*. http://intex.univ-fcomte.fr/downloads/Manual.pdf

Simões, Guilherme A., 1993. *Dicionário de Expressões Populares Portuguesas*. Lisbon: D. Quixote.