6th INTEX Workshop
Sofia, Hungary, May 28-30, 2003

Mapping, filtering and measuring impact of ambiguous words in Portuguese [*]

Jorge Baptista
Universidade do Algarve – FCHS
Laboratório de Engenharia da Linguagem – CAUTL – IST

Luís Faísca
Universidade do Algarve – FCHS
Núcleo de Estudos de Representações Sociais – Alberto Caeiro

{jbaptis,lfaisca}@ualg.pt

Abstract

This paper deals with ambiguous simple words of Portuguese. The Portuguese dictionary of simple inflected words contains (DELAF) 936.215 entries, from which there are 889.986 different inflected forms. It is possible to obtain the full list of ambiguous inflected forms (43.126), that is, word forms belonging to different categories and/or lemmas: capital,A/N/N (capital). We may consider A/N/N an ambiguity class. There are 137 ambiguity classes. Each ambiguity class presents a certain level of ambiguity (Amb) that corresponds to the number of lexical entries associated to each ambiguous form (again, for class A/N/N Amb=3). Based on this information it is possible to map how ambiguity affects the lexicon. Using the frequency information associated to the list of tokens of a large corpus (the *CETEMPÚBLICO* corpus, with 200 million words), it is possible to calculate how ambiguity affects real texts. Combining the two types of information, it is possible to devise and evaluate different strategies to reduce lexical ambiguity.

1. Mapping ambiguity in the Portuguese DELAF

The Portuguese DELAF (v. 2), built by the LabEL team[1], has been publicly available since 2002. It contains 936,215 entries, consisting of 889,986 different inflected forms. There are 43,127 different ambiguous inflected forms, that is, word forms belonging to different categories and/or lemmas, which correspond to 89,356 DELAF ambiguous entries[2]. For example, the inflected word *capital* (capital) is ambiguous because it has three entries in the DELAF[3]:

`capital,capital.A:ms:fs`	(as in *pena capital*, 'death penalty')
`capital,capital.N:fs`	(the first city of a country)
`capital,capital.N:ms`	(funds for investement)

[*] This paper was first presented to the *6th INTEX Workshop*, held in Sofia, Hungary, May 28-30, 2003. Research for this paper was partially funded by Fundação para a Ciência e a Tecnologia.

[1] http://label.ist.utl.pt/ .

[2] Mean ambiguity is then 1.0518 analysis per word form in the DELAF, and 2.0719 for each ambiguous word.

[3] Notations: The codes for grammatical categories are transparent: A=Adjective, ADV=Adverb, CONJ=Conjunction, DET=Determiner, INTERJ=Interjection, N=Noun, PFX=Prefix, PREP=Preposition, PRO=Pronoun, V=Verb. Ambiguity classes (cf. §1.1.) are designated by the sequence of grammatical categories, in alphabetic order, connected by the underscore (in this case, **A_N_V**), in bold, whereas ambiguity types (cf.§1.3) were represented in the same way, but using the slash (e.g. **A/N**).

It is possible to obtain the full list of different, inflected, ambiguous forms with the grammatical categories associated to them[4]. The three entries above are thus factorized as:

```
capital,A_N_N
```

About 10 % of Portuguese DELAF entries are ambiguous words, which correspond to 4.846 % of the DELAF different inflected forms.

1.1. Ambiguity class (AC)

One can consider **A_N_N** an ambiguity class (**AC**). There are 137 ambiguity classes in the DELAF. The number of ambiguous word forms per ambiguity class is very uneven (Table 1):

Table 1.
Distribution of ambiguity classes (AC) per number of ambiguous word forms (WF).

WF	AC	WF	AC	WF	AC	WF	AC
1	56	12	4	37	1	1,060	1
2	20	13	3	47	1	1,324	1
3	6	16	1	51	1	6,422	1
4	6	19	1	64	1	7,248	1
5	4	20	2	82	1	24,318	1
6	3	21	1	86	1		
7	3	25	1	97	1		
8	5	27	1	198	1		
10	2	29	1	684	1		
11	1	35	1	842	1		

There are 56 ambiguity classes (40.8%) with only one word form. The quantitatively more important ambiguity classes are (Table 2):

Table 2.
Quantitatively more important ambiguity classes (AC). WF = number of word forms per ambiguity class; %WF = percentage of total number of word forms (43,127).

AC	WF	%WF	AC	WF	%WF
A_V_V	21	0.049	DET_N	86	0.199
ADV_N	25	0.058	A_N_N_V	97	0.225
A_N_V_V	27	0.063	N_N_V	198	0.459
A_A_N_N	29	0.067	A_N_N	684	1.586
INTERJ_N	35	0.081	N_N	842	1.952
A_A_N	37	0.086	V_V	1,060	2.458
A_A	47	0.109	A_N_V	1,324	3.070
PREPXDET_PREPXPRO	51	0.118	N_V	6,422	14.891
DET_PRO	64	0.148	A_V	7,248	16.806

The 7 largest ambiguity classes represent 97.15 % of the ambiguous word forms of the DELAF; the **A_N** ambiguity class alone contains more than 56 % of the total of the ambiguous word forms.

[4] This was done using a simple PERL program built by M. Silberztein (2003); see Annex.

1.2. Ambiguity level (Amb)

Each ambiguity class corresponds to a certain ambiguity level (**Amb**). For instance, for the **A_N** class *Amb=2*, while for **A_N_V** class *Amb = 3*, and so on. Distribution of ambiguity classes by the ambiguity level of their word forms is shown below (Table 3):

Table 3.
Distribution of ambiguity classes (AC) by ambiguity level (Amb). WF = number of word forms per ambiguity class; %WF = percentage of total number of word forms (43,127).

Amb	AC	%AC	WF	%WF
2	40	29.197	40,331	93.517
3	50	36.496	2,532	5.871
4	29	21.168	224	0.519
5	16	11.679	38	0.088
6	2	1.460	2	0.005
Total	137	100.000	43,127	100.000

The majority of the ambiguous word forms show *Amb=2* (93.5 %), or *Amb=3* (5.87%), but the number of ambiguity classes with *Amb=2* is slightly less (40) than that with *Amb=3* (50). The most ambiguous word forms present *Amb=6*, but there are only two words of this kind, each one belonging to a different ambiguity class. The distribution of the most important ambiguity classes with *Amb=2* is shown below (Table 4):

Table 4.
Ambiguity classes (AC) with ambiguity level Amb=2. WF = number of word forms per ambiguity class; %WF = percentage of total number of word forms (43,127).

AC	WF	%WF
A_DET	10	0.023
N_PFX	11	0.026
ADV_V	13	0.030
NTERJ_V	19	0.044
A_ADV_	20	0.046
ADV_N	25	0.058
INTERJ_N	35	0.081
A_A	47	0.109
PREPXDET_PREPXPRO	51	0.118
DET_PRO	64	0.148
DET_N	86	0.199
N_N	842	1.952
V_V	1,060	2.458
N_V	6,422	14.891
A_V	7,248	16.806
A_N	24,318	56.387

Five ambiguity classes with *Amb=2* stand out as the most prominent: **N_N**, **V_V**, **N_V**, **A_V** and **A_N**. Together, these five classes constitute about 92.5 % of the ambiguous words of DELAF, with **A_N** ambiguity class alone representing more than half.

1.3. Ambiguity types (AT)

It is also possible to consider different ambiguity types, i.e., *n*-tuples of categories involved in different ambiguity classes. For example, in ambiguity class **A_N_V** there are three (binary) ambiguity types: **A/N**, **A/V** and **N/V**, besides the (ternary) **A/N/V** ambiguity type. An ambiguity type involves different ambiguity classes. For example, here are the 25 ambiguity classes where the ambiguity type **A/V** may be found (Table 5):

Table 5.
Ambiguity type **A/V**. AC = ambiguity class, **WF** = number of word forms per ambiguity class; **%WF** = percentage of total number of word forms (43,127); **%AT** = percentage of total number of word forms involving this ambiguity type (8,774); **Amb** = ambiguity level.

AC	WF	%WF	%AT	Amb
A_A_N_N_V	5	0.012	0.057	5
A_A_N_V	8	0.019	0.091	4
A_ADV_DET_N_PRO_V	1	0.002	0.011	6
A_ADV_DET_N_V	1	0.002	0.011	5
A_ADV_N_PRO_V	1	0.002	0.011	5
A_ADV_N_V	6	0.014	0.068	4
A_ADV_N_V_V	1	0.002	0.011	5
A_ADV_V	4	0.009	0.046	3
A_CONJ_N_V_V	1	0.002	0.011	5
A_CONJ_V	1	0.002	0.011	3
A_DET_N_PRO_V	2	0.005	0.023	5
A_INTERJ_N_V	5	0.012	0.057	4
A_INTERJ_N_V_V	2	0.005	0.023	5
A_INTERJ_V	1	0.002	0.011	3
A_N_N_N_V	12	0.028	0.137	5
A_N_N_V	97	0.225	1.106	4
A_N_N_V_V	2	0.005	0.023	5
A_N_PFX_V	1	0.002	0.011	4
A_N_PREP_V_V	1	0.002	0.011	5
A_N_V	1,324	3.070	15.090	3
A_N_V_V	27	0.063	0.308	4
A_PFX_V	1	0.002	0.011	3
A_V	7,248	16.806	82.608	2
A_V_V	21	0.049	0.239	3
A_V_V_V	1	0.002	0.011	4
Total	8,774	20.345	100.000	

As we can see, the ambiguity class **A_V** alone contains most of the word forms presenting this ambiguity type (82.6%), followed by class **A_N_V** (15%). The remaining classes are quantitatively less important. On the other hand, the **A/V** ambiguity type involves ambiguity classes were all other categories are represented.

1.4. Using ambiguity information with DELAF

The three types of information on ambiguity – **Amb** (ambiguity level), **AC** (ambiguity class) and **AT** (ambiguity type) – can be added to the DELAF: it is possible to build a priority DELAF of ambiguous words where this information is given, as in the following entries of *capital*:

```
capital,capital.A+Amb=3+AC=A_N_N+AT=A/N+AT=N/N:ms:fs
capital,capital.N+Amb=3+AC=A_N_N+AT=A/N+AT=N/N:fs
capital,capital.N+Amb=3+AC=A_N_N+AT=A/N+AT=N/N:ms
```

This information could then be used on disambiguation grammars, as we will see below.

2. Measuring the impact of ambiguity in a Portuguese corpus

2.1. Preliminaries.

In order to see how ambiguity affects real texts, we have applied the Portuguese DELAF to a large corpus of Portuguese journalistic text, the *CETEMPúblico* corpus[5], using INTEX (v. 4.31) [6]. Here are raw results from lexical analysis of this text (Table 6):

Table 6.
Lexical analysis of *CETEMPúblico* (Part 01) corpus using DELAF

Tokens	12,830,305	(179,248 different)
Simple words	9,705,387	(179,194 different)
Npr+.fst		57,275 (not considered in this paper)
ERR (unknown tokens)		15,175 (not considered in this paper)
DLF	8,960,421 tokens	83,690 different inflected forms 103,045 entries

The corpus contains about 12.8 million tokens (9.7 million simple words). About 8,96 million simple words were identified by the DELAF (92.324% of the text words). The simple word dictionary of the text (DLF) contains 103,045 entries and 83,690 different inflected forms (46.67% of the corpus different simple words).

A finite-state transducer (FST)[7], allows the (tentative) tagging of proper nouns (*N+Nprop*; Silberztein 2000:123), i.e. simple words beginning in capital letter, followed by another letter, and not previously identified by the DELAF (Fig.1):

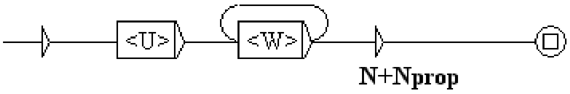

Fig. 1. ***Npr+.fst***

In this FST, <U> stands for any uppercase letter from the alphabet, while <W> stands for any letter. There are 57,275 different tokens that may be (tentatively) tagged in this way (31.963% of the corpus different simple words), otherwise, they would not be recognised and would then be considered unknown tokens (ERR). The remaining 15,175 different words (8.468% of the corpus different simple words) were left as unknown (ERR). For the purposes of this paper, the list of unknown tokens, ERR, and the list of tokens tagged by the *Npr+.fst* were not taken into consideration. We will

[5] http://www.linguateca.pt/cetempublico/. This corpus with over 200 million words, is divided in 20 fragments. For the purpose of this paper we used the first fragment, which is a text of 59,790 Kb.
[6] http://www.nyu.edu/pages/linguistics/intex. M. Silberztein (1993, 2000).
[7] In previous versions of INTEX the same result was obtained by other means.

thus focus on the remaining words that form the dictionary of simple words of the text (DLF).

2.2. The DLF of the *CETEMPúblico* (Part 01) corpus

The distribution of the DLF lexical entries by ambiguity level is shown below (Table 7):

Table 7.
Distribution of the simple word dictionary of the *CETEMPúblico* (Part 01) corpus by ambiguity level (Amb). DIF = number of different inflected forms of the DLF; %DIF = percentage of total number of forms of the DLF (83,690); LE = number of lexical entries of the DLF; %LE = percentage of total number of lexical entries of the DLF (103,046); FRQ = frequency, i.e., number of tokens; %FRQ = percentage of total number of tokens (8,960,421); PROD = Amb*FRQ; %PROD = percentage of total PROD (17,691,627).

Amb	DIF	%DIF	LE	%LE	FRQ	%FRQ	PROD	%PROD
1	66,539	79.507	66,539	64.572	3,989,179	44.520	3,989,179	22.548
2	15,195	18.156	30,390	29.492	2,782,676	31.055	5,565,352	31.458
3	1,746	2.086	5,238	5.083	1,058,253	11.810	3,174,759	17.945
4	173	0.207	692	0.672	690,790	7.709	2,763,160	15.618
5	35	0.042	175	0.170	437,961	4.888	2,189,805	12.378
6	2	0.002	12	0.012	1,562	0.017	9,372	0.053
Total	**83,690**	**100.000**	**103,046**	**100.000**	**8,960,421**	**100.000**	**17,691,627**	**100,000**

In the first place, it is interesting to verify that almost 80 % of the DLF different inflected forms (**DIF**) are non-ambiguous and represent 64,57 % of DLF entries[8].

As it is now possible to associate the **DIF** list to the frequency list of the corpus' tokens, we can see that non-ambiguous words represent about 44.52% of the tokens. This constitutes a first approximation to the degree of ambiguity of the corpus' words, for it means that more than a half (55.48%) of the corpus different tokens is ambiguous.

However, we proposed (Baptista e Faísca 2001) another measure for evaluating the **impact** of ambiguity in a text. The weight of an ambiguous word to the global ambiguity of a text may be viewed as the product of its ambiguity level (**Amb**) by its frequency (**FRQ**). In the absence of a better name, we plainly called it **product** (**PROD**). This measure seems more adequate than mere frequency, for it is related with the number of analysis that must removed from the text in the process of reducing its ambiguity. Any gain in ambiguity reduction can be calculated as the difference between the total PROD (17,691,627) and the remaining PROD after the disambiguation process has been carried out. The base-line would be a figure corresponding to the total disambiguation of the texts words, that is, the sum of their frequencies (8,960,421).

Considering this **PROD** measure, it is now clear that even if non-ambiguous words are a substantial part of a text's tokens, ambiguous words contribute significantly (77.452%) to the global ambiguity of a corpus such as this. Here we find something that can be viewed as a surprise: while frequency values associated to each ambiguity

[8] The average ambiguity of the DLF inflected forms is 1.231 (2.129 for the ambiguous words only; compare with values in note 2).

level are quite different, the impact of ambiguity classes with *Amb=3* to *5* is relatively similar, even if ambiguity classes with *Amb=2* reflect a greater impact.

2.3. The DLF of the tokens' frequency list of the CETEM*Público* corpus

We tried to verify if there were significant changes when, instead of just *Part 1* (with less than 10 million words), we would consider the ambiguity of the entire CETEM*Público* corpus (that contains about 200 million words). For this, we used the frequency list of the corpus[9]. The distribution of tokens by ambiguity level is shown in Table 8:

Table 8
Distribution of the simple word dictionary of the CETEM*Público* corpus by ambiguity level (Amb). DIF = number of different inflected forms in the DLF; %DIF = percentage of total number of forms in the DLF (157.989); LE = number of lexical entries; %LE = percentage of total number of lexical entries (186.239); FRQ = frequency, i.e., number of tokens in the corpus; %FRQ = percentage of total number of tokens (159.899.990); PROD = Amb*FRQ; %PROD = percentage of total PROD (17,691,627).

Amb	DIF	%DIF	LE	%LE	FRQ	%FRQ	PROD	%PROD
1	132,473	83.850	132,473	71.131	69,931,637	43.735	69,931,637	22.094
2	23,068	14.601	46,136	24.772	49,910,297	31.213	99,820,594	31.537
3	2,202	1.394	6,606	3.547	20,947,691	13.100	62,843,073	19.854
4	208	0.132	832	0.447	11,655,834	7.289	46,623,336	14.730
5	36	0.023	180	0.097	7,425,415	4.644	37,127,075	11.730
6	2	0.001	12	0.006	29,116	0.018	174,696	0.055
Total	**157,989**	**100.000**	**186,239**	**100.000**	**159,899,990**	**100.000**	**316,520,411**	**100.000**

It the first place, while the number of tokens of the entire corpus is 17.8 times larger than that of Part 1, the number of different word forms and the number of lexical entries is just about twice as large. Still, even with such a large corpus, DIF only constitute 17,752% of the entire DELAF different inflected forms, while DE represents 19,893% of DELAF lexical entries Secondly, when we compared the difference between the percentages for DIF, LE, FRQ and PROD by ambiguity level among the two corpora (Table 9), we find that they were, for the most part, minimal:

Table 9
Difference between the percentages of DIF, LE, FRQ and PROD by ambiguity level in the entire CETEM*Público* (CP) and only in Part 1 (P1).

Amb	%DIF (CP-P1)	%LE (CP-P1)	%FRQ (CP-P1)	%PROD (CP-P1)
1	+4.34251	+6.55864	-0.78539	-0.45412
2	-3.55498	-4.71953	+0.15845	+0.07886
3	-0.69223	-1.53594	+1.29050	+1.90935
4	-0.07535	-0.22526	-0.41955	-0.88804
5	-0.01921	-0.07335	-0.24421	-0.64824
6	-0.00073	-0.00556	+0.00121	+0.00219

[9] http://www.linguateca.pt/. Due to different atomization used by the authors of this list, we only used the frequency information associated to simple words and ignored all sequences of words connected by hyphen (e.g. verb-clitic pronoun combinations), that they considered as a single token. Furthermore, unknown tokens were not considered here, including *Nprop* (see §2.1). The resulting list contains 157.989 different simple words, which correspond to about 160 million tokens (that is, a corpus 17,845 times larger and a word list longer than t those of Part 1). Notice also that this list includes the word forms of Part 1.

The only significant differences were: the percentage of non-ambiguous different inflected forms (DIF) and of lexical entries (LE) in the DLF of the entire *CETEMPúblico* is slightly larger than that of *Part 1*. However, the percentage of FRQ and of PROD is smaller, therefore, the impact on the corpora global ambiguity of ambiguous words with *Amb=2* or *3* is slightly larger (2.8%), especially for words with *Amb=3* (1.91%).

It is thus reasonable to assume that most observations regarding ambiguity in Part 1 will also hold for the entire *CETEMPúblico* corpus.

3. Disambiguation grammar of past participle in compound tenses with *ter* (to have): a case study.

We begin by explaining that in Portuguese, the past participle, noted <V:K> in the DELAF and abbreviated as **K** in this paper, is an invariable form of the verb, only used in compound tenses with temporal auxiliary *ter* (to have) and (more rarely, and quickly becoming obsolete) *haver* (to there be)[10] as in:

O João cumprimentou a Maria
John greeted Mary
O João (*tinha +havia*) *cumprimentado a Maria*
John has greeted Mary

There are 11,374 K in the Portuguese DELAF. In many cases (5,301), K is ambiguous with words of other categories, mainly with adjectives. In fact, the ambiguity type A/V:k affects 5,182 word forms, which represents 97.755% of K ambiguous forms and constitutes 59,061% of all ambiguity classes with ambiguity type A/V.

In view of identifying the compound tense, it is necessary to disambiguate K. Usually, general-purpose disambiguation grammars, such as the one shown below (Fig.2) only use the grammatical categories available in the DELAF.

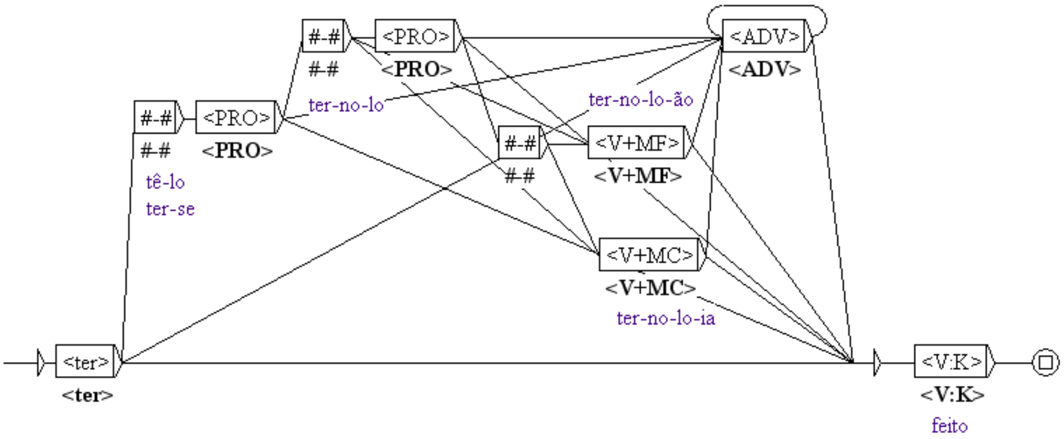

Fig. 2 Disambiguation grammar *Ter_Vk.fst*

[10] The compound tense with *haver* was not considered in this paper. The so called absolute participle construction is, in true, a reduced adjectival clause. A list of past participles that cannot be used as adjectives is given in Casteleiro 1981:150-154.

This grammar can be viewed as a first approach, obviously very incipient, to the compound past tense disambiguation problem.

When we apply this grammar to the CETEMPúblico (Part 1), we obtain 27,327 matches from the 145,998 occurrences of ambiguous K word forms. Naturally, the next step is to evaluate the results of this grammar. In this sense, we think that ambiguity information can be useful to identify the cases where the grammar could be improved or to detect problems in the dictionary. Results are shown in Table 10, sorted by ambiguity level:

Table 10
Ambiguous <V:K> in the DELAF. Amb=ambiguity level; AC = Ambiguity class; DIF=Different inflected forms (in the DELAF); WF-P1 = different word forms (in the corpus); FRQ-P1=frequency (in the corpus); WF-fst=different word forms retrieved by the FST; FRQ-fst=frequency. For each column, we indicated the percentages, based o the total of that column.

Amb	CA	DIF	%	WF-P1	%	FRQ-P1	%	WF-fst	%	FRQ-fst	%
	(A/V:K)	DELAF	DIF		WF-P1		FRQ-P1		WF-fst		FRQ-fst
1	[V]	6,073	53.40	313	10.06	8,796	6.02	170	11.24	7,765	28.42
2	A_V	4,647	40.86	2,322	74.61	68,605	46.99	1,125	74.36	13,937	51.00
	N_V	111	0.98	43	1.38	17,643	12.08	10	0.66	600	2.20
	V_V	3	0.03	0	0.00	0	0.00	0	0.00	0	0.00
3	A_ADV_V	2	0.02	2	0.06	798	0.55	1	0.07	5	0.02
	A_N_V	509	4.48	409	13.14	43,811	30.01	195	12.89	4,238	15.51
	A_V_V	10	0.09	6	0.19	81	0.06	2	0.13	2	0.01
	N_V_V	3	0.03	3	0.10	12	0.01	0	0.00	0	0.00
4	A_INTERJ_N_V	3	0.03	3	0.10	3,560	2.44	3	0.20	189	0.69
	A_N_V_V	8	0.07	8	0.26	888	0.61	4	0.26	309	1.13
	A_V_V_V	1	0.01	1	0.03	42	0.03	1	0.07	2	0.01
5	A_CONJ_N_V_V	1	0.01	1	0.03	902	0.62	1	0.07	66	0.24
	A_N_PREP_V_V	1	0.01	1	0.03	860	0.59	1	0.07	212	0.78
Total		11,372	100.00	3,112	100.00	145,998	100.00	1,513	100.00	27,325	100.00

In order to evaluate the performance of the grammar, we added to the table the non-ambiguous K entries of the DELAF. For each ambiguity level we indicate the ambiguity classes involved. The number of different inflected forms in the DELAF concerns only K forms. We calculated their total frequency in the corpus and the number of occurrences matched by the grammar.

The impact of K forms in the total PROD of *CETEMPúblico* (17,691,627; see §2.1) is only 1.934%, but as one can see, it involves 145,998 occurrences.

a) Amb=1 (non-ambiguous K forms)
Theoretically, the non ambiguous K forms (*Amb=1*) should had been all be matched by the grammar. In fact success rate is high (88.279%) but 1,031 occurrences of K forms were not matched by the grammar[11].

Looking through the concordances of non-ambiguous K not detected by the grammar, we found many forms that can also function as adjectives, in predicative construction

[11] There are 607 occurrences from the 170 word forms that were detected by the grammar that do not appear in the context described by the grammar. 143 word forms (occurring 240 times) were not detected by the grammar. Adding 607 to 424 referred to above, we obtain 1031, i.e., the difference between the K of Part 1 and the K detected by the grammar.

with copulatives *ser* and *estar* and even in attributive position next to a noun. These forms were not marked as adjectives in the DELAF:

```
Segundo a jovem, Emmet Roy foi acometido de dores
está fundeado na Baía de Cascais
```

In many cases, theses adjectives are attested in adnominal attribute position:

```
90 dias num barco fundeado a 500 metros de Monróvia
```

Some non-ambiguous K appear in the <*haver*> <V:K> compound tense, not described here :

```
muitas ambulâncias haviam afluído ao local
```

There are some K forms that are ambiguous with proper nouns. This is the case of *Chiado* (a commercial area in downtown Lisbon; 158 occurrences) and *Granado* (a proper (family) name; 28 occurrences):

```
queremos recuperar o prestígio do velho Chiado
a resposta (oral) que dei ao jornalista António Granado
```

Of course, ambiguity between K forms and proper names were not considered.

b) Amb > 1 (ambiguous K forms)

The number of matches of ambiguous K forms with Amb>1 is very low (an average of 14%). This comes naturally from the fact that the more a word has other categories, the less likely is it to appear as a K in a compound tense. Other disambiguation grammars should be built in order to solve some of the predicative and attributive uses of adjectives belonging to the **A/V:k** ambiguity type.

Even for the matched occurrences, success rate is not satisfactory, because the grammar eliminates correct analysis, i.e., it produces significant **silence**. For instance, there are some N of the **N/V:k** ambiguity type that can have *ter* as a support verb. Ambiguity class A_INTERJ_N_V is one of these cases. It contains three forms: *cuidado*, *obrigado*, *sentido*. Looking to the concordances where these forms appear, it was possible to determine the silence produced by the grammar:

Table 11
Ambiguous <V:K> in the DELAF: Ambiguity class **A_INTERJ_N_V**. Number of forms per category matched by the grammar. %=success rate, i.e. percentage of forms correctly tagged as K.

A_INTERJ_N_V	A	INTERJ	N	V:K	Total	%
cuidado	0	0	45	2	47	4,26
obrigado	0	0	0	30	30	100,00
sentido	0	0	55	57	112	50,89
Total	0	0	100	89	189	47,09

The success rate varies considerably: while *cuidado* was incorrectly tagged most of the times, with *obrigado* it was just the opposite; in the case of *sentido*, about half of the tags were correct.

However, this may not be always the case. With class **A_N_V_V**[12], in spite of the fact that there is potential N analysis, the grammar shows a 100% success rate.

Considering that any disambiguation grammar should avoid producing silence, it is possible to increase precision. This can be done by reducing the scope of the grammar using the information of ambiguity class. In most of the cases where correct analysis were eliminated, this was due to the fact the word was not a <V:K>, but a <N>. If, instead of just using the tag <V:K>, we use the same grammar but with tags stating the ambiguity class, leaving out only the classes were an analysis as <N> is potentially possible, it should be possible to obtain higher success rates. This was in fact the case: for the **A_V** class alone, success rate is very high (around 99 %).[13]

For the remaining ambiguity classes, the use of ambiguity information to identify problematic ambiguous word forms can help devising different strategies to constraint the general grammar shown above and to improve its performance.

4. Exotic words

4.1. A filter for ambiguous exotic words

There are many ambiguous words in the DELAF that present some analysis that are very unlikely to occur in texts. For example, the words *uma* and *umas*, among other analysis (as DET), can be tagged as inflected forms of an extremely rare verb *umar* (an intransitive verb said about wood that gets wet). The word *uma* belongs to class DET_DET_V (Amb=3), while *umas* is a DET_V (Amb=2). These words appear 82,302 and 664 times in the corpus, which results in a very high impact in text global ambiguity: PROD=248,234. If we remove the V analysis from the DELAF, using a filter, we obtain substantial reduction of PROD to 165,268.

The general purpose of building lexical filters is to reduce the noise produced during lexical analysis of texts, but maintaining a low and controlled level of silence, in order to improve the following steps in the text's processing. It is only a facultative tool, to be used with care, surely. But it is also a useful tool, especially when it is meant to be a part of applications were high efficiency is paramount to the objective of 100% correctness (at the current state of the art, this goal is still a mirage).

In a previous paper (Baptista & Faísca 2001), we have used frequency information associated to the ambiguous words of CetemPúblico (Part 1) to identify the 560 most frequent ambiguous exotic word forms[14] and to produce a filter that eliminates those

[12] A_N_V_V class: 8 DELAF entries, 8 different inflected forms in the text's DLF, corresponding to 888 occurrences (0.61%), 4 DIF matched by the grammar (*dito, lido, sobrado* e *valido*) corresponding to 309 occurrences (1,13%).

[13] The remaining classes (A_V_V and A_V_V_V) have only very few elements (10 and 1, respectively) and show a very low number of matches (4), therefore success rate (100%) is inexpressive. On the other hand, words of class A_ADV_V, with only two elements, and also with low number of matches (5, all of *demasiado*) were always incorrectly tagged as K, for these were all instances of the adjective (4 occurrences) or the adverb (1 occurrence).

[14] More precisely, we selected exotic words from ambiguous word forms with the larger impact in the corpus global ambiguity, with PROD≥200.

exotic analyses. The 17,047% ambiguity reduction thus obtained[15] may be viewed as an important contribution to the disambiguation task in the general processing of texts.

This approach can be seen as complementary with the approach that consist of structuring the lexicon by levels (Garrigues 1992). Furthermore, it presents the advantage that only some of the inflected forms of a lexical entry (the ambiguous forms) are affected.

It is possible to use ambiguity information to guide the discovery of exotic words and to help constructing different types of filters. Two examples are given below.

4.2. ***X-áveis***

Many second person-plural past-imperfective verb forms are ambiguous with *A-vel* plural forms:

`amáveis,amar.V:I2p`	('you-pl used to love')
`amáveis,amável.A:mp:fp`	('amiable, friendly')

This is another case of a A/V ambiguity type. The use of the second-plural person is becoming obsolete in European Portuguese. The general idea is that we would not produce any significant silence if we would remove these <V:I2p> analysis from the DELAF, especially in this kind of journalistic texts such as the *CETEMPúblico*.

There are 1,611 ambiguous *X-áveis* word forms. Only 4 of them do not involve A/V ambiguity type. The remaining 1,607 are distributed as follows (Table 12):

Table 11
X-áveis word forms with ambiguity type **A/V**. AC=Ambiguity class; DELAF=different inflected word forms in the DELAF; WF-P1= different inflected word forms in the corpus; FRQ=frequency; PROD= product (FRQ*Amb).

Amb	AC	DELAF	WF-P1	FRQ	PROD
2	A_V	1601	164	1113	2226
3	A_N_V	6	5	2304	6912
		1,607	169	3417	9138

In (almost[16]) 100% of the **A_V** cases, the <V:I2p> analysis is incorrect:

`a União Europeia já fez progressos` **`consideráveis`**	(considerable)
`as vossas condições de vida aqui serão` **`deploráveis`**	(deplorable)
`não só permite adiar todas as medidas` **`desagradáveis`**	(unpleasant)
`As sondagens são contraditórias e pouco` **`fiáveis`**	(reliable)
`Tratando-se de matérias muito` **`inflamáveis`**	(inflamable)
`aconteceram as cenas mais` **`lamentáveis`**	(regrettable)

[15] This was calculated as the difference between the corpus' global ambiguity before the application of the filter (PROD_0) and the residual ambiguity left by the filter (PROD_1), compared with the baseline corresponding to the flat text, where all ambiguity would have been removed (i.e. the sum of the frequency of all word forms of the corpus (FRQ): (PROD_0 – PROD_1)*100/(PROD_0 – FRQ).

[16] The only exception found was the word *veneráveis* (venerable, A:mp:fp), with 4 occurrences, two of them used as nouns while the other were adjectives. However, notice that this word was not considered N in the DELAF.

médio ou longo prazo haverá resultados **palpáveis**	(palpable)
com resultados considerados **razoáveis**	(reasonable)
hábitos alimentares naturais e **saudáveis**	(healthy)
estão ao serviço dos grupos mais **vulneráveis**	(vulnerable)

For the **A_N_V** words also, the <V:I2p> analysis is incorrect. However, is not possible to fully disambiguate between the N or the A analyses, except for the case of adjective *ultrapassáveis* ('possible to be by-passed'), for which we can not conceive a natural N construction.

The impact of ambiguity in the corpus resulting from the elimination of the verb analysis would represent a PROD of 5,721, which corresponds to a reduction of PROD of 3,417 (37.393% of total PROD before reduction).

4.2. ***X-ais***

There are 662 ambiguous *X-ais* word forms in the DELAF. Only 262 of these words appear in the corpus, totalling 75,388 occurrences. A single word (*mais*, **ADV_N**, 'plus/more') is responsible for 43,499 (57,7%) of these occurrences. The remaining words belong to 16 different ambiguity classes, but only 11 are in fact present in the corpus, representing 31,889 occurrences. From these, there are 299 word forms presenting the A/V ambiguity type, of which 137 appear in the corpus, totalling 14,120 occurrences (that is, 44.279% of all ambiguous *X-ais* word forms, excluding *mais*). All verb forms correspond to <V:P2p> for verbs ending in *–ar* or <V:S2p:Y2p> for verbs ending in *–ir*.

Table 12
X-ais word forms with ambiguity type **A/V**. AC=Ambiguity class; DELAF=different inflected word forms in the DELAF; WF-P1= different inflected word forms in the corpus; FRQ=frequency; PROD= product (FRQ*Amb).

Amb	AC	DELAF	WF-P1	FRQ	PROD
2	A_V	222	84	5,306	10,612
3	A_N_V	65	44	4,498	13,494
	A_V_V	2	0	0	0
4	A_N_N_V	8	8	3,797	15,188
	A_N_V_V	1	0	0	0
5	A_N_N_N_V	1	1	519	2,595
		299	137	14,120	41,889

Most of these V analyses correspond, in fact, to exotic words. They represent a PROD of 14,120 (33,708% of the initial PROD, before filtering).

The fact that we can pin-point ambiguous words via their ambiguity level (Amb), ambiguity class (AC) or ambiguity type (AT) makes it possible to ascertain which words are more or less likely to occur in texts. Such intuitions can be confirmed on large corpora in order to decide which would be more reasonable to include in filters. Ambiguity grammars may include this new type of tags.

5. Conclusion

In this paper, we mapped the ambiguity of words of the Portuguese DELAF and evaluated its impact on a large corpus. Linguistically oriented disambiguation techniques can profit from the ambiguity information retrieved in this manner. It is possible to use this kind of information to trace precisely the situations where a disambiguation grammar is most efficient or, otherwise, shows poor success rate. It can also help to maintain dictionaries, making easier to find lacunae or eventual mistakes. Finally, it may guide us to determine general rules to produce lexical filters.

References

Baptista, J; Faísca, L.; 2001. Um filtro para palavras ‘exóticas’ frequentes do Português. *Seminários de Linguística* 4: 65-86. Faro: U.Algarve.

Casteleiro, J. M.; 1981. *Sintaxe transformacional do adjectivo. Regime das construções completivas*. Lisboa: INIC.

Garrigues, M.; 1992. Dictionnaires hierarchiques du français. Principes et méthode d’extraction. Langue Française 96 :88-100. Paris: Larousse.

Silberztein, M.; 1993. *Dictionnaire électroniques et analyse automatique de textes. Le système* INTEX. Paris : Masson.

Silberztein, M.; 2000. *Intex Manual*. Paris : LADL/ASSTRIL.

ANNEXE

```
# !/usr/local/bin/perl
# Max Silberztein © 2003
# perl program that displays category ambiguities
# for each form of a DELAF/DELACF dictionary.
#
# C:> perl display_ambiguities.perl <mydelaf.dic >result.txt
#
# ATTENTION: this perl program uses \r\n as NEWLINE separators.
#

$lastform="";
$lastinfo="";

while (<>)
{
   # remove \n and \r's (rather than chomp)
   s/[\n\r]//g;

   # remove all syntactic and semantic codes (e.g. +zzz)
   s/\+[^+:]*//g;

   # computes the $form and the $category
   ($form,$info1)= split (/,/,$_);
   ($lemma,$info2) = split (/\./,$info1);
   ($category) = split (/:/,$info2);

   if ($form eq $lastform) # same form as previous line
   {
      # add the current category to the previous list of categories
      $lastinfo = $lastinfo . "/" . $category; # add the current category to the list
   }
   else # a new form
   {
      if ($lastform ne "") # the first $lastform is empty
      {
         # prints out the $lastform with the list of categories
         printf ("%s,%s\r\n",$lastform,$lastinfo);
      }
      $lastform = $form;
      $lastinfo = $category;
   }
}

printf ("%s,%s\r\n",$lastform,$lastinfo);
```