Evaluating CETEMPúblico, a free resource for Portuguese

Diana Santos
SINTEF Tele og Data
Postboks 124, Blindern
N-0314 Oslo, Norway
Diana.Santos@informatics.sintef.no

Paulo Rocha
Departamento de Informática
Universidade do Minho
PT-4710-057 Braga, Portugal
Paulo.Rocha@alfa.di.uminho.pt

Abstract

In this paper we present a thorough evaluation of a corpus resource for Portuguese, CETEMPúblico, a 180-million word newspaper corpus free for R&D in Portuguese processing. We provide information that should be useful to those using the resource, and to considerable improvement for later versions. In addition, we think that the procedures presented can be of interest for the larger NLP community, since corpus evaluation and description is unfortunately not a common exercise.

1 Introduction

CETEMPúblico is a large corpus of European Portuguese newspaper language, available at no cost to the community dealing with the processing of Portuguese.[1] It was created in the framework of the Computational Processing of Portuguese project, a government funded initiative to foster language engineering of the Portuguese language.[2]

Evaluating this resource, we have two main goals in mind: To contribute to improve its usefulness; and to suggest ways of going about as far as corpus evaluation is concerned in general (noting that most corpora projects are simply described and not evaluated).

In fact, and despite the amount of research devoted to corpus processing nowadays, there is not much information about the actual corpora being processed, which may lead naïve users and/or readers to conclude that this is not an interesting issue. In our opinion, that is the wrong conclusion.

There is, in fact, a lot to be said about any particular corpus. We believe, in addition, that such information should be available when one is buying, or even just browsing, a corpus, and it should be taken into consideration when, in turn, systems or hypotheses are evaluated with the help of that corpus.

In this paper, we will solely be concerned with CETEMPúblico, but it is our belief that similar kinds of information could be published about different corpora. Our intention is to give a positive contribution both to the whole community involved in the processing of Portuguese and to the particular users of this corpus. At the moment of writing, 160 people have ordered (and, we assume, consequently received) it[3]. There have also been more than four thousand queries via the Web site which gives access to the corpus.

We want to provide evaluation data and describe how one can improve the corpus. We are genuinely interested in increasing its value, and have, since corpus release,[4] made available four patches (e-mailing this information to all

[1] *CETEMPúblico* stands for "Corpus de Extractos de Textos Electrónicos MCT / Público", and its full reference is http://cgi.portugues.mct.pt/cetempublico/

[2] See http://www.portugues.mct.pt/

[3] Although we also made available a CQP (Christ et al., 1999) encoded version in March 2001, the vast majority of the users received the text-only version.

[4] The corpus was ready in July 2000; the first copies were sent out in October, with the information that version 1.0 creation date was 25 July 2000.

who ordered the corpus). We have also tried to considerably improve the Web page.

We decided to concentrate on the evaluation of version 1.0, given that massive distribution was done of that particular version[5]. Web access to the corpus (Santos and Bick, 2000) will not be dealt with here. Note that all trivial improvements described here have already been addressed in some patch.

2 Short technical description

As described in detail in Rocha and Santos (2000) and also in the FAQ at the corpus Web page, CETEMPúblico was built from the raw material provided by the Portuguese daily newspaper *Público*: text files in Macintosh format, covering approximately the years 1991 to 1998, and including both published news articles and those created but not necessarily brought to print. These files were automatically tagged with a classification based on, but not identical to, the one used by the newspaper to identify sections, and with the semester the article was associated to. In addition, sentence separation, and title and author identification were automatically created. The texts were then divided in extracts with an average length of two paragraphs. These extracts were randomly shuffled (for copyright reasons) and numbered, and the final corpus was the ordered sequence of the extract numbers.

To illustrate the corpus in text format, we present in Appendix A an extract that includes all possible tags with the exception of <marca>.

3 General evaluation

We start by commenting on the distribution process, and then go on to analyse the corpus contents and the specific options chosen in its creation.

Let us first comment on the **distribution options**. While this resource is entirely free (one has just to register in a Web page in order to receive the corpus at the address of one's choice), several critical remarks are not out of place:

First of all, when publicizing the resource, it was not clear for whom the CD distribution was actually meant: Later on, we discovered that many traditional linguists ordered it just to find out that they were much better off with the on-line version.

Second, more accompanying information in the CD would not hurt, instead of pointing to a Web page as the only source: In fact, the assumption that everyone has access to the Web while working with CETEMPúblico is not necessarily true in Portugal or Brazil.

Finally, we did not produce a medium-size technical description; in addition to the FAQ on the Web page, we provided only a full paper (Rocha and Santos, 2000) describing the whole project, arguably an overkill.

About the **corpus contents**, several fundamental decisions can – and actually have, in previous conferences or by e-mail – be criticized, in particular the use of a single text source and the inclusion of sentence tags (by criteria so far not yet documented). Still, we think that both are easy to defend, since 1) the time taken in copyright handling and contract writing with every copyright owner strongly suggests minimizing their number. And 2) although sentence separation is a controversial issue, it is straightforward to dispose of sentence separation tags. So, this option cannot really be considered an obstacle to users.[6]

We will concentrate instead on each annotation, after discussing the choice of texts and extracts.

3.1 Extract definition and choice

Looking at the final corpus, it is evident that many extracts should be discarded or, at least, rewritten. We tried to remove specific kinds of "text", namely soccer classifications, citations from other newspapers, etc., but it is still possible to detect several other objects of dubious interest in the resulting corpus.

In fact, using regular expression patterns of the kind "existence of multiple tabs in a line ending in numbers", we identified 5270 extracts having some form of classification, as well as 662 extracts with no valid content.

[5] We have no estimate of how many users have actually succeeded, or even tried, to apply the patches made available later on. We have just launched a Web questionnaire in order to have a better idea of our user community.

[6] Since extract definition is based on paragraph and not sentence boundary, the option of marking <s> boundaries has no other consequences.

Now, it is arguable that classifications of other sports (e.g., athletics and motor races), solutions to crossword puzzles, film and book reviews, and TV programming tables, just to name a few, should have been extracted on the same grounds presented for removing soccer. Our decision was obviously based on a question of extent. (Soccer results are much more frequent.) However, we now regret this methodological flaw and would like to clean up a little more (as done in the patches), or add back soccer results.

Another problem detected, concerning the extract structure, was our unfortunate algorithm of appending titles to the previous extract, just like authors, instead of joining them to the next extract. This means that 4.8% of the extracts end with a title in CETEMPúblico. (9.6% end with an author.)

3.2 Spurious repetitions

The worst problem presented by the CETEMPúblico corpus is the question of repeated material. (Incidentally, it is interesting to note that this is also a serious problem in searching the Web, as mentioned by Kobayashi and Takeda (1999).) Repeated articles[7] can be due to two independent factors:

- parallel editions of the local section of the newspaper in the two main cities of Portugal (Lisboa and Porto)
- later publication of previously "rejected" articles

In addition to manually inspecting rare items that one would not expect to appear more than a few times in the corpus (but which had higher frequency than expected), we used the following strategies to detect repeated extracts:

1. Record the first and last 40 characters of each extract, in a hash table, as well as their size in characters. Then fully compare only the repeated extracts under this criterion.
2. Using the Perl module MD5 (useful for cryptographical purposes), we attributed to each extract a checksum of 32 bytes, and recorded it in a hash table. Repeated extracts have the same checksum, but it is extremely unlikely that two different ones will.

The results obtained for exactly equal extracts are displayed in Table 1 for both methods.

Another related (and obviously more complicated) problem is what to do with quasi-duplicates, i.e. sentences or texts that are almost, but not, identical. An estimate of the number of approximately equal extracts, obtained with the 40 character-method but with relaxed size constraints (10%) yields some further 15,665 possibly repeated extracts. It is not obvious whether one can automatically identify which one is the revised version, or even whether it is desirable to choose that one. We have, anyway, compiled a list of these cases, thinking that they might serve as raw material for studying the revision process (and to obtain a list of errors and their correction).

Kind	Different extracts		Extracts to remove	
	40chr	MD5	40chr	MD5
twice	45,046	44,188	45,046	44,188
3 times	1,493	1,401	2,986	2,802
4 times	301	271	903	813
5 times	68	63	272	252
6-10	83	81	552	548
> 11	31	31	643	880
Total	47,022	46,035	50,402	49,483

Table 1. Overview of exact duplication

3.3 Title and author identification

In the CETEMPúblico corpus, newspaper titles and subtitles, as well as author identifications, have been marked up as result of heuristic processing. In Rocha and Santos (2000), a preliminary evaluation of precision and recall for these tasks was published, but here we want to evaluate this in a different way, without making reference to the original text files.

Given the corpus, we want to address precision and error rate (i.e., of all chunks tagged as titles, how many have been rightly tagged?, and how many are wrong?). We reviewed manually the first 500 instances of <t>[8], of which 427 were undoubtedly titles, a further 4 wrongly tagged authors, and at least 15 belonged to book or film reviews, indicating

[7] Repeated **sentences** can also occur in the lead and in the body of an article, and (in the opinion section) to highlight parts of an article.

[8] In the 15th chunk of the corpus. This aparently naïve choice of test data does not bias evaluation, since the extracts are randomly placed in the corpus and do not reflect any order of time period or kind of text.

title, author and publisher, or director and broadcasting date, etc.

We then looked into the following error-prone situation: After having noted that several paragraphs in a row including title and author tags were usually wrong (and should have been marked as list items instead), we looked for extracts containing sequences of four titles / authors and manually checked 200. The precision in this case was very low: Only 38% were correctly tagged. Of the incorrect ones, as much as 34% were part of book reviews as described above. This indicates clearly that we should have processed special text formats prior to applying our general heuristic rules.

Regarding recall, we did the following partial inspection: We noted several short sentences ending in ? or ! (a criterion to parse a text chunk as a full sentence) that should actually be tagged as titles. We therefore looked at 200 paragraphs with one single sentence ending in question or exclamation mark containing less than 8 words, and concluded that 41 cases (20%) could definitively be marked as titles, while no less than 85 of these cases where questions taken from interviews. Most other cases were questions inside ordinary articles.

As far as authors are concerned, the phrase *Leitor devidamente identificado* ("duly identified reader", used to sign reader's letters where the writer does not wish to disclose his or her identity) was correctly identified only in 78% of the cases (135 in 172). In 17% of the occurrences, it was wrongly tagged as title.

From a list of 500 authors randomly extracted for evaluation purposes, only 395 (79%) were unambiguously so, while 8 (1.5%) could still be considered correct by somehow more relaxed criteria. We thus conclude that up to 21% of the author tags in the corpus may be wrongly attributed, a figure much higher than the originally estimated 4%.

Among those cases, foreign names (generally in the context of film or music reviews, or book presentations) were frequently mistagged as authors of articles in Público, a situation highly unlikely and amenable to automatic correction. Figure 1 is an example.

```
<a> Contos Assombrosos </a>
<a> Amazing Stories </a>
<a> De Steven Spielberg </a>
<t> Com Kevin Costner, Patrick Swayze e Sid Caesar  </t>
```

Figure 1. Wrong attribution of <a> and <t>

3.4 Sentence separation

In addition to paragraph separation coming from the original newspaper files, CETEMPúblico comes with sentence separation as an added-value feature.

Now, sentence separation is obviously not a trivial question, and there are no foolproof rules for complicated cases (Nunberg, 1990; Grefenstette and Tapainanen, 1994; Santos, 1998). So, instead of trying to produce other subjective criteria for evaluating a particularly delicate area, we decided to look at the amount of work needed to revise the sentence separation for a given purpose, as reported in section 4.2.

But we did some complementary searches for cases we would expect to be wrong whatever the sentence separation philosophy. We thus found 6,358 sentences initiated by a punctuation mark (comma, closing quotes, period, question mark and exclamation mark, respectively amounting to 4053, 410, 1607, 227 and 61 occurrences), as well as a plethora of suspiciously small sentences, cf. Table 2.

Sentence size	Number of sentences	Error estimation
one	14,783	100%
two	55,121	53%
three	70,909	20%

Table 2. Too small sentences

Sentence separation marks some sentences as fragments (<s frag>); in addition, the <li> attribute was used to render list elements. We are not sure now whether it was worthwhile to have two different markup elements.

<s frag>	63,122
	113,540
<t>	687,720
<a>	263,269

Table 3. Number of cases of non-standard <s>

Finally, the sentence separation module also introduces the <marca> tag to identify meta-characters that are used for later coreference (eg. in footnotes). The asterisk "*" was marked as such in CETEMPúblico, but not inside author or title descriptions, an undesirable inconsistency.

3.5 Extraneous characters

An annoying detail is the amount of strange characters that have remained in the corpus after font conversion, such as non-Portuguese characters, hyphens, bullet list marking, and the characters < > instead of quotes.

It is straightforward to replace these with other ISO-8859-1 characters or combinations of characters, as was done with dashes and quotes.[9] Only the last line of Table 4 requires some care, since *É* is a otherwise valid Portuguese character that should only be replaced a few times.

Character	Action	Number
Ð	non-breaking hyphen	856
Ï	use oe	246
tab stop	remove/replace by " "	50,312
control character	eliminate extract	53,631
character 0x95	(?)	40,665
<	use <	1,283
>	use >	1,232
É	replace by ...	3,167

Table 4. Occurrence of extraneous chars

3.6 Text classification

CETEMPúblico extracts come with a subject classification derived from (but not equal to) the original newspaper section. Due to format differences of the original files, only 86% of the extracts have some classification associated. The others carry the label ND (not determined).

We evaluate here this classification, since for half of the corpus article separation had to be carried out automatically and thus chances exist that errors may have crept in.

The first thing we did was to check whether repeated extracts had been attributed the same classification. Astonishingly, there were many differences: of the 47,002 cases of multiple extracts, 10,872 (23%) had different categories, even though only in 2% of the cases none of the conflicting categories was ND.

Another experiment was to look at well-known polysemic or ambiguous items and see whether their meaning correlated with the kind of text it was purported to be in. We thus inspected manually several thousand concordances dealing with the following middle frequency words[10]: 201 occurrences of *vassoura* (broom; last vehicle in a bicycle race); 124 of *passador* (sieve; drug seller; emigrant dealer); 314 of *cunha* (wooden object; corruption device); 599 of *coxa* (noun thigh; adjective lame); 205 of *prego* (nail; meat sandwich; pawnshop); 145 of *garfo* (fork; biking); 5505 of *estrela* (star; filmstar; success); 375 of *dobragem* (folding; dubbing; parachuting and F1 term); 573 of *escravatura* (slavery).

We could only find two cases of firm disagreement with source classification (in the two last mentioned queries). This is not such a good result as it seems, though, since it can be argued that subject classification is too high level (society, politics, culture) to allow for definite results.

4 Corpus in use

The best way to evaluate a corpus resource is to see how well it fares regarding the tasks it is put to. We will not evaluate concordancing for human inspection, because we assume that this is a rather straightforward task for which CETEMPúblico is useful, especially because it requires direct scrutiny. Obviously, human inspection and judgement make the results more robust.

4.1 Proper name identification

One of the authors developed proper name identification tools (Santos, 1999) prior to the existence of CETEMPúblico. We ran them on this corpus to see how they worked.

We proceeded in the following way: We inspected manually the first 1,000 proper names obtained from CETEMPúblico and got less then 4% wrong, i.e., over 96% precision.

Size	Number
One word	26,518
Two words	15,512
Two words and *de*	4,623
Three words	2,132
Three words and *de*	2,354
Four words	201
Four words and *de*	583
>= five words	359
problems [11]	383

Table 5. Size distribution of proper nouns

[9] Note that it is not always possible to have a one-to-one mapping from MacRoman into ISO-8859-1.

[10] Glosses provided are not exhaustive.

[11] This category encompasses “deviant” proper names, mainly including foreign accents and numbers, irrespective of proper name length.

Then, we computed the distribution of the 52,665 proper nouns identified by the program (23,401 types) on the first million words of the corpus as shown in Table 5, and inspected manually those 1,017 having a length larger or equal than four words. Of these 88% were correct and 6.5% were plainly wrong. Cases of merging two proper names and cases where it was easy to guess one missing (preceding or following) word accounted each for approximately 5% of the remaining instances.

While use of CETEMPúblico allowed us to uncover cases not catered for by the program, it also illuminated some potential[12] tokenization problems in the corpus, namely a large quantity of tokens ending in a dash (21,455 tokens, 6,458 types) or in a slash (7313 tokens, 4530 types), as well as up to 132,455 tokens including one single parenthesis (28,466 types).

4.2 Treebank building

The first million words of CETEMPúblico was selected for the creation of a treebank for Portuguese (Floresta Sintá(c)tica[13]), given that its use is copyright cleared and the corpus is free.

The treebank team engaged in a manual revision of the text prior to treebank coding, refining sentence separation with the help of syntactically-based criteria (Afonso and Marchi, 2001). We have tried to compute the amount of change produced by human intervention, which turned out to be a surprisingly complex task (Santos, 2001).

This one million words subcorpus contained 8,043 extracts.[14] Assuming that the first million is not different from the rest of the corpus, the results indicate an estimate of 17% of the corpus extracts in need of improvement.

Looking at sentences, 2,977 sentences of the 42,026 original ones had to be re-separated into 4,304 of the resulting 43,271. Table 6 displays an estimate of what was actually involved in the revision of sentence tags (percentages are relative to the original number of sentences). The "Other" category includes changes among the tags <t>, <a>, <li> and <s>.

<s>-addition	1,481-1,872	3.52-4.24%
<s>-removal	612-115	1.46-2.65%
Other	550	1.3%

Table 6. Revision of <s> tags

4.3 Spelling checker evaluation

One of the first and most direct uses of a large corpus is to study the coverage, evaluate, and especially improve a spelling checker and morphological analyser.

Our preliminary results of evaluating Jspell (Almeida and Pinto, 1994) as far as type and token spelling is concerned are as follows: Among the 942,980 types of CETEMPúblico, 574,199 were not recognized by the current version of Jspell (60.4%), amounting to 3.07% of the size of the corpus. A superficial comparison showed that CETEMPúblico contains a higher percentage of unrecognized words, both types and tokens, than other Portuguese newspaper corpora. Numbers for a 1.5-million word corpus of *Diário do Minho* (a regional newspaper) and for a 4-million word corpus of a political party newspaper are respectively 26.5% and 25.41% unrecognized types and 2.26% and 1.67% unrecognized tokens. These numbers may be partially explained by *Público*'s higher coverage of international affairs, together with its cinema and music sections, both bringing an increase in foreign proper names[15].

Description	Tokens	Types
Foreign first names	130	125
Portuguese first names	19	16
Foreign surnames	216	208
Portuguese surnames	35	34
Foreign organizations	50	45
Portuguese organizations	26	23
Foreign geographical[16]	48	48
Portuguese geographic	28	28
acronyms	81	77
foreign words	171	161
Portuguese foreign words[17]	26	25

[12] Different tokenizers may have different strategies, but we assume that these will be hard cases for most.

[13] See http://cgi.portugues.mct.pt/treebank/.

[14] Numbered from 1 to 8067, since version 1.2 was used, and therefore 24 invalid extracts had been already removed. In addition, the treebank reviewers considered that further 129 should be taken out.

[15] The percentage of unrecognized tokens varies from 4.8% for culture to 2.0% for society extracts.

[16] We classify as Portuguese or foreign the word, not the location: thus, *Tanzânia* is a Portuguese word.

[17] That is, words routinely used in Portuguese but which up to now have kept a distinctly foreign spelling, such as *pullover*.

words missing in dict.	101	98
incorrectly spelled[18]	36	36
others	33	32
total	1,000	956

Table 7. Distribution of "errors"

We investigated the "errors" found by the system, to see how many were real and how many were due to a defficient lexical (or rule) coverage. Table 7 shows the distribution of 1,000 "errors" randomly obtained from the 12th corpus chunk.

The absolute frequencies of the most common spelling errors in CETEMPúblico is another interesting evaluation parameter. Applying Jspell to types with frequency > 100 (excluding capitalized and hyphenated words), we identified manually the "real" errors. Strikingly, all involved lack or excess of accents. The most frequent appeared 840 times (*juíz*), the second one (*saíu*) 659, and the third (*impôr*) had 637 occurrences. Their correctly spelled variants (*juiz*, *saiu*, *impor*) appeared respectively 11896, 9892 and 5125 times.

5 Comparison with other corpora

One can find excellent reports on the difficulties encountered in creating corpora (see e.g. Armstrong et al. (1998) and references therein), but it is significantly rarer to get an evaluation of the resulting objects. It is thus not easy to compare CETEMPúblico with other corpora on the issues discussed here.

For example, it was not easy to find a thorough documentation of BNC[19] problems (although there is a mailing list and a specific e-mail address to report bugs), nor is similar information to be found in distribution agencies' (such as LDC or ELRA) Web sites.

It is obviously outside the scope of the present paper to do a thorough analysis of other corpora as well, but our previous experience shows that it is not at all uncommon to experience problems with characters and fonts, repeated texts or sentences, rubbish-like sections, wrong markup and/or lack of it. All this independently of corpora being paid and/or distributed by agencies supposed to have performed validation checks. The same happens for corpora that have been manually revised.

As regards sentence separation, Johansson et al. (1996) mention that proofreading of the automatic insertion of <s>-units was necessary for the ENPC corpus, but they do not report problems of human editors in deciding what an <s> should be. Let us, however, note that ENPC compilers were free to use an <omit> tag for complicated cases and, last but not least, were not dealing with newspaper text.

6 Concluding remarks

This paper can be read from a user's angle as a complement to the documentation of the CETEMPúblico corpus. In addition, by showing several simple forms of evaluating a corpus resource, we hope to have inspired others to do the same for other corpora.

While the work described in this paper already allowed us to publish several patches, improve our corpus processing library and contribute to new versions of other people's programs, namely Jspell, our future plans are to do more extensive testing using more powerful techniques (e.g. statistical) to investigate other problems or features of the corpus. In any case, we believe that the work reported in this paper comes logically first.

Acknowledgements

We are first of all grateful to the *Público* newspaper (especially José Vítor Malheiros, the responsible for the online edition) for making this resource possible. We thank José João Dias de Almeida for several suggestions, the team of Floresta Sintá(c)tica for their thorough revision of the first million words, Stefan Evert for invaluable CQP support, and Jan Engh for helpful comments.

References

Susana Cavadas Afonso and Ana Raquel Marchi. 2001. Critérios de separação de sentenças/frases, cgi.portugues.mct.pt/treebank/CriteriosSeparacao.html

J.J. Almeida and Ulisses Pinto. 1994. Jspell – um módulo para análise léxica genérica de linguagem natural. *Actas do Congresso da Associação Portuguesa de Linguística* (Évora, 1994), www.di.uminho.pt/~jj/pln/jspell1.ps.gz.

[18] Including one case of lack of space between two words, *suacontribuição.*

[19] British National Corpus. http://info.ox.ac.uk/bnc/

Susan Armstrong, Masja Kempen, David McKelvie, Dominique Petitpierre, Reinhard Rapp, and Henry S. Thompson. 1998. Multilingual Corpora for Cooperation. In Antonio Rubio et al. (eds.), *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998), Vol. 2, pp.975-80.

Oliver Christ, Bruno M. Schulze, Anja Hofmann and Esther Koenig. 1999. The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual, Institute for Natural Language Processing, University of Stuttgart http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual

Gregory Grefenstette and Pasi Tapanainen. 1994. What is a word, What is a sentence? Problems of Tokenization. *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX'94)*, pp. 79-87

Stig Johansson, Jarle Ebeling and Knut Hofland. 1996. Coding and aligning the English-Norwegian Parallel Corpus. In Karin Aijmer, Bengt Altenberg & Mats Johansson (eds.), *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies (Lund, 4-5 March 1994)*, Lund University Press, pp.87-112.

Mei Kobayashi and Koichi Takeda. 1999. Information retrieval on the web: Selected topics. IBM Research, Tokyo Research Laboratory, IBM Japan, Dec. 16, 1999.

Geoffrey Nunberg. 1990. *The linguistics of punctuation*. CSLI Lecture Notes, Number 18.

Paulo Alexandre Rocha and Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In Graça Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, (São Paulo, 19-22 November 2000), pp.131-140.

Diana Santos. 1998. Punctuation and multilinguality: Reflections from a language engineering perspective. In Jo Terje Ydstie and Anne C. Wollebæk (eds.), *Working Papers in Applied Linguistics* 4/98. Oslo: Department of Linguistics, Faculty of Arts, University of Oslo, pp.138-60.

Diana Santos. 1999. Comparação de corpora em português: algumas experiências. www.portugues.mct.pt/Diana/download/CCP.ps

Diana Santos. 2001. Resultado da revisão do primeiro milhão de palavras do CETEMPúblico c gi.portugues.mct.pt/treebank/RevisaoMilhao.html

Diana Santos and Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. In Maria Gavriladou et al. (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), pp.205-210.

Appendix A. Example of an extract

```
<ext n=1914 sec=nd sem=93b>
<p> <s>Produção da Hammer.</s>
<s>Um episódio da II Guerra Mundial, um caso de heroísmo, quando toda uma companhia é destruída no Norte de África.</s> </p>
<li>THE STEEL BAYONET de Michael Carreras com Leo Glenn e Kieron Moore</li>
<li>Grã-Bretanha, 1957, 82 min</li>
<li>Canal 1, às 15h15</li>
<p><s>Um ex-presidiário esforçadamente em busca de regeneração (Nicolas Cage) e a mulher, uma honesta e voluntariosa polícia (Holly Hunter), querem formar família mas descobrem que não podem ter filhos e decidem raptar um bebé.</s>
<s>O cinema dos irmãos Coen sempre atraiu críticas de «exibicionismo» e «fogo-de-artifício».</s>
<s>Esta comédia desbragada, que de uma só vez faz um curto-circuito com as referências à banda desenhada, ao burlesco ou à série «Mad Max», é o tipo de objecto que mais evidencia o que os detractores dos Coen considerarão um «exercício de estilo».</s>
<s>«Arizona Junior», concorde-se, é uma obra que exibe um gozo evidente pelas proezas do trabalho de câmara e Nicolas Cage, Holly Hunter ou John Goodman têm a consistência de figuras de cartão.</s>
<s>Mas nem por isso se deve ignorar estarmos perante um dos universos mais paranóicos do cinema actual.</s> </p>
<t>RAISING ARIZONA de Joel Coen com Nicolas Cage, Holly Hunter e John Goodman</t>
<t>EUA, 1987, 97 min</t>
<a>Quatro, às 21h35</a> </ext>
```