Providing Internet Access to Portuguese Corpora: the AC/DC Project

Diana Santos, Eckhard Bick

SINTEF Telecom and Informatics
Postboks 1024 Blindern, N-0314 Oslo, Norway
Diana.Santos@informatics.sintef.no, lineb@hum.au.dk

Abstract

In this paper we report on the activity of the project Computational Processing of Portuguese (*Processamento computacional do português*) in what concerns providing access to Portuguese corpora through the Internet. One of its activities, the AC/DC project (*Acesso a corpora/Disponibilização de Corpora*, roughly "Access and Availability of Corpora") allows a user to query around 40 million words of Portuguese text. After describing the aims of the service, which is still being subject to regular improvements, we focus on the process of tagging and parsing the underlying corpora, using a Constraint Grammar parser for Portuguese.

General Background

The project Computational Processing of Portuguese (CPP) is financed by the Portuguese Ministry of Science and Technology in order to foster R&D on the computational processing of Portuguese and eventually lead to the availability of state-of-the-art products and services in Portuguese in the new information age. The AC/DC project, a subactivity of CPP, fits in under the more general goal of tackling the problem of lack of available and usable resources for research and evaluation.

The main lines of activity are:

1. Creation of publicly available resources
2. Redistribution of already available resources
3. Cataloguing the area
4. Evaluation of particular fields

The AC/DC project falls mainly under the second class of activities, but insofar that it adds value to existing corpora, it can be viewed as contributing equally to the first goal.

The AC/DC project

The AC/DC project stemmed from the wish to have in a comparable form for all so far publicly available corpora of the Portuguese language. After identification and clearance of the relevant copyright issues, the corpora were encoded in the IMS Corpus Workbench (Christ et al., 1999), to which a Web interface was added. Technical and scientific reasons for the choice of the underlying corpus system have already been described in (Santos, 1998); an additional political advantage was the fact that this system runs under Linux, a non-proprietary system.

The AC/DC project has so far consisted of two phases. The first phase identified and brought to a common format the textual resources already available, providing a Web-based service of Portuguese corpora. The second phase improved the information associated with these corpora by parsing them with a broad-coverage parser for Portuguese, encoding the result in the same workbench, and serving them in the same overall service.

Overview of the first phase

The first phase of the project resulted in five different corpora available for search on the Web since September 1999, on http://cgi.portugues.mct.pt/acesso/. This service has been regularly updated and improved since, and the corpora featured by the service – reflecting its status on 23 March 2000 – are summarized in Table 1.

Paragraph and sentence separation were computed automatically for every corpus; for some corpora, other parts were also explicitly encoded by means of structural attributes, such as titles, footnotes and/or author identification.

Corpus Identification	Size in words (k)	Size in sentences
Natura/Público	6,242	225,088
ENPCpub	72	4,371
Natura/Diário do Minho	2,110	91,203
ECI-EBR	718	43,866
ECI-EE	26	776
NILC/São Carlos	33,618	2,195,056
FRASESPP	16	594
FRASESPB	19	652
Total	**42,821**	**2,561,606**

Table 1: Portuguese corpora

With the exception of the last two, which are sentence corpora instead of text corpora and whose original goal was parser performance testing, none of these corpora have been compiled by the authors nor under the framework of the AC/DC project. Rather, our project started by amassing all resources already publicly available – in quite distinct formats and revision status, incidentally. (We hope to report on the creation of a larger resource soon, but this lies outside the scope of the present paper.)

Let us outline here the process followed in the first phase of the AC/DC project (a more detailed description of the whole process can be found in (Santos, to appear)):

After getting the corpora as text files in ISO-8859-1 format, Perl programs were written that

- Cleaned the input
- Added structural tags (such as titles, parts, captions, signatures and the like)
- (Assembled the whole corpus in case it happened to be distributed among several files)
- Separated sentences and paragraphs
- Tokenized properly the result (a quite complex process indeed)
- Counted (and documented) the resulting object

Then, the corpus objects were encoded in the IMS workbench, and installed in the Web server together with

the corresponding HTML documentation, semi-automatically generated. For each corpus, several counts were done, and a quantitative overview of all corpora produced in table format. It would be extremely tedious, and error-prone, to change the values by hand every time a change was introduced in the programs.

Comparing this service with the one reported (Santos, 1998) for the Oslo Corpus of Bosnian Texts (OCBT), the main differences are:

- no user identification is required, i.e., there's no additional layer of bureaucracy imposed on those who want to query the corpus;
- no restrictions are made in terms of size of results or of query (since the corpora are freely available);[1]
- there is no parsing of the user queries in addition to the one done by CQP (no attempt to correct the user, or to help him/her apart from very obvious cases).

Even though the last feature may seem to be a step backwards, it is our belief that corpus users who require sophisticated queries have to be able to pose them, so that, in the long run, they have to consult the CQP user's manual (Christ et al., 1999) and work their way through the possibilities offered. Adding a layer of "simplification" is just replacing one query language with another, which is not our goal. Although a new query language may be considered more user-friendly by some users, following such method may lead to compromising significantly the system's expressive power, as is the case of the Norwegian tagged corpus, a project (see http://www.tekstlab.uio.no/norsk/bokmaal) which used OCBT's underlying Web interface but provided a menu-based query language on top of it, which restricted, to a large extent, the system's original capabilities.

Introduction to the second phase

The second phase of the project aims at enriching the aforementioned corpora with morphological, PoS and syntactic annotation. To do this, automatically, we use a robust Constraint Grammar based tagger-parser which - to our knowledge - is the most developed system for Portuguese to date.

The first two corpora were annotated and made available on the Web on February 2000. In Table 2, we provide a rough quantitative overview of their constitution.

Parsed corpus	NATPANOT	EBRANOT
Sentences	224,500	43,500
Words	6,250,000	709,000
Nouns	1,311,000	141,500
Verbs	770,000	112,500
Adjectives	353,000	40,000
Adverbs	319,700	48,000
Proper nouns	541,000	31,600
Contractions	495,300	43,100

Table 2: Annotated corpora

A CG parser for Portuguese

The present section introduces the CG multi-level parsing system used in the annotation project, and presents some statistics of its performance applied to the Portuguese corpora in question.

The parser uses a lexicon base of about 50.000 lexemes for its morphological analysis, and performs context governed rule based disambiguation at successive levels of analysis, including word class, inflexion, dependency syntax, valency instantiation and some experimental polysemy resolution. Processing speed is ca. 500 words/sec on a Pentium II based Linux system, when all annotation levels are included.

Rules are expressed in the Constraint Grammar formalism (Karlsson, 1995), using the CG2 variant (Tapanainen, 1997). Following CG tradition, modular word based tags are used on all levels, and the system's grammar is implemented by adding or removing individual tags or sets of tags in a context dependent way. Usually, the whole sentence is used as a rule scope window, providing for a much richer context than is used in most probabilistic or automated learning systems. All in all, the grammar comprises of about 8000 rules, of which 6000 are used in the present annotation task. The fact that the final parse is created in a reductionist way, and the *last surviving* reading is regarded as correct, guarantees a high degree of robustness, especially when comparing to PSG type systems based on rewriting rules.

In previous evaluations (Bick, 1996 and 2000), at near 100% disambiguation, the system achieved correctness rates of over 99% for PoS and 96-97% for syntax, when analysing free running text. So far, tests suggest that performance and robustness are fairly stable across a variety of written text types, for both Brazilian and modern European Portuguese. Pilot evaluations for the analysis of transscribed speech and historical texts indicate that the same system can handle even non-standard text types when allowing for a drop in syntactic performance of a few percentage points. Provided a fairly standard (or filtered) orthography, PoS tagging suffered no substantial decrease in performance. Not least in the present corpus annotation task, robustness has been a key factor, due to considerable text type differences between individual sub-corpora, and the incorporation of some speech and dialectal data (ECI-EBR). At the time of writing, only some pilot evaluation of parser performance variation had been done on the first couple of the AC/DC corpora. Correctness percentages relate to parser word token numbers (excluding punctuation).

Parsed corpus	NATPANOT	EBRANOT
sample size	4127 words	2592 words
PoS correct	99%	99.3%
Syntax correct	96.3%	96.8%

Table 3: Parser performance

The syntactic annotation paradigm

Though it can be used to generate syntactic tree structures (as in the grammar teaching system at http://visl.hum.sdu.dk), the parser internally handles syntax as (flat) dependency grammar, using syntactic function tags with directional dependency markers. On the clause level, @SUBJ> and @<SUBJ, for instance, mark subjects (or – in the case of groups – *heads* of subjects), the arrows indicating the position of the governing verb (i.e. pre- and post-positioned subjects, respectively). At group level, dependency arrow heads are marked for head

[1] Except for one corpus, which has the appropriate restrictions concealed in the Web interface.

type: @>N is a prenominal modifier (to be combined with PoS information, like 'adjective', 'determiner' etc.), and @P< is the argument of a preposition. Clause function is marked by adding a second, "external", tag on clause header words (relatives, interrogatives and subordinating conjunctions) or non-finite verb chain headers (in clauses without a header word). These tags indicate both syntactic form and function: @#FS-<ACC, for instance, means a finite subclause which functions as a direct (accusative) object. Together, these syntactic dependency tags represent overall structure the same way a mobile is built – every word knows its head, and derives its "outer function" from this head. In the example below, the prenominal (@>N) 'os' attaches to a subject head (@SUBJ>) 'problemas', which again "knows" its head, the finite main verb (@FMV) 'são'. The whole subclause is "represented" by 'que' which carries two tags, the - internal - subordinator tag (@SUB), and the – external – object tag (@#FS-<ACC), which in turn attaches the clause to the top main verb (@FMV) 'sabe'.

Sabe	[saber] V PR 3S IND	@FMV
que	[que] KS	**@#FS-<ACC@SUB**
os	[o] <art> DET M P	@>N
problemas	[problema] N M P	@SUBJ>
são	[ser] V PR 3P IND	@FMV
graves	[grave] ADJ M/F P	@<SC

Since syntactic tags, in this scheme, are word based, they combine with other – morphological – tags in a natural way, and can easily be searched for with the same corpus searching tools.

The annotation process

Given the existence of the previously described parser, it was an obvious choice to use it in the AC/DC project. The second phase of this project happened thus as a collaborative effort between the two sites Oslo and Århus. It proceeded as follows:

From the corpus encoded in the first phase of the AC/DC project in Oslo, a pure text version was re-created and sent to Århus to be analysed by the parser, which was therefore free to do whatever tokenization it would find appropriate. The result, in the format returned by the parser – after some filtering of parser-internal information – was sent back to Oslo, where it would then undergo considerable restructuring in order to:

1. get back to the original tokenization.
2. prepare for encoding in the IMS workbench.

The fact that this task is done in a collaborative way, and that neither of the authors wanted to make fundamental changes to their respective systems for the purpose of this cooperation, resulted in a far more complicated process than is generally reported in the literature. Let us explain the reasons for doing it this way:

First, Bick's parser is a general purpose system, not especially designed to annotate corpora in any specific format. In fact, some of the many applications it has been used for (such as language teaching (Bick, 1997), machine translation (Bick, 2000) and lexicography) are generally considered more challenging than corpus annotation by itself. One would not, therefore, expect the parser to need to suffer considerable changes for the present application.

Second, the corpora and their prior processing were carefully considered, and there are several reasons why it would not be appropriate to change them:

- We aimed at the most neutral tokenization, namely using only spaces and punctuation, as well as a list of abbreviations, as the sole source for the process. This makes tokenization repeatable, easy to document, and theory-neutral. That tokenization of real text is an error-prone complex process (as documented e.g. by Grefenstette and Tapanainen, 1994) can be easily seen in the fact that a percentage as high as 6% of the elements in some of the corpora (excluding punctuation proper) included punctuation marks (i.e., dots, hyphens, slashes, commas, etc.).
- The corpora are intended to be employed, among other uses, as a way of comparing different systems as far as the annotation they provide is concerned. It would not do to bias any aspect of such comparison through the use of a particular parser's (in this case, Bick's) choices.

On the other hand, it would have been a bad idea to force more than sentence separation on a parser designed for handling running text, and this is why the corpora were transformed into "running text form" before submitting them to the parser. The presence or absence of spaces between punctuation and lexical material, for instance, contains a great deal of structural information which is exploited by the parser, but often lost in standard corpus mark-up where all punctuation is isolated and angle-bracketed. Also, the run-time splitting of contractions into individual "words" (like 'em+uma' for 'numa'), and the intermediate introduction of polylexical units (like 'em_vez_de' or 'do_que') considerably facilitates the recognition of rule context patterns, and thus, assignment of *syntactic* function tags.

We show one example of the parser result, followed by the way it was re-encoded for the AC/DC project.

```
<p par=1>
<s>
Há      [haver]    V PR 3S IND VFIN @FMV
casos   [caso]     N M P @<ACC
jurídicos [jurídico] ADJ M P @N<
que     [que] <rel> SPEC M/F S/P @SUBJ> @#FS-N<
são     [ser] V PR 3P IND VFIN @FMV
como    [como] <rel> <prp> ADV @COM @#AS-<SC
as      [a] <artd> DET F P @>N
cerejas [cereja] N F P @AS<
$.
</s>
<s>
O       [o] <dem> DET M S @APP
de      [de] PRP @N<
Otelo   [Otelo] PROP M/F S/P @P<
$,
por=exemplo    [por=exemplo] ADV @ADVL
$.
</s>

<p par=1>
<s>
Há      haver    V       PR_3S_IND_VFIN  FMV
casos   caso     N       M_P     <ACC
jurídicos jurídico ADJ    M_P     N<
que     que      SPEC_rel M/F_S/P SUBJ>_#FS-N<
são     ser      V       PR_3P_IND_VFIN  FMV
```

```
como      como      ADV_rel_prp   0     COM_#AS-<SC
as        a         DET_artd      F_P   >N
cerejas   cereja    N             F_P   AS<
.         .         PU            0     PONT
</s>
<s>
O         o         DET_dem M_S   APP
de        de        PRP           0     N<
Otelo     Otelo     PROP          M/F_S/P P<
,         ,         PU            0     PONT
por       por=exemplo             ADV   0     ADVL
exemplo   por=exemplo             ADV   0     ADVL
.         .         PU            0     PONT
</s>
```

On the parser side of the process, two filter interfaces had to be crafted. An input filter, used to restore running text, removed and "stored" corpus meta tags (<par =357>, </s> etc.), and normalized punctuation to ordinary text standard (e.g. "-quotes). After analysis, a (more complex) output filter was used to remove all valency and semantic tags and, in the case of derived words, to create standard lexical base forms from the internally used root base forms and affix information. Also, possible orthographical changes introduced by the parser in its search for lexicon matches (Luso-Brazilian variation, spelling and accentuation irregularities) were reversed in order to maintain maximal corpus fidelity.

What was *not* filtered in the present project, were the actual tags, making the annotated corpora compatible with live CG style analyses (as, for instance, at http://visl.hum.sdu.dk). Given the size, modularity and granularity of the combined tag sets from *all* parsing levels, it would, however, be feasible to create filters for a wide variety of different (less detailed) tag sets at a later stage, as has been repeatedly shown when co-operating with other teaching or tagging projects.

The annotation result

Let us explain in more detail the conversions done in some cases, and their motivation.

Clitic processing

Clitic processing is always a vexing problem in the automatic analysis of Portuguese, due to the mesoclitics (*afirmá-lo-ei,* 'state-it-I will') and the phono-graphical changes required by the clitics (*pu-la = pus+a,* 'I put her').

In a corpus context, the problem is to keep both the used form and the information of which "canonical", non-cliticized, form it corresponds in a way that allows easy search for both. There is no simple solution to this problem, as the discussion of several alternatives will hopefully convince the reader:

- "Restoration" of the underlying forms (such as *afirmarei-o*) would destroy the actual text, in addition to creating non-Portuguese (i.e., ungrammatical) sequences. This would undermine the most important motivation for consulting corpora in the first place: the need for authentic text material.
- Separation by the hyphen would create morphemes that cannot occur in isolation (such as *afirmá* or *ei*); and would increase ambiguity of the participant forms where there is none (e.g., in the case of *como-a* ('I eat it'), a verb form followed by an accusative personal pronoun would be transformed into *como* and *a*, both highly ambiguous word forms in Portuguese).
- Marking that the forms did not occur in isolation by leaving hyphens in both sides, finally, would have both the disadvantage of not preserving the text and of not letting one look for similar words (e.g., in *Vi o rapaz* and *Vi-o*, one would have *Vi* and *Vi-* for the first person of the verb *ver*).
- Finally, leaving the verb and the clitic as a single token, in addition to not letting forms decide on similar words, as in the previous case, makes the information rather compact, in that the classification of one token has to carry both the features pertaining to the verb and the ones pertaining to the clitics(s).

This last choice is, however, the easiest to accomplish and the one chosen in the *first* phase of the AC/DC project. Its advantages are that the form is preserved, and that the information on the smaller constituents is then provided by the parser in a second phase[2]. It is also the option that makes counting easiest: a word is determined solely by graphical means.[3]

Compounds/MWE treatment

The same rationale applies to the treatment of multiword expressions. While wishing to maintain the information provided by the parser on compounds, we do not want to lose a parser independent tokenization strategy.

On can distinguish three kinds of cases considered as one token by the parser:

1. What is called in traditional Portuguese grammar "locuções", i.e., several words working as a grammatical unit, such as *a partir de*, *no entanto*, *por trás de*, etc.
2. Idioms and fixed phrases with no morphological variation, such as *por exemplo*, *um pouco*, *de pé*, *castelo de cartas*, *seja como for*, *fora de si*, *por minha causa*, *não há como*, *horas seguidas*, etc.
3. And all sorts of compound proper nouns, e.g. *Rua São Justino*, *Seu Carlos*, or *Auto da Compadecida*.

Even though the difficulty of making this sort of decisions varies according to the kind, in neither case is this identification an error-free process. One cannot, therefore, rely blindly on the parser output. Nor can one expect to be able to decide without help of any parser. It is also obvious that different parsers, grammar theories and lexicons may drastically differ in such classification decisions. In (Santos, 1990), it was even suggested that the definition of MWEs is actually application dependent.

We have thus decided to provide each form as a distinct token, while keeping available the result of the parser's processing by encoding the whole compound as the lemma for each form, as in the example *por exemplo* in the previous example, or in the following MWEs featuring the word *horas* ('hours').

```
horas     horas=de=ponta  N     F_P   P<
```

[2] Syntactically, in any case, regarding a verb+clitic construction or a preposition+determiner contraction as one (functional) unit is awkward, since the very notion of syntactic constituents contradicts graphical word boundaries in these cases.

[3] One can, of course, also count which words have hyphens and, of these, which are most probably verbs with clitics. A full-fledged parser may still be required to decide in a few cases, though, especially when there are typos in the material.

```
de      horas=de=ponta N     F_P   P<
ponta   horas=de=ponta N     F_P   P<

horas   horas=seguidas N     F_P   <ADV
seguidas horas=seguidas N    F_P   <ADV

Horas   horas=a=fio    ADV   0     ADVL>
a       horas=a=fio    ADV   0     ADVL>
fio     horas=a=fio    ADV   0     ADVL>
```

The encoded result shows that not all tokens in the corpus are individually classified. While this is possibly irrelevant for "members" of a proper noun, it may be disturbing for other uses of the corpus. However, since these cases form a closed list, automatic addition of subanalyses for MWEs is a feasible solution:

```
Horas   hora    N     F_P   ADVL>
a       a       PRP   0     N<
fio     fio     N     M_S   P<
```

Contractions

Contractions in Portuguese are cases where a preposition and a determiner (article or pronoun) are contracted into a single word form, with no ortographical marking (examples are *dela*, *comigo*, *pro*, *pelas*, *do*, respectively de + ela, com + eu, para + o, por + as, de +o). The parser transforms these items into their constituents, and sets a morphological flag.

Consistent with our approach in the two previous cases, we restore the contractions and add the corresponding attributes.

Summing up, in addition to our wish not to modify the original text, one important reason why we undergo all this trouble is that alternative analyses require different tokenization, in each of the three cases discussed:

- Contractions: *deste* (a verb form or the contraction *de* + *este*); *pelo* (a singular noun or the contraction *por* + *o*); *consigo* (a verb form or the contraction *com* + *si*)
- MWEs: *mais valia* (adverb plus verb or complex noun); *a favor de* (preposition noun preposition or complex expression).
- Clitics: *tem-nos* (the clitic is *nos* (first person plural) or underlying *os* (third person object masculine pronoun).

The extent to which tokenization is different in the two systems is surprisingly large, as Table 4 proves beyond doubt.

Processing stage	**EBRANOT with and without punctuation**		**NATPANOT with and without punctuation**	
Original version	884,729	661,395	857,742	696,918
Parser's output	889,580	669,162	878,351	719,700
Contr. merging	846,723	626,305	816,170	657,520
MWE expansion	878,586	658,168	845,113	686,463
Clitics merge	872,563	652,146	843,372	684,723

Table 4: Tokenization size

Putting together verbs with enclitics accounts for a reduction of 0.9% in EBRANOT and 0.25% in the first part of NATPANOT. The corresponding shrinking for contractions is 6.4% and 8.6% respectively. On the other hand, the expansion of MWE into several tokens contributes to raises in the number of tokens of 5.1% and 4.4%. Even though the aim of all this processing is the restoration of the original tokenization, we still have differences in the number of tokens, the reasons for which are currently being investigated.

All in all, compared to the tokens returned by the parser, 12% to 14% of the final tokens, excluding punctuation marks, are new. This is an interesting measure, in our opinion, since it shows how unreliable measures of performance (e.g. errors per "words") can be when they are compared numerically without taking into account the tokenization assumptions involved.

Encoding in IMS-CWB

We do not intend to provide here more than some general clues as to the use of the IMS-CWB. Readers of this section are encouraged to read elsewhere (Christ et al., 1999, Christ, 1998) on the capabilities and internals of this corpus system. But for those who already use it, it might be relevant to motivate some of our choices.

Positional attributes

Different annotation levels were encoded as four different positional attributes: lemma; part of speech (N, V, ADJ, DET, etc. plus a combination thereof in the case of multiword expressions; together with some subcategories returned by the parser, as in DET_poss or ADV_dem_quant_komp); morphological information (like gender, tense, etc.); and functional information. Neither morphological nor functional information is necessarily unique, so the several pieces are concatenated, separated by underscores, in order to provide one value for the corresponding attribute. Thus, "M_P" stands for Masculine Plural, and "PRD_#AS-<ADVL" means that the word functions as a role predicator and indicates an absolute clause which has the function ADVL in the main clause.

In some undecidable cases the parser uses morphological portmanteau-tags. Here, alternative or undefined options are marked with a slash (e.g. "S/P" means singular **or** plural). For its syntactical annotation, the parser resolves remaining ambiguity by progressively more heuristic rule levels.[4] Provisionally, for the task at hand (a corpus search interface), tag ambiguity was set to be zero.

One of the obvious advantages of the physical separation between corpus and annotation provided by the IMS-CWB is that the very same corpus can have POS1, LEMA1, etc. for this parser's output, and POS2, LEMA2, etc. for another. As mentioned above, we intend to annotate the same corpora with different parsers and taggers, which will allow a user to look for systematic differences between systems or problematic areas in general for the parsing of Portuguese.

The Web interface

[4] In fact, some syntactic ambiguity can - in flat dependency grammar - be expressed by using only *one* tag: @N<, for instance, means a postnominal constituent, but underspecifies just how many nouns to the left the attachment head is to be found (e.g. 'o @>N homem @? com @N< a @>N bicicleta @P< da @N< China @P<').

For the moment, the Web interface is simply a window into CQP – the corpus query processor – with some trivial possibilities added

- the semicolon is not required when only one command is involved
- quotes are not obligatory when only one token is involved.

The only substantial addition is the possibility of asking for the distribution of a regular expression as a simple query.

As far as restrictions are concerned, the user cannot change the corpus s/he is querying. Likewise, s/he cannot, for obvious reasons, rely on the use of local corpora.

Given that the parsing of the corpora is still work in progress, we keep the non-annotated and the annotated versions of each corpus distinct. We expect to merge them when 100% tokenization agreement is achieved.

We present here some examples of the query power allowed by our service, due to the combined advantages of using this parser, the IMS corpus workbench and the particular Web interface. One can look for (see documentation on our Website for the actual syntax):

- Objects of the verb X
- Verbs which have as object Y
- Preposition X occurring *not* within a proper noun
- Nouns having pre-modifying *and/or* post-modifying adjectives
- Verbs in the conditional
- Words not in the lexicon
- Adjectives forming part of (complex) proper nouns
- Forms being used both as verbs and as nouns in the corpus, but more frequently as verbs

Evaluation

Evaluation of the usefulness of the service – or more especially of the second phase – can be done according to different axes

- number of visits and successful queries
- accuracy/recall of the queries (which is obviously also dependent on parser performance)

Due to lack of sufficient information – the documentation is being written in parallel with writing the present paper, and no general announcement has yet been made regarding the second phase of the AC/DC project – it is too early to study access patterns. We intend to measure the usefulness of the query result for particular queries by making three different counts:

1. how many examples one would have to look at if the corpus was not tagged, compared to the ones found
2. how many cases found were actually right
3. how many cases were missing

The first measure can actually be the most important in an interactive service where one can refine one's queries and try alternative query options at once.

Future work: a third phase?

The most obvious need after creating this service, and the resources it serves, is to supply enough secondary documentation and teaching material so that would-be corpus users can exploit to their satisfaction a reasonable part of the tools provided. Such secondary material could include guided tours, a discussion of alternative grammatical or tagging approaches, FAQ-lists, or a regular teaching interface for university students.

In terms of content, the service could be improved by proof-reading some of the automatically annotated corpora, tagging other text kinds (speech data, historical data) or using the parser to provide graphical syntactic tree structure annotation (cf. http://visl.hum.sdu.dk).

Although manual annotation of the corpora is a possibility, we believe it is better first to engage in a detailed analysis of the parser's performance, with subsequent documentation of its strengths and weaknesses. In the long run, an improved parser would allow faster annotation proofing, so *manual* corpus annotation should not be seen as an end in itself.

Finally, other work falling under the scope of the AC/DC project, no longer necessarily connected with the second author, is to engage in the same kind of collaborative process with other parsers, resources and systems available for Portuguese, and compare the results.

References

Bick, Eckhard. (1996). Automatic parsing of Portuguese. In Proc. Second Workshop on Computational Processing of Written Portuguese (Curitiba, 23-25 October 1996) (pp. 91--100).

Bick, Eckhard. (1997). Internet Based Grammar Teaching. In E. Christoffersen & B. Music (Eds.), Datalingvistisk Forenings årsmøde 1997 - DALF '97 (pp. 86--106). Kolding.

Bick, Eckhard. (1998). Structural Lexical Heuristics in the Automatic Analysis of Portuguese. In B. Maegaard (Ed.), Proc. 11th Nordic Conference on Computational Linguistics, Nodalida '98 (pp. 44--56). Copenhagen.

Bick, Eckhard. (2000). The Parsing System "Palavras" – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Århus.

Christ, Oliver. (1998). Linking WordNet to a Corpus Query System. In J. Nerbonne (Ed.), Linguistic Databases (pp.189--202). Stanford: CSLI Publications.

Christ, O., Schulze, B. M., Hofmann, A., & Koenig, E. (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. University of Stuttgart, March 8, 1999 (CQP V2.2).

Grefenstette, G. & Tapanainen, P. (1994). What is a word, What is a sentence? Problems of Tokenization. In Proc. 3rd International Conference on Computational Lexicography, COMPLEX'94 (pp. 79--87).

Karlsson, Fred, et. al. (1995). Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text. Berlin: Mouton de Gruyter.

Santos, Diana. (1990). Lexical gaps and idioms in Machine Translation. In H. Karlgren (Ed.), Proceedings of COLING'90, Vol 2 (pp.330--335). Helsinki.

Santos, Diana. (1998). Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts. In Rubio et al. (Eds.), Proceedings of The First International Conference on Language Resources and Evaluation, Vol. 1 (pp.475--481). Granada.

Santos, Diana. (to appear). Comparação de corpora em português: algumas experiências. In T. Berber Sardinha (Ed.), Língua Portuguesa no Computador, São Paulo.

Tapanainen, Pasi. (1996). The Constraint Grammar Parser CG-2. Publication No. 27. Helsinki: Department of General Linguistics, University of Helsinki.