LINGUATECA’S INFRASTRUCTURE FOR PORTUGUESE AND HOW IT ALLOWS THE DETAILED STUDY OF LANGUAGE VARIETIES

DIANA SANTOS

ABSTRACT

In this paper I present briefly Linguateca, an infrastructure project for Portuguese which is over ten years old, showing how it provides several possibilities to study grammatical and semantic differences between varieties of the language.

After a short history of Portuguese corpus linguistics, presenting the main projects in the area, I discuss in some detail the AC/DC project\(^1\) and what is called the AC/DC cluster (encompassing other related corpus projects sharing the same core). Emphasizing its potential for language variation studies, the paper also (i) describes CONDIVport’s integration as an impetus for new capabilities, and (ii) provides a sketch of newly added functionalities to AC/DC.

[1] AN INFRASTRUCTURE FOR PORTUGUESE LANGUAGE TECHNOLOGY

In line with the main audience of the Linguateca project, there have already been several descriptions of Linguateca as an infra-structure for Portuguese, in Portuguese (Santos, 2009), as well as substantial reporting\(^2\). However, an international audience has only seen traces and scattered references so far, so this paper intends to fill this gap in what corpus resources are concerned.

It all started in 1998, with a small project (1998-1999) for preparing the future of the computational processing of the Portuguese language hosted by SINTEF, as an area to be taken specially good care of in the future science and technology programme (for the White book on Science and Technology created by the then Ministry of Science and Technology in Portugal), and wrote a small memo to be discussed publicly by all interested parties (Santos, 1999a).

\(^1\) The name stands for Acesso a Corpores, Disponibilização de Corpores (roughly: access to corpora, making corpora available), and is meant to signal that it should both benefit users – granting them access; and corpus owners: helping them to make their corpora widely available. See www.linguateca.pt/ACDC/

\(^2\) More than 500 items in the publication list at http://www.linguateca.pt
After the discussion, and given that several projects had been started (catalogue, publications catalogue, and some corpus services), one more year was granted, that prepared the ground for what later became Linguateca.

Linguateca was conceived as a three-axed initiative to foster R&D in the computational processing of the Portuguese language, with relevant work on (i) information dissemination, (ii) resource creation, and (iii) organization of evaluation initiatives.

PORTUGUESE TEXT CORPORA

Assuming that an international audience is probably generally unaware of what has been done in Portuguese corpus processing, I will attempt here a short presentation of the field, with special emphasis on what is offered by Linguateca.

A brief history

As far as I know, corpus compilation for Portuguese started during the 1960s with the Português Fundamental (Bacelar do Nascimento et al., 1984, 1987), a project shaped after and inspired by the Français Fondamental (Gougenheim et al., 1964). Strict criteria for documenting authentic usage in oral contexts all over the country were used, and a significant number of documents of spoken Portuguese (from 1971 to 1974) was recorded, transcribed and analysed at the Centro de Linguística da Universidade de Lisboa, see Bacelar do Nascimento (2001). The work of this team has continued ever since with the compilation i.a. of the large Corpus de Referência do Português Contemporâneo, CRPC³ (Bacelar do Nascimento, 2000), as can also be appreciated in the recent papers on the comparison of African varieties of Portuguese (Bacelar do Nascimento et al., 2008a,b).

Several degrees of latitude and longitude further, the NURC project (Cal-lou, 1999) was taking place in Brazil, aiming to describe the oral and educated language⁴ in five major Brazilian cities (Recife, Salvador, Rio de Janeiro, São Paulo and Porto Alegre), being thus a five-headed project. Started in 1970, it produced different oral corpora and different research lines, as can be better appreciated in the overview by Varejão (2009). In NURC-RJ, comparative oral corpora of the decades 1970s and 1990s were deployed, and it is currently connected with the project Para uma História do Português do Brasil, PHPB⁵, including also written materials since the XVIth century. In Recife, the project was extended to address conversation analysis, while in Porto Alegre it merged with the VARSUL project (Menon et al., 2009).

References:

[4] Norma Urbana Culta, see e.g. http://www.letras.ufrj.br/nurc-rj/
Outside a Portuguese-speaking countries context, Brigham Young University (US) researchers were interested in electronically available Portuguese material, having created the Borba-Ramsey corpus\(^6\), a subset of which was later included in the European Corpus Initiative (Thomson et al., 1994) and has since 1999 been browsable also through AC/DC. We can also mention Portext (Maciel, 1997) in France, the English-Norwegian Parallel Corpus in Norway (Oksefjell, 1999) and the VISL (Bick, 1997) project in Denmark, as early providers of Portuguese texts searchable on the web. Castilho et al. (1995) mention John Uriagereka from Maryland as having proposed a joint database for Portuguese and Gallician as early as 1991. From the same source we also learn that in 1993 there was already a corpus project in Mozambique, led by Perpétua Gonçalves.

As to the specific comparison of different varieties of Portuguese, there are at least six corpus-based projects that deserve mention here: The Tycho Brahe project (Galvés, 2009), VARPOR (Brandão & Mota, 2003), PEPB (Peres & Kato, 2004), Corpus do Português (Davies & Ferreira, 2006-), Banco do Português (Berber Sardinha, 2007), and CONDIVport (Soares da Silva, 2010). Early corpus-based work can be found in Barreiro et al. (1996); Wittmann et al. (1995).

For further information and historical overviews on Portuguese corpora – of which the pointers presented are just a small part, since many other corpora have come to light during the last decade – see Bacelar do Nascimento et al. (1996), Oksefjell & Santos (1998), Berber Sardinha (1999), Davies (2008), Berber Sardinha & Almeida (2008), Santos (2009) and Varejão (2009), as well as, of course, Linguateca’s resource catalogue.

What I would like to stress here, before introducing the AC/DC project in the next section, is: when it started back in 1998, there were no services on the web that allowed a linguist or an engineer to query a Portuguese corpus. Also, the few available corpora for download had very different formatting, encoding, and conceptual organization, so that their content was hard to compare and required a lot of processing to be used simultaneously, as explained in Santos (1999b) as initial motivation for AC/DC.

[2.2] The AC/DC cluster

As devised in 1998-1999, AC/DC had as its main purpose to make a large number of corpus resources available on the web with a unified and simple interface that allowed people to interact with corpora without requiring physical access to institutions or software installation (at that time, there was no such thing for Portuguese). Later on we also considered as Linguateca’s task

\[\text{[6]}\] Named after the corpus compilers, Francisco Borba and Myriam Ramsey.
to create resources that were lacking, such as a large newspaper text corpus, CETEMPúblico, which was also included in the AC/DC service.

As a service to the (Portuguese-language processing) community, every corpus owner or developer could make use of AC/DC to serve his corpora, and we have in fact tried to contact everyone and make the offer explicit, for modern Portuguese. In some cases, however, the offer was turned down (or simply ignored), for reasons that ranged from copyright problems to the desire of the particular groups to develop their own solutions. We note, however, that no requirement of exclusivity was ever made by Linguateca: on the contrary, our own corpora, notably CETEMPúblico (Santos & Rocha, 2001), were also distributed by the Linguistic Data Consortium (LDC) and by Mark Davies for some time. So, one of the most used corpus of Portuguese, the NILC corpus, was given access by AC/DC although many other solutions to make it available were created as well by NILC (Aluísio et al., 2004).

Other related (resource) services provided by Linguateca were then developed as, in a way, an outgrowth of the basic AC/DC services, and I refer to this extended set as the AC/DC cluster, including the Floresta Sintáctica treebank (Afonso et al., 2002; Freitas et al., 2008) – the first treebank for Portuguese, COMPARA (Frankenberg-Garcia & Santos, 2003) – a large manually revised Portuguese-English fiction parallel corpus, and CorTrad (Tagnin et al., 2009) – a parallel (multi-version and multi-genre) corpus. These other resources have further tools, parts, and interfaces, which will not be dealt with here, and were created in cooperation with other researchers and projects.\footnote{See \url{http://www.linguateca.pt/Floresta/}, \url{http://www.linguateca.pt/COMPARA/}, and \url{http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html} for more information.}

\section{Studying variation and language varieties with the AC/DC cluster}

I start by a presentation of the available material, then present the browsing of CONDIVport, which was compiled for variational analysis, and finally present new functionalities for corpus-based discovery that are currently under test in the AC/DC project.

\subsection{The initial and obvious data gathering}

In order to be able to compare and study varieties and variation, one has to have materials that represent them. So, the first and obvious requirement is to have plenty of material, so that one can take a “language bath”, and immerse in language use in different countries, times and social classes. While this seems easy and obvious, in practice it isn’t. In fact, what most people have in terms of electronic corpora is opportunistically gathered in nature, and Linguateca’s offer is no exception.
TABLE 1: Genre distribution in the AC/DC cluster, as of July 2010

<table>
<thead>
<tr>
<th>Genre</th>
<th>Size in words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Narrative fiction</td>
<td>17,208,056</td>
</tr>
<tr>
<td>General newspaper</td>
<td>246,112,499</td>
</tr>
<tr>
<td>Specialized newspaper</td>
<td>6,367,807</td>
</tr>
<tr>
<td>Informative, technical</td>
<td>4,489,043</td>
</tr>
<tr>
<td>Oral</td>
<td>500,811</td>
</tr>
<tr>
<td>Other or not classified</td>
<td>5,067,371</td>
</tr>
</tbody>
</table>

In Table 1, the AC/DC material is roughly quantified under the genre parameter. Of course genre is a very elusive category, and a really thorough study of Portuguese genre is still unavailable, so under “informative, technical” different subcategories were joined such as essay, encyclopedic and textbook material, as well as email on librarianship. Also, thematic newspaper corpora were classified as “specialized newspaper”, while local and (party) political newspapers were considered “general newspaper”. As to the “other” category, it joins at least (e-mail) spam, EU calls, business letters, legal documents and web texts, especially blogs.

Table 2 presents the material in terms of language variety.

TABLE 2: Variety distribution in the AC/DC cluster, as of July 2010

<table>
<thead>
<tr>
<th>Language variety</th>
<th>Size in words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>76,802</td>
</tr>
<tr>
<td>Brazil</td>
<td>64,878,821</td>
</tr>
<tr>
<td>Portugal</td>
<td>215,377,125</td>
</tr>
<tr>
<td>Unknown</td>
<td>723,626</td>
</tr>
</tbody>
</table>

Finally, just for the fiction material, Table 3 presents the distribution per decade in the last two centuries. Since three of the sources concern parallel corpora, let me clarify that only the material in Portuguese is counted (and the dates for the translation concern the publication of the translation, not of the original). For more details see the corresponding project pages. Note also that literary text of which the exact sources are not known (such as those included in some multi-genre corpora in AC/DC) is not included.

In addition to the textual material, special sentence separation and tokenizer modules for Portuguese were developed in AC/DC, and all data were parsed by PALAVRAS (Bick, 2000), offering lemma, part of speech, morphological information (such as tense form, gender, number, pronoun case, diminutive, augmentative and superlative degree) and syntactical function (in a ver-
TABLE 3: Temporal distribution of literary texts in the AC/DC cluster (tokens) (July 2010)

<table>
<thead>
<tr>
<th>Decade</th>
<th>Vercial</th>
<th>COMPARA</th>
<th>ENPC</th>
<th>CorTrad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1800</td>
<td>207,473</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1810</td>
<td>252,599</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1820</td>
<td>229,116</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1830</td>
<td>53,110</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1840</td>
<td>323,622</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1850</td>
<td>89,258</td>
<td>11,302</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1860</td>
<td>591,702</td>
<td></td>
<td>22,053</td>
<td></td>
</tr>
<tr>
<td>1870</td>
<td>511,453</td>
<td></td>
<td>18,766</td>
<td></td>
</tr>
<tr>
<td>1880</td>
<td>666,540</td>
<td></td>
<td>84,549</td>
<td></td>
</tr>
<tr>
<td>1890</td>
<td>304,846</td>
<td></td>
<td>17,055</td>
<td></td>
</tr>
<tr>
<td>1900</td>
<td>543,050</td>
<td></td>
<td>29,937</td>
<td></td>
</tr>
<tr>
<td>1910</td>
<td>377,369</td>
<td></td>
<td>21,840</td>
<td></td>
</tr>
<tr>
<td>1920</td>
<td>328,588</td>
<td></td>
<td>5,943</td>
<td></td>
</tr>
<tr>
<td>1930</td>
<td>103,136</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1940</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1950</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1960</td>
<td></td>
<td></td>
<td>17,802</td>
<td></td>
</tr>
<tr>
<td>1970</td>
<td></td>
<td></td>
<td>160,240</td>
<td></td>
</tr>
<tr>
<td>1980</td>
<td></td>
<td></td>
<td>256,423</td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td></td>
<td></td>
<td>764,942</td>
<td>72,389</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td>25,074</td>
<td>98,806</td>
</tr>
</tbody>
</table>

sion of dependency structure constraint grammar developed for Portuguese by Eckhard Bick, including also some discourse-related features such as topic and focus and some semantic information). As discussed in Inácio & Santos (2006), some of the material in the AC/DC cluster has been manually revised, as to their text and to their annotation, but most of it has not (after all, AC/DC encompasses more than 280 million words, or ca.16 million different sentences).

In addition to having developed our own AC/DC format as a transduction of PALAVRAS output format, we have also started to add semantic information in some domains, using a simple lexicon-driven approach followed by human rule writing for correction and improvement of both precision and recall, as described in Silva & Santos (2009); Santos & Mota (2010).

The distribution of the colour domain can be appreciated in Figure 1, where both the density of colour tokens and types is shown. As far as I know, this is the largest semantically annotated corpus, which has undergone human revision, currently available. (Although colour annotation of the largest corpora
LINGUATECA’S INFRASTRUCTURE FOR PORTUGUESE AND HOW IT ALLOWS THE DETAILED STUDY OF LANGUAGE VARIETIES

has not yet been fully revised.)

![Graph](image)

**Figure 1:** The semantic field of colour in AC/DC as of July 2010: circles describe types, triangles tokens. Colour density is defined as 10,000 times the ratio of colour words (tokens or types) compared to all words (tokens or types) in the corpus.

[3.2] Support for formal variational linguistics

In addition to providing an “electronic bookshelf”, or a web distribution window, to any group or project that is willing to have us making their corpus or resource available, the AC/DC project may also develop specific facilities for the resources it (re)distributes, if they display new features.

This happened with the CONDIVport corpus – we started by simply making it available through the web as a regular AC/DC member, but soon we understood the interest in providing support for more complex models of (on-line) linguistic research: Given that CONDIVport was compiled to study the convergence and divergence of national varieties of Portuguese, under the framework initially developed by the *Quantitative Lexicology and Variational Linguistics group* in Leuven[^8], it had, in addition to three specific themes (soccer, fashion

and health), texts from three different time periods, from Brazil and from Portugal. In addition, as an integral part of the methodology, a list of terms in the two first of these themes had also been compiled.

For foundations and critical discussions of the methodology, I redirect the reader to Geeraerts et al. (1999); Geeraerts & Grondelaers (1999); Speelman et al. (2003); Soares da Silva (2010). Here, I will only provide concrete examples of how the process goes: First, one gathers a set of formal onomasiological profiles\(^9\) for key concepts in a given area – let us take clothing as an example: key concepts may be blusa (roughly “blouse”) or saia (roughly “skirt”). Their onomasiological profile is a set of lexical items which the linguist classifies (in context) as belonging to this class. So, as an example, the CASACO F (“female overcoat”) profile has been found to be: blazer, bêizer, casaco, casaquinho, casaquinha, manteau, mantô, paletó, paletot (Soares da Silva, 2008a, page 66).

Together with their frequencies, these profiles allow the researcher to compute several measures such as uniformity, and relative uniformity – an example from soccer (Soares da Silva, 2008b, page 28) is presented in Table 4, concerning the profile of a special kind of soccer player, and how the several words used to represent it occur in the 50’s, 70’s and 2000’s – and thereafter draw conclusions as to vocabulary trends and convergence/divergence among the varieties at stake (P for Portugal, B for Brazil).

<table>
<thead>
<tr>
<th></th>
<th>P50</th>
<th>B50</th>
<th>P70</th>
<th>B70</th>
<th>P00</th>
<th>B00</th>
</tr>
</thead>
<tbody>
<tr>
<td>atacante</td>
<td>101</td>
<td>8,8</td>
<td>119</td>
<td>36,6</td>
<td>50</td>
<td>13,6</td>
</tr>
<tr>
<td>avançado</td>
<td>820</td>
<td>71,6</td>
<td>3</td>
<td>0,9</td>
<td>175</td>
<td>47,4</td>
</tr>
<tr>
<td>avante</td>
<td>0</td>
<td>0,0</td>
<td>159</td>
<td>48,9</td>
<td>0</td>
<td>0,0</td>
</tr>
<tr>
<td>dianteiro</td>
<td>220</td>
<td>19,2</td>
<td>22</td>
<td>6,8</td>
<td>74</td>
<td>20,1</td>
</tr>
<tr>
<td>forward</td>
<td>1</td>
<td>0,1</td>
<td>17</td>
<td>5,2</td>
<td>0</td>
<td>0,0</td>
</tr>
<tr>
<td>ponta-de-lança</td>
<td>3</td>
<td>0,3</td>
<td>5</td>
<td>1,5</td>
<td>70</td>
<td>19,0</td>
</tr>
</tbody>
</table>

U = 16,9 \quad U' = 0,6 

U = 28,8 \quad U' = 0,8 

U = 10,1 \quad U' = 0,4

This is a morose process that requires classification of a large number of corpus instances (all occurrences of the forms above). Only after all those decisions have been taken can the measures be computed and compared.

Now, one of the advantages of making the underlying corpora (annotation) available to other researchers is that other people can then inspect the individ-

\[^9\] From Speelman et al. (2003), onomasiological variation concerns “different terms used to refer to the same entity”, while formal onomasiological variation requires that no conceptual change is at stake, and therefore does not include cases like hyperonyms or hyponyms which are also frequently used about the same referent in discourse. The authors themselves are aware that this is not easy to distinguish for all corpus instances, though.
lingua teca’s infrastructure for portuguese and how it allows the detailed study of language varieties

...ual classifications, search for the classes and the specific contexts of occurrence, and even provide feedback or corrections if needed. A similar point has been made in Santos & Oksefjell (1999) in what concerns parallel corpora.

This allows for both a wider dissemination of the original research, and a better quality control through communication with one’s peers. Both aims are included in Linguateca’s mission for the computational processing (and study) of the Portuguese language.

It is thus currently possible to ask, in addition to the occurrence or distribution of the forms included in the profiles, for an entire profile, or for the profile distributions themselves. That is, how many cases of the members of the profile CASACO appear by date/decade, or variety.

We have also used the initial profiles compiled in CONDIVport as a seed to compiling larger sets of fashion-related lexical items, thus “colouring” the different corpora also with clothing information.10

3.3 New capabilities in the AC/DC interface
Several capabilities newly added to the AC/DC interface deserve mention here:

- Human validation of corpus illustration sentences for semantic relation evaluation (the VARRA service, developed in connection with yet another subproject in Linguateca, PAPEL11, whose goal was to create a free lexical ontology for Portuguese based on an existing general dictionary);

- Comparison of two search expressions, inspired by the CorpusEye search system (Bick, 2004), to compare explicitly two distributions;

- Reuse of a pattern database, inspired by the search system of Davies & Ferreira (2006-) and based on the capabilities of the underlying CWB system (Schulze, 1996; Evert, 2009).

These will pave the way for yet further developments in the AC/DC cluster, some of which can be mentioned here as natural extensions, namely (i) the possibility to include (tailored) synonym search as an option, following e.g. Christ (1998); and (ii) search by subject matter through concept nets.

Illustration sentences
Although their wealth of real, in context, examples is generally accepted as one of the basic advantages of corpora, as opposed to laboriously crafted ones (by

[10] Whether the use of semantic domains and ontology-based classifications is also useful for variation analysis is something that will have to be ascertained empirically.
a lexicographer or textbook author), it is not easy to come up automatically with good examples from a corpus, as pointed out by Kilgarriff et al. (2008).

Even harder did we find the task of illustrating, or validating, semantic relations between words in context, as we wished to do for PAPEL, whose relations between words (and not word senses) had been produced automatically and were thus in need of human validation (Gonçalo Oliveira et al., 2009, 2010).

We have thus developed an AC/DC-based service to help us achieve two related purposes: (i) find out the best patterns to validate and/or discover semantic relations in text, and (ii) develop clearer insights into the semantic fabric of Portuguese, while at the same time improving a public-domain semantic resource. As is common practice in Linguateca, we offer this as a service to the community, which means that everyone can use it to develop or evaluate their own resources.

Comparison of two phenomena
Although one could already perform a comparison by doing two (or more) searches in AC/DC on a row and then comparing the results, this capability provides an easier way by aligning the results on two sides of the same screen. Since we have been doing similar things within DISPARA for a long time now, cf. the quantitative wrapup function in Santos (2002), it seemed appropriate to offer this also in a monolingual corpus context.

Reuse of a pattern database
Again, this is not new in the sense that in other services offered by Linguateca, namely Águia (Santos, 2003), use was made of a set of patterns to query complex treebank structures in the Floresta project, but this feature had never been integrated in the main service interface, which relied mainly in direct e-mail answers to users asking us how to produce complex queries.

Now we have created an option of loading previous queries/commands into the search space, which, although possibly slowing down the corpus system, will also provide higher expressivity. It remains to be seen how much of this will in fact be reused/employed by power users of the AC/DC services.

Acknowledgements
Linguateca has throughout the years been jointly funded by the Portuguese Government, the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC, UMIC and FCCN.

I would like to thank the remaining members of the AC/DC team, Paulo

Rocha, Luís Costa, Rosário Silva and Cristina Mota for the joint work, the corpus owners for letting us grant access to them on the web, and all users who have requested features or suggested improvements.

Eckhard Bick's long-standing collaboration with his PALAVRAS parser has been the single most important factor for AC/DC's success near its users.

Thanks also to Tony Berber Sardinha and Violeta Quental for relevant information concerning the history of Brazilian corpora, to Fernanda Bacelar do Nascimento for relevant references, to Augusto Soares da Silva for the introduction to quantitative lexicology methodology, to Cristina Mota for pertinent comments on a draft version and, last but not least, to the VARRA team (Cláudia Freitas, Hugo Gonçalo Oliveira and Violeta Quental).

REFERENCES


LINGUATECA’S INFRASTRUCTURE FOR PORTUGUESE AND HOW IT ALLOWS THE DETAILED STUDY OF LANGUAGE VARIETIES


Davies, Mark & Michael Ferreira. 2006-. Corpus do português. 45 milhões de palavras, sécs. XIV-XX. http://www.corpusdoportugues.org.


