OSLa J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 113–128. (ISSN 1890-9639)
http://www.journals.uio.no/osla

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

DIANA SANTOS
SINTEF, Oslo & FCCN, Lisboa

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 Corpos, Disponibilização de Corpos* (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.

[2] 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.

[2.1] *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*, CRPC[3] (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 (Callou 1999) was taking place in Brazil, aiming to describe the oral and educated language[4] 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*, PHPB[5], 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).

Outside a Portuguese-speaking countries context, Brigham Young University

[3] "Reference Corpus of contemporary Portuguese", see http://www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_crpc.php

[4] Norma URbana Culta, see e.g.http://www.letras.ufrj.br/nurc-rj/

[5] "for a history of Brazilian Portuguese", http://www.letras.ufrj.br/phpb-rj/

(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 (Galves 2009), VARPORT (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 to create resources that were lacking, such as a large newpaper text corpus, CETEMPúblico (Santos & Rocha 2001), 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

[6] Named after the corpus compilers, Francisco Borba and Myriam Ramsey.

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, 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.[7]

[3] 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.

[3.1] *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.

In Table 1 on the facing page, 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

[7] See http://www.linguateca.pt/Floresta/, http://www.linguateca.pt/COMPARA/, and http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html for more information.

TABLE 1: Genre distribution in the AC/DC cluster, as of July 2010

Genre	Size in words
Narrative fiction	17,208,056
General newspaper	246,112,499
Specialized newspaper	6,367,807
Informative, technical	4,489,043
Oral	500,811
Other or not classified	5,067,371

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

Language variety	Size in words
Africa	76,802
Brazil	64,878,821
Portugal	215,377,125
Unknown	723,626

Finally, just for the fiction material, Table 3 on the following page 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, aumentative and superlative degree) and syntactical function (in a version 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

TABLE 3: Temporal distribution of literary texts in the AC/DC cluster (tokens) (July 2010)

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

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 on the next page, 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 has not yet been fully revised.)

[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

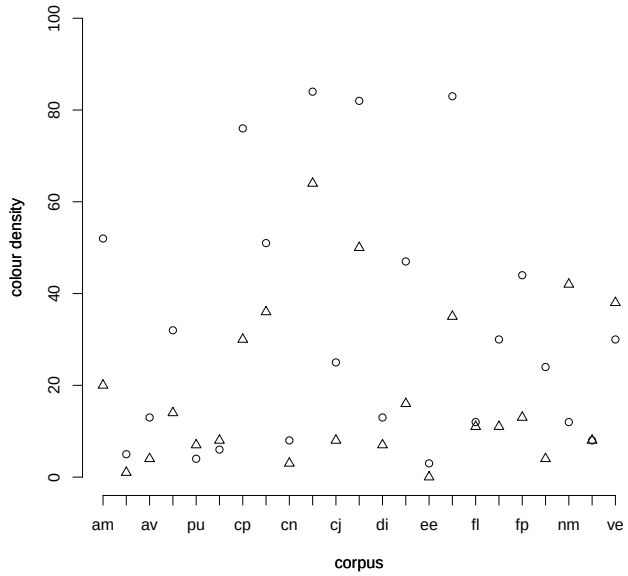

FIGURE 1: The semantic field of colour in AC/DC as of July 2010: circles describe types, triangles tokens. Colour density is defined as 10,000 times the ratio of colour words (tokens or types) compared to all words (tokens or types) in the corpus.

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 Leuwen[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

[8] See http://wwwling.arts.kuleuven.ac.be/qlvl/

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*, *blêizer*, *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 4: Absolute and relative frequencies and absolute and relative uniformity U e U' of the AVANÇADO onomasiological profile

Avançado	**P50**		**B50**		**P70**		**B70**		**P00**		**B00**	
atacante	101	8,8	119	36,6	50	13,6	208	73,8	42	9,7	658	96,2
avançado	820	71,6	3	0,9	175	47,4	0	0,0	240	55,4	0	0,0
avante	0	0,0	159	48,9	0	0,0	31	11,0	0	0,0	23	3,4
dianteiro	220	19,2	22	6,8	74	20,1	2	0,7	38	8,8	0	0,0
forward	1	0,1	17	5,2	0	0,0	0	0,0	0	0,0	0	0,0
ponta-de-lança	3	0,3	5	1,5	70	19,0	41	14,5	113	26,1	3	0,4
	U = 16,9		U' = 0,6		U = 28,8		U' = 0,8		U = 10,1		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 individual 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.

[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.

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, PAPEL[11], 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 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

[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.

[11] See http://www.linguateca.pt/PAPEL/

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[12], 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

[12] See http://www.linguateca.pt/acesso/varra.php

Gonçalo Oliveira and Violeta Quental).

REFERENCES

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002. Floresta Sintá(c)tica: a treebank for Portuguese. In Manuel Gonzalez Rodrigues & Carmen Paz Suarez Araujo (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, 1698–1703. Paris: ELRA.

Aluísio, Sandra, Gisele Montilha Pinheiro, Aline M.P. Manfrin, Leandro H.M. de Oliveira, Luiz C. Genoves Jr & Stella E. O. Tagnin. 2004. The Lácio-Web: Corpora and tools to advance Brazilian Portuguese language investigations and computational linguistic tools. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1779–1782.

Bacelar do Nascimento, Maria Fernanda. 2000. O corpus de referência do português contemporâneo e os projectos de investigação do Centro de Linguística da Universidade de Lisboa sobre variedades do português falado e escrito. In E. Gärtner, C. Hundt & A. Schönberger (eds.), *Estudos de gramática portuguesa (I)*, 185–200. Centro do Livro e do Disco de Língua Portuguesa. Biblioteca Luso-Brasileira.

Bacelar do Nascimento, Maria Fernanda. 2001. Les études portugaises sur la langue parlée. In M. H. A. Carreira (ed.), *Travaux et documents, les langues romanes en dialogue(s)*, vol. 11, 209–221. Université Paris 8.

Bacelar do Nascimento, Maria Fernanda, Antónia Estrela, Amália Mendes, Luisa Pereira & Rita Veloso. 2008a. African Varieties of Portuguese: Corpus Constitution and Lexical Analysis. In *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics - LULCL II 2008*.

Bacelar do Nascimento, Maria Fernanda, M. L. G. Marques & Maria Luísa Segura da Cruz. 1984. *Português Fundamental: Vocabulário e Gramática*. Centro de Linguística da Universidade de Lisboa.

Bacelar do Nascimento, Maria Fernanda, M. L. G. Marques & Maria Luísa Segura da Cruz. 1987. *Português Fundamental: Métodos e Documentos*, vol. 2. Centro de Linguística da Universidade de Lisboa.

Bacelar do Nascimento, Maria Fernanda, Luisa Pereira, Antonia Estrela, José Bettencourt Gonçalves & Sancho Oliveira. 2008b. Aspectos de unidade e diversidade do português: as variedades africanas face à variedade europeia. *Veredas* 9. 35–69.

Bacelar do Nascimento, Maria Fernanda, Maria Celeste Rodrigues & José Bettencourt Gonçalves (eds.). 1996. *Actas do XI Encontro Nacional da associação portuguesa de linguística (Lisboa, 2-4 de Outubro de 1995), vol I: Corpora*. Lisboa, Portugal: APL/Colibri.

Barreiro, Anabela, Luzia Helena Wittmann & Maria de Jesus Pereira. 1996. Lexical differences between European and Brazilian Portugueses. *INESC Journal of Research and Development* 5(2). 75–101.

Berber Sardinha, A. P. 1999. Beginning Portuguese corpus linguistics: exploring a corpus to teach Portuguese as a foreign language. *DELTA* 15(2). 289–299.

Berber Sardinha, Tony. 2007. History and compilation of a large register-diversified corpus of Portuguese at CEPRIL. *The Especialist* 28. 211–226.

Berber Sardinha, Tony & Gladis Maria de Barcellos Almeida. 2008. A Lingüística de Corpus no Brasil. In Stella E. O. Tagnin & Oto Araújo Vale (eds.), *Avanços da Lingüística de Corpus no Brasil*, 17–40. Editora Humanitas/FFLCH/USP.

Bick, Eckhard. 1997. Internet Based Grammar Teaching. In Ellen Christoffersen & Bradley Music (eds.), *Proceedings of Datalingvistisk Forenings Årsmøde 1997 i Kolding*, 86–106. Handelshøjskole Syd, Institut for Erhvervssprog og Sproglig Informatik.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus University, Aarhus, Denmark.

Bick, Eckhard. 2004. Looking at the Floresta Sintá(c)tica with a CorpusEye: A user-friendly cross-language search interface. URL http://www.linguateca.pt/documentos/floresta-corpuseye_en.pdf.

Brandão, Silvia Figueiredo & Maria Antónia Mota (eds.). 2003. *Análise contrastiva de variedades do português: primeiros estudos*. In-Fólio.

Callou, Dinah. 1999. O projecto NURC no Brasil: da década de 70 à década de 90. *Linguística* 11. 231–250.

Castilho, Ataliba Teixeira de, Gisele Machline de Oliveira e Silva & Dante Lucchesi. 1995. Informatização de acervos da língua portuguesa. *Boletim da Abralin* 17. 143–151.

Christ, Oliver. 1998. Linking WordNet to a Corpus Query System. In John Nerbonne (ed.), *Linguistic databases*, CSLI lecture notes, 189–202. CSLI Publications, CSLI Stanford.

Davies, Mark. 2008. New Directions in Spanish and Portuguese Corpus Linguistics. *Studies in Hispanic and Lusophone Linguistics* 1. 149–186.

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

Evert, Stefan. 2009. The CQP Query Language Tutorial. URL http://cwb.sourceforge.net/temp/CQPTutorial.pdf.

Frankenberg-Garcia, Ana & Diana Santos. 2003. Introducing COMPARA, the Portuguese-English parallel translation corpus. In Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds.), *Corpora in Translation Education*, 71–87. Manchester: St. Jerome Publishing.

Freitas, Cláudia, Paulo Rocha & Eckhard Bick. 2008. Floresta Sintá(c)tica: Bigger, Thicker and Easier. In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, 216–219. Berlin/Heidelberg: Springer Verlag.

Galves, Charlotte. 2009. Padrões rítmicos, domínios prosódicos e modelagem probabilística em corpora do português. URL http://www.tycho.iel.unicamp.br/tycho/prdpmp/projetofinal.pdf.

Geeraerts, Dirk & Stefan Grondelaers. 1999. Purism and fashion. French influence on Belgian and Netherlandic Dutch. *Belgian Journal of Linguistics* 13. 53–68.

Geeraerts, Dirk, Stefan Grondelaers & Dirk Speelman. 1999. *Convergentie en divergentie in de nederlandse woordenschat*. Amsterdam: Meertens Instituut.

Gonçalo Oliveira, Hugo, Diana Santos & Paulo Gomes. 2009. Relations extracted from a Portuguese dictionary: results and first evaluation. In Luís Seabra Lopes, Nuno Lau, Pedro Mariano & Luís M. Rocha (eds.), *New Trends in Artificial Intelligence, Local Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA 2009)*, 541–552.

Gonçalo Oliveira, Hugo, Diana Santos & Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática* 2(1). 77–93.

Gougenheim, G., R. Michéa, P. Rivenc & A. Sauvageot. 1964. *L'élaboration du français fondamental*. Paris: Didier.

Inácio, Susana & Diana Santos. 2006. Syntactical Annotation of COMPARA: Workflow and First Results. In Renata Vieira, Paulo Quaresma, Maria da Graça

Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*, 256–259. Berlin/Heidelberg: Springer Verlag.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell & Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX 2008*. Barcelona.

Maciel, Carlos. 1997. Textes et textes juridiques dans la Base de Données Textuelles PORTEXT. In *Secondes Journées Internationales de Terminologie (Actes du colloque, Le Havre, 14-15 octobre 1994)*. Le Havre.

Menon, Odete Pereira da Silva, Edson Domingos Fagundes & Loremi Loregian-Penkal. 2009. The VARSUL database. *Linguistik online* 38(2). 13–21.

Oksefjell, Signe. 1999. A Description of the English-Norwegian Parallel Corpus: Compilation and Further Developments. *International Journal of Corpus Linguistics* 4(2). 197–216.

Oksefjell, Signe & Diana Santos. 1998. Breve panorâmica dos recursos de português mencionados na Web. In Vera Lúcia Strube de Lima (ed.), *III Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR'98)*, 38–47.

Peres, João Andrade & Mary Aizawa Kato. 2004. Studies in the Comparative Semantics of European and Brazilian Portuguese. Special issue of *Journal of Portuguese Linguistics*, 3 (1).

Santos, Diana. 1999a. Computational processing of Portuguese: working memo. URL http://www.linguateca.pt/branco/white_paper.html.

Santos, Diana. 1999b. Disponibilização de corpora de texto através da WWW. In Palmira Marrafa & Maria Antónia Mota (eds.), *Linguística Computacional: Investigação Fundamental e Aplicações. I Workshop sobre Linguística Computacional da APL, FLUL, Maio de 1998*, 323–335. Lisboa: Colibri / APL.

Santos, Diana. 2002. DISPARA, a system for distributing parallel corpora on the Web. In Nuno Mamede & Elisabete Ranchhod (eds.), *Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002)*, 209–218. Berlin/Heidelberg: Springer-Verlag.

Santos, Diana. 2003. Timber! Issues in treebank building and use. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language: 6th International Workshop,*

PROPOR 2003. Faro, Portugal, June 2003, 151–158. Berlin/Heidelberg: Springer Verlag.

Santos, Diana. 2009. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamatica* 1(1). 25–59.

Santos, Diana & Cristina Mota. 2010. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, 1437–1444. European Language Resources Association.

Santos, Diana & Signe Oksefjell. 1999. Using a parallel corpus to validate independent claims. *Languages in Contrast* 2(1). 117–132.

Santos, Diana & Paulo Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 442–449.

Schulze, Maximilian Bruno. 1996. *MP User's Manual*. Institutt für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Silva, Rosário & Diana Santos. 2009. Arco-íris: notas sobre a anotação do campo semântico da cor em português. URL http://www.linguateca.pt/acesso/ArcoIris.pdf. First version: 25 June 2009.

Soares da Silva, Augusto. 2008a. Integrando a variação social e métodos quantitativos na investigação sobre linguagem e cognição: para uma sociolinguística cognitiva do português europeu e brasileiro. *Revista de Estudos Linguísticos* 16(1). 49–81.

Soares da Silva, Augusto. 2008b. O corpus CONDIV e o estudo da convergência e divergência entre variedades do português. In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*, 25–28. Linguateca.

Soares da Silva, Augusto. 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman (eds.), *Advances in cognitive sociolinguistics*, 41–83. Mouton de Gruyter.

Speelman, Dirk, Stefan Grondelaers & Dirk Geeraerts. 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37. 317–337.

Tagnin, Stella E. O., Elisa Duarte Teixeira & Diana Santos. 2009. CorTrad: a multiversion translation corpus for the Portuguese-English pair. *Arena Romanistica* 4. 314–323. [The 28th International Conference on lexis and grammar, Bergen, Norway, 30 September - 3 October 2009].

Thomson, H., S. Armstrong-Warwick & D. McKelvie. 1994. Data in your language: The ECI Multilingual Corpus 1. In *Proceedings of the International Workshop on Shareable Natural Language Resources (Nara, Japan, 10-11 August 1994)*. Institute of Science and Technology.

Varejão, Filomena de Oliveira Azevedo. 2009. O português do Brasil: Revisitando a História. *Cadernos de Letras da UFF – Dossiê: Difusão da língua portuguesa* 39. 119–137.

Wittmann, Luzia, Tânia Pêgo & Diana Santos. 1995. Português do Brasil e de Portugal: alguns contrastes. In *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística*, 465–487. Lisboa: APL/Colibri.

AUTHOR CONTACT INFORMATION

Diana Santos
Department of Literature, Area Studies and European Languages
Faculty of Humanities, University of Oslo
P.O.Box 1003 Blindern
N-0315 Oslo
Norway
d.s.m.santos@ilos.uio.no