The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools

Sandra Aluisio[1], Gisele Montilha Pinheiro[1], Aline M. P. Manfrin[1], Leandro H. M. de Oliveira[1], Luiz C. Genoves Jr.[1] and Stella E. O. Tagnin[2]

[1]NILC/ICMC-USP, [2]FFLCH-USP
Núcleo Interinstitucional de Lingüística Computacional (NILC),
ICMC-University of São Paulo, CP 668, 13560-970 São Carlos, SP, Brazil,
FFLCH – DLM, University of São Paulo, Av. Prof. Luciano Gualberto, 403, 05508-900 - São Paulo – SP, Brazil
sandra@icmc.usp.br, gisele@nilc.icmc.usp.br, aline@nilc.icmc.usp.br, leandroh@nilc.icmc.usp.br,
genoves@nilc.icmc.usp.br, seotagni@usp.br

Abstract

In this paper we discuss the five requirements for building large publicly available corpora which geared the construction of the Lácio-Web corpora and their environments: 1) a comprehensive text typology; 2) text copyright clearance, compilation and annotation scheme; 3) a friendly and didactic interface; 4) the need to serve as support for several types of research; 5) the need to offer an array of associated tools. Also, we present the features that make Lácio-Web corpora interesting and novel as well as the limitations of this project, such as corpora size and balance, and the non-inclusion of spoken texts in the project's reference corpus.

1. Introduction

The BNC corpus was built and the ANC corpus is being built through successful consortiums among dictionary publishers and/or software companies interested in language processing, British and American government, respectively, and academic researchers who came together to build publicly available large POS annotated reference corpora for English. The BNC requires a non-exclusive, nontransferable license to use the corpus for linguistic research purposes and/or the development of language products. It is not permitted to exploit the BNC commercially nor to publish any part of it although it is permitted to commercially explore the results of research carried out using the corpus. As for the ANC, the data has been distributed almost freely for non-commercial research purposes since its first release, while commercial use will be limited to members of the ANC Consortium for five years after the first corpus installment, which happened in the fall of 2003. In regard to copyrights, both projects asked their founding members to make their data available to the project. Strong consortiums of this kind to produce basic and important resources like corpora for Brazilian Portuguese language are unknown in Brazil. On one hand, this fact is understandable as Portuguese is not the lingua franca for science and business and Corpus Linguistics is still incipient in several Brazilian Linguistic Departments which are only now starting to work with electronic corpora and corpus-based Computational Linguistic tools. On the other hand, Portuguese is the mother tongue of approximately 200 million people (Brazil, Portugal, Angola, Cabo Verde, Guiné Bissau, Mozambique and S. Tomé and Príncipe) and the sixth most spoken language in the world today. Moreover, it is important to note that the two language variants with the greatest number of users – European Portuguese (EP) and Brazilian Portuguese (BP) – differ on the phonological, lexical, morphological and syntactic levels (Wittmann et al. 95) suggesting a real need to build different corpora for them, similar to what happened for British and American English for which we now have the BNC and the ANC. Besides, Brazilian users have awakened to the benefits online corpora and their associated tools can produce, forcing present corpora projects to take into consideration the real expectancies of a variety of target groups, like: academic and commercial groups; expert and lay users; and linguistic and computational linguistic researchers. This was the scenery when the Lácio-Web (LW) project was launched in early 2002.

LW is a 30 month-project and is being developed at the University of São Paulo, Brazil, in a joint venture among NILC[1], IME[2] and FFLCH[3]. It tries to fill the gap with regard to large linguistic resources and computational linguistic tools for BP (Aluísio et al 2003a; Aluísio et al 2003b). It is composed of 1) a contemporary synchronic reference corpus of BP written texts called Lácio-Ref; 2) Mac-Morpho, a gold standard portion from Lácio-Ref, which was manually-validated for morpho-syntactical tags; 3) an automatically-annotated portion of the Lácio-Ref with lemmas, POS and syntactic tags ; 4) a deviation corpus containing non-revised texts (Lácio-Dev); 5) and parallel and 6) comparable Portuguese-English corpora called, respectively, Par-C and Comp-C.

LW is innovative in that it aims at a) compiling different types of freely accessible corpora for both non-expert and expert users, b) training several POS taggers, which are freely available on the Web, on a very large manually annotated corpus of contemporary BP texts using a carefully designed tagset specifically created for this task, and c) making available different tools, which have been developed at NILC, for each corpora, such as automatic term extraction (ATE) methods and sentence and word aligners.

The LW corpora were designed for users pursuing both theoretical and practical linguistics studies, and developing both computational linguistics tools and applications like taggers, parsers, grammar checkers, natural language processing information retrieval methods and automatic summarizers. What especially distinguishes it from other corpora projects is its proposal to work as a benchmark to evaluate computational linguistic tools like POS taggers, ATE methods, sentence and word aligners, and grammar checkers. This can be accomplished with five of its corpora: Mac-Morpho (to evaluate POS

[1] http://www.nilc.icmc.usp.br/nilc/index.html
[2] http://www.ime.usp.br/
[3] http://www.fflch.usp.br

taggers), Comp-C and parts of Lácio-Ref (for ATE methods), Par-C (for alignment methods) and Lácio-Dev (to assess grammar checkers).

Concerning issues for building publicly available large corpora there are five which we wish to focus on in this paper as they geared the construction of the LW corpora and their environments: 1) the design of it's a comprehensive text typology; 2) text copyright clearance, compilation and annotation scheme (both for the header and the linguistic annotation); 3) the design of the interface (to meet the needs of both lay and expert users and to allow for subcorpora creation); 4) need to serve as support for several types of research; 5) provision of an array of associated tools. We will discuss them in Sections 3 to 7 after comparing LW with related corpora projects, in Section 2.

2. Related Corpora Projects

LW can be considered a specially designed project among Brazilian Portuguese (BP) corpora projects. However, it is a short-termed project (30 months) with 17 active members, most of them not working full-time on the project. This has certainly affected corpus size and balance and accounts for the non-inclusion of spoken texts in the reference corpus.

Considering the eleven major BP corpora projects presented in the poster section at the III Meeting of BP Corpora[4] that was held at Unicamp, Brazil in November 2003, only two of them include written and spoken texts: BANCO DE PORTUGUÊS[5] and VARPORT[6]. Longer (and successful) corpora projects for Brazilian Portuguese have addressed either the spoken language (e.g NURC-RJ[7], NURC-SP[8] and VARSUL[9]) or the written language making texts only partially available due to copyright restrictions, e.g. the written corpus "Usos do Português" from the State University of São Paulo, at Araraquara, which gave birth to three dictionaries: a dictionary of frequency, a dictionary of verbs and one of contemporary Brazilian Portuguese usage (Borba, 2002); and a grammar of Brazilian Portuguese usage; and the NILC Corpus which was built to support the development of a grammar checker for Brazilian Portuguese named ReGra. While the BANCO DE PORTUGUÊS, a huge monitor corpus with 240 million tokens, includes written and spoken texts it is also not publicly available due to copyright restrictions.

With regard to manual POS annotation, to the best of our knowledge, there are only two small Brazilian corpora which were used to train statistical taggers: (i) the 20,982-word Radiobrás Corpus and (ii) the 104,966-word corpus built from NILC's corrected text base covering 3 genres (news, literature and textbooks) (Aires et al, 2002). While the Tycho Brahe project[10] is a 1 million word corpus built in Brazil, which was manually annotated for POS, it is composed of 16th to 19th century EP texts.

The requirement for complete texts to be made publicly available aims at meeting the needs of non-academic users or textual linguists and discourse analysts for whom partial texts are not enough. Again, this distinguishes LW from the BNC, which does not make full texts available. This is the main purpose of LW´s general usage corpus, the Lácio-Ref.

Expert users, including computational linguistic researchers, have been taken into consideration within the LW project in the building of a) Mac-Morpho, a gold standard portion from Lácio-Ref, comprising 1,2 million words, which was manually-validated for morpho-syntactical tags and of b) an automatically-annotated portion of the Lácio-Ref with lemmas, POS and syntactic tags used by the parser Curupira developed at NILC. For those researchers interested in analyzing spontaneous writing a deviation corpus will be created composed of non-revised texts (Lácio-Dev). Finally, LW also contains multilingual corpora for those researchers interested, for instance, in terminology: parallel and comparable Portuguese-English corpora called, respectively, Par-C and Comp-C. The Lácio-Ref and Mac-Morpho versions which were made available in the first release of LW corpora will be better described in Section 6.

3. Design of LW Text Typology

When designing the text typology for the project we had the main corpus, Lácio-Ref, in mind. All the other corpora, except for Mac-Morpho, will also receive a header with bibliographic and cataloging data for text typology. The Mac-Morpho texts do not contain headers as it is a part of Lácio-Ref and its purpose is to train POS taggers. However, as it is composed of newspaper articles from Folha de São Paulo[11], its texts are named according to the section of the newspaper they belong to followed by publication date, thus providing a minimal reference for them.

The design of both the Lácio-Ref corpus and its text typology have been based on corpus linguistics principles (Sinclair and Ball, 1996) and on important corpora projects, e.g. ANC, BNC and Czech National Corpus (CNC)[12]. We have also tried to overcome some flaws in the typologies used in the written parts of the latter two (BNC and CNC) as they would prevent us from aiming at a wider contingent of potential users for the Lácio-Ref corpus. For example, Burnard (2002) points out some drawbacks in handling the diversity of the materials in the BNC "which requires a clearer and better agreed taxonomy of text types than currently exists and better access facilities for subcorpora as users frequently ask for a collection of texts of type X" (Burnard, 2002:68).

The Lácio-Ref corpus classifies its texts into four distinct categories: genre, text type, domain and medium. The definition and structure of **genre** and **text type** categories are detailed below. The other two can be found in the site of the project[13].

The **genre** of a text captures its *communicative intention* and its *discourse character*. That is, it classifies the community in which the text is used and the human

[4] See the poster with Corpora Projects in Brazil at www.nilc.icmc.usp.br/iiiencontro/iiiencontro.htm

[5] lael.pucsp.br/corpora/bp/conc/index.html

[6] www.letras.ufrj.br/varport

[7] www.letras.ufrj.br/nurc-rj/

[8] www.fflch.usp.br/dlcv/nurc/

[9] www.cce.ufsc.br/~varsul

[10] http://www.ime.usp.br/~tycho/corpus

[11] www.folha.uol.com.br/folha

[12] ucnk.ff.cuni.cz/english/index.html

[13] www.nilc.icmc.usp.br/lacioweb/english/index.htm

activities that make it relevant. By convention, we use a super genre, the literary super genre, as a set of genres. The genres and subgenres of LW text typology are: scientific; reference (with encyclopedic, lexicographic, terminological subgenres); Informative (journalistic subgenre); Law (jurisprudence, legal subgenres); Prose (biography, short story, novel subgenres); Poetry and Drama. Furthermore genres can be told apart by the text types usually associated with each of them. For example, articles, theses, and projects are associated with the Scientific genre; encyclopedias, dictionaries, and lexicons are classified as Reference; news reports, and editorials as Informative; legislation and codes as Law, textbooks, cooking recipes as Instructional; and letters, memos, and manuals belong to the Technical Management genre.

The structure of a text defines its **type**, that is, the text components and the way they are assembled together, its lexicon, its syntax, its adequacy to the main theme, etc. In LW, there is an open list of text types which can be enlarged as the need arises. The current version contains 39 entries[14], e.g. Article, Contract, Chronicle, Law, News Reporting, Text-book, Agreement, Manual, Report, Dictionary Entry, Recipe, Letter, Editorial, Memo, Essay.

This 4-type text typology (genre, text type, domain and medium) allows for very powerful searches by enabling the creation of specific subcorpora for different types of research. The related interface is presented in Section 5 and examples of research it allows in Section 6.

4. Text Copyright Clearance, Compilation and Annotation

For texts to be included in one of the corpora they must meet the following requirements: a) have an authorization issued by the authors; b) be in files of the appropriate naming convention and format, and c) have XML-headers with the correct classification of its contents.

Regarding copyright, clearance difficulties in obtaining it has prevented us from gathering a balanced corpus within the duration of the project for, unlike other corpora created by a consortium of publishers like the ANC and BNC, we started with no repository of copyrighted texts of our own. However, it is worth noticing that we have so far been very successful in obtaining authorization[15] from publishers of newspapers, magazines, informative and scientific periodicals, and scientific books; and authors of theses and academic papers, as well as donations of electronic versions literary books in the public domain, which will enable us, to start the design of a balanced corpus for modern Brazilian Portuguese in the near future.

Once copyright has been cleared, texts go through a three phase process: **Compilation-Formatting**: which consists of selecting the texts, storing them in electronic media using text format, and formatting them in accordance with the original text (e.g. if the text contains a table, figure or another graphic object an XML tag will be inserted in its place indicating the absent object); **Coding**: the text file names take in consideration the text's genre, media and other characteristics that allow for the automatic construction of subcorpora; **Identification**: a header is inserted in the file containing annotations describing usual bibliographic information (title, author, date of publication, editor, language, etc.) and cataloging data (size of file, sample type, gender of authors and the categorization of domain, genre, textual type and media type, i.e. the 4-type text typology information).

5. Interface Design

To meet the requirements of various types of users like linguists, computational linguists as well as laymen who may be interested in looking for examples of BP usage, we have designed an interface in which the main page introduces the project and the internal pages (accessed after registration) are each related to a specific LW corpus allowing users to focus on their objective. The registration allows us to keep a record of the visits to the website and of the searches carried out with the Lácio-Ref corpus, which will enable us to improve the project. Besides, the internal pages have help buttons with explanation regarding the choices to be made in order that laymen users can fell comfortable with the environment.

Besides, the internal pages have help buttons with explanations regarding the choices to be made in order for non-expert users to feel comfortable with the environment.

Each LW corpus has its set of pages with its associated tools (see Section 7). Specifically, Lácio-Ref offers three types of searches (simple, advanced and customized) to create a study subcorpus, which can be downloaded in two versions: one with an XML header containing bibliographic and cataloguing data, another with title, subtitle, author and the raw texts. The latter version was meant to be used with the concordancer and frequency counters which were specially designed to treat BP multiword proper names.

The simple search is recommended when the user wishes to obtain a rather generic corpus because the combination of parameters only takes into consideration: medium, super genre (for the literary genre) and text genre. In the advanced search, the search possibilities first take into consideration medium, super genre and genre. But they also take into consideration subgenre as well as other specific data for each textual genre (e.g. for the Informative genre, the name of the periodical and its section; for the Scientific genre, the author's name, and for the Literary super genre, the author's name and the title of the book). The customized search explores header fields not taken into consideration by the simple and the advanced searches, being useful for sophisticated searches such as type of sample, type of text, domain and sub domain, type of authorship, author's sex and all bibliographic references.

6. Support for Several Types of Investigations

The first release of the LW project took place on January 20, 2004 making two corpora available: a version of the Lácio-Ref for research and construction of subcorpora, and the downloadable MAC-MORPHO.

The Lácio-Ref version of the first release contains **4,156,816 words** The corpus is constituted of five genres of texts (informative, scientific, prose, poetry and drama), various types of texts (news reports, articles, stories, letters, etc.), several domains (Education, Engineering,

[14] www.nilc.icmc.usp.br/lacioweb/english/typology.htm

[15] See one example of authorization letter in the Project site.

Politics, among others) and some media (magazine, Internet, book, etc.). Possible studies with this corpus are: language description at the lexical, syntactic, semantic and discursive levels; lexicographic and terminological studies; automatic text categorization, etc.

Upon registration the user is granted an area to store the results of his/her research using the Lácio-Ref. In other words, to store the subcorpus built according to his/her research criteria. For example: "all texts about Fashion", "all editorials from the X magazine or Y newspaper", "all short stories by author Z" or "all dissertations on Zootechnology"[16].

Mac-Morpho contains **1,167,183 words** of journalistic texts extracted from ten sections of the daily newspaper Folha de São Paulo, 1994. These texts have been tagged by Eckhard Bick´s parser "Palavras"[17], mapped onto Lácio-Web's tagset and manually revised for POS-tags. MAC-MORPHO is available for download in two versions: one adequate for linguistic research with frequency counters and concordancers[18]; another for training taggers. This corpus allows for studies involving development and evaluation of POS-taggers; language description, etc.

7. Corpora and their Associated Tools

The LW Project was designed to provide simple tools for linguistic research like concordancers and frequency counters and more elaborate ones like POS taggers, sentence and word aligners, and ATE tools, all of which are derived from research projects at NILC. These tools are didactically associated with each corpus as one of the uses we have foreseen for the project is that its corpora can be queried by students in Corpus Linguistic (CL) courses. This is an innovative use, to the best of our knowledge, and can contribute to make CL methods better known in Brazil. The tools and the associated corpora are: **Lácio-Ref´s subcorpora**: frequency counters (which include a preliminary treatment for multiword proper names), concordancer to find the immediate context to the left and to the right of the search word, and POS-taggers; **Mac-Morpho**: concordancer and POS-taggers (Brill´s TBL, MXPOST and Treetagger) and a combination of them to increase tagger precision; **Lácio-Dev**: frequency counters, concordancer, lemmatizers and POS-taggers; **Comp-C**: automatic term extraction methods[19]; **Par-C**: sentence[20] and Word[21] aligners; **Automatically-annotated portion of the Lácio-Ref with lemmas, POS and syntactic tags**: concordancer.

8. Conclusions and Further Work

Like all projects LW has also its limitations. We did not form a consortium like the BNC and the ANC did at the early stages of the project because we preferred to develop it as an academic initiative to learn basic lessons. It may develop into a consortium in the future as there is plenty of work to pursue in such an ambitious corpus project. The Lácio-Web has been funded for 30 months only (at the moment). Therefore, there are no immediate goals regarding the number of words to constitute the Lácio-Ref, as we depend upon authorization for inclusion of the texts and a have a small budget to pay researchers who annotate the corpus. It is one of our future projects to balance the Lácio-Ref corpus but this requires a systematic linguistic study. The Mac-Morpho corpus, however, has been completed.

At the site of the project, the sources for which we are already obtained permission are listed.

Acknowledgements

We are grateful for the grant of the governmental agency Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq) which made this project possible.

References

Aires, R. V. X., Aluísio, S. M., Kuhn, D. C. S., Andreeta, M. L. B., Oliveira Jr., O. N. (2000). Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. In Proceedings of SBIA (pp. 20--22). Atibaia/SP, Brazil.

Aluísio, S.M., Pinheiro, G., Finger, M., Nunes, M.G.V., Tagnin, S.E.O. (2003a) The Lácio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In Proceedings of Corpus Linguistics (pp. 14--21). Lancaster, UK.

Aluísio, S. M.; Pelizzoni, J. M.; Marchi, A. R.; Oliveira, L. H.; Manenti, R.; Marquiafável, V. (2003b). An account of the challenge of tagging a reference corpus of Brazilian Portuguese. In Proceedings of PROPOR (pp. 110--117). Faro, Portugal.

Borba, F. S. (2002) Dicionário de usos do Português do Brasil. São Paulo, SP: Editora Ática.

Burnard, L. (2002). Where did we Go Wrong? A Retrospective Look at the British National Corpus. In Ketterman, Bernhard, & G. Marko (Eds.) Teaching and Learning by doing Corpus Analysis. In Proceedings of the Fourth International TALC (pp. 51-70). Amsterdam, NY: Rodope.

Sinclair, J. and Ball, J. (1996) Preliminary Recommendations on Text Typology. EAG-TCWG-TTYP/P.

Wittmann, L. Pego,T. & Santos, D. (1995) Português do Brasil e de Portugal: alguns contrastes. In Actas do XI Encontro da Associação Portuguesa de Lingüística (pp 465-487). Lisboa, Portugal.

[16] The publishers of Folha de São Paulo restricted the copyright for its texts - users are not allowed to create a subcorpus containing all the 1994 issues.

[17] visl.hum.sdu.dk

[18] There are two concordancers available on the project’s site.

[19] http://www.nilc.icmc.usp.br/nilc/projects/termextract.htm

[20] http://www.nilc.icmc.usp.br/nilc/projects/pesa.htm

[21] http://www.nilc.icmc.usp.br/nilc/projects/PEWA.htm