The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation

Sandra M. Aluísio[*#], Gisele M. Pinheiro[#], Marcelo Finger[†],
Maria das Graças V. Nunes[*#], Stella E. O. Tagnin[‡]

[*]*ICMC – DCCE, University of São Paulo, CP 668, 13560-970 São Carlos, SP, Brazil*
[#]*Núcleo Interinstitucional de Lingüística Computacional (NILC), ICMC-USP, CP 668, 13560-970 São Carlos, SP, Brazil*
[†]IME – DCC, *University of São Paulo,* Rua do Matão, 05508-090 *São Paulo – 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*, *mfinge@ime.usp.br*, *gracan@icmc.usp.br*, *seotagni@usp.br*

1. Introduction

The pioneering balanced Brown Corpus launched in 1964, annotated reference corpora, such as Suzanne and the Penn Treebank and the balanced mega British National Corpus (BNC)[1], to cite only a few, have helped both the development of English computational linguistic tools and English corpus linguistics. Portuguese, on the other hand, still requires a lot of work for building the basic resources to develop linguistic research based on and driven by corpus. Portuguese is the mother tongue of approximately 200 million people (Brazil, Portugal, Angola, Cabo Verde, Guiné Bissau, Mozambique and S. Tomé e Príncipe) and is the sixth most spoken language in the world today. The two language variants with the greatest number of users – European Portuguese (EP) and Brazilian Portuguese (BP) – differ in phonological, lexical, morphological and syntactical levels (Wittmann et al. 95) suggesting a real need to build corpora considering both of them.

For European Portuguese (EP), the project AC/DC[2] has compiled several corpora of non-literary (e.g. journalism) and literary texts (poetry, prose, plays) mainly from 16th to 19th centuries. These corpora are available both in raw format and annotated with lemma, POS and associated attributes, and syntactic tags, being the main purpose of this compilation to raise the quality of Portuguese NLP. Another example for EP is the Corpus de Referência do Português Contemporâneo (CRPC)[3], which has been under construction since 1988, and contains excerpts of several types of written discourse (literary, journalism, technical, scientific, didactic, economy, legal, parliamentary, etc.) and oral discourse from the main variants of Portuguese. Its main goal is to establish an on-line representative sample collection of general usage contemporary Portuguese accessible to anyone interested in engaging in theoretical and practical studies or applications. Both projects include texts from the Brazilian Portuguese and are valuable resources although there is a lot of work to be done with regard to text classification based on genre and text typology to better balance the variants included. On the other hand, Brazilian Portuguese (BP) corpora to date have mainly addressed spoken language; some written language corpora are only partially available. These corpora were used by specific projects, particularly for the production of dictionaries, and some are not publicly available due to copyright restrictions. As for their application, they have generally been used for specific linguistic studies such as sociolinguistic and phonetic-phonological research as well as historical linguistic studies (e.g NURC-RJ[4] and NURC-SP[5], PHPB[6] and VARSUL[7] projects); lexicography (e.g. the written corpus "Usos do Português"

1 info.ox.ac.uk/bnc/
2 www.linguateca.pt/
3 www.clul.ul.pt/sectores/projecto_crpc.html
4 www.letras.ufrj.br/nurc-rj/
5 www.fflch.usp.br/dlcv/nurc/
6 www.letras.ufrj.br/phpb-rj/
7 www.cce.ufsc.br/~varsul

from the State University of São Paulo, at Araraquara, which gave birth to three dictionaries: a dictionary of frequency (Biderman, 2001), a dictionary of verbs (Borba, 2001) and one of contemporary Brazilian Portuguese usage (Borba, 2002); and a grammar of Brazilian Portuguese usage (Neves 2000)) and literary studies (e.g. the NUPILL project[8], which made available classical Brazilian literary books for teaching and literary studies). There are also small specialized written language corpora compiled by researchers or research groups in order to assess the performance of NLP systems. Such corpora are generally not available either and their results are not reproducible in principle.

The Lacio-Web project, a two-year project launched in early 2002, tries to fill this gap as it aims at compiling corpora which are freely accessible for both non-expert users interested in the Brazilian Portuguese language and expert users who pursue theoretical and practical linguistic studies and develop computational linguistics tools (e.g. taggers, parsers, sentence and word aligners, automatic term extraction tools, and automatic summarizers) and applications such as computer systems for natural language information retrieval, machine translation and grammar checking. Lacio-Web (LW) is being developed at the University of São Paulo (USP) under the auspices of the governmental agency Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (CNPq).

The LW project comprises six corpora: 1) a reference corpus called *Lacio-Ref*; 2) *Mac-Morpho*[9], a gold standard portion from Lacio-Ref, comprising 1,1 million words, which was manually-validated for morpho-syntactical tags; 3) an *automatically-annotated portion of the Lacio-Ref* with lemmas, POS and syntactic tags which are used by the parser Curupira developed at NILC[10]; 4) a deviation corpus composed of non-revised texts (*Lacio-Dev*); 5) and parallel and 6) comparable Portuguese-English corpora called, respectively, *Par-C* and *Comp_C*. The corpora will be available on the WWW for download. We will also develop a web-based interface for access to the corpora to meet several users´ needs. For this purpose we will consider the Project of Korpus 2000 (Andersen et al., 2002) as its corpus interface was designed with non-expert user needs in mind.

It is worth mentioning that the design of the Lacio-Ref corpus and its text typology have been based on corpus linguistics principles (Sinclair and Ball, 1996) and on important corpora projects, e.g. American National Corpus (ANC) (Filmore et al. 1998; Ide & Macleod, 2001; Ide et al. 2002), BNC and Czech National Corpus (CNC)[11]. We have also tried to overcome some flaws in the typologies used in the written parts of the latter two (BNC and CNC) which would prevent us from broadening the set of potential users of the LW corpora. The construction of LW corpora builds upon the previous experience at NILC in the ad hoc compilation of a 35 million-token corpus named Corpus Nilc (CN). This paper details the corpora being created (Section 3), presents the rationale for LW corpora (Section 4) as, in Brazil, there is an urgent need for corpora (both annotated and raw) constructed according to corpus linguistics principles. It also compares and contrasts the development of both corpora (LW and CN) emphasizing the lessons learned in the process (Section 2).

2. Lessons learned from developing and critically analyzing CN

The CN corpus was built to support the development of a grammar checker for Brazilian Portuguese named ReGra (Martins et al., 1998). Specifically, CN was designed to inform linguistic studies for the development of ReGra and to provide data for performance testing. The construction of ReGra started in 1993 and it has been improved since then under the auspices of both a Brazilian governmental agency and a Brazilian software company. CN is an opportunist corpus built in an ad hoc manner: its text selection was carried out on demand and its text classification was based on a very particular typology suitable for testing the grammar checker.

At the beginning of the LW Project, we counted on being able to feed a major part of the CN into the Lacio-Web, thus, reducing the cost of its construction. However, after a detailed analysis of the CN subcorpora we found major problems regarding: classification, the number of texts in certain subcorpora, sample size (the main criterion in CN was full texts), grouping and formatting, documentation and copyright. These items will be detailed below together with an indication of the cost involved to correctly include the CN

8 www.cce.ufsc.br/~nupill

9 Details about its tagset, the annotation process, including the results of the inter-annotator agreement evaluation and linguistic problems faced in developing Mac-Morpho can be found at www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

10 In a follow up project, we intend to do a manual revision of it. The resulting Treebank will serve, for example, to improve the parser itself and to train statistical parsers. Details about Curupira can be found at www.nilc.icmc.usp.br/nilc/tools/curupira.html

11 ucnk.ff.cuni.cz/english/index.html

texts in LW.

2.1 Text classification

As CN was built on demand, its text classification became problematic. The texts were divided into three classes, driven by the purpose of the corpus, namely: a) corrected texts, b) uncorrected texts and c) semi-corrected texts, i.e. texts published in books or journals e.g., unrevised texts, and text revised by advisors e.g., respectively. Inside these classes, the texts were grouped in an ad hoc fashion, either by domain (or subject), or genre or textual type. This irregular text classification prevents a user from recovering all texts on "sports" because only some of them were classified according to domain. One reason for this is that the texts in CN were not prepared to be automatically retrieved by a user. On the other hand, as we intend to design powerful and customized search tools for LW´s users, the first task after defining LW corpus typology was to define its text typology which has a 4-orthogonal category typology to classify texts (genre, textual type, medium and subject) and information about authorship and the publication of texts. Another benefit a sound text typology (driven by linguistic or commonly accepted external criteria) brings along is the possibility to better control the insertion of material into a corpus in order to achieve better representativeness in corpus design.

2.2 Number of texts

Some CN subcorpora are under-represented, i.e. they consist of a small number of texts. For example, the "Technical and Scientific" subcorpus has only a few samples of theses and incomplete dissertations, most of them from the Computer Science domain. We can say that CN benefits terminology studies in Computer Science but it is not representative regarding other technical and scientific domains. On the other hand, the design of the LW corpora is based on a 4-orthogonal category typology to organize the texts and we will endeavour to complete each category in each corpus (see Section 4 for details on this procedure). This set-up will enable the development of several types of tools and linguistic researches. Automatic text categorization, for example, is a type of research which would benefit both the advance of the research area itself and the organization of the texts in a typology. Moreover, in the near future we will be able to perform linguistic analysis to evaluate the representativeness of our newborn corpora following Biber´s recommendation to proceed in a cyclical fashion of analysis, design, compilation and analysis again (Biber, 1993).

2.3 Sample size

Some text samples deviated from the standard followed by CN which was to include only full texts. For example, some only have a few chapters of a book, other have only excerpts from the beginning, middle and end part of a whole document. In the LW Project, we are following the criterion of inclusion of sequential[12] full texts to allow for research on text structure and summarization, for example. However, if copyright issues demand otherwise we will have this fact annotated in the text header. This may happen with textbooks, for example.

2.4 Grouping and formatting

Issues regarding ad hoc grouping such as, to group several small texts from a same class in just one text causes several problems for the compilation of the header. It would be difficult to import these groups of texts to LW since some information on them has been lost and we intend to edit a header for each and every text from the corpora. This grouping has occurred in encyclopedia entries and small articles from newspapers. Also, CN has not keep text formatting e.g. sentence and paragraph marks for many texts. As we consider this annotation important it would be costly to recover this information.

2.5 Documentation and copyright

12 The newspaper corpora in AC/DC Projects had to have their texts scrambled, which prevents several types of research of being conducted.

Annotating a text with a header which provides internal and external information on the texts was not taken seriously in the compilation of CN. Some texts have the traditional information on authorship and publication details but nothing is said about its domain (subject) or genre and text type; others do not have any header at all. The LW will try to correct this flaw and will invest large amounts of effort to edit and encode a header following largely accepted standards (see Section 4). With regard to copyright, there was no effort to secure rights to include material in the CN as it was designed to support the development of an application. LW, however, has other design criteria as it will be freely available on the WWW. One of the most difficult problems in building a corpus is surely to get permission to include texts from copyright holders as there is a lot of correspondence involved before we manage to get copyright clearance. We are using similar contact and permission letters provided by BNC and ANC to include texts in LW corpora. Although we do not have a consortium of commercial members to feed our corpora we have managed to include relevant newspaper, magazines and books in our corpora.

In summary, CN was of great value to the development of ReGra and with regard to LW, CN is ready to contribute with the legislation (110.571 tokens), journalism (25,167.436 tokens) and literary texts (1,761.373 tokens).

3. The corpora of the Lacio-WEB Project

A common concern in recent studies is to provide precise guidelines concerning the typology of a corpus. Atkins, Clear & Ostler (1992) discussed contrast-based parameters that allow identification of different types of corpora: *synchronic vs. diachronic corpus*; *closed vs. open-ended corpus*; *full text vs. sample vs. monitor corpus*; *general vs. terminological*; single vs. *parallel-2 vs. parallel-3 vs. …,* etc. Some criteria for classifying corpora have also been provided by Cathy Ball, in a tutorial about concordances and corpora[13], on top of the parameters mentioned above: *balanced vs. opportunist vs. pyramidal corpus*, and *plain vs. annotated corpus*. A similar work on such parameters was carried out by Berber Sardinha (2000), who distinguished corpora according to their purpose: *reference vs. study vs. training and testing.*

The corpora comprised by this project have a varied composition and can be classified according to the 8 parameters cited above. Regarding purpose, the *Lacio-Ref* corpus is a reference corpus; the *Lacio-Dev* – a deviation corpus composed of unrevised texts – was meant for training and testing applications, such as grammar checkers; and the *Mac-Morpho* corpus is closed and serves as a training and testing corpus for NLP tools, such as POS taggers. As far as design is concerned, the *Mac-Morpho* corpus and the *automatically-annotated portion of Lacio-Ref* are of annotated type, while the *Lacio-Ref* is a plain corpus. For the parameter single vs. parallel, the *Par-C* is of the parallel-2 type; and for the general vs. terminological parameter, *the Comp_C* is of the terminological type.

All the corpora are synchronic, presenting BP language from 1900 onwards. Lacio-Dev, Par-C and Comp_C are opportunistic corpora. The specificities of each of these corpora are discussed below, together with their status in terms of completion and encoding. We distinguish two types of encoding: i) header and sentence markup and ii) annotation for linguistic phenomena such as lemmatization, POS and syntactic tags. Our header editor includes an XML compliant header, according to the international standard adopted in the Translational English Corpus (TEC)[14] from University of Manchester's Institute of Science and Technology (UMIST). A tool named SPLITTER, developed at NILC, will be used to split the texts into sentences.

1) Lacio-Ref Corpus: contains texts from various *genres* (e.g., literary and its subdivisions, factual, informative, scientific, law), *textual types* (e.g. article, manual, research project, letter, biography), *subjects* (e.g., politics, environment, life style, sports, arts, religion etc.); and *medium of distribution* (e.g., books, internet texts, cd-rom material, newspapers and magazines, etc.). The four-category typology cited above (genre, textual type, medium and subject) will be used to allow for specific searches on the corpus. Also, the texts may be searched by authorship and other publication details.

All pieces of text are authentic with identified sources (see Section 4) and are, preferentially, full texts. Tools will also be available to users for obtaining a statistical description of the texts, in terms of the text size (in number of words, pages or kbytes). *Status*: the corpus is being compiled (it already contains material contributed from the CN); the edition of the headers has not been started.

13 www.georgetown.edu/faculty/ballc/corpora/tutorial.html

14 www2.umist.ac.uk/ctis/research/TEC/tec_home_page.htm

2) Lacio-Dev Corpus: most texts (in a total of 516.840 tokens) will be imported from the CN which contains a subcorpus comprised of unrevised texts of varied subjects produced by undergraduate students and prospective students attempting to enter the University. This Corpus can be used to assess the performance of grammar checkers for BP, as there is a need to ensure that the checkers are able to detect linguistic inadequacies during the tests. *Status*: edition of the headers has not been started.

3) Mac-Morpho Corpus: this 1.1 million-word Corpus is composed of a collection of randomly selected texts from several issues of Folha de São Paulo (1994)[15], a major Brazilian newspaper, which ensures high quality contemporary Brazilian Portuguese from different authors and domain. The manual validation and correction process was carried out on the morphosyntactic tagging of the texts performed by the parser PALAVRAS[16]. The corpus contains structural markers for sentences. Besides being annotated by the XML-compliant format proposed by the Advisory Group on Languages Engineering Standards EAGLES (see Section 4) it will be available in annotators' format (one word per line followed by its tag) which is appropriate for training and evaluating POS tagging methods. *Status*: the compilation and manual annotation have been completed; both the header edition and the corpus encoding have not been started.

4) a portion of Lacio-Ref automatically-annotated with lemmas, POS and syntactical tags: as opposed to Mac-Morpho, this corpus will consist of a varied selection of text genres. It will be annotated by the parser Curupira and in the near future we hope to carry out a manual revision of it. The resulting Treebank will have several uses: to improve the parser itself, to train statistical parsers, to perform more accurate searches, etc. *Status*: compilation and header edition have not been started.

5) Par-C Corpus: because of its opportunistic behavior, this corpus will be enlarged from time to time. In this initial phase of the LW project Par-C is composed of 65 pairs of authentic academic parallel texts (abstracts) in Computer Science contributed from Project PESA[17] which aims at evaluating sentence alignment methods. The corpus was divided into two groups: one comprising 65 pairs of authentic (non-revised) texts; another with the same 65 pairs revised by a human translator (pre-edited corpus). They were named CAT and CPT, respectively. CAT has 416 BP sentences and 439 English sentences. CPT has 418 BP sentences and 431 English sentences. *Status*: compilation and alignment completed; edition of headers has not been started.

6) Comp-C Corpus: it is also an opportunistic corpus which can be used to evaluate term extraction methods as well as for other linguistic researches. In this initial phase it comprises English-Portuguese comparable texts contributed by the Project COMET (Tagnin 2001, 2002). These texts are technical, scientific and marketing-related, amounting to 300,000 words in each language. The corpus was compiled by students in the Diploma in Translation course at FFCLH/USP in order to build glossaries[18]. *Status*: compilation completed; edition of the headers to be started.

4. Issues in corpus creation tackled in the Lacio-Web Project

4.1 Accessibility of the corpora

As mentioned before, one of our main goals in constructing the LacioWeb corpora is to create a benchmark for Computational Linguistics in Brazilian Portuguese for tasks such as POS tagging, parsing, text alignment, and term extraction. For that purpose, texts must be provided with several sets of hand-validated linguistic annotation e.g. lemmas, POS, and syntactic tags. Additionally, in order to maximize its use, a corpus should be encoded according to largely agreed standards and must be available to each and every researcher in general. Therefore, the texts of the LacioWeb corpora project will be made freely available on the Internet, followings the initiative of several recently created corpora such as: the Tycho Brahe Corpus of Historical Portuguese[19], which also provides POS- and parsing-annotated texts and the COMPARA parallel corpus for Brazilian and European Portuguese texts[20]. With regard to providing standardized encoding for the resources we are

15 http://www1.folha.uol.com.br/fsp/
16 http://visl.hum.sdu.dk/visl/
17 www.nilc.icmc.usp.br/nilc/projects/pesa.htm
18 http://www.fflch.usp.br/citrat
19 www.ime.usp.br/~tycho
20 www.linguateca.pt/COMPARA/

following the design of the ANC Corpus[21] which will use the specifications of an XML-compliant format of the Corpus Encoding Standard XCES (Ide et al., 2000). In our case, the texts will be available for download from the project's home page[22]. It will be redesigned to allow powerful searches and friendly access to non-expert users who may use LW corpora as an extension of dictionary consultation to obtain authentic examples of real word usage in BP.

4.2 Exhaustivity versus selectivity

One of the basic tenets of Science is the reproducibility of results. In the field of Computational Linguistics, when an algorithm is tested against a non-public corpus the result is not reproducible in principle. It may well be the case that other researchers, applying the same algorithms to other corpora, may find distinct results. Without applying the test to a common, public corpus, these discrepancies may never be reconciled. From this point of view, any method, algorithm or measurement performed over a non-public corpus cannot be considered definitively tested. Moreover, no comparison of two distinct computational linguistic methods which perform the same task can be made unless they are tested against the same set of data. The reason is clear.

One method may be better suited for one kind of text, while the other may be better suited for some other kind of data. Two researchers may test both methods on their private data and arrive at different conclusions. Again, the only way to solve this dispute is to apply both methods to the same piece of public data. And if different pieces of data yield different results, this is an important piece of information that one will not have unless the data is public. This fact has guided us in the construction of the Lacio-Ref, in particular in the guidelines for text gathering and classification[23]. This leads us to the next topic: the exhaustivity of the texts to be included, mainly, in Lacio-Ref. By exhaustivity we mean the broad and varied coverage of the most common categories used to classify texts such as: authorship, text size, and the four-category typology, namely genre (literary − subcategorized into poetry, drama, prose − and non-literary − subcategorized into factual, informative, academic and legislation), textual type (e.g. article, manual, research project, letter, biography), medium (e.g. newspapers, books, ads, journals, magazines) and subject (or domain). The genre typology instances can be further subcategorized. This is a differential regarding reference corpus design as they are generally built in an ad hoc manner. Linguistic aspects of texts are usually only provided for corpora which are designed to be balanced, but in the hope of broadening the set of potential users of the Lacio-Ref such information will be provided for its texts.

Each text classification according to these categories and others related with text publication will be documented in the text's header via XML-annotations in a format inspired by the TEC corpus header annotation. Choosing to be exhaustive instead of selective is another way to broaden the set of potential users of the corpus and the possibilities of linguistic analysis. We are not limited to including only a specific genre, e.g. informative with newspaper data, as followed by the CRPC project, cited in the Introduction, with regard to the Brazilian Portuguese variant. We also follow a much more detailed genre typology than those used in the BNC written corpus and the CNC SYN2000. Additionally we use a 4-orthogonal category typology (genre, type of text, medium and domain) as commented before while these other corpora focus mainly on domain. For example, BNC classifies its texts by medium, domain and time, reducing its genre typology to two classes: informative writings which are subcategorized by domain and imaginative writings which are subcategorized by literary and creative works. SYN2000 follows the same genre typology as the BNC but subdivides imaginative texts into poetry, drama, fiction, other and transitional types and the informative into journalism and technical and specialized texts. The latter is again subcategorized by domain as in the BNC. This does not mean that we are unaware of the difficulties to classify the texts following a more detailed taxonomy but this enterprise is worth trying as it can provide a better search tool for the user.

4.3 Representativeness and balance

Availability and exhaustivity were two criteria that superseded representativeness and text balancing in the selection of texts for the Lacio-Ref corpus. Problems of copyright would prevent us from obtaining a balanced corpus within the two-year duration of the project for, unlike other corpora created by a consortium of publishers like the ANC, we started with no repository of texts of our own[24]. However, we have so far been

21 http://americannationalcorpus.org/

22 www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

23 We tried to adhere as much as possible to the EAGLES recommendations on text typology.

24 It is important to note that although we have the CN its texts don´t have copyright clearance, therefore they are not ready

very successful in obtaining authorization from newspaper and magazine publishers as well as donations of electronic versions of public domain books, which will enable us, in the near future, to start the design of a balanced corpus of modern Brazilian Portuguese which will be included in the set of corpora of the Lacio-Web project. After the initial text gathering and classification we will be in a position to follow Biber's classical recommendation (1993, abstract), "The actual construction of a corpus would then proceed in cycles: the original design based on theoretical and pilot-study analyses, followed by collection of texts, followed by further empirical investigations of linguistic variation and revision of the design."

5. Conclusion and future work

In Brazil, there is an urgent need for corpora constructed according to corpus linguistics principles regarding its text typology, encoded according to largely accepted standards, and freely available. In this paper we presented the Lacio-Web project which aims at compiling freely accessible corpora for both non-expert users interested in the Brazilian Portuguese language and expert users who pursue theoretical and practical linguistic studies and develop computational linguistics tools. Its purpose is not only to raise the quality of Portuguese NLP but also popularize the use of corpora for layman who may be interested in using them as an extension of dictionary consultation to obtain authentic examples of real word usage in BP. After the first year of the project we have already corrected the morphosyntactic annotation of the 1,1 million-word corpus MAC-Morpho, gathered several texts for the Lacio-Ref, critically analyzed the CN corpus which can also provide material for the Lacio-Ref. We are almost ready to release the first version of the bilingual corpus Par-C and the Comp-C. It is worth mentioning that the header must be included in the entire corpora before the release. In a new project, we intend to pursue the balancing of Lacio-Ref to make it more useful for research.

References

Andersen, M.S., Asmussen, H. & Asmussen, J. 2002 The project of Korpus 2000 going public. In Proceedings of Euralex 2002, pp 291-299.

Atkins, S., Clear, J. & Ostler, N. 1992 Corpus design criteria. Literary and Linguistic Computing 7: 1-16.

Biber, D. 1993 Representativeness in corpus design. Literary and Linguistic Computing 8: 1-15.

Biderman, M.T.C. 2001 Dicionário de freqüências do português brasileiro contemporâneo. In: Martins Fontes (ed), Teoria Lingüística. São Paulo, pp 335-348.

Borba, F. S. 1991 Dicionário Gramatical de Verbos. São Paulo, UNESP.

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

Fillmore, C., Ide, N., Jurafsky, D., and Macleod, C. 1998 An American National Corpus: A Proposal. In Proceedings of the First International Language Resources and Evaluation Conference, Granada, Spain, pp 965-70.

Ide, N., Reppen, R., Suderman, K. 2002 The American National Corpus: More Than the Web Can Provide. In Proceedings of the Third Language Resources and Evaluation Conference (LREC), Las Palmas, Canary Islands, Spain, pp 839-844.

Ide, N., Macleod, C. 2001 The American National Corpus: A Standardized Resource of American English. In Proceedings of Corpus Linguistics 2001, Lancaster UK. Available in:
www.cs.vassar.edu/faculty/ide/pubs.html

Ide, N., Bonhomme, P., Romary, L. 2000 XCES: An XML-based Standard for Linguistic Corpora. In Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece, pp 825-30.

for inclusion in the Lacio-Ref.

Martins, R.T.; Hasegawa, R.; Nunes, M.G.V.; Montilha, G.; Oliveira Jr., O.N. 1998 Linguistic issues in the development of ReGra: a Grammar Checker for Brazilian Portuguese. Natural Language Engineering 4(4): 287-307.

Neves, M. H. M. 2000 Gramática de Usos do Português. São Paulo, UNESP.

Sardinha, T.B. 2000 Lingüística de Corpus: Histórico e Problemática (Corpus Linguistics: History nd Problematization), DELTA 16(2): 323-367.

Sinclair, J. and Ball, J. 1996 Preliminary Recommendations on Text Typology. EAG-TCWG-TTYP/P, June 1996. Available in: www.ilc.pi.cnr.it/EAGLES/texttyp/texttyp.html

Tagnin S.E.O. 2002 *Corpora and the Innocent Translator: how can they help him.* In Lewandowska-Tomaszczyk, Barbara & Marcel Thelen (eds.) Translation and Meaning - Part 6 - Proceedings of the Lodz Session of the 3rd International Maastricht-Lodz Duo Colloquium on 'Translation and Meaning", held in Lodz, Poland, 22-24 September, 2000, Maastricht: Universitaire Pers Maastricht, pp 489-496

Tagnin S.E.O. 2001 COMET – A Multilingual Corpus for Teaching and Translation. In: PALC '01 – International Conference on Practical Applications in Language Corpora, Lodz, Polônia, To appear in the Proceedings.

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, Lisboa, Portugal, pp 465-487.