the ESPecialist, vol. 28, nº 2 (211-225) 2007 ISSN 0102-7077

# HISTORY AND COMPILATION OF A LARGE REGISTER-DIVERSIFIED CORPUS OF PORTUGUESE AT CEPRIL[1]
# História e Compilação de um Corpus Grande e Diversificado de Português no CEPRIL

Tony Berber Sardinha* (Pontifical Catholic University of São Paulo (PUC-SP), São Paulo, Brasil)

## Abstract

*In this paper I describe the Bank of Portuguese, a large register-diversified corpus of Brazilian Portuguese, which is held at CEPRIL (Center for Language Research, Information and Resources) at Pontifícia Universidade Católica de São Paulo (Pontifical Catholic University of São Paulo, Brazil). The aim is to provide details of its nature, history, current state, as well as of issues related to its planning, development and future prospects. With nearly 230 million words, it is currently one of the largest corpora of Portuguese. The corpus started off as a collection of texts in hard copy and then turned into an electronic collection built around smaller corpora that were collected by individual researchers. Later on, other large subcorpora were added, such as a newspaper collection. There are problems with the corpus, such as register imbalance (the newspaper section is much larger than the others), lack of access to its full contents outside of the university, and the need for updating its contents.*

**Key-words**: *corpora; DIRECT; CEPRIL; Corpus Linguistics.*

## Resumo

*Neste trabalho, apresento o Banco de Português, um corpus grande e variado de português brasileiro, que é armazenado no CEPRIL (Centro de Pesquisa, Recursos e Informação em Linguagem) da Pontifícia*

[1] An earlier version of this paper appeared as a DIRECT Working Paper.

* I want to thank the two anonymous reviewers for their comments. I also want to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brasília, Brazil) for the financial support under grants #350455/2003-1, #307307/2006-9 and #400574/2007-1.

*Universidade Católica de São Paulo. O objetivo é descrever em detalhe sua natureza, história, estado atual, bem como discutir questões relacionadas ao planejamento e desenvolvimento futuro do corpus. Com cerca de 230 milhões de palavras, é atualmente um dos maiores corpora de português do mundo. O corpus era inicialmente uma coleção de textos em papel e mais tarde tornou-se um corpus eletrônico, à medida que corpora eletrônicos foram sendo disponibilizados por pesquisadores. Subseqüentemente, coletâneas maiores foram adicionadas, como a de um jornal diário. Há problemas com o corpus, como a falta de equilíbrio entre os subcorpora, a impossibilidade de acesso ao corpus completo fora da universidade e a presença de conteúdo desatualizado.*

**Palavras-chave**: *corpora; DIRECT; CEPRIL; Lingüística de Corpus.*

## 1. Introduction

In this paper I describe the Bank of Portuguese, a large register-diversified corpus of Brazilian Portuguese, which is held at CEPRIL[2] (Center for Language Research, Information and Resources) at Pontifícia Universidade Católica de São Paulo (Pontifical Catholic University of São Paulo, Brazil). The aim is to provide details of its nature, history, current state, as well as of issues related to its planning, development, and future prospects. I have been involved in designing, collecting and keeping the corpus since the beginning; currently I am deputy coordinator of the DIRECT Research Team[3] in charge of maintaining the corpus. The Bank is only one of the many features of the DIRECT Research Team, whose theoretical background includes Systemic Functional

[2] For more on the history of CEPRIL and the Project that originated it (The Projeto Nacional Ensino de Inglês Instrumental em Universidades Brasileiras), visit http://www.pucsp.br/pos/lael/cepril/cepril-info.php and http://www.pucsp.br/pos/lael/cepril/workingpapers/. The original meaning of the acronym was Centro de Pesquisas, Recursos e Informação em Leitura.

[3] http://www2.lael.pucsp.br/direct and http://plsql1.cnpq.br/dwdiretorio/pr_detalhe_prod_tipo_pesq? strPNroIdCNPq=5017174447204095&strPTipoProducao= Demais%20trabalhos&strPTipo=Tipo

Linguistics, Discourse Analysis, Genre Analysis as well as applied corpus analysis. The DIRECT Research Team is headed by Prof. Leila Barbara.

Today there are several corpora of Portuguese available (Pinheiro, Oliveira, Tagnin & Aluísio, 2003), both online and offline. Online corpora include Lácio-Web (http://www.nilc.icmc.usp.br/lacioweb/), the Tycho-Brahe Corpus of Historical Portuguese (http://www.ime.usp.br/~tycho/corpus/) and the Corpus do Português (diachronic, http://www.corpusdoportugues.org/). Examples of offline corpora are the CRPC, Corpus de Referência do Português Contemporâneo (http://www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_crpc.php), and the Corpus de Araraquara (Borba, 2004).

The Bank has a long history, having been founded in the early 1990's. Since its inception, it has been a source of data for corpus-based studies for several projects in Applied Linguistics. With nearly 230 million words, it is currently one of the largest corpora of Brazilian Portuguese in existence (cf. Pinheiro, Oliveira, Tagnin & Aluísio, 2003).

The corpus is named a 'bank' for a number of reasons. Firstly, this brings out its character as a 'store' of (treasured) material, which can be increased or reduced, depending on the circumstances. Secondly, the term 'corpus' is normally reserved for those collections which have been gathered with the purpose of being a source of linguistic analysis. Although the sole aim of the Bank of Portuguese is to be a source of linguistic analysis, historically it has not been used as a single source nor as the main source for analysis. Traditionally, people have used smaller portions of the corpus, normally restricted to one genre, and have carried out different analyses. In addition, when portions of the corpus were not the main focus of the research, the whole corpus was used as a reference wordlist for keyword extraction using WordSmith Tools. Only more recently has the bank been used as a single corpus, mostly in large scale lexical analysis.

Access to the corpus is granted by password via World Wide Web, at the address http://lael.pucsp.br/corpora/bp. Users can view the texts and run concordances on a one- million-word sample but downloading the texts is not permitted. Access is made available to researchers connected to the DIRECT research team, which hosts the corpus (more details below). A sample of the corpus is made available to the public at the site http://lael.pucsp.br/corpora/bp/conc.

The Bank of Portuguese is not a monitor corpus. A monitor corpus needs to be updated frequently (for example every year) in order to represent the current usage in the language. This was feasible up to 2000, when data was supplied to us in great quantity in yearly batches, but these supplies have ceased.

The Bank does not claim to be a representative corpus of the Portuguese language. The corpus was not designed to have balanced proportions of the many registers and genres that occur in the language, nor is it concerned with keeping its contents up to date. Although replacing old contents in the corpus with newer material is in the plans, decisions such as the maximum desirable age for texts still need to be made. At present, two sections have data above 10 years of age (since publication date):

| Section of the corpus | Tokens over 10 years of age | |
|---|---|---|
| Conversation | 2.478.515 | 78% |
| Literature | 1.488.195 | 100% |

**Table 1 – Corpus data over 10 years of age (since publication)**

The Bank of Portuguese has grown in importance over the years, so much so that today it is the most important product of the DIRECT research team. This is different from other major corpora around the world, which have as their main focus, at the outset, the design and compilation of a corpus. There are perhaps two main types of corpora in this respect. The first are corpora which have been designed to be independent of any one particular research project. This is the case of the Brown corpus and the BNC. These corpora have been used by researchers from different institutions with a wide range of interests. The other group is corpora which were developed with the aim of achieving some other purpose rather than corpus itself, that is, the corpus was not the end-product of the whole enterprise. Corpora such as the Bank of English and The Bank of Portuguese fit into this category. It is interesting to note design similarities across corpora of the same kind. Both the BNC and Brown (and FROWN, LOB, FLOB, etc.) are all sampling corpora, which have been designed according to very explicit principles and which are then made available to the research community. On the other hand, corpora such as the Bank of English and the Bank of Portuguese are open, organic, ever growing corpora, none of which is made available to the large public in their entirety outside the hosting project. In view of these similarities, it is perhaps fair to assume that the intended use of the corpus predetermines the design of the corpus to a large degree.

## 2. Composition

The Bank of Portuguese is a large, cumulative, open collection of texts and oral transcriptions. Its current size (2003) is the following:

| | Register | Tokens | Types |
|---|---|---|---|
| Written | Newspaper and magazine | 223,371,280 | 582,372 |
| | Literature | 1,488,195 | 72,481 |
| | Academic | 502,438 | 29,537 |
| | Business | 225,607 | 16,175 |
| Spoken | Conversation, meetings, classes, phone talk, interviews | 3,178,883 | 49,989 |
| | Total | 228,766,402 | 607,392 |
| | Files | 1550 | |

**Table 2 – Composition of the Bank of Portuguese**

The corpus contains texts written and spoken in Brazilian Portuguese, with very few exceptions, mostly in newspapers, which may include texts not written by Brazilians. Newspapers may also carry translations and other kinds of texts whose origin is uncertain. Nevertheless, every effort is made to avoid non-Brazilian varieties of Portuguese. There is no estimate of the number of texts that were not written by Brazilian speakers, as the nationality of authors is in many cases very difficult to determine.

As the table above shows, most of the corpus is made up of newspaper and magazine material, followed by conversation and literary texts. It is also evident that the Bank of Portuguese is not a balanced corpus. The decision to keep the corpus unbalanced has been a conscious one and grew out of several considerations: (a) the corpus started as a collection of small text corpora, gathered as part of individual research projects; (b) producing a balanced corpus is a major undertaking, which requires funds that the hosting research team (DIRECT) has never been able to allocate; (c) until recently, the main aim of the corpus was to serve as a depository of smaller corpora, mainly in the area of business, which were to be analysed on their own, rather than the corpus as a whole; (d) the main use for the whole corpus has typically been that of providing baseline word frequency data for comparison with smaller genre specific corpora (through WordSmith Tools KeyWords). These points will be explored in more detail further below.

Newspaper texts are in greater number because of the relative ease with which they can be acquired. Most of our newspaper materials have been obtained directly from the publisher, free of charge. The same material had been published on CD-ROM and sold at newsstands, but it had been encoded in such a way that it was accessible through the programs shipped with the CD. Since raw ASCII text was needed, the CDs could not be used as source. But even if the files could be converted to ASCII, permission would still need to be asked, and so the decision was made to approach the publishers and obtain the ASCII files directly from them.[4]

The conversation section of the corpus contains several types of interaction, such as dialogs, speeches, classes, meetings, and so on. Most of it came from sociolinguistic projects which intended to map

[4] We are grateful to Carlos Kauffmann, at Folha de São Paulo, for his cooperation.

language use around the country. The literature texts came from web sources. Academic texts were collected partly from web sources and partly directly from publishers.

The business section is a little different from the rest in that it contains texts that were collected by DIRECT team members themselves as part of their individual investigations. These include several kinds of genres, such as letters, annual business reports, meetings, bids, leaflets, emails, and so on.

As Hunston (2002: 31) notes, commenting on COBUILD's Bank of English, having newspaper texts in large amounts 'sacrifices what is desirable to what is feasible'. That is, it would be desirable to balance the amount of newspaper texts with other kinds of text, but most of the time this is not possible. This can be a major problem in closed corpora such as the BNC, but in open corpora such as The Bank of English or the Bank of Portuguese, this does not have to be a serious drawback, since:

> *the hope is that once the corpus is of a substantial size the relevant figures can be checked and efforts made to collect data from under-represented groups, so that balance, where it is possible, is achieved after the corpus is (partially) complete, rather than from the outset*. (Hunston, 2002: 29).

In that way, no data needs to be wasted. Data is too valuable to discard simply because it does not fit into the proportions established by the corpus planners. Furthermore, the very notion of determining these proportions is a matter for debate, as Hunston (2002 :28) warns: 'where the proportions … are unknowable, attempts to be representative tend to rest on little more than guesswork.'

## 3. History

As mentioned before, the history of the Bank of Portuguese goes a long way, back to 1990, at the start of the DIRECT (Development of International Research for English, Commerce and Technology) Research Team. The DIRECT Research Team  was created as a joint effort between the Pontifical Catholic University of São Paulo (Brazil) and the University of Liverpool (United Kingdom). The origins of The Bank of Portuguese can be traced back to the 'Banco de Textos' (Text Bank), a collection of paper documents that was part of the DIRECT research team, and which was held at our Applied Linguistics Postgraduate Program in the early 1990's. The collection comprised mostly business texts, such as annual business reports, letters and brochures.

The immediate concern was with the digitalisation of the Text Bank. This proved harder than first imagined. At first, material was captured into electronic format by typing. Optical character recognition was less than perfect at the time (1990's) and was soon given up. Since DIRECT was a joint binational team, with the University of Liverpool, some of the DIRECT staff and students went on exchange trips to England and thus had access to better equipment. During those trips, researchers tried to scan as much material as possible, in addition to getting more printed material at local businesses.

This coincided with a nationwide effort to turn existing language archives into electronic databases. In 1993, a meeting was held with the aim of establishing links between different research teams around Brazil which had a corpus or some other type of text archive, with a view to joining these and turning them into a large national electronic collection. This resulting collection would then be operated by the Brazilian Linguistics Association (ABRALIN). These plans, however, never took effect. The meeting was successful though in that it brought together several leading researchers in a range of different areas (lexicography, language teaching, sociolinguistics, historical linguistics, literature, etc.). It may also have had an influence on those involved in the sense of providing an initial momentum for the digitalisation of their collections (Castilho, Oliveira e Silva & Lucchesi, 1995).

But progress was slow and by today's standards very little actual data was digitalized. This initial battle to obtain electronic corpora has taught us not to take for granted the vast amounts of data already in electronic format that are made available today.

Two text banks were envisaged at DIRECT, one of English and one of Portuguese. The English Bank was the first one to be set up, because at DIRECT our initial focus was on the use of English as a

Foreign Language in the workplace, mostly in native – non-native communication. Its Portuguese counterpart came shortly after. The English bank initially developed faster than the Portuguese one, mainly due to the input of millions of words of Guardian newspaper texts, supplied by Michael Scott.

Michael Scott launched the initial versions of the programs that would later become WordSmith Tools in the early 1990's, and the DIRECT team followed their development, trying them out and giving feedback. The programs were quickly embraced by DIRECT team members, who learned how to use them and adopted them as their main (and for some, the only) software for corpus analysis and access. This meant that most investigations, up to this day, have centered around tasks which WordSmith Tools can perform, such as concordancing, wordlisting and key word extraction. Conversely, other text operations which the software cannot perform with ease (such as the computation of co-occurrence statistics) or which it cannot perform at all (such as tagging and annotation) have not been really taken on at DIRECT, with few exceptions. The Bank of Portuguese is used as a large reference corpus of Portuguese for keyword analysis still today.

## 4. Design

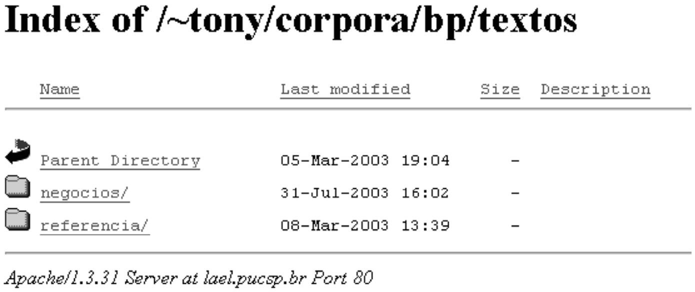
Strictly speaking, the Bank of Portuguese has never been designed as a whole. As mentioned above, the corpus came into being as a result of joining together several different smaller corpora. Each of the smaller corpora was collected following the interests of a particular researcher or team of researchers. This flexible approach to corpus compilation may be termed 'opportunistic', since material is added if and when it becomes available. As such, it contrasts with other corpora, such as the British National Corpus or the Brown Corpus, which have been compiled from a set of previously defined linguistic and social demographic variables. These allowed the planners to decide beforehand which registers and genres would become part of the corpus as well as establishing which weight each component would have (Hunston, 2002; Krishnamurthy, 2002).

The Bank of Portuguese has grown mostly out of the gathering of small corpora (by small is meant less than 50 thousand words). These were made available to us by different individuals, research groups and students at our MA and PhD programs at the Applied Linguistics Postgraduate Program. The role played by students is an important one in helping keep the corpus growing. They are asked to hand over the corpora used in their projects as soon as their research is completed. Not only does this help keep more texts coming into the corpus, but it also establishes a joint sense of ownership, which is essential for the survival of the corpus. This is needed in practical terms since DIRECT, which hosts the corpus, has no funds allocated to cover the expenses of maintaining the corpus itself.

## 5. Access

Access to the Bank of Portuguese is presently made available in two ways. The first is through the Windows network in the Applied Linguistics Postgraduate Program building. This sort of access is restricted to the local machines on the premises of the program.

The other type of access is via WWW at the address http://lael.pucsp.br/corpora/bp. Access is granted to registered users only, and these include DIRECT members, students and staff at the Applied Linguistics Postgraduate Program. Access needs to be restricted as



**Index of /~tony/corpora/bp/textos**

```
   Name                      Last modified       Size  Description

   Parent Directory          05-Mar-2003 19:04     -
   negocios/                 31-Jul-2003 16:02     -
   referencia/               08-Mar-2003 13:39     -
```

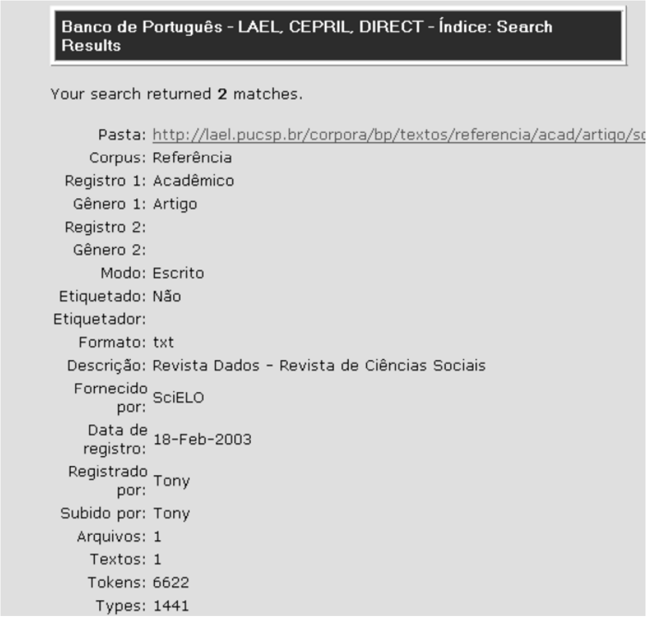
*Apache/1.3.31 Server at lael.pucsp.br Port 80*

**Figure 1 – WWW access to the corpus**

DIRECT does not have copyright permission to distribute the texts. The screenshot below illustrates the initial window on the site.

As the picture shows, users access the corpus by browsing folders. The top division is ‘negocios’ (business) and ‘referencia’ (reference). The business section constitutes what might be called the core corpus, the oldest part of the corpus, which started to be collected in the early 1990’s. The reference section comprises all of the other kinds of text (see table 2). There is no ‘spoken’ and ‘written’ main division in the directory structure. Users can reach spoken and written texts by looking at the labels of each directory.

More detailed information about each directory, such as its source, date, genre and so on is available through the Bank of Portuguese index, a perl database which is accessible on the web. The picture below shows a screenshot of the index displaying the search results for a specific academic text:



Banco de Português - LAEL, CEPRIL, DIRECT - Índice: Search Results

Your search returned **2** matches.

| | |
|---|---|
| Pasta: | http://lael.pucsp.br/corpora/bp/textos/referencia/acad/artigo/sc |
| Corpus: | Referência |
| Registro 1: | Acadêmico |
| Gênero 1: | Artigo |
| Registro 2: | |
| Gênero 2: | |
| Modo: | Escrito |
| Etiquetado: | Não |
| Etiquetador: | |
| Formato: | txt |
| Descrição: | Revista Dados - Revista de Ciências Sociais |
| Fornecido por: | SciELO |
| Data de registro: | 18-Feb-2003 |
| Registrado por: | Tony |
| Subido por: | Tony |
| Arquivos: | 1 |
| Textos: | 1 |
| Tokens: | 6622 |
| Types: | 1441 |

**Figure 2 – Corpus index**

Each record contains information such as the following: the location of the file in the directory structure (with a clickable link), the major corpus division (business or reference), up to three genre labels, mode (written or spoken), whether the file has been tagged, the tagger (if any) which has been used, a description of the file consisting of its source and/or title, date it was entered in the database, the number of texts in the file (as some files contain more than one text, for instance newspapers, in which case an approximate number is supplied), and the number of tokens and types. Each record of the index must be completed manually, one for each file. Because of the large number of files in the bank, the index has not been completed yet.

## 6. Tagging and annotation

Text files in the Bank of Portuguese are in 'raw' format, that is, they have not been edited, and tags have not been inserted to represent sentences, paragraphs or other typographical conventions. Spoken texts have simple identification tags for speakers.

The Bank of Portuguese has not been annotated. Information about each file contents is presented through the perl index (see above) and not in headers.

The Bank of Portuguese has not been fully tagged yet, apart from a small portion of half a million words, which was used as training data for an implementation of the QTAG part of speech tagger (Mason, 1997), and two million words which have been tagged through VISL[5] (Bick, 1996).

## 7. Future prospects

Several actions are envisaged for the future, including:

(a) Increasing the size of under-represented registers. Although balancing the corpus has never been a concern, clearly some registers should become larger in order to counteract the effect of the newspaper section.

[5] http://visl.hum.sdu.dk/Linguistics.html

(b) Obtaining more texts off the web. Presently, the World Wide Web provides access to a large amount of different texts, including newspapers, magazines, academic articles, interview transcripts, business reports and many other genres which may be collected automatically through offline web browsers (such as HTTrack).

(c) Broadening the range of registers, particularly by including new emerging text types such as blogs and chats as well as other digital forms of communication such as emails.

(d) Replacing older texts. Recycling is necessary so that the corpus remains contemporary, otherwise it may no longer represent current usage and may become a diachronic corpus. Permanence may affect representativeness (Hunston, 2002), in cases where to be representative means to stay current with the latest developments in language. Deciding on an 'age limit' is not straightforward. Around 2001, the Bank of English replaced older material from 1985-1990 (Krishnamurthy, 2002). This suggests a 10-year limit for texts to remain in the corpus. But this clearly depends on the kind of text. Arguably fiction literature does not age as fast as newspapers do.

(e) Finishing the index. As shown above, the online perl index records the information for each file in the corpus. This is a valuable source of information to users, who can query the database (for register or genre, for instance) and select parts of the corpus, thus creating specialized subcorpora that suit their needs.

(f) Allowing broader access. At present, access is restricted to users on the LAEL network in the university. The plan is for a DVD version of the corpus to be launched so that more people use the corpus without having to come to CEPRIL.

## 8. Final remarks

The Bank of Portuguese has been built over the years with the help of a large team of researchers and students. It is part of DIRECT, a research team whose aim is to describe business and professional discourse. The bank has always been an important part of the team efforts, but recently it took on a more central role.

The importance of personal contacts in obtaining data cannot be overemphasized. Approaching companies or newspapers through institutional channels has more often than not proven unsuccessful, but approaching them through people researchers had contact with has been much more fruitful. The approach to compilation is flexible and may be called both 'opportunistic' (any available material is added) and 'participative' (project members contribute their own individual corpora to the Bank once their research is complete). This helps to keep the corpus growing in diversity. Publishers in Brazil are still quite averse to distributing data to Corpus Linguistics research, despite all the guarantees provided by the university and by the project.

The shape and organization of the Bank of Portuguese has been determined by its users over the years rather than by a specific team of corpus planners. Most of the research undertaken at DIRECT has had a text focus rather than a language focus (Scott, 2000; Scott & Tribble, 2006). This meant that researchers were not interested in investigating language as a whole but in analysing their individual corpora for textual and discoursal features. As a result, the corpus grew out of a collection of independent business corpora. Later, because of a need to compare these corpora with a norm, the bank took on the role of a reference corpus. This enabled users to extract keywords via WordSmith Tools.

Due to the development of Corpus Linguistics in Brazil and in Portugal, the Bank of Portuguese is likely to grow in importance. Above all, electronic corpora of Portuguese will certainly receive growing recognition, as more people realize that a large electronic corpus is a very valuable asset in today's world.

Recebido em: 08/2007; Aceito: 10/2007.

## References

BICK, E. 1996 Automatic parsing of Portuguese. Paper presented at the Workshop on Computational Processing of Written Portuguese, Curitiba, Brazil. Available online at http://beta.visl.sdu.dk/~eckhard/postscript/curitiba.ps.

BORBA, F.D.S. 2004 *Dicionário UNESP do Português contemporâneo (UNESP dictionary of contemporary portuguese)*. UNESP.

CASTILHO, A.T.; OLIVEIRA E SILVA, G.M.D. & LUCCHESI, D. 1995 Informatização de acervos da língua portuguesa. *Boletim da ABRALIN,* **17:** 143-151.

HUNSTON, S. 2002 *Corpora in Applied Linguistics*. Cambridge University Press.

KRISHNAMURTHY, R. 2002 The Bank of English past, present, and future: corpus size, composition, annotation, and software. Talk given at the 2nd ILASH Half-Day Workshop on "Computational Language Resources", University of Sheffield, February 8th 2002.

MASON, O. 1997 QTAG-A Portable Probabilistic Tagger Computer Software. (Version 1). Birmingham: University of Birmingham. Available online at http://www.english.bham.ac.uk/staff/omason/software/qtag.html.

PINHEIRO, G.M.; OLIVEIRA, L.H.M.D.; TAGNIN, S. & ALUÍSIO, S. 2003 Mapeamento de projetos de corpora no Brasil. Paper presented at the III Encontro de Corpora, Unicamp.

SCOTT, M. 2000 Focusing on the text and its key words. IN: L. BURNARD & T. MCENERY (eds.), *Rethinking language pedagogy from a corpus perspective - Papers from the Third International Conference on Teaching and Language Corpora* (pp. 103-122). Peter Lang.

SCOTT, M. & TRIBBLE, C. 2006 *Textual patterns - key words and corpus analysis in language education*. John Benjamins.

*Tony Berber Sardinha received a BA in English from the Catholic University of São Paulo, Brazil, an MA in Applied Linguistics from the same university and a PhD from the English Department of the University of Liverpool (UK). He is a researcher with CNPq (Brazilian National Research Council) and CEPRIL (Center for Research, Resources and Information on Language), an Adjunct Professor with both the Linguistics Department and the Graduate Program in Applied Linguistics, Catholic University of São Paulo. He was recently a visiting scholar in Corpus Linguistics at Northern Arizona University (USA) and his research interests include Corpus Linguistics, Applied Linguistics, Language Teaching, Business Discourse, Metaphor, Forensic Linguistics, Computer Programming, and Web Design and Tools Development.* tony@corpuslg.org.