D.E.L.T.A., Vol. 15, Nº 2, 1999 (289-299)

Beginning Portuguese Corpus Linguistics: Exploring a Corpus to Teach Portuguese as a Foreign Language*

(Iniciando a Lingüística do Corpus do Português: Explorando um Corpus para Ensinar Português como Língua Estrangeira)

A. P. Berber Sardinha *(Pontifícia Universidade Católica de São Paulo)*

Abstract: The study reports the results of the exploration of a machine-readable corpus of Brazilian Portuguese. The corpus was collected from news distributed on the Internet. The news items themselves consisted of excerpts from newspaper stories and TV transcripts. The focus of the paper is on the description of selected language features needed for the production of teaching materials for private Portuguese classes in Britain. Several lexical and grammatical items are described using corpus linguistics tools in what amounts to pioneering work on corpus analysis of Portuguese. The paper concludes that guidance provided by existing reference materials such as textbooks, grammars and dictionaries are inadequate since these sources are not based on samples of authentic language.

Resumo: O presente trabalho apresenta os resultados da exploração de um corpus eletrônico de português do Brasil. O corpus foi coletado a partir de notícias distribuídas na Internet pela Radiobrás. As notícias foram retiradas de reportagens de jornais e de transcrições de notícias de TV. A ênfase do trabalho é a descrição de algumas características lingüísticas necessárias para a produção de materiais para aulas particulares de português oferecidas na Grã-Bretanha. Ao apresentar a descrição de vários itens lexicais e gramaticais dentro do paradigma da lingüística do corpus, o trabalho oferece uma contribuição pioneira no sentido de iniciar a lingüística do corpus do português. O trabalho conclui que o tipo de suporte disponível em materiais de referência existentes como livros de curso, gramáticas e dicionários tendem a ser inadequados para o aluno de português como língua estrangeira já que eles não se baseiam em amostras autênticas de linguagem como aquelas proporcionadas por um corpus eletrônico.

Key Words: Corpus Linguistics; Teaching Portuguese as Foreign Language; Corpus-based description of Portuguese.

* Earlier versions of this paper were presented at the Teaching and Language Corpora conference (TALC96), University of Lancaster, Lancaster, UK, August 1996, and at the Language Research Unit, University of Birmingham, UK, March 1997.

PALAVRAS-CHAVE: Lingüística do Corpus; Ensino de Português como Língua Estrangeira; Descrição do Português baseada no Corpus.

0. Introduction

The aim of the study presented here is to report on an initial exploration of a corpus of Portuguese which was compiled in the mid 1990's in the University of Liverpool, the Corpus of Brazilian Media Portuguese (CBMP). The corpus was used for assisting in the creation of materials for teaching Portuguese as a foreign language to private students in Britain. Nevertheless, the focus of the present paper is on the reporting of the description of Portuguese, and not on presenting the materials used in the classes. The students were adults who wanted one-to-one private tuition in Brazilian Portuguese for a range of purposes. The classes were not part of any teaching program associated with the University of Liverpool. The information obtained by analysing the corpus for was used to illustrate, expand on and even question the information provided by reference materials such as grammars, textbooks and dictionaries. The emphasis throughout was in obtaining authentic evidence for particular teaching points.

The paper is also concerned with extending the kinds of analyses developed for the exploration of corpora of English to the analysis of Portuguese. As such, the project reported here can be seen as fitting in a small body of pioneering research devoted to the compilation and description of corpora of Portuguese. Although other corpora of Portuguese have been around for some time (e.g. Borba-Ramsey Corpus in ACL, 1994; Contemporary Corpus of Portuguese in ELRA 1998; PORTEXT Corpus in Maciel, 1997), the corpus introduced in this paper is the first one in the available literature which was used for teaching Portuguese as a foreign language using the methodology of corpus linguistics (e.g. Kennedy, 1998; McEnery e Wilson, 1996; Sinclair, 1991). The kinds of analysis carried out included collocation and induction of patterns of cooccurrence and extraction of word frequency information. Another important feature is that the corpus was made available to the research community through the Internet for some time, which meant that different researchers in various parts of the world used it for purposes other than teaching. Hoey (1996), for example, used a subset of the CBMP to compare the usage of 'reason' and 'razão'. The corpus has also been indexed on numerous web pages devoted to corpus linguistics as the only corpus of Portuguese.

1. The CBMP corpus

Machine-readable corpora of Portuguese are recent. The Borba-Ramsey corpus is perhaps the first corpus of Portuguese which was made available to a large audience; it was published on CD-ROM in 1994. In 1995 the newspaper 'Folha de S. Paulo' published its first full edition on CD-ROM, which, although not strictly a corpus, can be used as corpus data.

In 1994, a project in Liverpool University was started which was aimed at collecting a corpus of contemporary Portuguese. The corpus was called 'CBMP', for 'Corpus of Brazilian Media Portuguese', and it was so named because it was made up of newspaper and magazine clippings and TV transcripts. These were distributed by e-mail to subscribers of an information service sponsored by Brazilian research funding agencies. The texts and the transcripts in the corpus were published or broadcast conventionally. The CBMP has 4.075.335 words, which places it on the 'small' end of the scale for present-day corpora. Nevertheless, it is larger than other much-cited corpora of English such as the Brown or LOB, and therefore it is not too small to provide useful information about language in use.

The texts included in the corpus are contemporary. This means that the language represented in the corpus is as close as possible to the language of the press in Brazil in the early 1990's. This also enables the corpus to be used as a source of texts for developing materials for language teaching. This is an important point since the availability of Brazilian newspapers and magazines abroad was very limited before the late 1990's, and therefore the CBMP was a source of fairly recent materials about Brazil.

In this paper, the focus will be on the exploitation of the corpus for the teaching of Portuguese as a foreign language. The main motivation for using a corpus rather than the existing materials for teaching Portuguese as a foreign language was that the latter were generally based on invented examples. Those involved in the lessons also recognized that the use of authentic materials was essential for language learning. In addition, previous reports on using corpora in language teaching, mainly through concordancing (explained in the next section), showed that exposing students to corpus material had important benefits (see next section).

The corpus was used in the preparation of materials in two main ways.

First, as a source of data for description of the Portuguese language, or at least, that variety which was included in the corpus (newspaper texts and television transcripts) prior to the preparation of teaching units. The description was carried out using computational techniques, and the aim was to describe features of the language which were relevant for the lessons. And second, as a source of examples to illustrate particular language points. This was done in a number of ways, including lists of examples, patterns, and concordances (see next section).

The lessons which were taught using corpus materials were individual classes offered to British English speakers in Liverpool. The paper will report on the description of several features of Portuguese which formed part of teaching units. For the most part, the corpus was used to describe aspects of the language which reference materials (textbook, grammar, and dictionary) did not deal with or dealt with unsatisfactorily. In the first part, the paper will provide a brief discussion on the usefulness of corpora in language teaching. In the second part, a description of selected language items will be offered.

2. The corpus in the classroom

One of the ways in which language samples from the corpus were presented to the students was through concordances. A concordance is a list of the occurrences of a given word (or words) in a corpus. The kind of concordance used here is that known as KWIC, or Key Word in Context. In this kind of concordance, the word searched for (the 'keyword') appears in the center of the listing surrounded by a portion of the text that occurred next to it in the corpus. By observing the kinds of words appearing near the keyword (the co-text), the analyst or the student can gain insights about collocations, or groups of words that tend to occur near each other.

The concordances were printed out on paper and used as worksheets. A number of activities was carried out using the concordances, but a detailed account of these is beyond the scope of this paper.

The main reason why concordances were adopted as a technique for exploring the corpus with the students was that they provide students with the opportunity to engage in discovery activities. It is argued that concordances have a positive impact on the learners, the teachers and on language learning itself (Johns, 1994). Learners become very effective researchers because concordances provide motivation for inquiry and

speculation. In addition, as soon as they start working with the data themselves, students they become active researchers instead of passive recipients of knowledge. As Johns (1994) aptly puts it, research is too valuable a tool to be left in the hands of researchers'. Similarly, teachers are no longer the fount of all knowledge, since they can resort to the corpus for answers. In trying to make sense of their data, students generate their own explanations which are arguably better learned than ready-made rules from the textbook. In this context, the role of the computer is that of 'informant not surrogate teacher' (unlike in CALL, for example) (Murison-Bowie, 1996: 39), that is, concordancers and computerized corpora are not seen as substitutes for the teacher; rather these elements are seen as tools to be used by the teacher with his or her students.

There were two main ways in which concordances were used in the classes. One was by following the inductive approach, which goes from the bottom up, that is, from inspection of the data up to a generalization. The other way was by using the deductive method, which goes from the top to the bottom, that is, from a rule or hypothesis to the data, ending with a revised hypothesis. Neither approach is without faults. The alleged problem with the inductive approach is that the student is not 'encouraged to (...) to test [their] conclusions' (Murison-Bowie, 1993: 46). The argument against a purely deductive method is that hypotheses can be confirmed or rejected wrongly due to a lack of evidence, since no corpus is complete, especially small ones such as the CBMP. In general, students tended to adopt an approach depending on their initial interests. If they had a hypothesis, then they would more naturally follow a deductive approach. In the absence of a working hypothesis, students would tend to take an inductive approach.

3. Exploration of the corpus

Two kinds of information were drawn on in the description of the corpus: frequency and collocation. For details about the frequency of words in the corpus, we employed the WordList tool in WordSmith Suite (Scott, 1996). WordSmith is a computer program that offers tools for the analysis of language in collections of texts (Berber Sardinha, 1996).

For collocations, we employed concordances. The concordances were generated using the Concord Tool in WordSmith. The Concord allows the user to obtain concordances easily and quickly. It also provides access to lists of collocates, or those words that occur near the keyword at a frequency determined by the analyst. The maximum distance between the keyword and

its co-text is the (collocation) span, measured in words. For example, a span of 3 words on either side of the node means that the words that are no further than three words to the left or to the right of the keyword are counted as collocates.

In this study, the collocation span varied from two to five words on either side of the node. The two-word span was tried first, and if this did not return at least 20 collocates of frequency 2 or higher, then the span was widened. The threshold of 20 collocates is an informal parameter used by corpus analysts (Kilgariff, 1998).

Once the collocates were obtained, a 'structure' was generated (cf. Francis and Hunston, 1996), which is a generalization about the usage of the search word based on its collocates. Structures were accompanied by examples. After that, a 'pattern' was produced, which is a more abstract generalization. By comparing patterns, it became possible to see more clearly what the similarities and differences between the words were. No statistical tests were carried out on the strength of the association among each search word and its collocates because the primary aim of the investigation was not to obtain final answers about patterns in Portuguese, but rather to allow those involved in the lessons to gain insights into the usage of words based on authentic evidence.

The fundamental notion applied to the analysis is that distinct senses are identifiable as distinct cooccurrence patterns (Sinclair, 1991). Hence, if the analyses revealed different patterns for the words under investigation, this would indicate different senses. Once these patterns had been specified, concordances showing these patterns were run and printed out.

Another guiding principle in the preparation of materials was the frequency of items in the corpus. It was felt that this information was relevant and should be passed on to the students. Information on frequency is not available to native speakers through introspection, and needs to be obtained from a corpus. As Sinclair and Renouf (1988: 151) comment, this is a feature common to users of any language:

> *'the human being, contrary to popular belief, is not well organized for isolating consciously what is central and typical in the language; anything unusual is sharply perceived, but the humdrum everyday events are appreciated subliminally'.*

4. Individual language items

In this section we present a sample of the analyses we carried out using the CBMP corpus. The individual analyses were prompted by questions asked by the students during the classes.

One of the features of Portuguese which caused the students trouble was the future tense. In Portuguese, the future can be formed either by inflecting the verb or by using an auxiliary verb plus an infinitive. The latter is called the periphrastic future and native speaker intuition tells us it is the most common form of the future in Brazilian Portuguese. However, even recently published grammars do not recognize this fact, giving more space to the inflected form (e.g. Mesquita, 1994); the periphrastic form is simply included as colloquial usage restricted to speech. As a result, when students resorted to grammars, they usually found they gave emphasis to the inflected future, while speakers use the periphrastic future.

The periphrastic future is formed by the verb 'ir' (conjugated as 'vou', 'vai', vamos', or 'vão') plus an infinitive. The form 'vai' is the 41st most common word across the corpus, with 8001 occurrences (0.2% of the corpus); significantly, of its 20 top collocates, 16 were infinitive verbs. This suggest that one of the main uses of the form 'vai' is to form the future, and not as an independent verb. The most frequent inflected future form is 'serão' which is 4 times less frequent than 'vai', with 2200 occurrences (0.1 % of the corpus). According to the corpus, then, the periphrastic future seems to be the most common future form, despite what the grammars say. As a result, we decided to emphasize the periphrastic future with our students. In the case of the future, then, frequency information was crucial in deciding which forms to teach, unlike in the case of the prepositions discussed above.

Another problem that the students faced relates to the verbs 'saber' and conhecer' which normally translate into English as 'to know'. Conventionally a distinction is made in bilingual dictionaries and coursebooks between 'to know something' and 'to know somebody'. If you know something, the verb can be either 'saber' or 'conhecer'; if you know how to do something, the verb is 'saber'; but if you know somebody, the verb is 'conhecer'. The problem with this rule is that it does not specify what the verb should be when it precedes 'something'. According to the rule, these two verbs are interchangeable when they mean 'to know something'. The decision was then taken to search the corpus for possible differences in complementation between the two verbs.

In the corpus, the verbs have different frequencies (see Table 1) which is a first indication that in contexts where it is meant 'to know something', 'saber' may be the unmarked choice. Nevertheless, the frequencies alone did not answer the question about the differences in complementation. A description of patterns was then carried out.

Rank	Item	Frequency
961	Saber	484
3361	Conhecer	108

TABLE 1: Rank and frequency of 'saber' and 'conhecer'

Position of collocate[1]	Saber	Conhecer
1st to the right	se (if), a (the-fem), do (of the-masc), como (how), o (the), que (that), da (of the-fem), onde (where), qual (which), quem (who)	o (the-masc), os (the-masc-pl), a (the-fem), as (the-fem-pl), um (a-masc), detalhes (details)
2nd to the right	detalhes (details), programa (program), projeto (project), resultados (results), numeros (numbers), objeçõoes (objections), parque (park), projetos (projects), realidade (reality), rendimento (yield)	opinião (opinion), oripem (origin), notícia (news), causa (cause), decisaes (decisions), declaração (statement), destino (destiny), detalhes (details), fraudes (frauds), número (number), resultado (result), ritmo (rhythm), rumo (direction)

TABLE 2: Frequencies and collocates of 'saber' and 'conhecer'

The first collocates to the right of the verb in Table 2 indicates that what distinguishes these two verbs is that 'saber' is followed by conjunctions, such as 'se' (if) and 'como' (how), and by the contracted preposition 'do'. What both verbs have in common is that they are both followed by articles like 'o' (the-masc) and 'a' (the-fem) which matches the translation 'know something'. What was needed next was to know which words followed these articles. Table 2 also displays the second collocates to the right which are nouns. There seemed to be an interesting trend here. Out of the 13 nouns that

[1] The minimum frequency of collocates is 2, except for the first ones to the right of 'saber'.

collocated with 'saber', 10 were singular: 'opinião', 'origem', 'notícia', 'causa', 'declaração', 'destino', 'número', 'resultado', 'ritmo', and 'rumo'. Based on this information, the patterns in Table 3 were produced. 'Saber' is followed by a preposition or a subordinate conjunction; both 'saber' and 'conhecer' are followed by an article, but 'saber' seems to be followed by a singular noun. Note that none of the patterns emerging from the corpus includes the traditional 'verb + Personal Noun' and 'verb + Infinitive verb' patterns which are commonly used to teach 'conhecer' and 'saber', respectively.

Pattern	Verb
Verb+de(prep)	saber
Verb+conj	saber
Verb+article	saber / conhecer
Verb+article + singular Noun	saber

TABLE 3: Patterns of 'saber' and 'conhecer'

The second person pronouns is another area about which the corpus supplied details. In the tables of conjugations found in grammars students normally come across the second person pronouns 'tu' and 'vós', but in most contexts across the country these pronouns have been replaced with 'você' and 'vocês'. Their corpus frequencies reflect this situation. 'Você' appears 112 times, and 'vocês' 44 times, whereas 'vós' appears only 8 times and 'tu' only 6 times. This information was used as evidence to persuade students to ignore 'tu' and 'vós' in conjugation tables since they would rarely come across these pronouns and the verb forms associated with them in authentic newspaper texts.

5. Final comments

In this paper, lexical and grammatical items were described through inspection of a corpus of Portuguese. In general, the information available from inspection of the corpus was in conflict with the information found in reference materials. This seems to be the case because Portuguese grammars, dictionaries and coursebooks have largely been based on intuition rather than on authentic data. When authentic examples are provided at all, these are inevitably from literary fiction, a variety which is still regarded as the norm, but which does not reflect the language used on daily basis in Brazil. The general conclusion is that guidance provided by existing reference materials is inadequate in that they fail to provide evidence of language in use.

Despite its restricted size, the CBMP has provided detailed evidence for various patterns. Significantly, evidence of this kind was not available to the native speaker teachers from intuition and had to obtained through inspection of the corpus.

There are several limitations which may be overcome in further research. First, the small size of the corpus. A larger corpus should provide evidence of a wider range of patterns while at the same time giving more details on the patterns which emerged so far. Second, the narrowing of the collocational span from five to two words in some cases may have limited the range of patterns that might actually exist. Finally, collocational significance was not computed for the collocates, which would have been instrumental in ruling out spurious associations. These limiting factors may have led to a simplification of the patterns obtained here, but whether this is true or not can only be attested through access to a larger corpus and more powerful computational tools. At any rate, these limitations do not compromise the findings since the aim of this exercise was to obtain evidence of language in use for pedagogical purposes rather than to describe a variety of the Portuguese language.

REFERÊNCIAS BIBLIOGRÁFICAS

ACL (1994) Borba Ramsey corpus. In: *European Corpus Initiative. Multilingual Corpus 1.* HRCR, University of Edinburgh, and ISSCO, University of Geneva.

AIJMER, K. & ALTENBERG, B. (eds.) (1991) *English corpus linguistics - Studies in honour of Jan Svartvik.* Longman, London.

ASTON, G. (1995) Corpora in language pedagogy: matching theory and practice. In: *Principle and practice in Applied Linguistics - Studies in honour of H G Widdowson.* (eds.: COOK, G.; SEIDLHOFER, B.) Oxford University Press, Oxford, 257-270.

BERBER SARDINHA, A. P. (1996) "Review of WordSmith tools." *Computers & Texts* **12**: 19-21.

COLLIER, A. (1993) Issues of large-scale collocational analysis. In: *English Language Corpora: Design, analysis and exploitation - Papers from the thirteenth International Conference on Englsh Language research on computerized corpora, Nijmengen 1992.* (eds.: AARTS, J.; DE HAAN, P.; OOSTDIJK, N.) Rodopi, Amsterdam/Atlanta,GA, 289-298.

EDWARDS, J. (1993) Survey of electronic corpora and related resources for

language researchers. In: *Talking data: Transcription and coding in discourse research.* (eds.: EDWARDS, J.; LAMPERT, M. D.) Earlbaum, London and Hillsdale, NJ, 263-310.

ELRA (1998) Contemporary Corpus of Portuguese. Information available on the Internet at http://www.icp.grenet.fr/ELRA/cata/tabtext.html.

FOLHA DE S. PAULO (1995) *Folha CD-Rom.* Folha de S.Paulo Newspaper, São Paulo, Brasil.

FRANCIS, G. & S. HUNSTON (1996) *Grammar patterns.* Vol.1: Verbs. Collins COBUILD, London.

HOEY, M. (1996) Cohesive words: A paper of consequence. In: *Words*. (ed.: SVARTVIK, J.) The Foundation Natur och Kultur, Lund, 71-90.

JOHNS, T. (1994) From printout to handout: Grammar and vocabulary teaching in the context of Data-driven learning. In: *Perspectives on pedagogical grammar.* (ed.: ODLIN, T.) Cambridge University Press, Cambridge, 293-313.

KENNEDY, G. (1998) *An introduction to Corpus Linguistics.* New York: Longman.

KILGARIFF, A. (1998) Email message, subject 'Size of representative corpus', date 21 August 1998, on CORPORA discussion list. Available on the Internet at http://nora.hd.uib.no/e-index-i.html.

MACIEL, C. (1997) La base PORTEXT à Nice. À propos d'une idée. Documento eletrônico disponível na Internet em http://lolita.unice.fr/~brunet/index.html.

MCENERY, T. & WILSON, A. (1996) *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

MESQUITA, R. M. (ed.) (1994) *Gramática da Língua Portuguesa.* (1st ed.) São Paulo, SP, Brasil: Saraiva.

MURISON-BOWIE, S. (1993) *MicroConcord Manual - An introduction to the practices and principles of concordancing in language teaching.* Oxford University Press, Oxford.

MURISON-BOWIE, S. (1996) Linguistic corpora and language teaching. *Annual Review of Applied Linguistics* **16**.

SCOTT, M. R. (1996) *WordSmith Tools.* Software for text analysis. Oxford University Press, Oxford.

SINCLAIR, J. (1991) *Corpus, concordance, collocation.* (Describing English Language Series.) Oup, Oxford.

SINCLAIR, J. McH. & RENOUF, A. (1988) A lexical syllabus for language learning. In: *Vocabulary and language teaching.* (eds.: CARTER, Ronald; MCCARTHY, M.) Longman, London, 140-160.

STUBBS, M. (1995) Corpus evidence for norms of lexical collocation. In: *Principle and practice in Applied Linguistics.* (eds.: COOK, G.; SEIDLHOFER, B.)

Oxford, Oxford University Press, 245-256.

_____ (1996) *Text and corpus analysis - Computer-assisted studies of language and culture*. Blackwells, Oxford.

THOMAS, E. W. (1969) *The syntax of spoken Brazilian Portuguese*. Vanderbilt University Press, Nashville, Tenn.