Meta
Journal des traducteurs
Translators' Journal

META

Frequency Dictionary of Contemporary Portuguese

Maria Tereza Camargo Biderman

Volume 41, Number 2, juin 1996

Traduction et terminologie au Brésil
Translation and Terminology in Brazil

URI: https://id.erudit.org/iderudit/002079ar
DOI: https://doi.org/10.7202/002079ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Camargo Biderman, M. T. (1996). Frequency Dictionary of Contemporary Portuguese. *Meta*, *41*(2), 275–278. https://doi.org/10.7202/002079ar

Article abstract

This article describes the compilation of a frequency dictionary of contemporary Brazilian Portuguese. It discusses some of the problems encountered and the solutions proposed, and introduces some useful tools that the author intends to develop from the dictionary.

FREQUENCY DICTIONARY OF CONTEMPORARY PORTUGUESE

MARIA TEREZA CAMARGO BIDERMAN
Universidade Estadual Paulista JMF (UNESP), São Paulo, Brazil

Résumé

L'auteur nous fait connaître son dictionnaire de fréquences du portugais brésilien contemporain. Elle discute des problèmes rencontrés et des solutions proposées et présente quelques produits utiles qui pourraient être créés à partir de ce dictionnaire de fréquences.

Abstract

This article describes the compilation of a frequency dictionary of contemporary Brazilian Portuguese. It discusses some of the problems encountered and the solutions proposed, and introduces some useful tools that the author intends to develop from the dictionary.

This dictionary is based on a *corpus of reference of 5,000,000 words* collected in printed texts, written in Brazilian Portuguese from 1950 to 1990. The corpus is drawn from:

1) literature (novels and short stories): 1,300,000 occurrences;
2) literature (drama): 600,000 occurrences;
3) science and technology: 1,300,000 occurrences;
4) journalism (newspapers and magazines): 1,300,000 occurrences;
5) speeches and discourse (political and religious): 500,000 occurrences.

At the moment, a group from our university is working on expanding the text database in order to reach *10,000,000 words*.

The size of this corpus is not large compared to the enormous database of English *Cobuild* of Birmingham, England, or the *Trésor de la langue française* of Nancy, France; nevertheless, it is currently the largest Portuguese database in eletronic form. Furthermore, it can be considered large enough to permit a diagnosis of the vocabulary of contemporary Portuguese. The frequency dictionary that preceded ours — produced at Stanford University in 1972 as a Ph.D. dissertation — collected only 500,000 occurrences of European Portuguese. This corpus was not only insufficient to represent our lexicon, but it also documented a period of the language that is no longer contemporary in synchronic terms. In fact, today language vocabularies are changing very fast: according to Rey-Debove, 10% of the vocabulary changes every 25 years.

We based our research on several lexicostatistic works:

— the five dictionaries of *Romance Languages* (Spanish, French, Rumanian, Italian and Portuguese) edited by Juilland and associates;
— Zampolli *et al.*: *Lessico di Frequenza della Lingua Italiana Contemporanea*;
— *American Heritage Word Frequency Book*;
— *Português Fundamental* (University of Lisbon).

We already have a first statistical version of the frequency of forms from the corpus. At present, we are working on the *lematization* of these forms. As we expected, a serious problem occurs with homographs. To deal with the ambiguity generated by homographs we began by listing all possible homonyms in Portuguese, collecting more than a

thousand items. We then proceeded by coding these homographs in our raw lexicostatical data on forms, in order to separate these forms afterwards with a view to disambiguating them.

We also collected in all Portuguese dictionaries and grammars more than 1,000 grammatical phrases of four different types: prepositional, adverbial, conjunctive and pronominal. These idiomatic expressions are lexical units that must be counted as such.

We are working on the creation of software that might help us in the complicated task of disambiguating complex lexical units, as well as homonyms. We have so far produced several programs that are helpful but inadequate. Moreover, because the volume of data to be processed is so great, much time and effort are required even with the assistance of computers.

The process of disambiguation is highly complex. Our team faces not only computational obstacles due to the complexity of the linguistic system, but also to the limitations of the computer, and the capacity and speed of its operating system. Considering the huge amount of data, we should have a faster system.

LEMATIZATION

The processing of the first lematization is being carried out as follows: 1) correction of spelling errors, since there was a considerable number of errors in the digital input of texts; 2) elimination of proper nouns and foreign words; 3) checking bizarre items in the databases to determine whether they are typing errors, idiosyncratic words used by a writer or other eccentric expressions; 4) codification of homomyms for further disambiguation; 5) preliminary lematization.

The third step is a singularly tiresome task, since it requires consulting the databases hundreds of times. The identification of the ambiguous forms could be time-consuming and painstaking if we do not succeed in developing functional software.

First we tried to identify ambiguous words by checking them with concordances generated by the computer where these homonyms are key-words. This method turned out to be very slow, which would seem to indicate that this will be an arduous and strenuous task.

Recently I started testing the software FOLIO VIEWS, which may be useful and may fulfil some of our needs, although it may require considerable work in terms of constant checking of the database.

To proceed with the *identification of lexical units* in the databases to create a frequency dictionnary a theoretical problem needs to be resolved: what is a word? Indeed what is a lexical unit in a language such as Portuguese? I have been dealing with this question for a long time, and have even published some articles and books on this issue. In short, it needs to be said that it is impossible to give a general answer to such a question, particularly a universal one that might apply to every language. The answer has to be relative. The ambiguous cases are those in which one may have one or more words, depending on the criteria considered to define a lexical group (a syntagm) or a complex word. One may apply two basic tests to determine if a word group is a single unit: 1) the insertion of any word between the written units that might be considered a lexeme or not; 2) the possible exchange of one of the words in the syntagm by a synonym or parasynonym. If both tests fail one can safely assume that the word group is categorized in the lexicon as a single unit.

I plan to examine and describe contemporary Portuguese lexicon: the general vocabulary currently used, the specific vocabulary of different genres, word-formation trends, etc. Hopefully, this descriptive analysis will show the structures of today's vocab-

ulary and the relative distribution of words in the different genres. The data analysed so far clearly demonstrate a strong influence of text genre and topic in explaining the presence or absence of some lexical items. The journalistic subcorpus showed the greatest heterogeneity, and consequently contained the lowest level of specialization.

I have also been scrutinizing *word-formation* in the corpus, particularly the prefixation process. In recent years, lexicologists have analysed this question, pointing out that there is a growing tendency to increase the use of prefixation in creating new words in Portuguese. Even though prefixation can be considered an old process for expanding vocabulary, suffixation used to be the main pattern of word creation in Portuguese. Apparently, there has been a change in this trend. Consider, for example, the use of the prefix *super-* followed by the adjective to designate the superlative degree of a quality, instead of using the suffixes *-issimo, -imo, -érrimo*, etc. as established by the pattern inherited from classical Portuguese. Some Greek, Latin or vernacular prefixes exhibit a very high frequency in the corpus such as: *anti-, des-, hiper-, pre-, semi-, sobre-, super-, re-, ultra-*. The vast majority of neologisms formed with these prefixes are not registered in the *Novo Dicionário da Língua Portuguesa* (1986), which is considered a kind of lexical norm for contemporary Portuguese in Brazilian society.

The long lists of neologisms often contain words whose formation conflicts with traditional word-formation patterns in the language. Although some of these formations are idiosyncratic of authors who wish to create a feeling of strangeness, such neologisms often violate not only the established norm, but also the very system of the language. The adverb *não* is being used as a prefixoid, a very unusual pattern for Portuguese vocabulary, although there is some evidence of this usage in the past (XVIth and XVIIth centuries). In my analysis of the data, I concluded that American English currently has a strong influence on Brazilian Portuguese, which may explain some of these eccentric neologisms.

I plan to develop some useful tools from this frequency dictionary:

1. A basic vocabulary of 3,000 words for teaching Portuguese in Brazilian primary schools, as well as for teaching Portuguese as a foreign language. Regarding this goal, I plan to compare our results with the word list of *basic Portuguese* (*Português Fundamental*) produced by the University of Lisbon in 1986. Although this research was based on an oral corpus and includes some features peculiar to European Portuguese, this contrastive analysis will provide useful insights for the teaching of both variants of our language to foreign learners.
2. An *index verborum* of the 50,000 most frequently used words to help create a contemporary dictionary of Brazilian Portuguese.
3. Word frequency lists to assist in the development of computational programs, such as spellers and font patterns for OCRs.
4. Finally, an analysis of the lexicon of contemporary Portuguese, as mentioned above.

REFERENCES

ALVES, I. M. (1978): "A formação de neologismos através da composição prefixal no vocabulário da imprensa brasileira contemporânea", *Estudos Lingüísticos*, Bauru, 2, pp. 212-224.

ALVES, I. M. (1984): "A integração dos neologismos por empréstimo ao léxico português", *Alfa* (Suplemento), São Paulo, 28, pp. 119-126.

BIDERMAN, M. T. C. (1978): *Teoria Lingüística, Lingüística matemática e computacional*, Rio de Janeiro, Livros Técnicos e Científicos.

CARVALHO, N. M. (1989): *Empréstimos Lingüísticos*, Série Princípios, São Paulo, Ática.

GUILBERT, L. (1975) : *La créativité lexicale*, Paris, Larousse.

GUILBERT, L. (1979) : *Néologie et lexicologie*, Paris, Larousse.

HOLANDA FERREIRA, A. B. (1986) : *Novo Dicionário da Língua Portuguesa*, 2ª ed., Rio de Janeiro, Editora Nova Fronteira.

PEREIRA, R. F. (1983) : *Neologismos na Mensagem Publicitária*, Assis, UNESP, ILHP, dissertação de mestrado.

RIO-TORTO, G. M. (1989): "Para uma teoria da formação das palavras em português", comunicação apresentada no XIX Congresso Internacional de Lingüística e Filologia Românica, Santiago de Compostela, Espanha.

SANDMAN, A. J. (1988): *Formação de Palavras no Português Brasileiro Contemporâneo*, Curitiba, Icone Editora.