THE CORPUS DO PORTUGUÊS AND THE ROUTLEDGE FREQUENCY DICTIONARY OF PORTUGUESE: NEW TOOLS FOR LEARNERS AND TEACHERS

Mark Davies[24]

Ana Maria Raposo Preto-Bay[25]

Abstract

In this paper we discuss two corpus-based resources of Portuguese that have recently become available, which hopefully are of real value to language learners. The first is the 45 million word Corpus do Português, which is freely available online and which offers many different types of searches that are oriented towards the language learner. These include searches by word, phrase, substring, lemma, and part of speech, as well as the ability to quickly and easily see the frequency across the major genres and time periods in the corpus. In addition, queries allow learners to compare the collocates of multiple words (to see the difference in meaning between the words), and between sections in the corpus (to see, for example, differences in word senses between genres). With one simple query, they can also see the frequency and distribution of all of the synonyms of a given word, and thus move far beyond a traditional thesaurus. The second tool is the recently-published Frequency Dictionary of Portuguese: Key Vocabulary for Learners (Routledge, 2007), which likewise has many features that are oriented towards language learners. Learners can easily find the most 5000 frequent lemmas in Portuguese, along with their English gloss, a sample sentence from the Corpus do Português, a translation of that sentence, and frequency and distributional information. In addition, there are "call-out boxes" with thematic vocabulary and frequency-based data on a wide range of grammatical phenomena that are often difficult for the Portuguese language learner.

Keywords: corpus-based, word frequency, dictionary, Portuguese, learners

The Corpus do Português

In terms of new resources available for learners of Portuguese, let us first turn to the *Corpus do Português*. This is a 45 million word corpus of Portuguese that was created by Mark Davies and Michael Ferreira (of Georgetown University) with generous funding from the United States National Endowment for the Humanities, and which was placed online in late 2006. The corpus contains 15 million words of historical Portuguese (1300s-1700s), 10 million words from the 1800s, and 20 million words from the 1900s. (It was the texts from the 1900s that served as the basis for the frequency dictionary.)

Of the 20 million words from the 1900s, two million words of text were taken from spoken Portuguese – either conversation (such as the *Linguagem Falada* project in Brazil or the *Projecto Corpus de Referência do Português Contemporâneo* from Portugal) or transcripts of interviews in newspapers and magazines. The written texts from the 1900s (18 million words) represent equally-sized sub-corpora from fiction, newspapers, and academic texts. In terms of the time period represented, virtually all of the texts from the 1900s are from 1970-2000, with the clear

[24] Mark Davies. Professor of Corpus Linguistics at Brigham Young University in Provo, Utah, USA. Co-creator of the Corpus do Português (www.corpusdoportugues. org), along with Michael Ferreira of Georgetown University. Co-author of the Frequency Dictionary of Portuguese: Core Vocabulary for Learners (Routledge, 2007) and the sole author of the Frequency Dictionary of Spanish: Core Vocabulary for Learners (Routledge, 2005). In addition to the Corpus do Português, the creator of several other large (100+ million word) online corpora, including the Corpus del Español, the Corpus of American English, BYU-BNC (the British National Corpus), and the TIME Corpus of Historical American English. Also the author of more than forty articles related to historical change and genre-based variation in syntax, as well as articles related to corpus design and construction.

[25] Ana Maria Raposo Preto-Bay. Assistant Professor in the Department of Spanish and Portuguese at Brigham Young University in Provo, Utah, USA. A native Portuguese, Ana graduated from the Universidade Clássica de Lisboa in English, German, and Modern Literature. She holds and MA in Linguistics from the University of Utah and a PhD in Instructional Psychology in the area of Language Acquisition from Brigham Young University. In addition to teaching Portuguese, she has taught in the Linguistics and English Departments where she worked in the areas of language acquisition and composition. In her research and published work she focuses primarily on issues dealing with literacy development— both in first and second language contexts—social and cognitive psychology, and foreign language teacher and program development.

majority being from the 1990s. Finally, we should mention that the corpus for this century was evenly divided between texts from Portugal and Brazil, for each of the four classes of texts just mentioned.

In addition to serving as the basis for the frequency dictionary, the corpus is also freely available online (http://www.corpusdoportugues.org), and it contains a number of features that make it quite useful for language learners. In terms of basic features, learners can search for any word, phrase, or grammatical construction. In less that one second, they can see the frequency in both Portugal and Brazil, in each of the four major genres (spoken, fiction, newspapers and magazines, and academic texts), and they can also see (based on the historical data) whether the construction is increasing or decreasing in usage. In addition to seeing the overall frequency of the word, phrase, substring, or grammatical construction (or any combination of these), they can also see the frequency and distribution of each matching form. For example, users could search for the lemma *cujo* 'whose' and see that it is clearly decreasing in usage and that it is the most frequent in the more formal registers (as with *whose* in English). Likewise, they could search for the passive ([ser] [vk*]: a form of *ser* 'to be' + a past participle), and in a matter of less than two seconds they would see that it is increasing in usage, and that it is the most frequent in the formal registers. Finally, it is possible to compare the relative frequency of essentially anything across different sections of the corpus. For example, users can find adjectives that are more common in fiction than in academic (*pensativo, lívido, abafado, aflito* 'thoughtful, livid, stuffy, afflicted') or vice versa (*norte-americano, nuclear, mundial, químico* 'North American, nuclear, worldwide, chemical'), or verbs that are more common in Brazil than Portugal (*retornar, descartar, ressaltar, brigar* 'to return, disagree, emphasize, fight') or vice-versa (*regressar, aperceber, calhar, sublinhar* 'to return, perceive, happen, emphasize').

Users can also easily obtain useful semantic information from the corpus, via collocates. At the most basis level, they can simply input a word (such as *espesso* 'thick'), click on 'Context', and then see the most frequent collocates (sorted by Mutual Information score, if desired). For example, in the case of *espesso* they would find *pelagem, fumo, cauda, névoa*, and *treva* 'hair, smoke, tail, mist, darkness'). In the search form they can also select one section of the corpus (e.g. Portugal or Brazil, or any of the four main genres, or any historical periods) and compare the collocates, which provide valuable insight into differences in meaning and usage between these sections. For example, by selecting 'Fiction' vs. 'Academic', learners can see that the collocates for forms of *duro* 'hard' are somewhat more figurative in fiction (*olhos, palavras, trabalho* 'eyes, words, work') while in academic they are more literal (*rochas, materiais, fibras* 'rocks, materials, fibers'). Finally, since users can compare collocates across historical periods, they can see how the usage is changing over time. For example, they can easily compare the collocates of *mulher* 'woman' in the 1800s and 1900s, and see that in the 1800s the emphasis was often on 'moral qualities' (*desgraçada, indigna, divina* 'fallen, unworthy, divine'), while in the 1900s the collocates mainly refer to a fairly prosaic social category (*jovem, sozinha,* negra 'young, single, black').

Two other features of the corpus are specifically oriented towards language learners. First, the interface allows users to input two words and compare the collocates of the words. This information provides valuable insight into the difference in meaning between the two words – probably much more than could be obtained from a dictionary. For example, users can compare the collocates of *romper* and *quebrar*, which could be difficult for native English speakers to differentiate, since they are both translated as 'to break'. The corpus shows that the most frequent collocates with *romper* but not *quebrar* are *marcha, grito, lábio, nuvem*, and *fogo* 'pace, shout, lip, cloud, and fire', while those with *quebrar* but not *romper* are *cabeça, perna, coisa, nariz*, and *monotonia* 'head, leg, thing, nose, and monotony'. Finally, users can use the corpus as a type of 'thesaurus on steroids' to compare the frequency, use, and distribution of the synonyms of a given word. For example, they simply enter '[=limpo]' to find twenty different synonyms of *limpo* 'clean (ADJ)', and they can see which are found more in formal or informal genres, in Brazil or in Portugal, and which are increasing or decreasing in usage over time. In summary, there are many types of queries that allow language learners to quickly and easily gain insight into usage in Portuguese, which would in most cases be far beyond their level of intuition in the second language, or beyond that of most language learning materials.

The Frequency Dictionary of Portuguese: Core Vocabulary for Learners

In addition to creating a corpus that was 'user-friendly' for learners and teachers of Portuguese, there was also a desire to create a published work that could be used as an integral part of Portuguese language classes. Recognizing the value of frequency-based materials for language learners, we decided to create a frequency dictionary of Portuguese.

There were a handful of printed frequency dictionaries of Portuguese (Brown 1951, Duncan 1972, Kelly 1970, Nascimento 1987, and Roche 1975), as well as two or three in electronic format on the web. Nevertheless, none of these was based on a large, balanced corpus of Portuguese (in other words, with texts from a number of different genres). Therefore, the goal was to create a dictionary that would contain the 5000 most frequent lemmas in Portuguese – based on the data from the Corpus do Português, and with a number of features that were specifically oriented towards the language learner. We worked on this dictionary in 2006 and 2007, and it was published in late 2007 as the *Frequency Dictionary of Portuguese: Core Vocabulary for Learners*, which was published by Routledge in late 2007.

Before discussing this particular corpus-based dictionary, however, we might first address the general question of the value of a frequency dictionary for language teachers and learners. Why not simply rely on the vocabulary lists in a course textbook? The short answer is that although a typical textbook provides some thematically-related vocabulary in each chapter (foods, illnesses, transportation, clothing, etc), there is almost never any indication of which of these words the student is most likely to encounter in actual conversation or texts. In fact, sometimes the words are so infrequent in actual texts that the student may never encounter them again in the "real world", outside of the test for that particular chapter.

While the situation for the classroom learner is sometimes difficult with regards to vocabulary acquisition, it can be equally as frustrating for independent learners. These individuals may pick up a work of fiction or a newspaper and begin to work through the text word for word, as they look up unfamiliar words in a dictionary. Yet there is often the uncomfortable suspicion on the part of such learners that their time could be maximized if they could simply begin with the most common words in Portuguese, and work progressively through the list.

The bottom line, then, is that frequency dictionaries can be a valuable tool for language teachers. It is often the case that students enter into an intermediate language course with deficiencies in terms of their vocabulary. In these cases, the teacher may often feel frustrated, because there doesn't seem to be any systematic way to bring less advanced students up to speed. With a frequency dictionary, however, the teacher could assign students to work through the list and fill in gaps in their vocabulary, and they would know that the students are using their time in the most effective way possible.

The Routledge Frequency Dictionary of Portuguese: Core Vocabulary for Learners (Davies and Preto-Bay, 2007) is designed to meet the needs of a wide range of language students and teachers. The main index contains the five thousand most common words in Portuguese, starting with such basic words as *o* and *de*, and quickly progressing through to more intermediate and advanced words. Because the dictionary is based on the actual frequency of words in a large 20 million word corpus (collection of texts) of many different types of Portuguese texts (fiction, non-fiction, and actual conversations), the user can feel comfortable that these are words that one is very likely to subsequently encounter in the "real world".

The following information is given for each of the 5000 entries in the dictionary:

rank frequency (1, 2, 3, ...), headword, part of speech, English equivalent, dialect, sample sentence, translation, range count, raw frequency total, indication of major register variation

As a concrete example, let us look at the entry for *bruxa* "witch":

4522 *bruxa* nf *witch*
A caça às bruxas é muitas vezes acompanhada de histeria – Witch hunts are often accompanied by hysteria
35 | 235 –ac

This entry shows that word number 4522 in the rank order list is [bruxa], which is a feminine noun [nf] that can be translated as "witch" in English. We then see an actual sentence or phrase from the corpus, which shows the word in context, as well as a translation of this sentence into English. The two following numbers show that the word occurs in 35 of the 100 equally-sized blocks from the corpus (i.e. the range count), and that this lemma occurs 235 times in the corpus. Finally, the notation [–ac] indicates that the word is much more common in the fiction

register than would otherwise be expected. In summary, then, each of the 5000 entries provides the language learners with information about the frequency of the word, its meaning (via the glosses and the sample sentence), and some indication of the distribution of the word in the different genres.

One of the criticisms of frequency dictionaries is that they are just "lists of words", and that there is no semantic grouping of any of the words. To address this criticism in part, we placed throughout the main frequency-based index are approximately thirty "call-out boxes", which serve to display in one list a number of thematically-related words. These include lists of words related to the body, food, family, weather, professions, nationalities, colors, emotions, verbs of movement and communication, and several other semantic domains.

In addition to vocabulary that is tied to a particular semantic category, however, we also focused on several topics in Portuguese grammar that are often difficult for beginning and intermediate students. For example, there are lists that show the most common diminutives, superlatives, and derivational suffixes to form nouns, the most common verbs and adjectives that take the subjunctive, which verbs most often take the "reflexive marker" *se*, which verbs most often occur almost exclusively in the imperfect and preterit, and which adjectives occur almost exclusively with the two copular verbs *ser* and *estar* or the semi-copular *ficar*. Finally, there are even more advanced lists that compare the use of nouns, verbs, adjectives, and adverbs across registers, and show which words are used primarily in spoken, fiction, newspapers, or academic texts. Related to this is a list showing which are the most frequent words that have entered the language in the past 100-200 years.

Aside from the main frequency listing, there are also indexes that sort the entries by alphabetical order and part of speech. The alphabetical index can be of great value to students who for example want to look up a word from a short story or newspaper article, and see how common the word is in general. The part of speech indexes could be of benefit to students who want to focus selectively on verbs, nouns, or some other part of speech. Finally, there are a number of thematically-related lists and lists related to common grammatical problems for beginning and intermediate students, all of which should enhance the learning experience. The expectation, then, is that this frequency dictionary will significantly maximize the efforts of a wide range of students and teachers who are involved in the acquisition and teaching of Portuguese vocabulary.

In summary, all of these features of the corpus-based frequency dictionary, as well as the *Corpus do Português* itself, represent linguistic tools that can greatly facilitate the learning of Portuguese by speakers of other languages.

References

Brown, Charles Barrett. 1951. *Brazilian Portuguese idiom list, selected on the basis of range and frequency of occurrence*. Nashville: Vanderbilt University Press

Davies, M., and A. Raposo Preto-Bay. 2007. *A Frequency Dictionary of Portuguese : Core Vocabulary for Learners*. London : Routledge.

Duncan, J.C. 1972. *A Frequency Dictionary of Portuguese Words*. Unpublished dissertation. Stanford University.

Kelly, J.R. 1970. "A computational frequency and range list of five hundred Brazilian Portuguese words". *Luso-Brazilian Review* 7:104-13

Nascimento, M. Bacelar do, et al. 1987. *Portugues Fundamental. Metodos e Documentos*. Lisbon, INIC.

Roche, Jean. (1975) *Sobre o vocabulário da poesia portuguesa*. Paris : Fundação Calouste Gulbenkian