A 61 Million Word Corpus of Brazilian Portuguese Film Subtitles as a Resource for Linguistic Research [*]

Kevin Tang

Abstract

This work documents the motivation and development of a subtitle-based corpus for Brazilian Portuguese, SUBTLEX-PT-BR, available at `http://crr.ugent.be/subtlex-pt-br/`. While the target language was Brazilian Portuguese, the methodology can be extended to any other languages with subtitles. A preliminary corpus comparison with a large conversational and written corpus was conducted to evaluate the validity of the corpus, and suggested that the subtitle corpus is more similar to the conversational than the written language. Future work on the methodology and the corpus itself is outlined. Its diverse use as a resource for linguistic research is discussed.

Keywords: corpus linguistics, subtitles, frequency, Brazilian Portuguese

1 Introduction

A linguistic corpus is a sample of a large population of a language. Traditionally, printed texts were used as they are a readily available sample of languages. Now with the increasing availability of digitally stored texts accessible via the Internet, corpus compiling is infinitely easier, and this has spawned new opportunities for corpus compilation, such as the Google n-gram corpus (Michel et al., 2011) – a digitalization of millions of books; Sharoff's internet corpus (Sharoff, 2006) which mines random stretches of connected text from the internet, and others; therefore written corpora are no longer hard to come by. A much more challenging problem is to find conversational word corpora. An ideal conversational corpus would record people's everyday life and their interactions in different situations and environment, but sadly these recordings would have to be transcribed by hand as speech recognition tools are not accurate enough. This would make the corpus very expensive to construct due to high labour costs for transcriptions. Having said that, there are now websites hosting and sharing films and TV subtitles of different languages, and since subtitles are essentially transcribed conversational text, they make a good and inexpensive source of conversational data.

In linguistics, these kinds of corpora provide us with a representation of language use, exposure and the lexicon of the people as a whole. Taking as an example one measurement in particular, token frequency plays a major role in linguistic literatures. For instance, token frequency is widely claimed to be a conditioning factor in alternations in the lexicon (Bybee, 1995, 2003; Huback, 2007; Coetzee & Kawahara, in press, 2013): lexical items that are used more often are more likely to undergo phonological processes. The motivation for building a corpus stemmed from our investigation in the case of the plural morphology of Brazilian Portuguese.

Brazilian Portuguese nouns ending in -Vw, e.g. [ˈsaw] ‘salt’, [ʒuɾ.ˈnaw] ‘newspaper’, [ˈmɛw] ‘honey’, [ˈsɛw] ‘sky’, [ʃɐ.ˈpɛw] ‘hat’, often, but not always, show an alternation in the plural, whereby the final glide becomes palatal, e.g. [ˈsajʃ], [ʒuɾ.ˈnajʃ], [ˈmɛwʃ], [ˈsɛwʃ] and [ʃɐ.ˈpɛjʃ] / [ʃɐ.ˈpɛwʃ]. In Becker, Clemens, and Nevins (2012) it is proposed that the con-

[*] I hereby thank Andrew Nevins for his encouragement and idea to construct such a corpus and Marc Brysbaert and Emmanuel Keuleers for their support and advice on creating this corpus, without whom, none of this would have been possible.

Name	Size	Licence	Language	Type
Corpus Brasileiro (Sardinha, 2009)	1 billion	Free	Brazilian	Written (90%) and Conversational (10%): text and transcripts of talk both online and offline
Corpus do Português (Davies & Ferreira, 2006)	45 million	Commercial	European and Brazilian	Written and Conversational
C-ORAL-BRASIL (Raso & Mello, 2012)	0.21 million	Commercial	Brazilian	Conversational: informal spontaneous speech

Table 1: Existing major Brazil Portuguese corpora

ditioning factor determining whether a noun will participate in such alternations is prosodic: monosyllables are preferentially protected, a trend confirmed in large-scale nonce word tasks. This motivated us to test whether token frequency would also be a predictor of alternation rates for existing words in the lexicon, given the claim in Huback (2007) that frequent words alternate more often in Brazilian Portuguese. To test this claim, a frequency corpus of Brazilian Portuguese was needed.[1]

After exploring the availability of corpora (a few major existing Brazil Portuguese corpora are listed in Table 1), it was found that although there were many existing corpora, they were either not free (Davies & Ferreira, 2006), were predominantly based on written registers (Sardinha, 2009) which is not ideal for psycholinguistic research, or too small to be representative of the language (Raso & Mello, 2012). The lack of a suitable corpus motivated us to construct our own corpus of conversational Brazilian Portuguese. A method of using film subtitles to construct frequency corpora, SUBTLEX, was developed by New, Brysbaert, Veronis, and Pallier, 2007 with French, and subsequently used for English (Brysbaert & New, 2009), Dutch (Keuleers, Brysbaert, & New, 2010) and many other languages. Crucially these SUBTLEX film subtitle frequencies have been proven to be excellent predictors of behavioural task measures. Following this method, we set out to compile a subtitle frequency corpus for Brazilian Portuguese.

2 Method

2.1 Collecting a corpus of subtitles

Subtitles are often provided alongside digital films, and are either written by professionals or by volunteers. Two different files are needed, the film file and the subtitle file which is played in synchronization with film files. There has been a trend of volunteers, often fans of certain TV series or films, creating subtitles either by translating the original subtitles or transcribing videos into their native languages. Numerous websites can now be found where these subtitle files are being uploaded every day and the number is only going to increase as more films and TV series come out.

One of the biggest subtitle websites, *Opensubtitles.org* which contains over 1,900,000 subtitles as of December 2012, was datamined for the present study. Websites such as this one

[1] The analyses of the alternation in Brazilian Portuguese are not presented in this paper. We found that the correlation between token frequency and alternations are epiphenomenal, and in fact depend on prosodic shape.

often impose a download limit per IP address per day to better manage their servers where these files are hosted, thereby immediately making the manual harvesting of the materials impossible. This limit was circumvented by changing IP addresses through public proxy servers. 25,303 zip files with films/TV series subtitles were self-identified as Brazilian Portuguese and were downloaded alphabetically.

2.2 Post-processing

2.2.1 Preliminary selection and cleaning. Each downloaded zip file contains one or two subtitle files (in different formats, e.g. .srt, .sub, and in multiple discs) and an information file. There are different kinds of subtitle formats, e.g. .srt, .sub, .txt etc. Only *.srt files were used, since they are the overwhelming majority, and this could also avoid potential duplicates from the different formats of the same subtitle files. This left us with 26,627 *.srt files.

The information file contains metadata about the subtitle files, such as the name of the author/uploader, filename, number of downloads, language, format, number of subtitle files, upload date, and release name. Some of this information could potentially be useful but due the diversity of its format, this was hard to extract, and therefore was not used in compiling the corpus.

Having selected the *.srt files, their format was explored in order to *clean* the file. A sample of what a raw subtitle file contains is shown below. It contains irrelevant information such as the subtitle line number and time indications, name and e-mail addresses of the translator and others.

1 00:00:01,820 –> 00:00:03,684 No século 23,...
2 00:00:03,685 –> 00:00:09,650 ... a cidade lunar Éden é o último refúgio da Humanidade.
3 00:00:10,110 –> 00:00:13,650 Um paraíso onde os últimos humanos...
4 00:00:13,651 –> 00:00:16,570 ... levam uma vida pacífica e sem alterações.
...
337 00:23:31,515 –> 00:23:34,015 Legendas: TheLoneGunners
338 00:23:34,016 –> 00:23:36,516 Garibada: MRRG
339 00:23:37,000 –> 00:23:40,142 www.OpenSubtitles.org

As we can see, some of this irrelevant information can be removed automatically such as the subtitle line numbers and time indications, but the rest has to be removed manually as there are no obvious indicators. Due to the sheer number of files we collected, this manual cleaning was not performed.

2.2.2 Removing duplicates. Perhaps the most important part of the post-processing is to remove duplicates, which could otherwise skew the overall frequency purely because there were more versions of subtitles available for popular films; therefore great care has been taken. Identifying duplicates is an interesting computational problem because these duplicates are actually not exact duplicates, but near-duplicates. One of the main causes is due to multiple translators translating the same film, therefore certain words, phrases and expressions could be translated differently while the majority of the vocabulary would remain the same.

Two techniques were employed to tackle this problem: 1) Kullback-Leibler divergence (D_{KL}) (Kullback & Leibler, 1951), a non-symmetric measure of the difference between two probability distributions. 2) K-means clustering algorithm (MacQueen et al., 1967), one of the simplest unsupervised learning algorithms that can perform clustering.

Firstly, the word frequency distributions were obtained for each *.srt file. Secondly, a pair-wise comparison was performed using Kullback-Leibler divergence (D_{KL}) on the distributions of each file against those of the rest of the files, e.g. for the first file to be checked for duplicates, 26,626 D_{KL} values were calculated. In general, duplicates will have very low D_{KL}. The cut-off D_{KL} value for being duplicates was manually calibrated and validated with manual inspections, to ensure that files were not misidentified as duplicates in many small scaled trials.

Amongst the duplicates, it was found that often duplicates could come in different forms due to 1) different translators (as mentioned above), 2) different release formats: a single release (1-disc) or in multiple parts (multiple-discs). To tackle this, firstly, the overall word frequencies of each of these duplicates were calculated and used to sort the duplicates into two clusters using the K-means clustering algorithm. This would create a cluster of files with higher frequencies and a cluster of files with lower frequencies. Secondly, the higher frequency group was chosen to capture the 1-disc file over the multiple-disc files, and if only the multiple-disc files were available, it would maximize the size of our corpus by picking the largest disc. Within the high frequency group, the file with the shortest tail (fewest words with a frequency of one) was then categorized as unique, as it was likely to contain fewer typos and other errors. The rest of duplicated files were then removed. Overall, 12,353 unique files were identified.

2.2.3 Removing Non-Portuguese files. Although the downloaded subtitle files were self-identified as Brazilian Portuguese, errors are often made by the uploaders who would sometimes include the original subtitles of the films with the translated versions in the zip package.

To filter out these non-Brazilian Portuguese files, we employed a language detection model (Shuyo, 2010). The model calculates language probabilities from features of spelling using a naïve Bayesian model with character n-gram, using language profiles generated from Wikipedia abstracts. It has an above-99% precision for 53 languages; however it does not distinguish between European and Brazilian Portuguese, as currently Wikipedia does not have that distinction.

The model filtered out 249 files, the remaining 12,104 files have a probability of at least 99.7% of being in Portuguese. Interestingly, some of the filtered files have two dominant probabilities, e.g. 80% Portuguese and 20% English. These are often subtitles with words which are not translated, e.g. sung lines in musicals. Such files are excluded from the final corpus.

2.3 Compiling and filtering

The 12,104 files were compiled to produce a corpus with a list of words, their frequency and contextual diversity (CD) (a measure of the number of subtitle files that a word has occurred in). A few filters were used to further clean our corpus:

- Filter out some web URLs and e-mail addresses.
- Filter out words that do not consist only of Brazilian Portuguese graphemes “áâãàéêíóôõúçüabcdefghijklmnopqrstuvwxyz”
- Filter out words with CD of 2 and below

Finally, this yielded a corpus with 61,609,241 tokens and 136,147 types.

3 Analysis: Validity with databases of conversational and written frequencies

Subtitles should in principle show a wide range of tenses, persons, and speech act types in dialogue, and therefore be more similar to a conversational corpus than a written corpus. In order to validate the subtitle corpus, two comparisons were made between subtitle frequencies and the frequencies from a conversational corpus and a written corpus. Corpus Brasileiro (Sardinha, 2009) was chosen as a test case as it is one of the largest existing corpora of Brazilian Portuguese. It is a collection of about a billion words, of which 9% is conversational, and the rest is written. The conversational subcorpus contains 83 million words, and the written subcorpus contains 750 million words. After removing non-alphabetic characters, such as full-stops and other non-word symbols, the conversational subcorpus left with 67 million word tokens remained with approximately 230,000 word types, and the written subcorpus left with 485 million word tokens with approximately 780,000 word types.

In the comparison with the conversational part, there were about 80,000 entries common to the subtitle corpus and Corpus Brasileiro. The Pearson's correlation between the two sets of frequencies (both log transformed) was .54, with a p-value < 0.0001***. Similarly, in the comparison with the written part, there were about 100,000 entries in common with a Pearson's correlation of 0.50 and a p-value < 0.0001***. Finally, the significance of the difference between two correlation coefficients was tested using the Fisher r-to-z transformation, it was found that $z = 10.63$, with a p-value < 0.0001***.

While the subtitle corpus significantly correlated with both the conversational and written corpora, the correlations were still significantly different from each other, which suggests that the subtitle corpus is significantly more similar to a conversational corpus than a written one in terms of global correlations.

4 Beyond token frequency

In this section, the potential roles a subtitle corpus can play as a linguistic resource, beyond token frequency, are discussed and aspects of the corpus and methodology which could be improved are highlighted.

4.1 Contextual diversity

As mentioned previously, contextual diversity is the number of contexts in which a word has been seen. In our corpus, the contexts would be films. A few studies have suggested that contextual diversity is better than word frequency in capturing word-naming and lexical decision times in terms of capturing more variances (Adelman, Brown, & Quesada, 2006; Brysbaert & New, 2009). To date, this has not been widely used in linguistics which currently prefers the use of word frequency (Bybee, 1995, 2003; Huback, 2007; Coetzee & Kawahara, in press, 2013). It would be beneficial to examine the use of this dimension in linguistic work.

4.2 N-grams, tagging, transcription, and metadata

The corpus could be beneficial for syntactic, morphological, phonological, and discourse analyses with the right processing. Due to copyright reasons, the full text of corpora cannot be made available verbatim. However by analysing which words occur most frequently with others, the limitations of not having the full text would mostly be lifted, and useful information on word meaning and usage could be found. The implementation of this could be done on-line using an web interface where users could submit their enquiries directly and the information would be computed and returned without exposing the full corpus text.

Additional information could be computed and added to the corpus. Firstly, Part of Speech (POS) and morphological tagging would provide more fine-grained distinctions between certain homographs and type and lemma frequencies. Secondly, phonetic transcriptions and syllabification could provide new insights into the context-sensitive distributions of sounds on a sentence-level, rather than in isolation. Lastly, corpus linguistics places great emphasis on specifying the range and distribution of the content. We can add in additional dimensions using the metadata in the information files which came with the subtitles; one could also control for genres, release dates of the films, original language and so forth.

4.3 Near-duplicates for semantics and translation studies

The complexity of translating subtitles is being researched from many angles, such as cross-cultural pragmatics (Bruti, 2006) and translation theories in general (Gottlieb, 2005). As mentioned, near-duplicates are mainly different translations of the same film. Such files would normally be removed for corpus compilation purposes, but they could potentially be an invaluable resource for those who work in translations, as they provide a wealth of translation samples of spontaneous speech. Furthermore, these variations in translations could potentially be used to develop semantic nets and improve machine translations (Fellbaum, 2010).

4.4 Technical improvements for automation and quality

A few technical aspects of the corpus creation need to be improved for full automation and better quality. Firstly, current cut-off points for Kullback-Leibler divergence are manually calibrated, which raises the risk of potentially removing too many or too few duplicates, as well as being manually exhausting. Secondly, distinct language modules for European and Brazilian Portuguese are needed for the language detection model to separate out any potential European Portuguese files. Thirdly, a strategy will be needed for removing translators' names and e-mail addresses, subtitle source websites and irrelevant information automatically. Lastly, the near-duplicate detection stage is the longest process in the entire compilation of the corpus. This is due to the fact that a pairwise comparison of files is not computationally efficient given a large number of files.

5 Conclusion

This study has outlined the methodology for constructing a subtitle corpus. A preliminary analysis of the corpus suggested that it represents a conversational corpus more than a written one. Different dimensions of the corpus, such as contextual diversity, collocations, syntactic and morphological tagging, were discussed. The potential uses of the by-product of the corpus, near-duplicates, were also suggested. A few technical improvements were highlighted for future work on a fully automated system for creating similar corpora.

Our creation of a very large subtitle corpus for Brazilian Portuguese, openly available and in a standardized format, will remain accessible as a potentially valuable resource for a number of researchers in adjacent fields. Different versions of the corpus with an interactive interface are available at `http://crr.ugent.be/subtlex-pt-br/`. With the growing number of films and TV-series coming out and being translated, many more releases of the corpus will become available, and will become increasingly representative of the language over time.

References

Adelman, J., Brown, G., & Quesada, J. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823.

Becker, M., Clemens, L., & Nevins, A. (2012). A richer model is not always more accurate: the case of French and Portuguese plurals. *Lingbuzz.* Retrieved from `http://ling.auf.net/lingBuzz/001336`

Bruti, S. (2006). Cross-cultural pragmatics: the translation of implicit compliments in subtitles. *The Journal of Specialised Translation*, *6*, 185–197.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, *10*(5), 425–455.

Bybee, J. (2003). *Phonology and language use.* Cambridge University Press.

Coetzee, A., & Kawahara, S. (in press, 2013). Frequency biases in phonological variation. *Natural Language and Linguistic Theory*, *31*(1).

Davies, M., & Ferreira, M. (2006). Corpus do português (45 million words, 1300s-1900s). *Available on-line at http://www. corpusdoportugues. org.*

Fellbaum, C. (2010). Translating with a semantic net: matching words and concepts. *Meaning in Translation*, *19*, 255.

Gottlieb, H. (2005). Multidimensional translation: semantics turned semiotics. *Challenges of Multidimensional Translation*, 33–61.

Huback, A. (2007). *Efeitos de freqüência nas representações mentais.* (Doctoral dissertation, Universidade Federal de Minas Gerais, Belo Horizonte).

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: a new measure for Dutch word frequency based on film subtitles. *Behavior research methods*, *42*(3), 643–650.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, *281-297*, p. 14). California, USA.

Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., … Orwant, J. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661.

Raso, T., & Mello, H. (2012). The c-oral-brasil i: reference corpus for informal spoken brazilian portuguese. *Computational Processing of the Portuguese Language*, 362–367.

Sardinha, T. (2009). The brazilian corpus: a one-billion word online resource. In *Proceedings of the corpus linguistics conference 2009 (cl2009),* (p. 88).

Sharoff, S. (2006). Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, *11*(4), 435–462.

Shuyo, N. (2010). *Language detection library for Java.* Retrieved from `http://code.google.com/p/language-detection/`