Cross-Linguistic Word Frequency Visualization for PT and EN

Mariana Shimabukuro[*] Christopher Collins[†]

Ontario Tech University

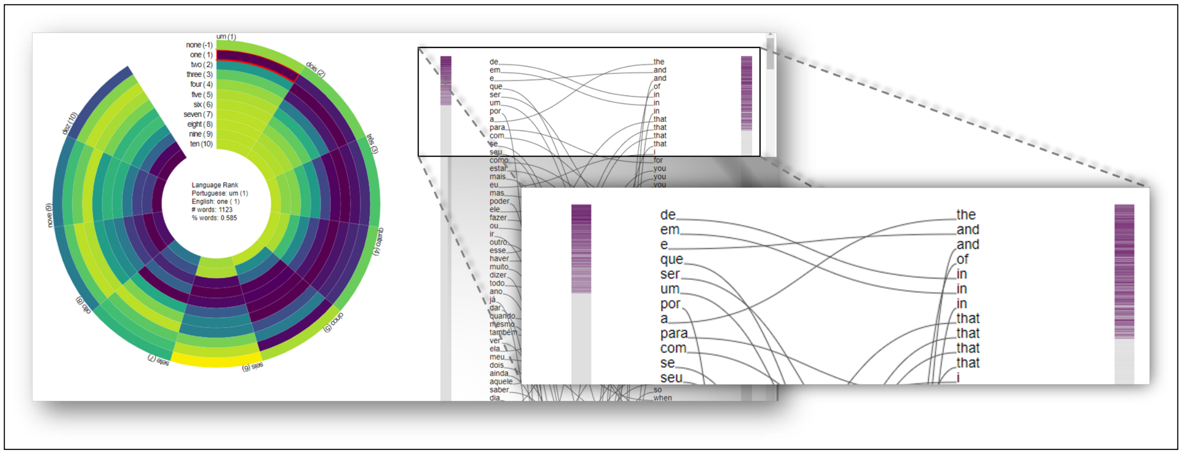

Figure 1: Circular heatmap and zoomed rank exploration view. The heatmap shows the distribution of words between Portuguese (PT) and their English (EN) translations in both language ranks. The side view shows words from a chosen cell and their rank location in the bars on the right (PT) and left (EN) side, e.g., the word “that” appears 4 times, indicating PT speakers tend to overuse “that”.

Abstract

In this project, we present a visualization for investigating the relationship between commonly used words in Portuguese and their translations in English. This cross-linguistic analysis can help us to understand English word choices made by Portuguese native speakers and the influence of language transfer effects. In this paper, we discuss how word frequency is commonly used as a resource for both textual and cross-linguistic analysis. Moreover, we briefly explain the data processing pipeline building on machine translation and word frequencies from large corpora. This research reveals interesting open questions related to linguistic visualizations and future directions for investigating language transfer effects.

Keywords: Linguistic visualization, language learning, cross-linguistic features.

1 Introduction

In the context of second language acquisition (SLA), English as a second language (ESL) and cross-linguistic effects (also known as transfer effects) can be measured using word frequency analysis. Word frequencies can be used to investigate over/under produced terms, word choices, and collocations in a target language and are often classified as transfer effects [4, 6–8]. James divides transfer effects into 2 broad categories [6]:

1. Grammatical errors: where clear grammatical rules are broken. Ex.: “I will *driven* to the airport next week.”
2. Acceptability errors: satisfies grammatical rules, but it is uncommon. Usually the sentence would sound odd, and after a split second a native speaker would understand that it is acceptable. This category includes errors related to over/underuse of words, word choice/form or collocation. Ex.: “This a strong/*powerful* tea.”, “I feel very *confusing* this morning.” or “I *ate* a medicine pill.”

Acceptability errors can be a source of embarrassment for non-native English speakers, and often come from the differences in word frequency across languages. For example, the word “polemic” is considered formal and rarely used in English, whereas its Portuguese translation, “polêmico,” is a very common word. Thus, a Portuguese speaker may choose to use the word “polemic” in English conversation, where “controversial” would be a more natural-sounding choice. We aim to create a visual interface to expose these distributional differences which may underlie acceptability errors.

Word frequency data is widely used in textual and linguistic analyses. Some of the common tasks that can be supported by word frequency analysis are: sorting based on relevance, clustering documents by topic, document summarization, document overview generation, and search recommendations [1, 3]. If we focus on language use, word frequency can be indicative of words that are used colloquially. For this reason, word frequency is also an important feature for studying language acquisition, such as early childhood, lexico-grammatical, and second language acquisition [2, 5].

This paper describes the data processing and design of a visualization that addresses the need for comparing word usage between two languages. We discuss future directions and challenges faced while building this tool and investigating word frequency as possible cause to acceptability transfer effect errors made by Portuguese (PT) native speakers (L1) into English (EN) as a second language (L2).

2 Methodology

This work builds on the work of Kochmar [7], which exposed word choice errors of English learners by identifying uncommon adjective-noun (AN) and verb-object (VO) combinations inherited from their L1 model into English. Kochmar’s model used the frequency of AN/VO combinations in both English and the L1‘s, along with the translation of L1‘s AN/VO combinations into English using Google Translate and dictionaries. We used this approach in the context of word usage between Portuguese as L2 and English as L2.

[*] e-mail: MarianaAkemi.Shimabukuro@uoit.ca
[†] e-mail: Christopher.Collins@uoit.ca

Table 1: PT_{EN} word list generated after translating the PT_{rank} list and mapping the EN_rank. This is a sample of the data used for the visualization.

		PT_{EN}		
PT_rank	**Word**	**Translation**	**EN_rank**	**N_rank**
1	o	o	5809	3438
2	de	in	5	4
3	em	in	5	4
...	...	...	...	...
21	este	this one	-1	-1
...	...	...	...	...
19200	voluntarismo	voluntarism	80097	12642

2.1 Data

We used two word frequency lists: (1) the **Corpus of Contemporary American English (COCA)** with the 500,000 most used terms extracted from 500 million words in documents dated between 1990 to 2017[1]; (2) the **Corpus of Portuguese/*Português*** word frequency list of the top 20,000 most used terms extracted from 20 million words from the 1900s[2].

2.2 Data Processing

The data processing steps we used to obtain the data are as follows:

1. Generate unique frequency ranks by ignoring the PoS tags in the word lists and merging their frequencies.
2. Use Google Translate API to translate each PT word into EN.
3. Map the translated words to their EN ranks.
4. Re-index the mapped EN ranks by ordering the translations using the *EN_rank* and restarting the new positions (*N_rank*) as shown in Table 1.

Step 4 was performed in order to reduce the gap difference between the rank range of the PT and EN words, 1-20k and 1-367k respectively. This data normalization was needed to improve the visual representation. In the end, we obtained 17,864 words (over 5K *EN_rank* duplicates) with *N_rank* ranging from 0 to approximate 12,000, 1,336 words in total were not found (*EN_rank* = -1).

3 Visualization Design

This section discusses the visualization and prototype design including two main views: a circular heatmap and a zoomed view.

3.1 Overview: Circular Heatmap

This overview visualization displays the distribution of words across the different rank levels between both languages in a circular heatmap shown in the left side of Figure 1. The slices each represent 1/10 of the words from the PT rank. The rings each represent 1/10 of the English rank range. Each cell is color-coded with the number of words contained in the intersection of the PT and EN rank range. The outer ring represents the translated words not found in the EN rank, ($EN_{rank} = -1$).

The heatmap in Figure 1 shows the data distributed along the spiral, which is equivalent to the diagonal of a rectangular matrix. The frequency was encoded and normalized per slice. On hover, we display more information about the cell overlaid on the center of the heatmap. When a cell is clicked the PT words and their translations are populated into the zoomed view to perform a lower level exploration (right side of Figure 1).

3.2 Zoomed View

This view has two main components: word lists and rank side bars. In the center there is a two column list of the selected PT (left) and their translated EN (right) words. Both columns are ordered by rank, respectively. Words are connected by a link indicating the pair. Beside the word lists, on the left we have a bar representing the PT rank and on the right, the EN rank bar.

These rank side bars encode all the PT words and EN translations positioned vertically and sorted by their respective ranks in each language. Each word is encoded by a short rectangle that altogether form the bars. The color of the rectangle represents the color of the cell that this word belongs to.

When the user interacts with the overview heatmap as shown in Figure 1, the pair of words and translations are displayed in the word lists. On top of that, all the words that do not belong to the current selection are filtered out and greyed out in the bars. This filter helps to put the the rank difference into perspective by focusing on the location of the words in the full rank represented by the side bars. This analysis is not possible by only using the word lists themselves.

4 Discussion and Future Work

We presented a prototype of possible ideas for visualizing word usage differences between a source and target language. Future work includes adding more filters (cell, slice and ring), and sorting options for the side bars and word pairs. A search feature would allow us to investigate the ranking position of specific words.

Besides visual improvements, the translation step can also be improved. The Google Translate API does not provide multiple translations, but our tool would benefit greatly from this missing information. In the future, we plan on using dictionary APIs instead of MT. Dictionaries can give multiple translations, which we will incorporate into the visualization design.

Another translation issue is inappropriate translations, in Table 1, the PT word "o" which means "the" in EN, was translated into the letter "o" instead. This mistranslation caused a gap that should not exist between the ranks. In Kochmar's work, machine translation worked best because she was translating collocations as in verb-object and adjective-noun combinations [7]. Multiple words instead of singles probably favors MT over dictionary fetching.

Even though we started with PT and EN, this visualization can be generalized to any language pair. As future work we will also encode semantic context using Word2Vec models to help learners decide on which translation to pick based on their specific scenario.

This project investigated the proposed methodology and we could both understand the problem and other possible solutions for analyzing the cross-linguistic effects between Portuguese and English. However, at this moment, we were not able to confirm the hypothesis of high frequent PT words translated into infrequent EN words are related to word choice/usage mistakes. For this, we plan an evaluation with English language learners.

References

[1] A. B. Alencar, M. C. F. de Oliveira, and F. V. Paulovich. Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):476–492, 2012.

[2] J. L. Bybee and P. J. Hopper. *Frequency and the emergence of linguistic structure*, vol. 45. John Benjamins Publishing, 2001.

[3] C. Collins, F. B. Viegas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *VAST, 2009.*

[4] A. A. Dinar. *Influence of the Semantic Aspects of Mother Tongue on Learning English as a Foreign Language*. PhD thesis, Sudan University of Science and Technology, 2016.

[5] T. Huckin et al. Second language reading and vocabulary learning. 1995.

[6] C. James. *Errors in language learning and use: Exploring error analysis*. Routledge, 2013.

[7] E. Kochmar. Error detection in content word combinations. Technical report, University of Cambridge, Computer Laboratory, 2016.

[8] M. Xhemaili. Literature review on the role of mother tongue in learning and teaching english for specific purpose. *J. Education and Practice*, 4(18):39–43, 2013.

[1] https://corpus.byu.edu/coca/

[2] https://www.corpusdoportugues.org/