Max Xie, Colin Hirschberg, Yuma Yamada, Sebastian Bissiri

LING 144

6 March 2022

Part of Speech of Code-Switched words between English and Mandarin

Introduction

In this project, we will be analyzing datasets in which code switching between English and Mandarin are used. The data includes instances of both directions of code-switching, from English to Mandarin and visa versa. We will cleave up each sentence into entirely English fragments and entirely Mandarin fragments in order to analyze the parts of speech of the initial switched words (i.e., the words that immediately follow code switching boundaries). More specifically, we are concerned with the words posterior to switches from the matrix language, which refers to the grammatical frame-determining language, into the embedded language (or inserted language), not words following switches back to the matrix language. Thus, presuming that the dominant language of the sentence coincides with the matrix language, the extent of our part of speech analysis encompasses only English inserted fragment-initial words in the Mandarin-dominated sentences and Mandarin inserted fragment-initial words in the English-dominated sentences. We will compute the frequencies of English nouns, adjectives, and verbs at this position for the Mandarin-dominated dataset and the frequencies of Mandarin nouns, adjectives, and verbs at this position for the

English-dominated dataset. This will lay the foundation for a comparison of part of speech-frequencies among initial switched words. This process will attempt to answer the questions hereafter listed.

Questions and Hypothesis

Formally, the question our research poses are the following:

1) At what parts of speech (a.k.a., lexical category) does Mandarin-English code-switching usually occur? More precisely, what parts of speech are the most frequent among words that immediately follow Mandarin-English code-switching points from the matrix language to the embedded language?

2) Does the identity of the matrix language or dominant language have any impact on the lexical

Our hypotheses, based on the background outlined in the following section, is as followed:

1) Code-switching should occur equally as often regardless of the part of speech of the switched word.

2) The matrix language has no bearing on the part-of-speech frequency distribution of the switched words that start embedded language fragments. This second hypothesis anticipates the same Mandarin-dominated results and English-dominated results.

Background

Code-switching is an interesting subject, and we decided to investigate this topic because it would allow us to try processing data in two different languages, which is a challenge that hasn't been set forth to us by our class as of yet. Not only that, but Chinese and English have completely different writing systems and characters, so using Python to filter and annotate this data, we thought, could be an interesting challenge. On top of that, our group member Max is a bilingual Mandarin and English speaker, so he has a special interest in this topic which contributed to us choosing to make it our topic.

Max had the intuition that nouns and adjectives would reign as the dominant part of speech in terms of representation compared to verbs in the code-switched initial word, based on his own experience as a code-switching speaker.

However, one source seems to indicate the opposite. An article called "A Comparison of Noun and Verb Retrieval" took twenty-one Mandarin-English bilinguals who considered themselves only native in Mandarin and twenty one English native monolinguals and got them to recall nouns in response to sketches of objects, and got them to recall verbs in response to sketches of actions (Li). Each noun or verb naming task was given a score for accuracy or inaccuracy, and the bilingual speakers scored higher for verbs than nouns in terms of retrieval rate, especially in their non-native language. This informs our hypothesis somewhat in that, perhaps it means that code-switching speakers may actually switch on verbs more often.

As a result of these two somewhat conflicting sources, we decided to make our hypothesis neutral, phrasing it to state that switched nouns, switched adjectives, and switched verbs are equally prevalent and that the matrix language will not influence the proportion of each part of speech in the switched words.

There are obviously many other sources on Mandarin-English code switching in general, but the source we found was the only one we found to be relevant to the questions we are asking in this project, as it relates to part-of-speech in code-switching data.

Dataset:

The data used in this project, two datasets of bilingual Mandarin and English speech, is all retrieved from SEAME-dev-set. SEAME is a Mandarin-English code-switching corpus replete with 192 hours of interview question-facilitated casual conversations. SEAME-dev-set contains two repackaged SEAME-derived test sets of mixed-language speech, one Mandarin-dominated and one Singaporean English-dominated. It should be noted that SEAME only reflects code-switching in Mandarin diasporan communities of Singapore and Malaysia, and this means our data is only retrieved from these populations. Any unexpected discrepancies between part of speech distribution in SEAME-dev-set and the noun and word retrieval experiment could result from differences in populations surveyed. Another possible discrepancy could occur from the fact that the word retrieval in SEAME transpired in conversation, each word being mentally prompted by neighboring words, unlike how word retrieval in the noun and word retrieval experiment occurred in response to pictures.

Data Filtration/Management:

This section will outline the steps we took to format and prepare our data for analysis. We began with one English-dominated test set (dataset1.txt), wherein

presumably Mandarin is the non-native language, and one Mandarin-dominated test set (dataset2.txt), wherein presumably English is the non-native language. These test sets were imported from Github user zengzp0912's repository. We converted both dataset1.txt and dataset2.txt to a csv with a NumberTag column and a sentence column, so they can be processed by the computer easier. We read the column under the header “sentence” into the list dataframe_1 for the dataset1.txt and the column under the header “sentence” into dataframe_2 for the dataset2.txt.

In a preprocessing step, all appearances of the symbol <v-noise>, which demarcates start of a recorded audio segment but is immaterial to code-switching, were filtered out by regex, and opening and closing quotes were eliminated from each sentence string without <v-noise>. For the purposes of part of speech-tagging, sentence strings must jettison their quotes, so that in the subsequent part of speech-tagging stage, the quotes do not erroneously get tagged as nouns, which would interfere with the noun counts. Although time did not permit the implementation of the removal of fillers (“ah,” “hah”, “okay”, or “lah”) in code by the replace() string method or sub() regex function, these will get labeled under a miscellaneous category X that will not tamper with any noun, adjective, or verb counts.

Our methods- Algorithms and Pseudocode:

This project will have 5 difference parts of analysis:

1. Read in datasets:

2. Dependencies import:

3. Sentence tokenization:

4. Part of speech tagging:

5. Graphing data:

1. Read in datasets:

The Google Colab notebook final_project.ipynb accommodates the bulk of our code. An algorithm centered on the ratio of the sum of English word lengths to the string length sorted the sentences from dataset1.txt and dataset2.txt into a set of English-dominated sentences, wherein the ratio exceeded or matched 55 percent, and a set of Mandarin-dominated sentences, wherein the ratio amounted to less than 55 percent. Although dataset1.txt and dataset2.txt purport themselves as English-dominated and Mandarin-dominated, respectively, dataset1.txt features a considerable number of Mandarin-dominated sentences while dataset2.txt is absolutely riddled with English-dominated sentences, so a conditional that compares the English-to-sentence length ratio against 55% is vital to effectively selectively group together English-major sentences as well Mandarin-major sentences. Once each sentence is identified as English-major or Mandarin-major, it is appropriately appended to either the list english_major or the list chinese_major.

2. Dependencies import:

In preparation to wield pandas, nltk (natural language toolkit) for English part of speech-tagging, and zh_core_web_md (a machine learning model geared towards Mandarin) for Mandarin part of speech-tagging, this program installed pandas installed nltk, and crucially installed spacy prior to downloading zh_core_web_md from spacy and loading zh_core_web_md to the Google Colab. Zh_core_web_md proves itself a very accurate word-tokenizer and part of speech-tagger according to Mandarin speaker Max Xie because zh_core_web_md was trained with immense quantities of human-annotated Mandarin text. Regex was also imported to later filter out <v-noise> by regex and discern language boundaries by regex.

3. Sentence tokenization:

A compartmentalizer function compartmentalizes (i.e., divides into a list of monolingual fragments) each sentence from english_major and from chinese_major, iterating through each word in any given sentence element of english_major or chinese_major and relying on regex for the classification of each word as English or Mandarin and for the placement of it in either the same compartment if no code switch precedes it or a new compartment it a code switch precedes it. A word-tokenizing function acts on the list output of the compartmentalizer, breaking down the monolingual fragments each into word tokens, so the word-tokenizing function spews out a list of sentence lists of word token strings and does so for both the English dominated and Mandarin-dominated sentences. In either case, regex was employed again to differentiate between English fragments, which nltk can tokenize, and Mandarin fragments, which spacy can tokenize.

4. Part of speech tagging:

Mathplotlibrary.pyplot

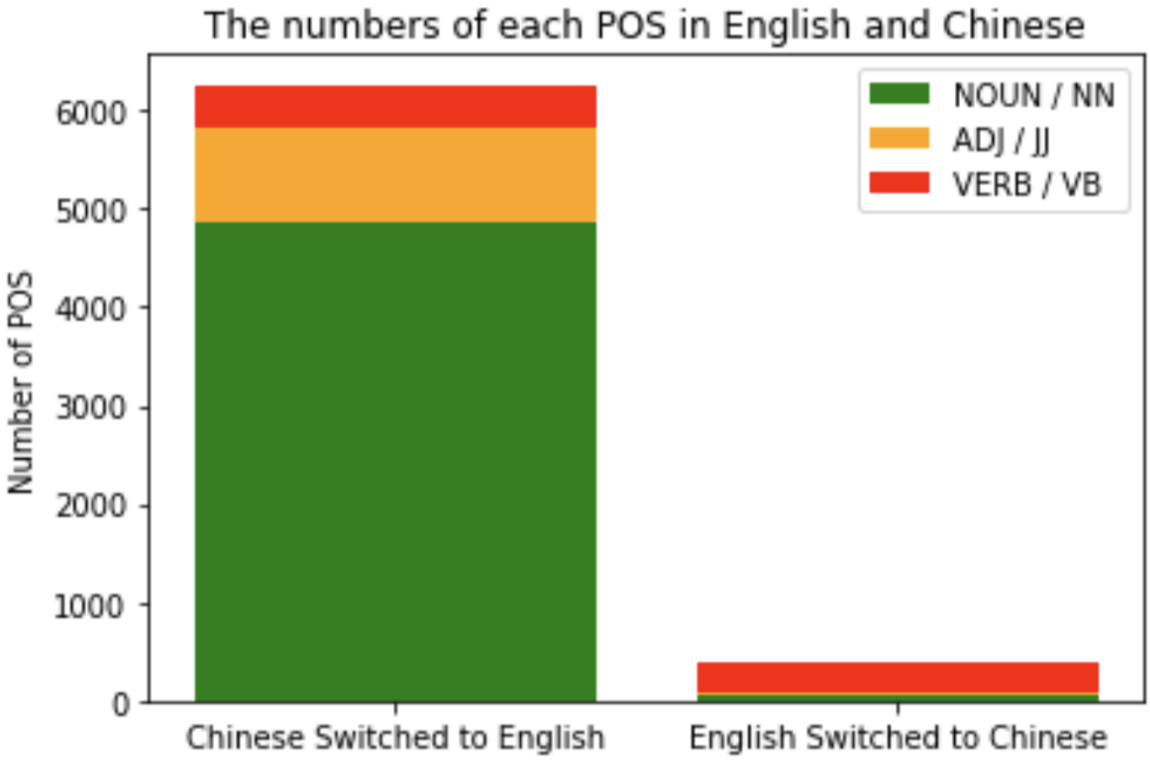
Numpy

A part of speech-tagging function marks, with abbreviated tags symbolic of their lexical categories, the word tokens of the output of word-tokenizing function, effectively giving rise to a list of sentence lists of word token-part of speech tag tuples.. Thus, enacting this function on both the Mandarin-dominated and English-dominated compartmentalized and word-tokenized sentences yielded the final data formats that functions processed to compute the number of initial switched English nouns, adjectives, and verbs as well as the number of initial switched Mandarin nouns, adjectives, and verbs. We designated the output of part of speech-tagging applied to the English-dominated sentence list of fragment lists of tokens as pos_tagged_en_major and the output applied to the Mandarin-dominated counterpart as pos_tagged_ch_major.

5. Graphing data:

Such functions that compute the number of initial switched English and Mandarin words of each part of speech were dubbed eng_number and chin_number. The function eng_number and chin_number were each called pos_tagged_en_major and pos_tagged_ch_major, respectively, to obtain two three-element lists of the numbers of initial switched nouns, adjectives, and verbs arranged in that order. Of course, one list encodes the counts for Mandarin lexical categories while the other signifies the counts for English lexical categories in that particular post-code-switching environment. Plus, only switches from the matrix language to the embedded language (i.e., Mandarin-to-English switches in the Mandarin-dominated dataset and English-to-Mandarin switches in the English-dominated dataset were counted.)

Results:



The outcome is astonishing: not only were the asymmetries between initial switched nouns and verbs within each language staggering, but also matrix language matters. Nouns comprise 4856 out of all 6257 switched English fragment-initial words while verbs account for 304 out of all 386 switched Mandarin fragment-initial words, so the matrix language or, alternatively framed, the language switched into has an impact.

Discussion:

The results partly confirm Max's intuitions that Mandarin speakers disposed to switch nouns to English than verbs when they are speaking Mandarin, but the prevalence of switched verbs over nouns in the Mandarin fragment-initial words in unexplainable and may offer some newfound insights. The picture naming project concluded that bilingual speakers have more immediate access to verbs than nouns, especially in their non-native language, so if Singaporean and Malaysian subjects who speak English more natively than Mandarin uttered those English-major sentences, this could explain it, but we have no information on the language preference of these speakers, only their age and gender demographics.

Conclusion:

While these findings are revelational in the conditioning of code-switching by lexical categories and on the impact of matrix language on the lexical categories of switched words, they only begin to suggest that these code-switching contributors exist. A similar study that computes a part of speech tally for switched fragment- initial words should be carried out, and the English-Mandarin code-switching study should be reconducted on bilingual speakers who are native in one but not the other to better assess the influence of sluggish noun retrieval and rapid verb retrieval rates in non-native languages that the picture naming experiment indicated.

References:

Lyu, D.-C., Tan, T.P., Chng, E.S., Li, H. “Mandarin–Code-Switching Speech Corpus in South-East Asia: SEAME”. *Language Resources and Evaluation*. 49(3). https://doi.org/10.1007/s10579-015-9303-x https://www.isca-speech.org/archive/pdfs/interspeech_2010/lyu10_interspeech.pdf

Li, Y, Faroqi-Shan, Y. Wang, M. (2019). “A Comparison of Noun and Verb Retrieval in Mandarin-English Bilinguals with English Monolinguals”. *Bilingualism: Language and Cognition*, 22(5), 1005-1028.