ASSESSING COLLOCATION BETWEEN 'QUE' AND SPANISH VERBS USING CONTINUOUS BAG OF WORDS

Lily Shi
MehtA+

Arjun Somnali
MehtA+

Phillip Guan
MehtA+

July 28, 2022

ABSTRACT

Throughout the development of the modern language, the use of complementizers, words used to link a subject and a complement clause, has been steadily decreasing. The Spanish complementizer "que", which translates to "that" in English, is one such word that Spanish speakers have been dropping out of their sentences. We attempt to study this phenomenon by performing analyses using a Continuous Bag of Words model (CBOW). We ran experiments with the model on two datasets from older and more recent time periods and developed a method to accurately find instances of que-drop and determine which words precede a que-drop the most often.

1 Introduction

In English and Spanish grammar, complementizers are words used to introduce complement clauses, including subordinate conjunctions, relative pronouns, and relative adjectives. Linguists have found that the use of complementizers has been decreasing steadily throughout the development of the modern language. In this paper, we investigate the decreasing use of the Spanish complementizer "que" in older texts compared to newer texts.

"Que" is usually used after verbs to introduce a description of an action or idea. When this complementizer is ommitted, it creates a *null complementizer*. In these cases, the missing complementizer is assumed to be present. For example, in the sentence "Espero que tengas un buen día" (I hope that you have a good day), the word "que" can be dropped while still retaining grammatical correctness, as in the modified sentence "Espero tengas un buen dia" (I hope you have a good day).

Our goal is to investigate this increase in que-drops.

2 Related Work

There has been some work done in the field of null complementizers. Riccelli (2018) [1] investigated variations of null and overt expressions of "que" between two Spanish dialects. Similar work has been done in other languages, such as Fukuda (2000) [2] and Liang (2022) [3] finding evidence of similar complementizer drops in Japaneses and French, respectively. Tagliamonte Smith (2005) [4] conducted a similar study with British English, finding that-drop in 91% of the cases that was expected. Additionally, Yoon (2015) [5] proposed that the interaction of the verb and the sentence contributed to the likelihood of que-drop. The process of studying the que-drop phenomenon may be tedious, and our research builds on these previous works by proposing a method to make these processes more efficient using machine learning.

3 Methodology

3.1 Dataset

We used two separate datasets: one containing old texts, and the other containing newer sentences. The first dataset consisted of eight Spanish books written in the 19th century that we extracted from Project Gutenberg. The second dataset was made of recent Spanish tweets collected from Twitter archives.

3.2 Preprocessing

Cleaning Dataset After importing the eight books and tweets, we began cleaning up the text. We removed irrelevant text such as footnotes and extra website formatting from the books, and usernames and emojis in the tweets. We then split each dataset into sentences and kept only the sentences that contained a "que". To futher clean the sentences, we removed the punctuation and comverted everything to lowercase.

Text to Tokens The model we used, called a Continous Bag of Words model, requires an input of two words preceeding and two words following the target word in order to predict the target word. We split each sentence into sequences of five words. Then, we could input the first two words and last two words into the model for it to predict the label of the middle word.

To convert the text into dense vectors of numbers, we simply mapped every word to a different index. We iterated through all the texts and made a dictionary containing each different word, giving each word an index starting from one such that the entire vocabulary could be matched to a dense vector of numbers.

Our word-to-index dictionary begins as follows:

'[UNK]': 0

'numérica': 1

'objeto': 2

'justo': 3

'defensiva': 4

We mapped the first value to an unkown word to use when we came across testing words that were not in the training vocabulary.

Using this dictionary, each five-word sequence can be converted into a five-number sequence. For example, the sequence "los hachazos hízome de los" could be changed to [29292, 15199, 20276, 9663, 29292] where the index of "los" is 29292, the index of "hachazos" is 15199, and so on.

3.3 Model

The model we used is called a Continuous-Bag-of-Words model (CBOW). The CBOW is a model architecture used by the Word2Vec model, which is a predictive deep learning based model that captures contextual and semantic similarity between words. Essentially, the Word2Vec is an unsupervised model which can take in massive textual corpora and generate dense word embeddings that represent the vocabulary of all possible words using the process described above. The dimensionality of the vectors created by Word2Vec are are much lower than the dimensionality of the sparse vectors created using traditional Bag of Words models.

The CBOW is one of two model architectures used by the Word2Vec. It is able to predict a target word in a sentence based on the surrounding context words.

Our model first had an embedding layer, which changes the size of the vector from the size of the entire vocabulary to just 100. Next we had two linear layers. The first linear layer had an activation function of ReLu, and the second had an activation function of Softmax.

The loss function punishes a machine learning model when it deviates too much from the desired results. Our loss function was negative log likelihood loss, as defined by:

$$l(\theta) = P(x_t) = -\sum_{i=1}^{n}(y_i \log \hat{y}\theta, i + (1 - y_i) \log(1 - \hat{y}\theta, i)) \tag{1}$$

We used a batch size of 64 and the optimizer Adam to achieve the best results.

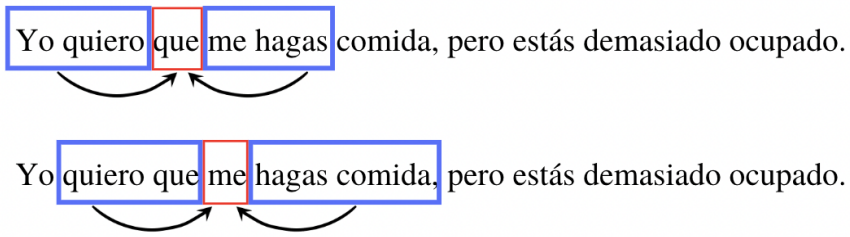

Figure 1: The sentence is split into consecutive 5-word sequences. In each sequence, the model takes in the first two words and the last two words, and predicts the center word. The model trains on the first 5-word sequence, shifts one word to the right, and keeps repeating this process until it reaches the end of the sentence.

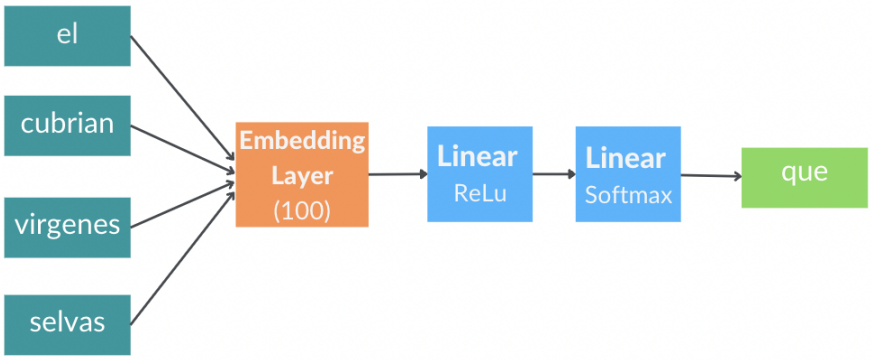

Figure 2: Our model takes in an input of four words, then runs it through an embedding layer and two linear layers.

4 Results

We tested our model on a dataset of recent Spanish tweets. Instead of predicting words that were already in the sentence, we predicted words that were not in the sentence. By doing this, we could find instances where “que” could have been used, but was not used. For example, the sentence "Ellos lamentan no estés muy contenta" is a sentence that contains a que-drop beacuse it could have been "Ellos lamentan que no estés muy contenta". When we test on the first four words, the model takes an input of "ellos", "lamentan", "no", and "estés". Theoretically, because the model has been trained on sentences containing "que", the model will predict "que" where there actually isn’t. In this way, we can identify all of the possible que-drops.

After finding all instances of que drop, we retained the words that preceeded and followed the "que". We then analyzed results by identifying which words preceded the que drop the most.

Word	Que dropped	Que Kept	Drop rate
Recordar	18	0	100
Ser	11	1	91.7
Cuento	33	11	75
Ver	23	8	74.2
Decir	22	21	51.2
Lo	328	367	47.2
Creer	19	34	35.8

Figure 3: We found the words where "que" was dropped the most. "Recordar" had the highest drop rate of 100%, and "ser" and "cuento" had high drop rates of 91.7% and 75% respectively.

5 Conclusion

Although we were limited by memory and dataset issues, our model can be significantly improved for future work by using larger datasets and running for more epochs. We believe our idea of using a CBOW model to detect que-drop may be very useful to researchers and linguists in the future, as it saves time for those who previously analyzed datasets by hand.

6 Division of Labor

We divided the work as follows:

- Arjun Somnali- Preprocessing books, coding model, training/testing model, creating poster
- Lily Shi- Preprocessing books, coding model, testing model, creating poster, writing paper
- Phillip Guan- Identifying texts from Project Gutenberg, helping create the poster, working on paper

7 Acknowledgements

We would like to thank Ms. Haripriya for providing this research opportunity with professors and significantly helping to debug and give structure to our code. We would also like to thank Mr. Bhagirath for helping write and debug our code, as well as improving morale the day of the presentation. We would like to thank Ms. Andrea and Ms. Anna for helping with our code. Finally, we would like to thank Prof. Adrian Ricelli and Isaac Ang for giving us ideas and structure to our project.

References

[1] Adrián Rodríguez Ricelli. Espero estén todos: The distribution of the null subordinating complementizer in two varieties of spanish. In *Language Variation and Contact-Induced Change*, pages 299–333, 2018.

[2] Minoru Fukuda. Complementizer drop and ip complementation in japanese. In *Kansas Working Papers in Linguistics*, pages 25:39–52.

[3] Yiming Liang, Pascal Amsili, and Heather Burnett. New ways of analyzing complementizer drop in montréal french: Exploration of cognitive factors. In *Language Variation and Change*, pages 359–385, 2022.

[4] Sali Tagliamonte and Jennifer Smith. No momentary fancy! the zero 'complementizer' in english dialects. In *English Language Linguistics*, pages 289–309.

[5] Jiyoung Yoon. The grammaticalization of the spanish complement-taking verb without a complementizer. In *Journal of Social Sciences*, page 11(3):338.