

Vocabulary Challenge

Enrique Gómez Cruz

March 14, 2025

Abstract

This work examines the vocabulary of the first volume of “Les Misérables” by Victor Hugo in three languages: French, Spanish, and English. The texts were sourced from Project Gutenberg and the Carlos Slim Foundation. The corpus underwent preprocessing to normalize text, remove special characters, and generate word frequency counts. Additionally, an attempt was made to construct parallel sentence pairs for machine translation learning, revealing difficulties in direct one-to-one mapping due to translation variations. A neural network trained with this dataset showed overfitting and poor generalization, but performance improved with a larger dataset of 120k sentence pairs. However, when applied to “Les Misérables”, translation quality significantly declined, highlighting the challenges of literary translation modeling.


1 Les Misérables

A vocabulary of the first volume of “Les Misérables” by Victor Hugo is extracted in three different languages: french, spanish and english. The english and french versions were taken from the [Project Gutenberg](#) website, and the spanish version from the [Carlos Slim foundation](#) website.

2 Corpus pre-processing

The spanish version came in PDF format, which had to be converted to plain-text. This was easily done by copy-pasting its contents into a plain text editor (Vim). The other versions were already in plain-text format.

The pre-processing pipeline which applies similarly to all three versions is the following:

1. Make every letter lowercase.
2. Convert to a blank space every character that is **not** amongst the following a-z, ç, áéíóú, àèìòù, äëïöü, âêîôû; and ñ (spanish version only). This conversion is a problem for english contractions—“don’t” gets converted to “don t”—so they need to be treated differently.
3. Replace every blank space character for a  (carriage return). This way every word is on its own line.
4. Delete every empty line.
5. The total word count is given by the total number of lines in the file, and the frequency of each unique word is obtained by sorting the file and then counting each consecutive identical lines.

Contractions in french and english like, “n’est” and “don’t” get converted to “nest” and “dont”, respectively after the special characters are removed.

There might exist a word-separator character that is not dash or underscore, but instead a unicode version of it. Special care must be taken if that is the case.

Every file transformation operation was done using regex within the Vim editor. The counting of words was done with the GNU utils `sort` and `uniq`.

3 Results

3.1 Word count

The english version of the book has the biggest word count and yet has the least amount of unique words, that is, its lexical density (unique words divided by total count) is the lowest, suggesting perhaps that the translation is of lower quality compared to the spanish version.

Table 1: Total word count and unique word count of Les Misérables corpus in french, spanish and english.

| Language | Count | Unique | Lexical Density |
|----------|---------|--------|-----------------|
| French | 119,003 | 11,991 | 0.100 |
| Spanish | 109,511 | 13,265 | 0.121 |
| English | 121,553 | 10,215 | 0.084 |

It is interesting to note the total mentions of different main characters in different translations. For example, Valjean is slightly more popular than Fantine in the french version, but the otherway around in the spanish version. It seems that the English version is a bit more faithful to the original text, at least in the number of mentions to the main characters. Words like “Fantine” and “Fantine’s” in the english version both added to Fantine’s popularity.

Table 2: Total mentions of several main characters.

| Character | French | Spanish | English |
|-----------|--------|---------|---------|
| Valjean | 196 | 192 | 196 |
| Fantine | 189 | 194 | 189 |
| Javert | 175 | 175 | 176 |
| Cosette | 61 | 58 | 61 |
| Éponine | 4 | 4 | 4 |

Table 3 and 4 show the most frequent and least frequent words in the vocabularies. It is to be expected that the most frequent words are conjunctions. The curious one single letter word in the least frequent words of the spanish version comes from the compound word “Pont-à-Mousson”.

Table 3: Most frequent words in french, spanish and english, with their respective count.

| French | Spanish | English |
|------------|------------|-------------|
| de (4456) | de (5327) | the (7974) |
| il (3187) | la (3918) | of (3912) |
| la (3030) | que (3505) | and (3138) |
| et (2948) | y (3123) | a (3128) |
| le (2540) | el (3081) | to (3039) |
| l (2377) | en (2836) | he (2496) |
| à (2318) | a (2489) | in (2317) |
| un (1806) | se (1633) | was (1971) |
| les (1529) | un (1601) | that (1736) |
| que (1348) | no (1498) | his (1536) |

Table 4: Least frequent words in french, spanish and english, with their respective count.

| French | Spanish | English |
|-----------------|-----------------|----------------|
| abdiquait (1) | abarcó (1) | abjure (1) |
| abbaye (1) | abarcar (1) | abjectness (1) |
| abat (1) | abandonaron (1) | ability (1) |
| abandonnons (1) | abandonaría (1) | abhorred (1) |
| abandonnée (1) | abandonara (1) | abdicates (1) |
| abandonne (1) | abandonamos (1) | abdicated (1) |
| abandonna (1) | abandonadas (1) | abbey (1) |
| abaissement (1) | abandonaba (1) | abate (1) |
| abaissé (1) | abadía (1) | abandons (1) |
| abaissaient (1) | à (1) | abandon (1) |

3.2 Zipf’s law

Zipf’s law is an empirical law stating that when a list of measured values is sorted in decreasing order, the value of the n th entry is often approximately inversely proportional to n . It is said that Zipf’s law applies best to natural language. Figure 1 shows that, indeed, the corpus in all three languages have a “Zipfian” distribution.

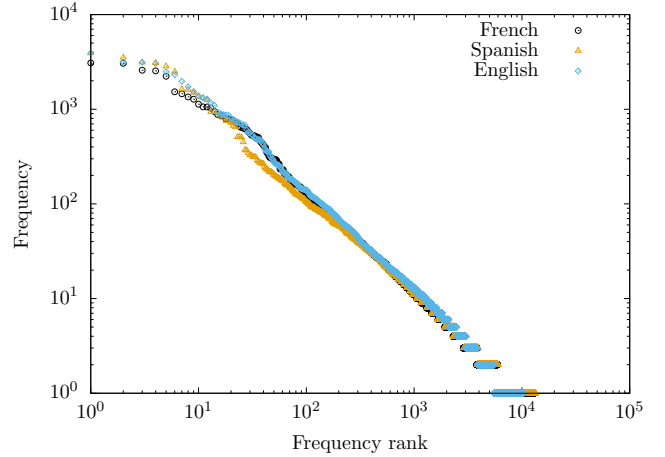


Figure 1: Word frequency sorted in decreasing order.

3.3 Learning to Translate

To learn how to translate, a set of source-target sentence pairs is required. In this case, the source is a sentence in French, and the target is its correct translation in Spanish.

Each sentence in the corpus was placed on its own line using regular expressions. A period, an exclamation mark, and a question mark served as line break indicators. Special care was taken to avoid breaking abbreviations such as "M." in Monsieur; therefore, only indicators that followed lower-case letters or special characters like “)” or “»” were used to determine line breaks.

Two different approaches were attempted to construct this dataset. The first involved mapping each sentence in the French and Spanish versions of the corpus on a one-to-one basis. However, this approach proved impractical due to the significant manual effort required – especially since the Spanish translation was not a direct, word-for-word rendering of the French text. The following sentence pair illustrates one of the issues in finding a one-to-one mapping:

La révolution survint, les événements se précipitèrent, les familles parlementaires décimées, chassées, traquées, se dispersèrent. M. Charles Myriel, dès les premiers jours de la révolution, émigra en Italie.

Sobrevino la Revolución, precipitáronse los sucesos; las familias parlamentarias, diezmadas, perseguidas, acosadas, se dispersaron, y el señor Charles Myriel, en los primeros días de la Revolución, emigró a Italia.

The second approach used Google translate to create the correct spanish translation. This was of course more practical.

The [example by Chollet](#) on the Keras page was used to train the network. However, the dataset size (7k sentences) was not sufficient for the network to learn effectively as seen in Figure 2, where there are clear signs of overfitting. The network’s effectiveness was measured using the ROUGE-1 and ROUGE-2 metrics, which count the number of matching unigrams and bigrams with the correct response, respectively. The network achieved a ROUGE-1 score of 0.505 and a ROUGE-2 score of 0.207.

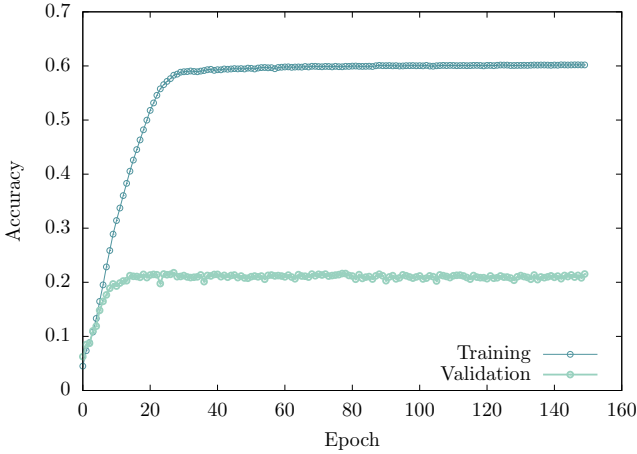


Figure 2: Accuracy of the network trained with the “Les Misérables” dataset.

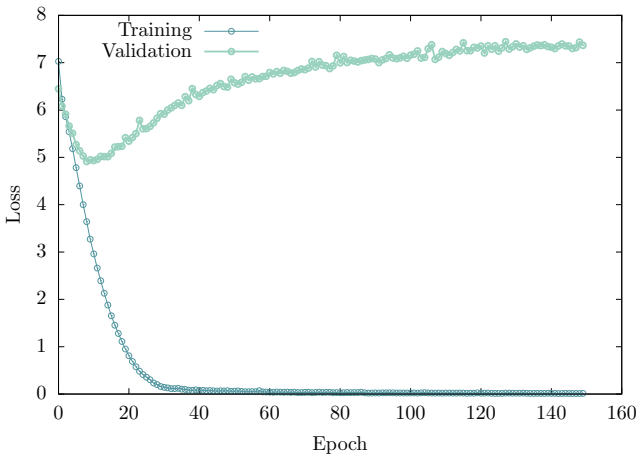


Figure 3: Loss of the network trained with the “Les Misérables” dataset.

The same experiment was conducted with a larger dataset of approximately 120k French sentences and their corresponding translations, this is a much bigger dataset compared to the 7k sentences extracted from “Les Misérables”. This yielded better results, with a maximum ROUGE-1 score of 0.687 and a ROUGE-2 score of 0.437. However, when the trained network was used to translate sentences from “Les Misérables”, the ROUGE-1 score dropped to 0.232 and the ROUGE-2 score to 0.003.

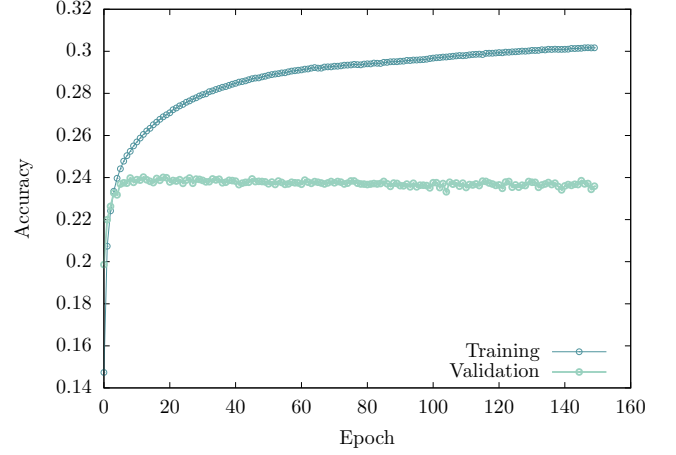


Figure 4: Accuracy of the network trained with the 120k french sentences dataset.

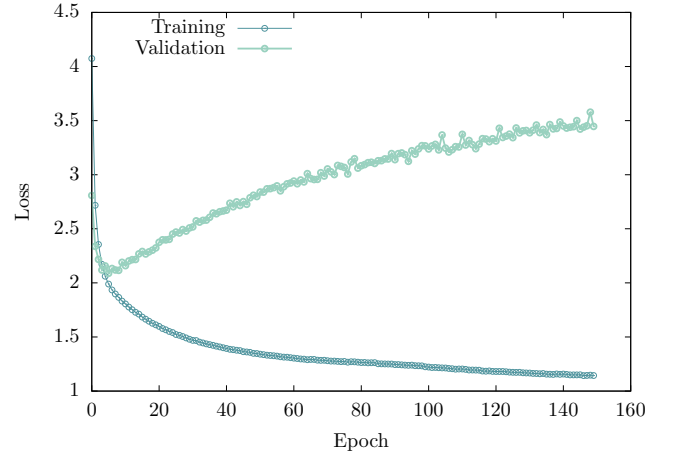


Figure 5: Loss of the network trained with the 120k french sentences dataset.