# Word Statistics - Report

Bruno Del Papa, Sigrid Trägenap and Charles Wilmot

*Frankfurt Institute for Advanced Studies*

**Abstract**

The frequency distribution of words and letters is a key object of study in statistical linguistics. The word frequency distribution follows a mathematical form known as *Zipf's law*. This report first tries to find this power-law distribution in different languages and to highlight variations between languages in terms of most common letters and most common used graphems. In a second part, languages are discriminated by their entropy in the context of letters, words or conditional entropy. Finally the possibility to reconstruct the history of a text (the bible) is explored.

## 1. Methods

### *1.1. Languages*

To ensure diversity in the analyzed texts and familiarity with the languages, two Romanic languages (French and Portuguese) and West Germanic languages (German and with special focus as *lingua franca*) were chosen. The Romanic languages and German offered the opportunity to analyze letters outside the standard English alphabet.

### *1.2. Analyzed corpus of texts*

Texts were chosen out of four different languages to create diversity in authors, periods, genres and style. Project Gutenberg `http://www.gutenberg.org/` provided public domain texts as `.txt` files, which were manually edited. Paragraphs that did not belong to the original texts excluded.

### *1.3. Set of Letters*

Since it was a goal of our project to analyze letter distributions in languages for which the classical Latin alphabet was not sufficient, we needed to create a custom set of letters covering all special characters in our chosen languages. This set of letters is based on the Unicode-block `latin1`:

abcdefghijklmnopqrstuvwxyz ßàáâãäåæçèéêëìíîï
ðñòóôõöøùúûüýþ āăąćĉċčďđēĕėęěĝğġģĥħ
ĩīĭįıijĵķĹĻĽL·LŁŃŅŇ’ŋ ōŏőœŕŗřśŝşšţťũūŭůűųŵ

### *1.4. Algorithms*

### *1.4.1. NLTK-Toolkit*

The Natural Language Toolkit `http://www.nltk.org/` is a platform for Python programs to work with human language data. It is already equipped with a selection of lexical resources and is able to tokenize strings into single words. The community-driven project is an important tool for computational linguistics. Since NLTK is focused on semantic processing of texts, the library was not able to deal with special characters and showed problems with tokenization of texts with intrinsic printing errors, such as the bible. Thus NLTK could only be used for parts of our analysis.

### *1.4.2. Custom Code*

To solve special problems with unicode characters and typesetting errors we implemented tools for letter and word frequency distributions in Python. The code can be found at: `https://github.com/charleswilmot/word_stats`.

### *1.5. Entropy*

In the context of text analysis, the entropy (denoted as $H$) is the expected value of the information contained in a message. A message can consist of different words, each formed by letters - both of which are described here as events. The information conveyed by one of these events is the negative logarithm of the probability of this event. For a given set of events $Z$ of magnitude $N$, entropy is maximized if each event has the probability $p(z) = \frac{1}{N}$.

$$H = \sum_{z \in Z} p_z \, I(p_z) = - \sum_{z \in Z} p_z \, log_2(p_z) \tag{1}$$

Since events can be understood as either words or letters in our context, entropy can be calculated as word-entropy or letter-entropy.

### *1.5.1. conditional entropy for words of length N*

Similar to conditional probabilities, a conditional entropy can also be formulated. $H_n$ expresses here the average information gained by a specific word of length $n$ when only words of length $n$ are considered.

$$H_n = \sum_{z \in Z^n} p(w|n) \, log_2(p(w|n)) \tag{2}$$

This conditional entropy can be understood as how unequal words of length $n$ are distributed.

## 2. Results

### *2.1. word frequency*

The word frequency distribution in all analyzed languages resembled power-law distributions, as expected from the *Zipf's law*. The English language showed small deviations from a pure power-law distribution, possibly explained by printing errors in the analyzed text (the bible), which could affect the set of words and their frequency.

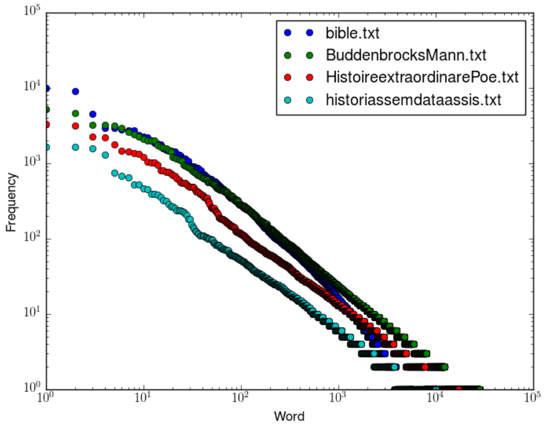


Figure 1: abundance of words with a certain frequency - Zipf's Law

### *2.2. letter frequency*

Languages show different sets of most common letters and syllables, which we analyzed here. Such classification of languages based on small text samples is potentially useful for language recognition softwares, which is a natural next step for the results we present here.

#### *2.2.1. single graphemes*

The following table shows the five most frequent graphemes (in this context letters) in different languages. The frequency of vowels and consonants differed between the languages, reflecting differences in the laguages' evolution. One should note that the letter *e* is the most frequent letter in almost all languages and, generally, vowel are among the most frequent graphemes.

| Language | most frequent letters (ordered by abundance) | | | | |
|---|---|---|---|---|---|
| English | e | t | h | a | o |
| French | e | s | a | t | n |
| German | e | n | i | r | t |
| Portuguese | a | e | o | s | r |

### *2.2.2. combination of graphemes*

To reconstruct the most frequent syllables in a language, we analyzed which letters were more likely to follow other letters in a given language. A common combination of graphemes is *qu*, which is also considered as digraph and was therefore not included in further analyses. The table below shows prominent combinations of graphemes in the analyzed languages. All of the languages analyzed could be identified by looking at the most frequent combinations of graphemes.

| Language | common combination of graphemes | | | | |
|---|---|---|---|---|---|
| English | th | ve | nd | ke | ju |
| French | on | je | wo | on | ks |
| German | ch | vo | zu | un | ge |
| Portuguese | fi | ll | co | ze | ko |

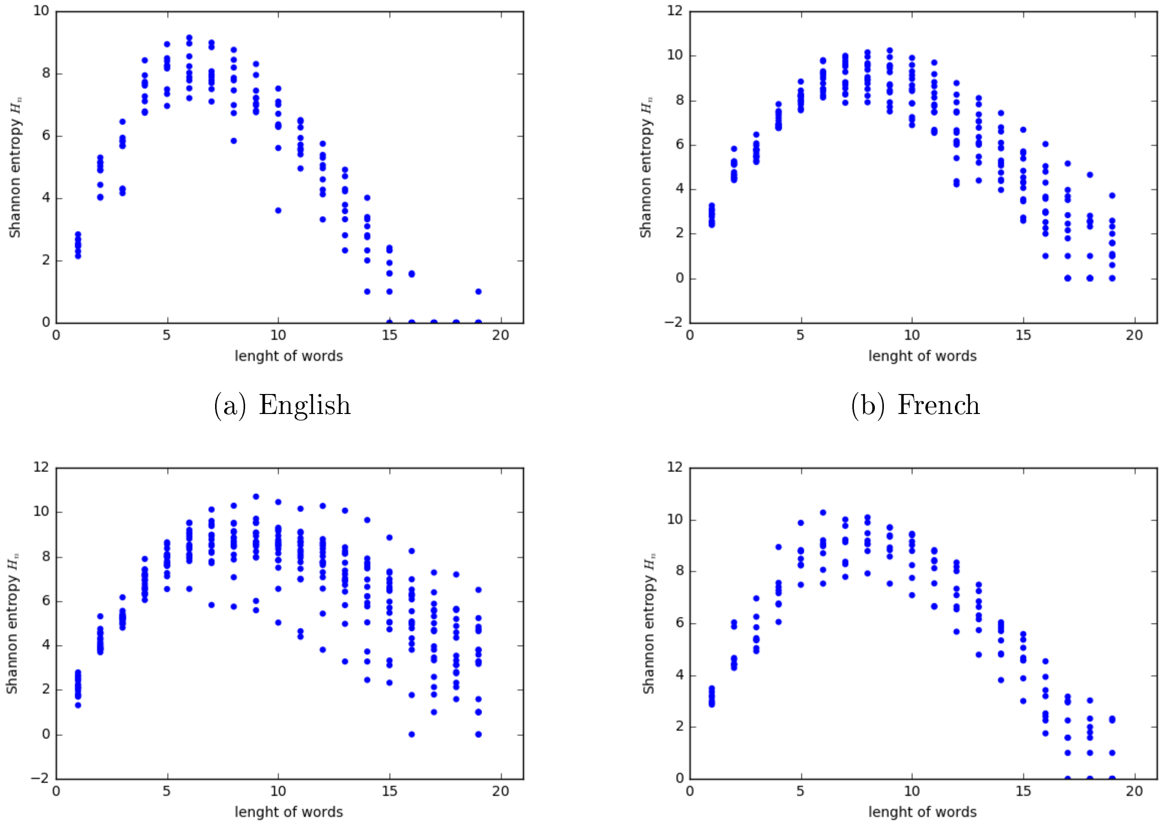
### *2.3. Entropy*

There was no significant difference in average entropy between different languages, which could be interpreted as no significant difference between the amount of information conveyed by different languages. The variation in general word entropy between different texts was in a higher magnitude than the variation of the average word entropy between languages.

| Language | Average word entropy [shannon] | SEM of word entropy |
|---|---|---|
| English | 8.764 | 0.41 |
| French | 9.624 | 0.22 |
| German | 9.440 | 0.41 |
| Portuguese | 9.203 | 0.83 |

### *2.3.1. conditional entropy - words of length n*

Languages differ in their distribution in word lengths and thus also show varying conditional entropies for word lengths. The conditional entropy differs between texts of a language, but in a general trend also clearly between languages. As Figure 2c shows, German has a tendency to more evenly distributed long words, which can be explained by the high abundance of compounds in German.



(a) English

(b) French

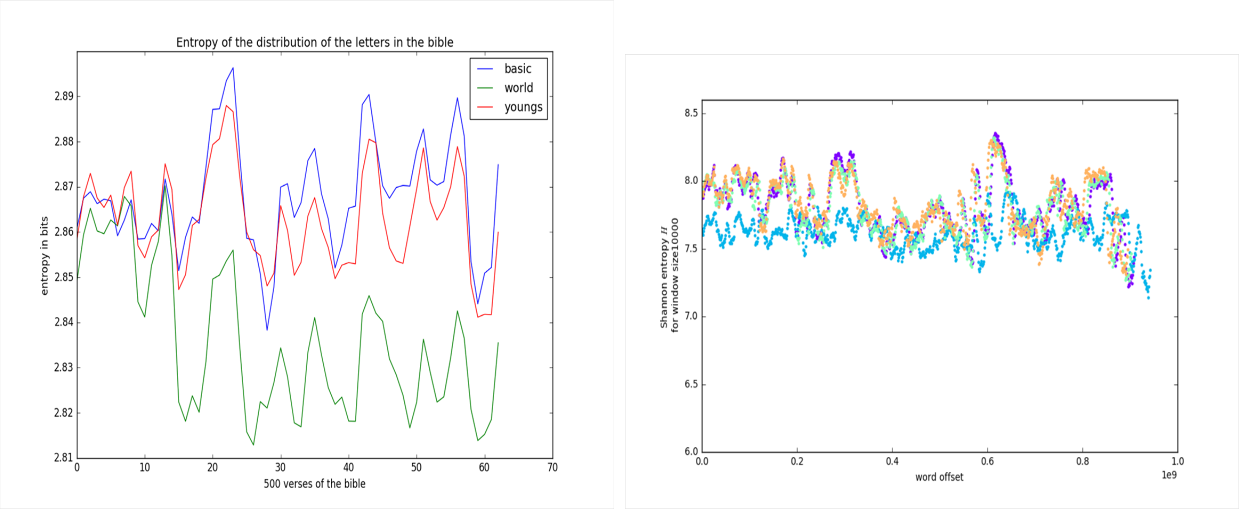
(c) German

(d) Portuguese

Figure 2: **Entropy for words with length $n$ differs between languages**
The conditional entropy for words of length n is shown for different texts (dots) in a language.
The most striking difference between languages is the high conditional entropy for words of length > 12 in German compared to all other languages. A low conditional entropy could resemble a situation were longer words are seldom used and if they are used, the same, specific word is often used. The German language has a high tendency to concatenate words to build new ones, hence creating a lexical richness in long words.
The English language has a peak at words with length 6, whereas German and French show a higher conditional entropy at words of length 8 and higher.

### *2.4. comparison of different text versions*

While looking for books to analyze, we found 8 versions of the Bible. As each is divided into books, themselves being written in the form of verses, it was simple to perform a linear comparison of the texts.
We first gathered the different versions and concatenated the books for each version, by preserving the order of the verses. Then we measured the statistic of the letters on overlapping windows.
Results show that the entropy of letters is almost constant all along the texts. They also showed a strong positive correlation, due to the fact that the different versions are still very similar. (See figure 3).



(a) Entropy of the letters for 3 different versions of the Bible.

(b) Entropy of the words for 4 different versions of the Bible

Figure 3: **Time-dependent entropy of texts allows reconstruction of textual history.** Word and letter entropy were calculated using a sliding on the text and calculation the entropy based on this sample of the text. The correlations in entropy between different versions is explained by textual similarity between the versions, although actual words or phrases are different - it is the same story told differently.

## 3. Outlook - Further Ideas

The four analyzed languages followed the *Zipf's law* and we were able to reproduce the most frequent letters and letter combinations with a custom-written code. This project could be expanded on the foundation of these preliminary analysis. Ideas for future studies are:

- train a classifier to recognize languages based on letter frequency or combination of letters
- analyze different literature genres in their conditional entropy. Are there authors or genres which have high information conveyed by long-words? Such analysis could unreavel hidden structures in grammar for the different languages
- Which parts of the bible (or any other text with multiple editions) were changed over time. Why does the entropy decrease or increase in certain verses (this could reflect repetition of certain words in verses, for example)?

## Appendix

*Simple description about the Bibles we compared*

- Basic : *Basic English, produced by Mr C. K. Ogden of the Orthological Institute, is a simple form of the English language which, with about 1,000 words, is able to give the sense of anything which may be said in English.*

- World : *The World English Bible (WEB) is a Public Domain (no copyright) Modern English translation of the Holy Bible, based on the American Standard Version of the Holy Bible first published in 1901, the Biblia Hebraica Stutgartensa Old Testament, and the Greek Majority Text New Testament.*

- Young : *Young's Literal Translation (YLT) was created by Robert Young who also compiled Young's Analytical Concordance. Produced and printed in the late 19th century, Young's Literal Translation was designed to assist students in the close study of the Biblical text by reproducing in English the Hebrew and Greek language and idioms in an exceedingly literal translation.*