Lab 1. Zipf's law of abbreviation and compression.

Introduction to Quantitative Linguistics

2022-2023 (2nd edition)

Ramon Ferrer-i-Cancho

March 18, 2023

In this lab session, we are going to practice on Zipf's law of abbreviation, namely the tendency of more frequent words to be shorter, as well as with its likely casual origin, namely compression.

1 Data preparation

Here you have to check if the law holds and investigate compression in a sample of 20 distinct languages from distinct families. The list should comprise languages that vary concerning the degree of inflection (the dataset will be reused in Lab 2, where that issue will become very relevant). For each language, you have to select a text, tokenize it conveniently and produce a plain text file with tree columns: the word form, the frequency of the word and its length. Texts can be selected from the Universal Declaration of Human Rights (UDHR) corpus or another parallel corpus of your choice (the use of non-parallel texts is not allowed). You can download the UDHR texts from `http://unicode.org/udhr/`. You can choose the tokenizer you wish but beware that you must be able to process languages from distinct families (we strongly recommend `spaCy` over `NLTK`). You are expected to use Python as programming language.

2 Data analysis

For data analysis, you are expected to use R as programming language by default. Taking the plain text files as input, you have to

- Investigate if the law of abbreviation holds in these languages, measuring the correlation between word length and word frequency and running a correlation test. We strongly recommend to use a Kendall tau correlation test because its relevance in the theory of optimal coding introduced during the theoretical lectures. We recommend using the function `cor.test`

from R stats. You have to investigate the effect of a Holm-Bonferroni correction on the p-values. We recommend using the function `p.adjust` from R stats with the method "holm".

- Measure the mean word length and compare it against the random baseline (L_r) and the minimum baseline (L_{min}).
- Measure degree of optimality of word lengths with $\eta = \frac{L_{min}}{L}$ and the optimality score

$$\Omega = \frac{L_r - L}{L_r - L_{min}} \tag{1}$$

  introduced in the theoretical lectures. For both scores, L_{min} has to be defined as the value of L that is obtained by reordering l_i so that L is minimized.

You have to generate the following materials

- A multipanel figure showing word length as a function of word frequency for a selection of at least 6 languages, each language in a distinct panel. For these six languages, we suggest that you arrange the panels as a matrix with three rows and two columns. Plots must be generated using `ggplot2` with facets in R (one facet for every language) or another tool of equivalent power and visual quality. In every panel, points correspond to words types. You must show the word forms of selected types. We strongly recommend using `ggrepel` in R.
- A table summarizing the results of the analysis of Zipf's law of abbreviation with the following columns: language name, family, tokens, types, τ and the uncorrected and corrected p-value of the Kendal τ correlation test.
- A table summarizing the results of the analysis of compression with the following columns: language name, family, L_{min}, L, L_r, η and Ω.
- Figure giving a visual into the degree of compression of languages. For instance, a figure showing L_r as function of L where every point is a language. You must show the names of (selected) languages. We strongly recommend using `ggrepel` in R. Think of how to take into account both L_r, L_{min} and L in a single plot.

2.1 Report

The report must contain at least the following sections

- Results. That section should contain the materials above and other results you may find relevant or useful.
- Discussion. A discussion of the results. Some possibilities on the law of abbreviation

  - Implications for the universality of the law of abbreviation.
  - Differences between languages. ...

  Some possibilities on compression:

  - The degree of compression of word lengths of languages as suggested by comparing L against L_r or L_{min} or as suggested by η or Ω.
  - Comparison with results in doi: 10.12775/3991-1.029, that were obtained with a distinct definition of the minimum baseline and a distinct optimality score (η).
  - Some speculations on which seem to be the most and the least compressed languages and why.

- Methods section. That sections contains any relevant details about the methods that you have used (do not hide important ideas or decisions in your code) or the parallel corpus that you have chosen. Some examples: decisions made in the preprocessing of the files, whether the correlation test is one-sided or two-sided and why.

2.2 Teaming

Work is done in pairs and two people cannot pair more than once for a lab session. If you think that you cannot satisfy these constraints ask the professor as soon as possible (before the deadline). Every team must work independently from other teams.

2.3 To deliver

One member of the team has to deliver the report and related materials as a single .zip file. The name of the .zip file has to indicate the names of the students (for instance, `name1_name2.zip`) and must contain

- A PDF with the report in the main directory.
- A folder called "text" with the source .tex file and the figures and tables used to produce the PDF of the report in the main folder.
- A folder called "code" with the code (R or Python scripts) that you have used to produce the materials.
- A folder called "data" with the plain text with the three columns. The name of each file must be of the form `language.txt` where *language* is the iso code of the language.

The .zip file has be sent to ramon.ferrer@upc.edu with the header "[IQL] Lab 1" before April 11 (the deadline is longer than two weeks because of Easter holidays).