Laws of word length

Guillem Bonet

June 2024

1 Introduction

As shown in multiple research over the years, languages do have some statistical regularities and patterns that are quantifiable and mathematically analyzable. These characteristics have multiple technical applications, and also, they can allow us to understand the origins of languages and the reasons they exist in the way we know them.

In this research project, we are going to analyze two predictions of optimal coding for studying word length. Firstly, we will look at the correlation between word length and frequency rank, and secondly at the correlation between word length and relative frequency.

These correlations have not been extensively studied, a study was produced for the English language [1] and another one for Spanish and Catalan [2]. In this work, we intend to extend this study to a larger set of languages with different characteristics and from different families.

2 Materials and Methodology

In order to analyze the two predictions on word length, we will use the dataset used in two recent studies [3] [4]. In addition, we will study word length as the length in characters and as the duration in spoken form.

Language	Family	Script	Length measure
English	Indo-European	Latin	Characters and duration
Spanish	Indo-European	Latin	Characters and duration
Catalan	Indo-European	Latin	Characters and duration
Arabic	Afro-Asiatic	Arabic	Characters and duration
Indonesian	Austronesian	Latin	Characters and duration
Turkish	Turkic	Latin	Characters and duration
Chinese	Sino-Tibetan	Multiple	Characters, strokes and pinyin characters
Tamil	Dravidian	Tamil	Characters and duration
Basque	Language isolate	Latin	Characters and duration

Table 1: Analyzed languages with their families and scripts.

The set of languages shown in Table 1 have been chosen as a wide representation of languages that encompass different families and script types. We will analyze the same languages as previous studies to validate our results and will extend it to the rest of the languages.

For most of the languages, we will analyze word length as character length and as word duration, except for the case of Chinese. For Chinese word length is measured in character length, strokes, and romanized characters (pinyin) length. Duration analysis was not possible due to lack of data.

Language	Data sources
English	CV & PUD
Spanish	CV & PUD
Catalan	CV
Arabic	CV & PUD
Indonesian	CV & PUD
Turkish	CV & PUD
Chinese	PUD (by characters, strokes and pinyin characters)
Tamil	CV
Basque	CV

Table 2: Analyzed languages with their data sources.

In most cases, the data is obtained from Common Voice Forced Alignments[5] (referred to as CV from now on) and Parallel Universal Dependencies[6] (referred to as PUD from now on). CV data contains information about frequencies by word, and words have specified their length both in length of characters and duration. Instead, PUD data has information about frequencies by word, but words only have their length specified by the length of characters not by duration. Table 2 displays each of the chosen languages and the data sources used for the analysis.

The analysis has been performed in 2 steps. Firstly, a preprocessing step with a Python script. This initial step takes the data [7] and converts it into a common format (except for some missing data for PUD) for both sources. We took the median duration rather than the average to reduce the effect of outliers. After the preprocessing is completed, the processing stage takes place, which is done using an R script.

The processing part iterates over the files, and for each file, it does 2 analyses, one using length data and frequency rank, and another one for length data and relative frequency. For the analysis of relative frequency, we multiplied the data by 1,000,000,000, therefore making it parts per million rather than a typical percentage to avoid undesired behavior when applying the log function.

The analysis consists, in the first place, of performing multiplicative binning on the data to avoid bias later on. Since the density of data varies a lot based on the length and frequency, applying linear regression directly on raw data would give a result that does not take into account the whole range and is only accurate in the parts where data is more dense. To perform the multiplicative

binning, we took equispaced points in the log scale for the x-axis and computed the median of all the y values that belong to the same bin (equispaced point) in the x-axis.

After the multiplicative binning, we obtained an unbiased representation of the data which allowed us to perform a Theil-Sen robust linear regression on the binned data logged on the x-axis.

In the case of languages that were in the CV dataset, which included word duration data in seconds, we also performed the same 2 analyses but using the duration as length instead of the length in characters. For each of these analyses, we printed the model's coefficients (slope and intercept) and the p-values, we also plotted the raw data with a few word labels, the binned data, and the linear regression line.

All the code used to do the analysis and generate the data which will be presented in the next section is available in a Github repository [8].

3 Results

In this section we will discuss the results, we will first look at them by language and then we will summarize all of the information.

3.1 English

The model results are summarized in Table 3. All the p-values indicate strong evidence against the null hypothesis with slightly different values for the different data sets.

Prediction	Source	P-value	Slope	Intercept
$l \sim \log i$ (character length)	CV	$< 2.2e-16$	0.635	1.25
$l \sim -\log p$ (character length)	CV	$< 2.2e-16$	-0.492	7.774
$l \sim \log i$ (duration)	CV	$< 2.2e-16$	0.055	0.045
$l \sim -\log p$ (duration)	CV	$< 2.2e-16$	-0.046	0.623
$l \sim \log i$ (character length)	PUD	$4.72e-13$	0.762	1.1
$l \sim -\log p$ (character length)	PUD	$2.9e-10$	-0.897	10.706

Table 3: Prediction results for English

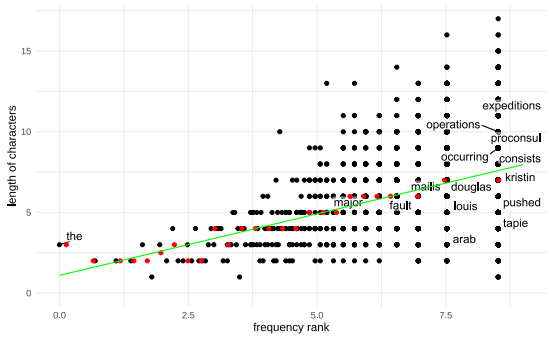


Figure 1: $l \sim \log i$ based on character length from PUD source plot for English

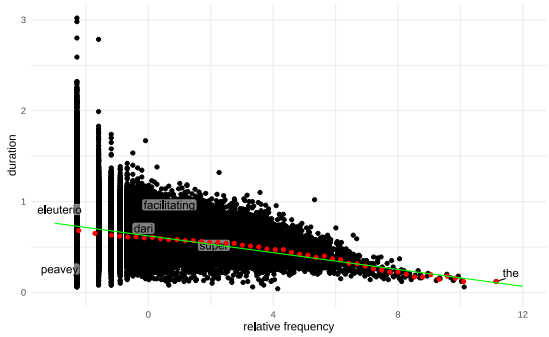


Figure 2: $l \sim \log i$ based on median duration from CV source plot for English

In Figure 1 and Figure 2, it can be seen how multiplicative binning is applied (in red) and robust linear regression is applied to the binned data to obtain the results.

3.2 Spanish

The model results are summarized in Table 4. As for English, the p-values indicate strong evidence against the null hypothesis, and the predictions are similar too.

Prediction	Source	P-value	Slope	Intercept
$l \sim \log i$ (character length)	CV	$< 2.2e-16$	0.655	1.725
$l \sim -\log p$ (character length)	CV	$< 2.2e-16$	-0.635	8.917
$l \sim \log i$ (duration)	CV	$< 2.2e-16$	0.052	0.086
$l \sim -\log p$ (duration)	CV	$< 2.2e-16$	-0.050	0.670
$l \sim \log i$ (character length)	PUD	$1.27e-14$	0.866	0.955
$l \sim -\log p$ (character length)	PUD	$1.31e-12$	-0.88	11.115

Table 4: Prediction results for Spanish

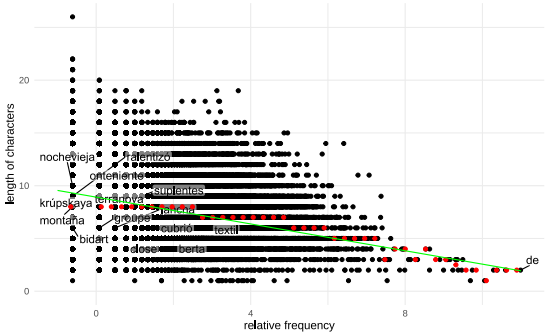


Figure 3: $l \sim -\log p$ based on character length from CV source plot for Spanish

In Figure 3, like for English, we can visually see the effects of the statistical analysis, but this time for $l \sim -\log p$ based on character length.

3.3 Catalan

The model results are summarized in Table 5. As for the previous two languages, the p-values indicate strong evidence against the null hypothesis, and the predictions are similar as well.

Prediction	Source	P-value	Slope	Intercept
$l \sim \log i$ (character length)	CV	$< 2.2e-16$	0.762	0.7
$l \sim -\log p$ (character length)	CV	$< 2.2e-16$	-0.693	9.09
$l \sim \log i$ (duration)	CV	$< 2.2e-16$	0.061	0.036
$l \sim -\log p$ (duration)	CV	$< 2.2e-16$	-0.057	0.698

Table 5: Prediction results for Catalan

3.4 Arabic

Arabic is the first language that had not been analyzed before in previous research [1] [2]. Once again, as can be seen in Table 6, the p-values indicate strong

evidence against the null hypothesis, and the predictions, although less than in the previous languages, look similar to the previous ones.

Prediction	Source	P-value	Slope	Intercept
$l \sim \log i$ (character length)	CV	$5.08e-14$	0.401	1.737
$l \sim -\log p$ (character length)	CV	$1.76e-12$	-0.476	6.938
$l \sim \log i$ (duration)	CV	$5.17e-15$	0.052	0.916
$l \sim -\log p$ (duration)	CV	$1.9e-12$	-0.064	0.86
$l \sim \log i$ (character length)	PUD	$1.58e-11$	0.635	0.917
$l \sim -\log p$ (character length)	PUD	$5.06e-11$	-0.726	8.714

Table 6: Prediction results for Arabic

3.5 Indonesian

Indonesian is the first language from the Austronesian family, and, like Arabic, regardless of being from a different language family than the previous, the p-values indicate strong evidence against the null hypothesis, and the predictions also look similar to the previous ones. The results can be seen in Table 7.

Prediction	Source	P-value	Slope	Intercept
$l \sim \log i$ (character length)	CV	$2.6e-11$	0.635	2.25
$l \sim -\log p$ (character length)	CV	$5.82e-11$	-0.663	10
$l \sim \log i$ (duration)	CV	$2.41e-13$	0.046	0.16
$l \sim -\log p$ (duration)	CV	$1.58e-11$	-0.047	0.716
$l \sim \log i$ (character length)	PUD	$6.21e-10$	0.545	3.071
$l \sim -\log p$ (character length)	PUD	$2.65e-06$	-0.635	10.25

Table 7: Prediction results for Indonesian

3.6 Turkish

Turkish is, once again, the first language of a new family, in this case, the Turkic. Similar to the previous ones, regardless of being from a different language family than the previous, the p-values indicate strong evidence against the null hypothesis, and the predictions also look similar to the previous ones as can be seen inTable 8.

Prediction	Source	P-value	Slope	Intercept
$l \sim \log i$ (character length)	CV	$< 2.2e-16$	0.762	1.5
$l \sim -\log p$ (character length)	CV	$< 2.2e-16$	-0.866	11.045
$l \sim \log i$ (duration)	CV	$< 2.2e-16$	0.052	0.102
$l \sim -\log p$ (duration)	CV	$< 2.2e-16$	-0.057	0.748
$l \sim \log i$ (character length)	PUD	$2.14e-08$	0.635	1.917
$l \sim -\log p$ (character length)	PUD	$3.16e-08$	-0.953	11.5

Table 8: Prediction results for Turkish

3.7 Chinese

Chinese is also a language from a new family, the Sino-Tibetan, and because of its representation, its case is different than previous languages. We analyzed the data basing character length on pinyin character length (the romanized version of Chinese characters), Chinese characters strokes, and Chinese characters length. This time around, the data looks different in some cases and even shows opposite trends as in previous languages. Nonetheless, the p-values still indicate strong evidence against the null hypothesis, which means that there is a statistical relationship. The data can be seen in Table 9.

Prediction	Source	P-value	Slope	Intercept
$l \sim \log i$ (pinyin character length)	PUD	0.0006232	-1.271	13
$l \sim -\log p$ (pinyin character length)	PUD	0.004781	3.15	-10.375
$l \sim \log i$ (character length)	PUD	$5.53e-05$	0.136	0.839
$l \sim -\log p$ (character length)	PUD	$5.2e-05$	-0.173	2.75
$l \sim \log i$ (strokes)	PUD	$4.23e-05$	-6.353	73.417
$l \sim -\log p$ (strokes)	PUD	0.0002516	26.408	-115.464

Table 9: Prediction results for Chinese

When looking at the data from Figure 4 and Figure 4, we can see the inverse correlation when word length is measured by pinyin character length. For $l \sim \log i$, the relationship is negative rather than positive as we had seen until now, therefore it is $l \sim -\log i$. The same thing happens for $l \sim -\log p$, which in this case appears to be $l \sim \log p$. Both show a significantly bigger slope than compared to the other languages.

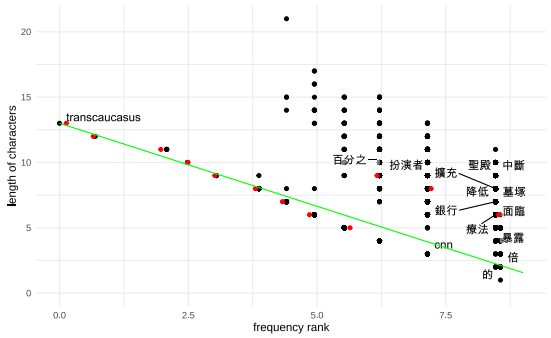


Figure 4: $l \sim \log i$ based on pinyin character length from PUD source plot for Chinese

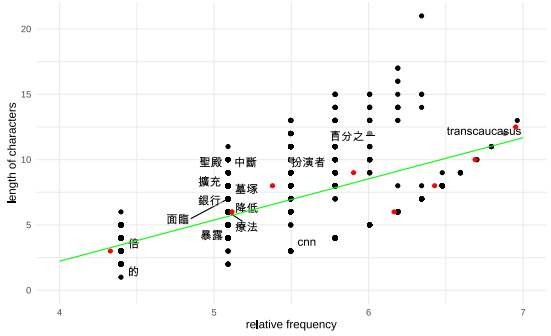


Figure 5: $l \sim -\log p$ based on pinyin character length from PUD source plot for Chinese

Instead, if we look at Chinese character length, the data is similar to the rest of the languages. But again, when looking at strokes in Figure 6 and Figure 7, the relationship is also opposed to the trend we had observed until now, and this time with much bigger slope and intercept values probably due to the greater amount of strokes in words compared to the number of Latin characters in words from the other languages.

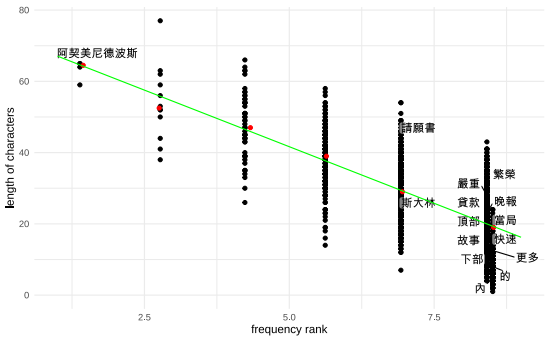


Figure 6: $l \sim \log i$ based on strokes from PUD source plot for Chinese

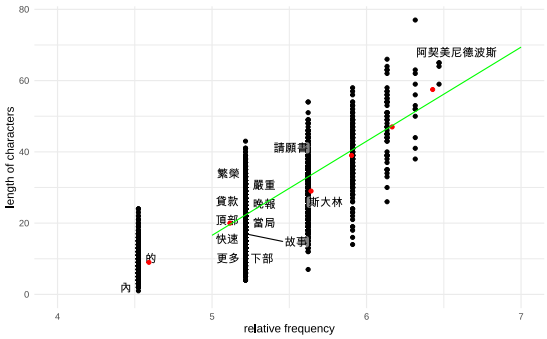


Figure 7: $l \sim -\log p$ based on strokes from PUD source plot for Chinese

3.8 Tamil

Given the case of Chinese, it is interesting to look at Tamil, which is another language that does not have a Latin script. As for the case of Arabic, we see that in Tamil the p-values indicate strong evidence against the null hypothesis, and the data also looks more similar to the majority of the analyzed languages. The data can be seen in Table 10.

Prediction	Source	P-value	Slope	Intercept
$l \sim \log i$ (character length)	CV	$8.06e-05$	0.52	3.068
$l \sim -\log p$ (character length)	CV	$3.97e-06$	-0.635	10.083
$l \sim \log i$ (duration)	CV	$2.79e-05$	0.044	0.323
$l \sim -\log p$ (duration)	CV	$4.9e-06$	-0.061	0.969

Table 10: Prediction results for Tamil

3.9 Basque

Basque is a different case since it is a language that does not have a family, but still, we can observe the same pattern as with the previous languages. The p-values indicate strong evidence against the null hypothesis, and the slopes and intercept follow the same pattern as the previous languages. The data can be seen in Table 11.

Prediction	Source	P-value	Slope	Intercept
$l \sim \log i$ (character length)	CV	$< 2.2e-16$	0.762	1.7
$l \sim -\log p$ (character length)	CV	$< 2.2e-16$	-0.794	10.646
$l \sim \log i$ (duration)	CV	$< 2.2e-16$	0.051	0.126
$l \sim -\log p$ (duration)	CV	$< 2.2e-16$	-0.055	0.74

Table 11: Prediction results for Basque

3.10 Summary

As can be seen in Table 12, all languages except for Chinese have a slope value between 0.4 and 0.8 and intercept between 0.7 and 3, so not only all of them have the same. In the case of Chinese, the characters version did follow the same trend with a smaller scope, but the the pinyin characters and strokes versions had an inverse trend.

Language	$l \sim \log i$ CV	$l \sim \log i$ PUD	$l \sim \log i$ Average
English	$0.635 * x + 1.25$	$0.762 * x + 1.1$	$0.7 * x + 1.175$
Spanish	$0.655 * x + 1.725$	$0.866 * x + 0.955$	$0.761 * x + 1.34$
Catalan	$0.762 * x + 0.7$	N/A	$0.762 * x + 0.7$
Arabic	$0.401 * x + 1.737$	$0.635 * x + 0.917$	$0.518 * x + 1.327$
Indonesian	$0.635 * x + 2.25$	$0.545 * x + 3.071$	$0.59 * x + 2.661$
Turkish	$0.762 * x + 1.5$	$0.635 * x + 1.917$	$0.7 * x + 1.709$
Chinese (pinyin)	N/A	$-1.271 * x + 13$	$-1.271 * x + 13$
Chinese (characters)	N/A	$0.136 * x + 0.839$	$0.136 * x + 0.839$
Chinese (strokes)	N/A	$-6.353 * x + 73.417$	$-6.353 * x + 73.417$
Tamil	$0.52 * x + 3.068$	N/A	$0.52 * x + 3.068$
Basque	$0.762 * x + 1.7$	N/A	$0.762 * x + 1.7$

Table 12: Summary of $l \sim \log i$ results for character length

For the case of duration as length, as can be seen in Table 13, all the slope values are between 0.04 and 0.06, and all intercept values are between 0 and 1. In this case, there is no data for Chinese, so the data appears to be very similar to the character length one, where all languages had a positive correlation with similar slope and intercept values.

Language	$l \sim \log i$ CV
English	$0.055 * x + 0.045$
Spanish	$0.052 * x + 0.086$
Catalan	$0.061 * x + 0.036$
Arabic	$0.052 * x + 0.916$
Indonesian	$0.046 * x + 0.16$
Turkish	$0.052 * x + 0.102$
Tamil	$0.044 * x + 0.323$
Basque	$0.051 * x + 0.126$

Table 13: Summary of $l \sim \log i$ results for duration

If we look at data for $l \sim -\log p$ based on character length in Table 14, we see that once again, all languages except for Chinese look similar with a slope between 0.4 and 0.95 and an intercept value between 6.9 and 11.1. For Chinese, we also observe that only the version that uses Chinese characters for word length gets a negative correlation like for the rest of the data but with different values, and the pinyin and stroke versions have an inverse correlation compared to the rest, in this case, a positive one.

Language	$l \sim -\log p$ CV	$l \sim -\log p$ PUD	$l \sim -\log p$ Average
English	$-0.492 * x + 7.774$	$-0.897 * x + 10.706$	$-0.695 * x + 9.24$
Spanish	$-0.635 * x + 8.917$	$-0.88 * x + 11.115$	$-0.758 * x + 10.016$
Catalan	$-0.693 * x + 9.09$	N/A	$-0.693 * x + 9.09$
Arabic	$-0.476 * x + 6.938$	$-0.726 * x + 8.714$	$-0.601 * x + 7.826$
Indonesian	$-0.663 * x + 10$	$-0.635 * x + 10.25$	$-0.649 * x + 10.125$
Turkish	$-0.866 * x + 11.045$	$-0.953 * x + 11.5$	$-0.91 * x + 11.273$
Chinese (pinyin)	N/A	$3.15 * x + -10.375$	$3.15 * x + -10.375$
Chinese (characters)	N/A	$-0.173 * x + 2.75$	$-0.173 * x + 2.75$
Chinese (strokes)	N/A	$26.408 * x - 115.464$	$26.408 * x - 115.464$
Tamil	$-0.635 * x + 10.083$	N/A	$-0.635 * x + 10.083$
Basque	$-0.794 * x + 10.646$	N/A	$-0.794 * x + 10.646$

Table 14: Summary of $l \sim -\log p$ results for character length

In the data for word length as duration shown in Table 15, we see that, once again, all languages follow the same pattern, all slope values are between 0.04 and 0.07 and all intercept values are between 0.6 and 1.

Language	$l \sim \log i$ CV
English	$-0.046 * x + 0.623$
Spanish	$-0.050 * x + 0.670$
Catalan	$-0.057 * x + 0.698$
Arabic	$-0.064 * x + 0.86$
Indonesian	$-0.047 * x + 0.716$
Turkish	$-0.057 * x + 0.748$
Tamil	$-0.061 * x + 0.969$
Basque	$-0.055 * x + 0.74$

Table 15: Summary of $l \sim -\log p$ results for duration

All of the plots generated can be found in Appendix A.

4 Discussion

From all the data that was presented, we can clearly say that there is a correlation between word length and the logarithm of the frequency rank and the logarithm of the relative frequency.

The $l \sim \log i$ correlation is positive except for the case of Chinese, which has mixed results depending on how word length is considered. For the case of the $l \sim -\log p$ correlation, we found it to be negative except for the case of Chinese, which, again, has mixed results.

Due to the nature of Chinese, it is challenging to analyze the language as it is unclear which measure of word length would be the best equivalent to those of Latin languages. For this reason, we believe it would be interesting to analyze Chinese with word duration and see if that provides more similar results to the rest of the languages.

For all of the languages, it was very surprising to see such consistent results with such small p-values in all the analyses, so it seems to have a logarithmic correlation which indicates an exponential relationship between word length and word frequency.

Although the language set that was studied is limited, it does offer a diverse representation of languages in terms of families and writing systems. We believe that the size of the analyzed dataset is optimal to see the results in detail without having too much, and aims to be a step into a larger analysis with a lot more languages which would go less in detail. The code used for this analysis can easily be used to analyze all the languages from the PUD and CV datasets.

We hope that this work will be useful for establishing global patterns in languages which can be useful for all types of language analysis.

References

[1] I. G. Torre, B. Luque, L. Lacasa, C. T. Kello, and A. Hernández-Fernández, "On the physical origin of linguistic laws and lognormality in speech," *Royal Society open science*, vol. 6, no. 8, p. 191023, 2019.

[2] A. Hernández-Fernández, I. G. Torre, J.-M. Garrido, and L. Lacasa, "Linguistic laws in speech: the case of catalan and spanish," *Entropy*, vol. 21, no. 12, p. 1153, 2019.

[3] S. Petrini, A. Casas-i Muñoz, J. Cluet-i Martinell, M. Wang, C. Bentz, and R. Ferrer-i Cancho, "The optimality of word lengths. theoretical foundations and an empirical study," *arXiv preprint arXiv:2208.10384*, 2022.

[4] ——, "Direct and indirect evidence of compression of word lengths. zipf's law of abbreviation revisited," *arXiv preprint arXiv:2303.10128*, 2023.

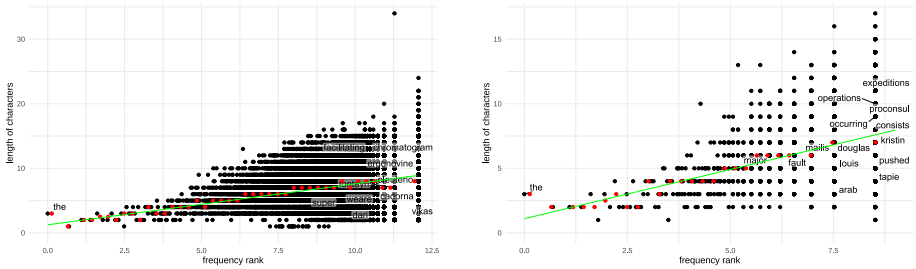
[5] JRMeyer, "Jrmeyer/common-voice-forced-alignments: Forced alignments for common voice." [Online]. Available: https://github.com/JRMeyer/common-voice-forced-alignments

[6] [Online]. Available: https://universaldependencies.org/

[7] S. Petrini, A. Casas-i Muñoz, J. Cluet-i Martinell, M. Wang, C. Bentz, and R. Ferrer-i Cancho, "Iql-course/iql-research-project-21-22: Research project of the iql 2021-22 course." [Online]. Available: https://github.com/IQL-course/IQL-Research-Project-21-22

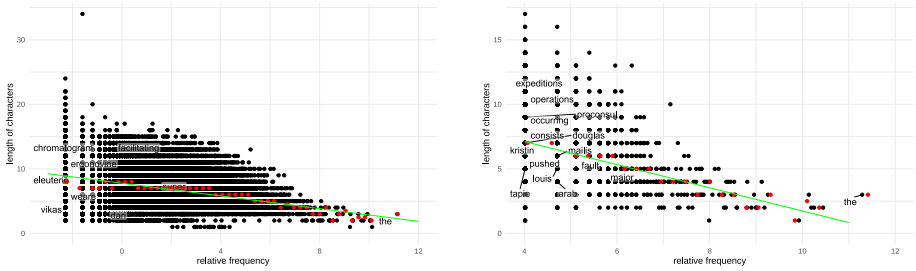
[8] G. Bonet, "Guillembonet/iql_final." [Online]. Available: https://github.com/Guillembonet/IQL_final

A Plots



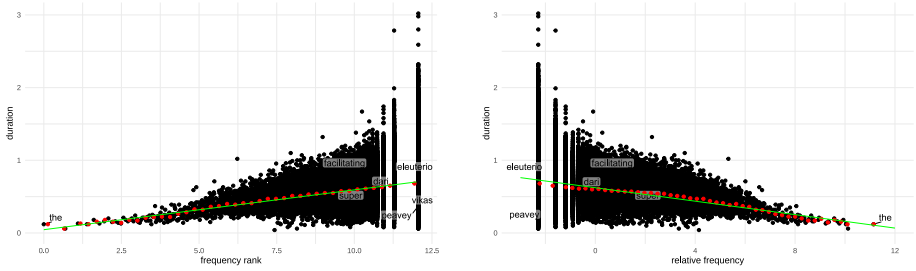
(a) CV dataset (b) PUD dataset

Figure 8: $l \sim \log i$ for character length as word length for English



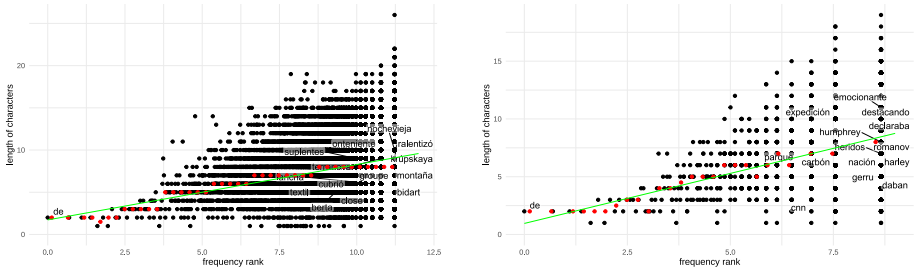
(a) CV dataset (b) PUD dataset

Figure 9: $l \sim -\log p$ for character length as word length for English



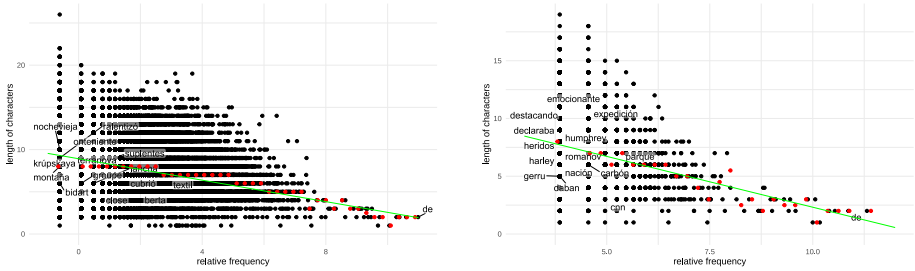
(a) $l \sim \log i$ (b) $l \sim -\log p$

Figure 10: Analysis with duration as word length for English from CV dataset



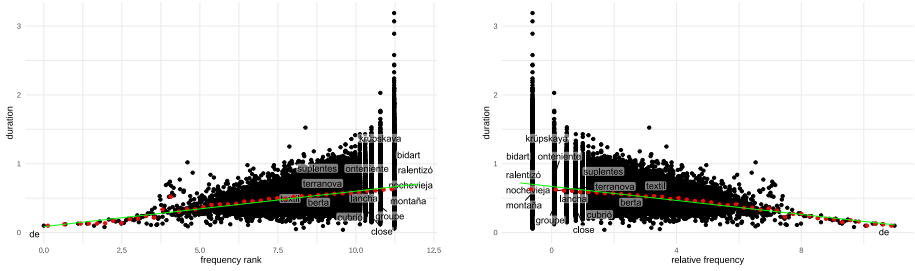
(a) CV dataset (b) PUD dataset

Figure 11: $l \sim \log i$ for character length as word length for Spanish



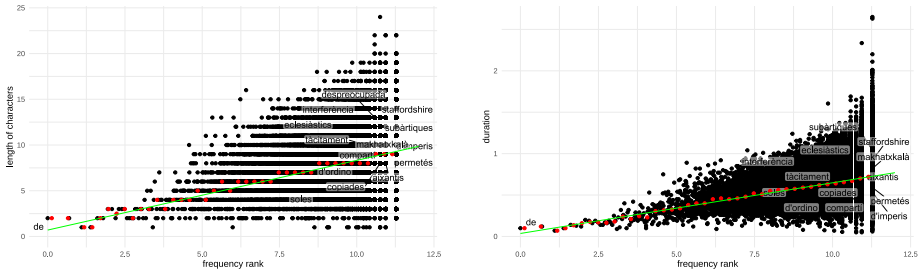
(a) CV dataset (b) PUD dataset

Figure 12: $l \sim -\log p$ for character length as word length for Spanish



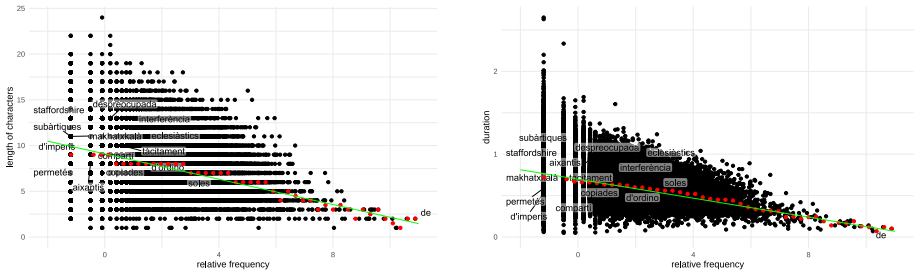
(a) $l \sim \log i$ (b) $l \sim -\log p$

Figure 13: Analysis with duration as word length for Spanish from CV dataset



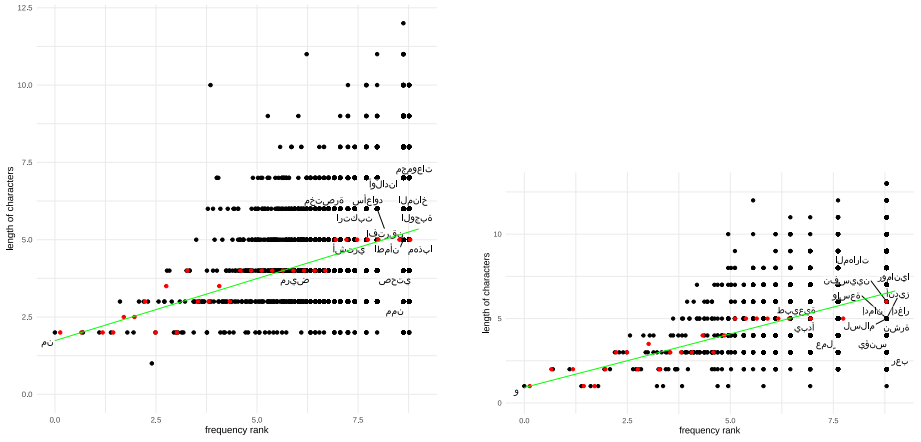
(a) Character length as word length (b) Duration as word length

Figure 14: $l \sim \log i$ from CV dataset for Catalan



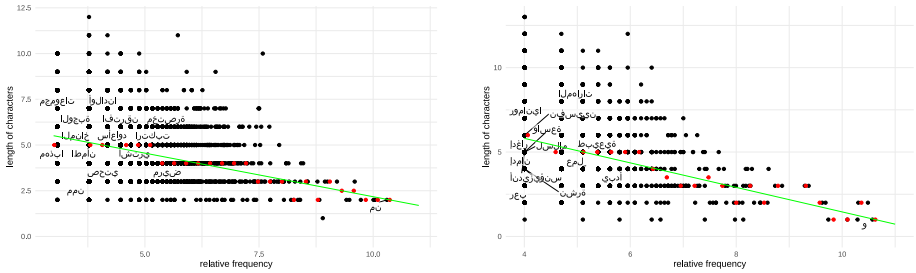
(a) Character length as word length (b) Duration as word length

Figure 15: $l \sim -\log p$ from CV dataset for Catalan



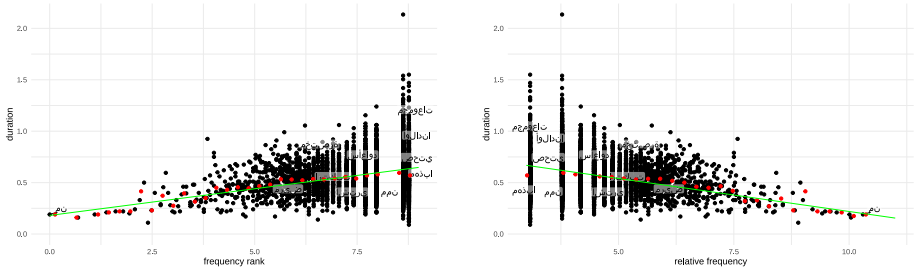
(a) CV dataset (b) PUD dataset

Figure 16: $l \sim \log i$ for character length as word length for Arabic



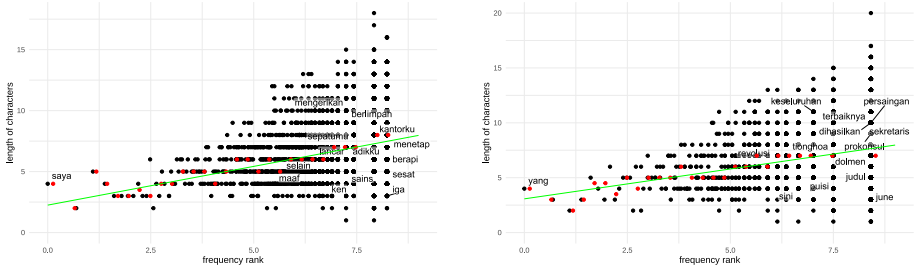
(a) CV dataset (b) PUD dataset

Figure 17: $l \sim -\log p$ for character length as word length for Arabic



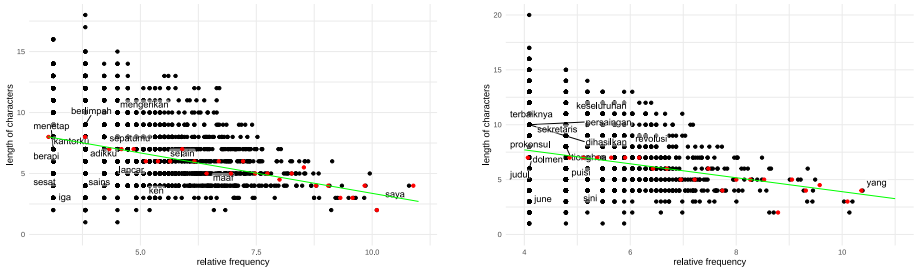
(a) $l \sim \log i$ (b) $l \sim -\log p$

Figure 18: Analysis with duration as word length for Arabic from CV dataset



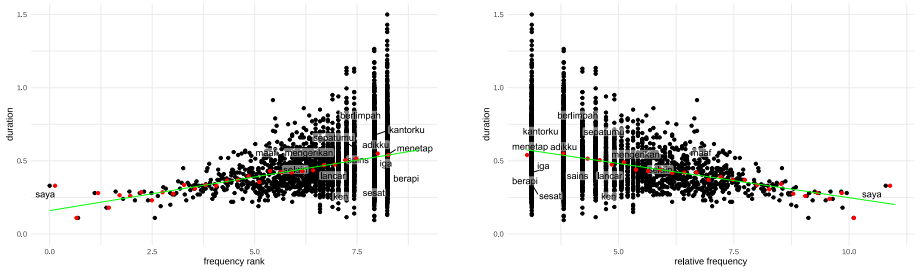
(a) CV dataset (b) PUD dataset

Figure 19: $l \sim \log i$ for character length as word length for Indonesian



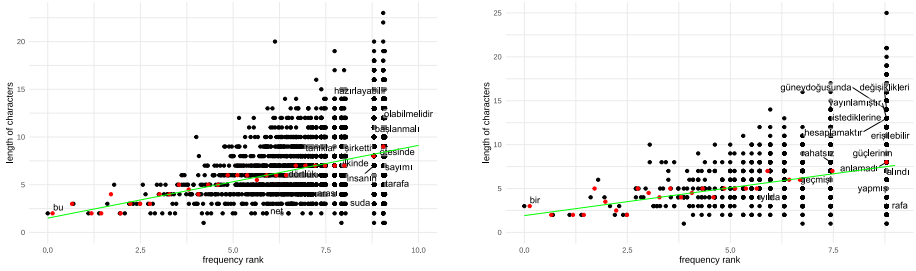
(a) CV dataset (b) PUD dataset

Figure 20: $l \sim -\log p$ for character length as word length for Indonesian



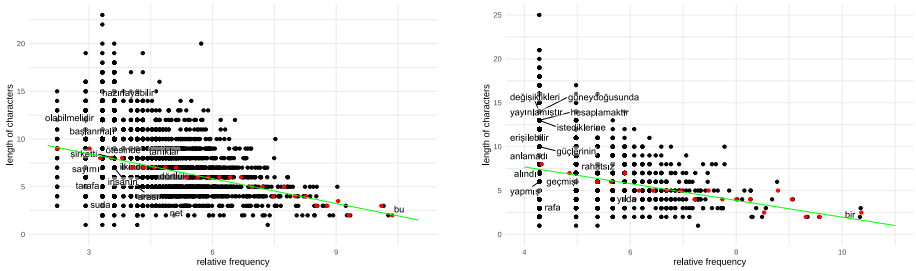
(a) $l \sim \log i$ (b) $l \sim -\log p$

Figure 21: Analysis with duration as word length for Indonesian from CV dataset



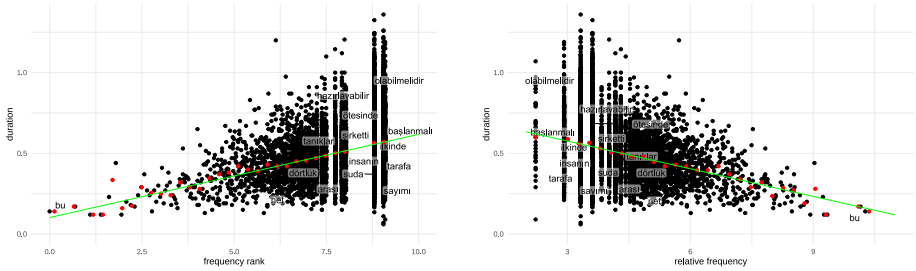
(a) CV dataset (b) PUD dataset

Figure 22: $l \sim \log i$ for character length as word length for Turkish



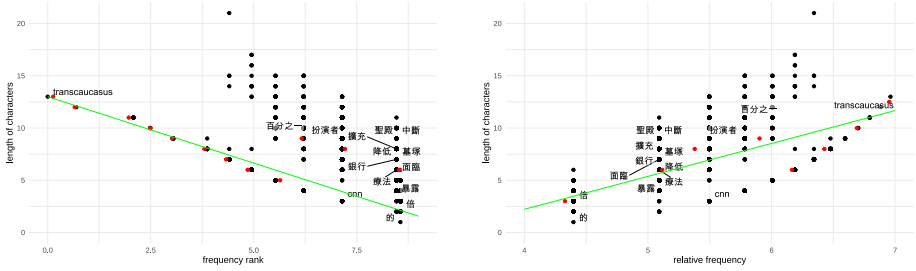
(a) CV dataset (b) PUD dataset

Figure 23: $l \sim -\log p$ for character length as word length for Turkish



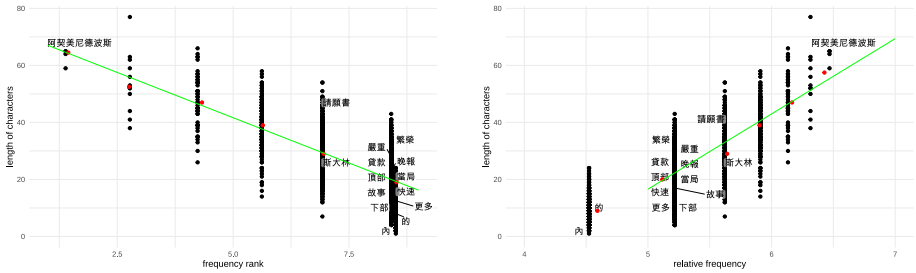
(a) $l \sim \log i$ (b) $l \sim -\log p$

Figure 24: Analysis with duration as word length for Turkish from CV dataset



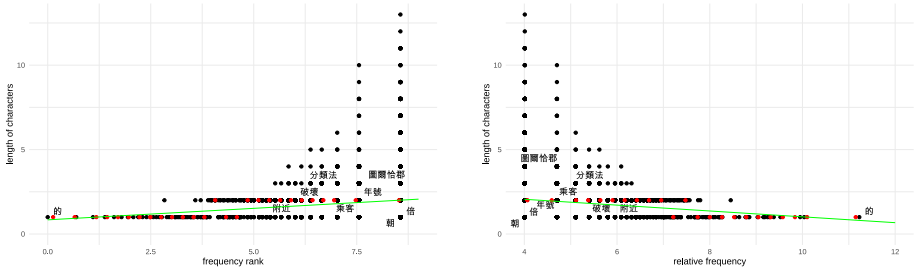
(a) $l \sim \log i$ (b) $l \sim -\log p$

Figure 25: Pinyin character length as word length for Chinese from PUD dataset



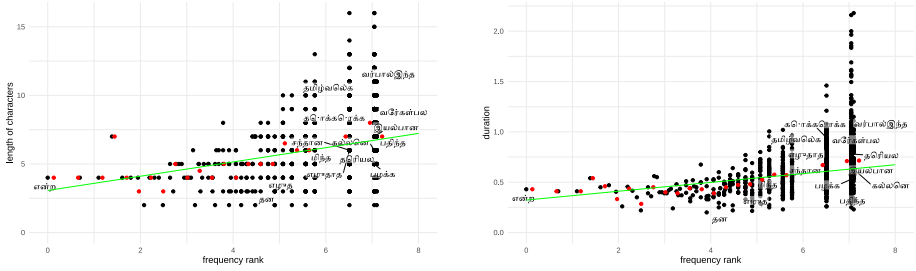
(a) $l \sim \log i$ (b) $l \sim -\log p$

Figure 26: Strokes as word length for Chinese from PUD dataset



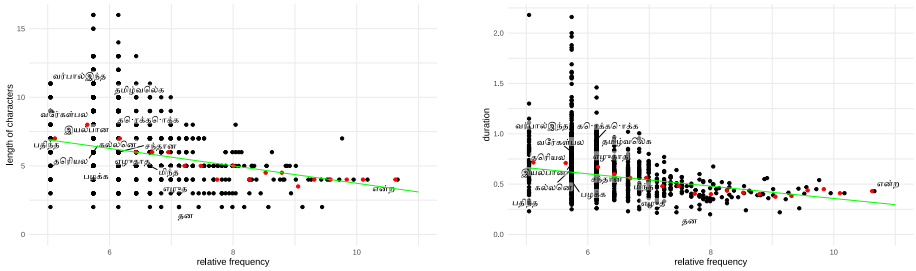
(a) $l \sim \log i$ (b) $l \sim -\log p$

Figure 27: Character length as word length for Chinese from PUD dataset



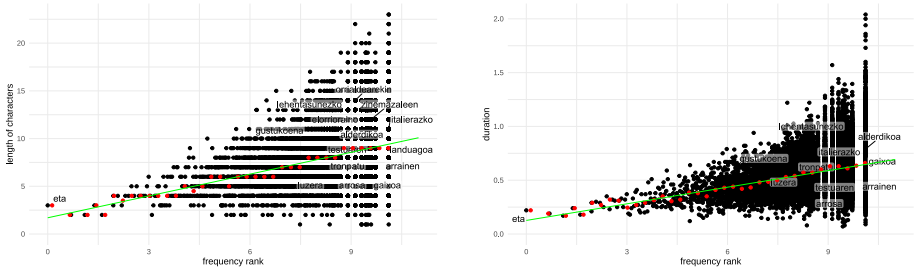
(a) Character length as word length (b) Duration as word length

Figure 28: $l \sim \log i$ from CV dataset for Tamil



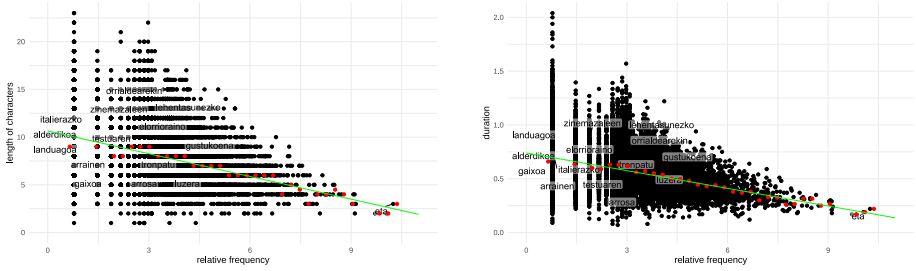
(a) Character length as word length (b) Duration as word length

Figure 29: $l \sim -\log p$ from CV dataset for Tamil



(a) Character length as word length (b) Duration as word length

Figure 30: $l \sim \log i$ from CV dataset for Basque



(a) Character length as word length (b) Duration as word length

Figure 31: $l \sim -\log p$ from CV dataset for Basque