CSN - Lab 2 - In-Degree Distribution in Syntactic Dependency Networks

David Kaindlstorfer, Sonia Petrini

October 13, 2021

1 Introduction

Syntactic Dependency Networks (SDNs) are spatial networks in which nodes are words and edges are syntactic relations. In directed SDNs, the *in-degree distribution* describes the syntactic links received by the dependent words from the heads. The goal of this report is to find the best model to describe the in-degree distribution of 10 languages from the Global SDN. Since the typical hypothesis is a **power-law**, we explore three nested models from the ***Zeta Riemann function family***, and compare them with two Null Models through Akaike Information Criterion (AIC). The comparison is made against the displaced Poisson and Geometric distributions, as they describe the degrees spectra of randomly generated graphs (Erdős–Rényi and Random Geometric Graph respectively).

2 Results

In this section we present the main results. First, we compute some descriptive statistics of the languages' in-degree distributions: results are shown in **Table 1**. As we can see with the help of **Table 2**, both the network size (N) and the maximum degree ($max\,k$) show a high variability across languages; nevertheless, the mean degree ($mean\,k$) varies in a very little range. This suggests a pattern concerning the average number of incoming syntactic links which is not dependent on network size, but is related to linguistic phenomena.

Table 1: Degree Distributions Descriptive Statistics

lang	N	max k	mean k	N/M	lang	N	max k	mean k	N/M
Arabic	21065	2249	3.35	0.30	*English*	29172	4547	6.86	0.15
Basque	11868	576	2.18	0.46	*Greek*	12704	1081	3.52	0.28
Catalan	35524	5522	5.75	0.17	*Hungarian*	34600	6540	3.10	0.32
Chinese	35563	7645	5.20	0.19	*Italian*	13433	2678	4.23	0.24
Czech	66014	4727	3.97	0.25	*Turkish*	20403	6704	2.31	0.43

Table 2: Descriptive Statistics Summary

N	max k	mean k	N/M
Min. : 11868	Min. : 576	Min. : 2.180	Min. : 0.1500
Mean : 28035	Mean : 4227	Mean : 4.047	Mean : 0.2790
Max. : 66014	Max. : 7645	Max. : 6.860	Max. : 0.4600

We try to capture these phenomena by modelling the degree distribution. We estimate the optimal parameters for the *Zeta Family* and the *Null Models* through Maximum **Likelihood Estimation (MLE)**, and display the resulting values in **Table 3**.

Table 3: Optimized Hyperparameters

language	λ	q	γ_1	γ_2	$kmax$
Arabic	3.22	0.30	2.104	2.103	2249
Basque	1.83	0.46	2.358	2.356	576
Catalan	5.73	0.17	1.922	1.92	5522
Chinese	5.17	0.19	1.886	1.884	7645
Czech	3.89	0.25	2.051	2.049	4727
English	6.85	0.15	1.8	1.795	4547
Greek	3.41	0.28	2.132	2.129	1081
Hungarian	2.93	0.32	2.335	2.335	6540
Italian	4.16	0.24	2.108	2.107	2678
Turkish	2.00	0.43	2.542	2.543	6704

Next, by plugging the right optimal parameters into each distribution, we can compute the AIC score through the usual formula, corrected for sample size. The results are shown in **Table 4**: for each language, the best model (i.e. the one with lowest AIC) is highlighted in blue.

Table 4: AIC Scores

language	$AIC_{Poisson}$	AIC_{Geom}	AIC_{Zeta}	AIC_{TrZeta}	AIC_{Zeta2}
Arabic	302184.78	86052.47	61866.43	61863.52	62038.80
Basque	77497.05	35696.57	27332.69	27332.46	28177.86
Catalan	1040690.22	188691.87	126824.09	126810.13	127036.04
Chinese	750771.36	181101.78	132437.65	132422.96	132926.95
Czech	1145494.45	295916.13	204825.59	204816.74	204964.02
English	860038.26	166198.97	120501.84	120456.79	121926.45
Greek	193412.07	53408.67	36278.26	36274.83	36438.94
Hungarian	549610.72	134828.14	81358.52	81360.22	83578.80
Italian	285083.34	62157.72	39280.32	39279.83	39398.20
Turkish	233280.08	64550.40	39935.05	39937.03	42675.52

For a clearer understanding of the relative distances between models, we shown the AIC differences (ΔAIC) in **Table 5**. These values represent how far is each model from the best one, thus they are computed as $AIC_{l,model} - AIC_{l,min}$ for each language l.

Table 5: AIC Differences

lang	$\Delta AIC_{Poisson}$	ΔAIC_{Geom}	ΔAIC_{Zeta}	ΔAIC_{TrZeta}	ΔAIC_{Zeta2}
Arabic	240321.25	24188.95	2.91	0.00	175.28
Basque	50164.59	8364.11	0.24	0.00	845.41
Catalan	913880.09	61881.73	13.96	0.00	225.90
Chinese	618348.40	48678.82	14.69	0.00	503.99
Czech	940677.71	91099.39	8.85	0.00	147.28
English	739581.47	45742.18	45.05	0.00	1469.66
Greek	157137.24	17133.83	3.43	0.00	164.11
Hungarian	468252.20	53469.62	0.00	1.71	2220.28
Italian	245803.51	22877.90	0.49	0.00	118.37
Turkish	193345.03	24615.35	0.00	1.98	2740.47

Finally, we are able to compare the estimated probabilities with the empirical ones by plotting both distributions in double logarithmic scale. In this section we only display two illustrative examples of the resulting plots, for Catalan and Hungarian, while the figures for the remaining languages can be found at the end of the report (Section 5).

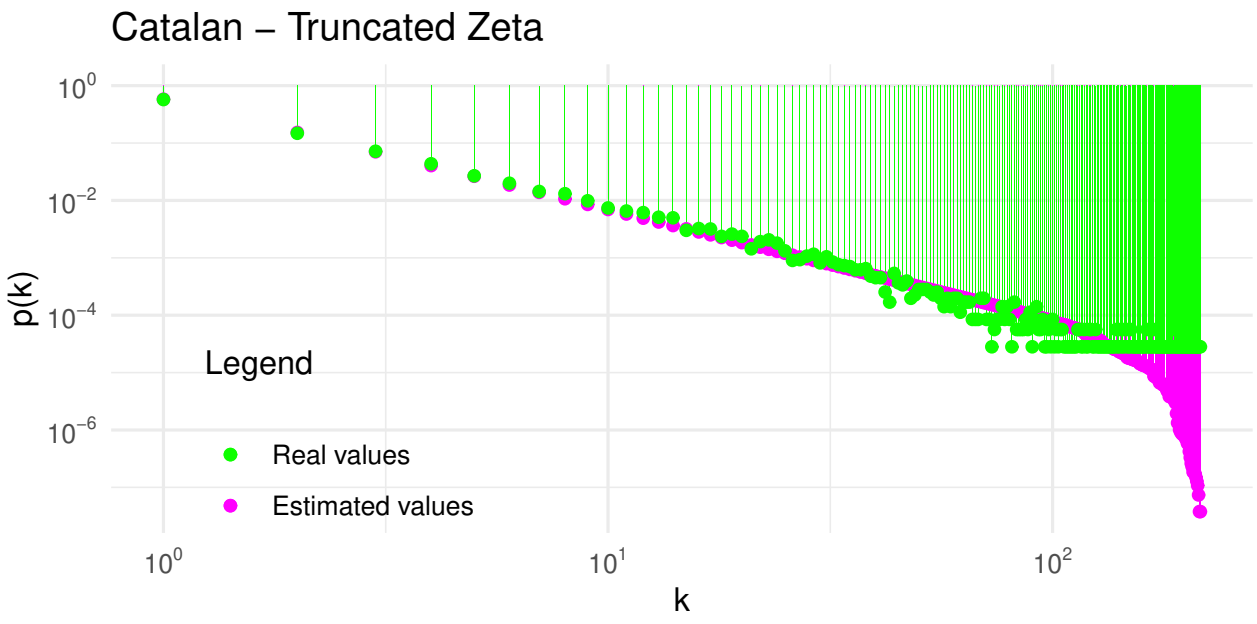

Figure 1: Real degree distribution and best model for Catalan. Best model is Truncated Zeta(1.92,5522). Log-Log scale

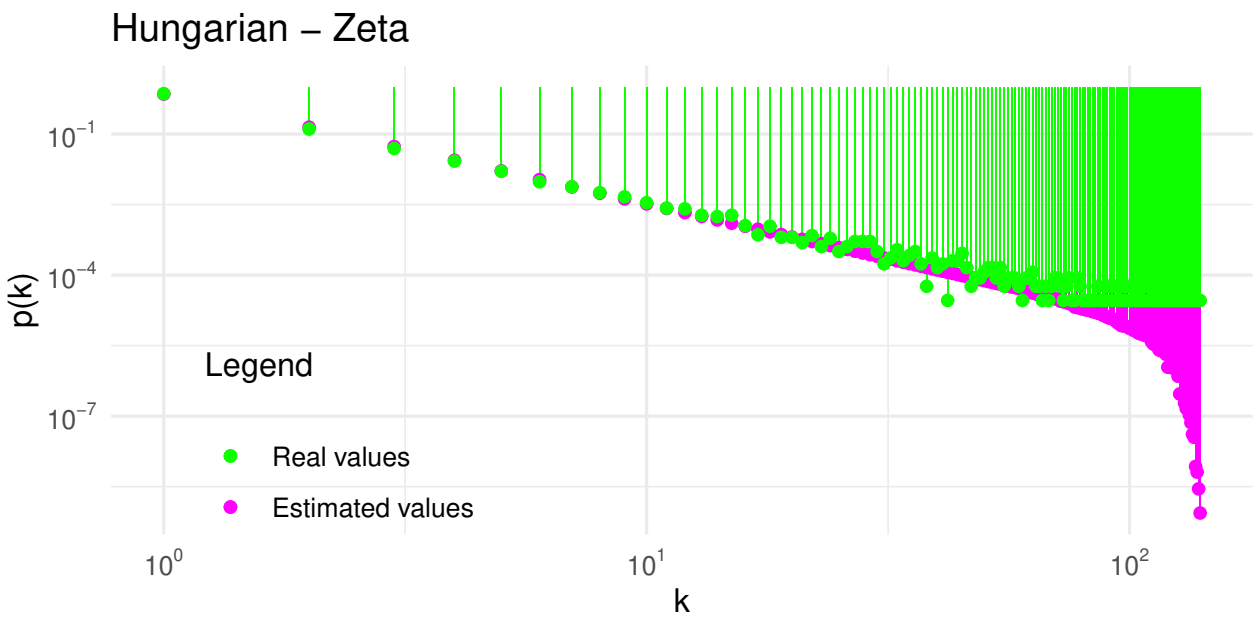

Figure 2: Real degree distribution and best model for Hungarian. Best model is Zeta(2.335). Log-Log scale

3 Methods

3.1 Maximum Likelihood Estimation

In the implementation of `mle()` in `R`, it is possible to specify a starting point for the exploration of the log-likelihood function, as well as upper and lower bounds. This is necessary, as the optimization requires the function to converge to some finite values, and we want to exclude from the search the possible parameters' values which would lead to undesired situations. Thus, we configure the estimators according to the following considerations:

- *Poisson*: we set the lower bound for λ to 0.0000001, in order to avoid dividing by 0 (starting point = M/N as suggested).
- *Geometric*: we impose $q \in]0, 1[$ to guarantee that $p(k) \in [0, 1]$ and it defines a probability (starting point = N/M as suggested).
- *Zeta*: since the Zeta Riemann function converges for $\gamma > 1$, we impose this constraint on the estimation of γ_1.
- *Truncated Zeta*: we set the maximum degree ($k\ max$) as a starting point, and we bound the search between 1 and N (number of nodes).

In addition to being very close to one another (more about this in the next section), γ_1 and γ_2 are also clearly fluctuating around the value of 2 (see **Table 3**). To make sure that this is not due to the choice of the starting value, we first set the start to 2 for both models, then to $\gamma_1 = 5$ and $\gamma_2 = 3$, and we compare the

obtained parameters. The resulting values are unchanged, meaning that 2 is indeed an ideal value for the *Zeta distribution family*.

3.2 Alternative Models

In addition to the Target models, we considered the suggested **Altman** distribution, and introduced a new possible model, with the following specification:

$$p(k) = \begin{cases} \frac{k^{-\gamma}}{H(\gamma, kmax)} & \text{if } k \leq t \\ \frac{1}{N} & \text{if } k > t \end{cases}$$

The proposed model is a Truncated Zeta distribution up to a certain value t (which should be approximately 100), and becomes constant to 1/N after it. This choice will be motivated in the following section. The computation of the likelihood and the implementation can be found in the code. Unfortunately, for both models the algorithm is unable to find a minimum: since these functions have additional parameters, and the proposed function also depends on the degree, the identification of the domain requires a scrupulous analysis.

4 Discussion

The Null Hypothesis of a random degree distribution was rejected for every language, in support of a *power-law model*. As shown in **Table 4**, the ***Zeta truncated*** model has consistently proven to be the best one, with the only exception of Hungarian and Turkish, for which the base ***Zeta*** provides a lower AIC. However, it is important to notice that the AIC scores for these two target models are very similar: we can better understand the distances between the models by looking at **Table 5**, reporting the *AIC Difference* between each model and the best one. There are three main aspects to take into account:

- *The distances within the Zeta family distributions are minimal.*
  In particular, the fixed parameter model Zeta(2) performs consistently worse than the models with one (Zeta) and two (Truncated Zeta) parameters, as expected by theory. Moreover, the difference between these two distributions does not seem to be significant: this is reasonable, since the estimated exponents (γ_1 and γ_2) are virtually the same, and $kmax$ is large, so that Zeta and Truncated Zeta will converge approximately to the same value (estimated parameters are shown in **Table 3**). As mentioned before, the values for γ_1 and γ_2 vary in a neighborhood of 2. Indeed, this value is often found empirically in the context of dependency networks [1] [2].

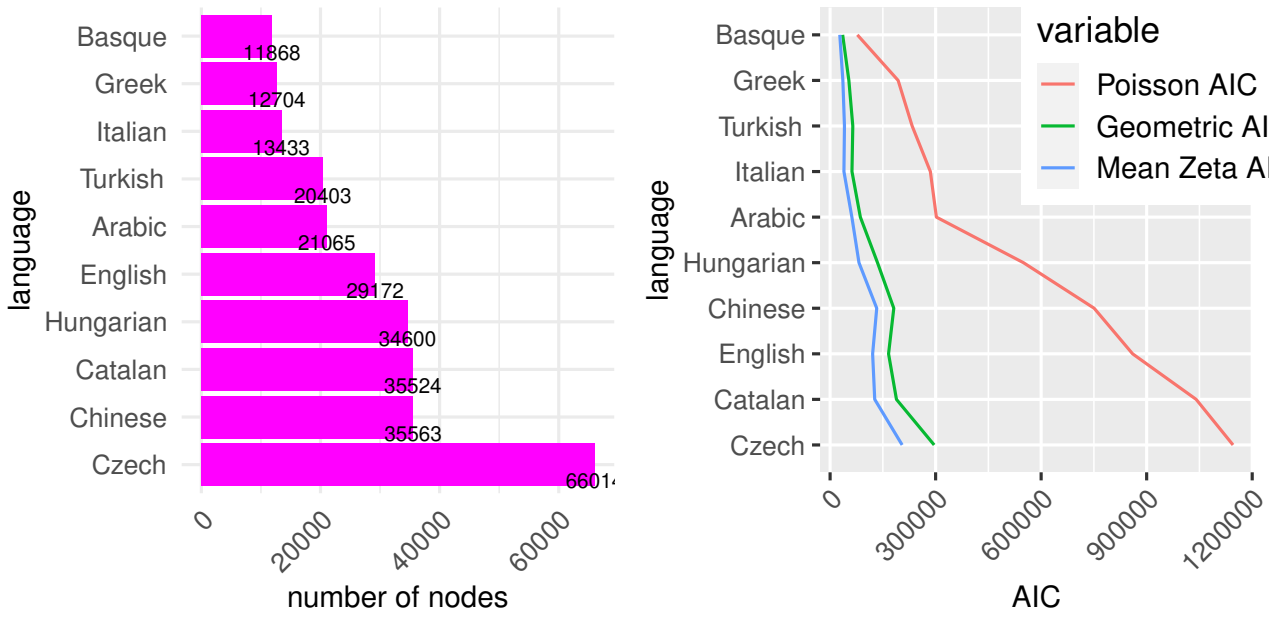

Figure 3: Real values and best model for

- *The Null Models' performance is considerably worse with respect to the Targets.*
  This is especially true for the Poisson distribution, whose AIC value is one order of magnitude higher than the one of the Geometric model. We can visualize this in **Figure 3** (right): by comparing the average performance of Zeta models with those of the two Null models, we see how the distance from the Poisson distribution is way higher for each language.

- *The magnitude of the distance between Null and Target models depends on the size of the network.*
  In **Figure 3** (left) we display the languages ordered by number of nodes (N), and we notice that, as N increases, the distance of the *Zeta family* from the Null models also increases, and this is especially true for the Poisson distribution. Indeed, we can roughly identify two groups of languages by setting a threshold at N=25000:

  - $N < 25000$: *Basque, Greek, Italian, Turkish, Arabic.*
    The AIC of the Target models is low, but very close to the AIC of the Nulls.
  - $N \geq 25000$: *English, Hungarian, Catalan, Chinese, Czech.*
    The AIC of the Targets is higher compared to the first group, but the Zeta Family and the Nulls are much further from each other.

  We can interpret the AIC difference as a measure of precision. In fact, the biggest the AIC Difference between Target and Null models, the highest the confidence in rejecting the Null Hypothesis. Again, this is a reasonable result, since we expect the accuracy of a parametric model to increase with the number of observations.

In **Figure 1** and **Figure 2** we display the resulting plots, in which the observed relative frequencies are compared with the estimated probabilities. We only show the distribution of the best model for each language

(the remaining figures are at the end of the document). As we can see, both Zeta and Truncated Zeta fit very well the data for low in-degrees, but as k increases they underestimate the probability of finding a node with degree k. This pattern is shown consistently by all the languages, and with roughly the same threshold at 10^2. However, we have to highlight that the degree distributions are heavily skewed to the left (the mean degree varies between 2.2 and 7.9, but the lower maximum degree, found in Basque, is 576), as we have also seen from the descriptive statistics at the beginning. This is related to the ***Dependency Distance Minimization*** principle: given that - even in long sentences - the syntactic distances tend to be small, this implies that the dependent words are generally positioned around their heads. Thus, the amount of links that a node can receive is bounded by physical distance, and is limited to the words located in its closest neighborhood. In particular, we notice that the real probability tends to converge to a minimum value, that is 1/N. In fact, since we assume that $p(0) = 0$, the least number of times a degree can appear is 1 out of N possible times. This observation has motivated the design of the alternative specification of the Zeta distribution, introduced above.

4.1 Conclusions

The in-degree distributions in the Global SDN have proven to approximate a power-law. We were able to reject the Null Hypothesis of a randomly generated network for every considered language: thanks to the Akaike Information Criterion we consistently identified as best model the ***Truncated Zeta distribution*** and the ***Zeta distribution***, with a very little distance between the two, but a very large distance from the Nulls. The estimated parameters for the *Zeta family* oscillate around 2, but the parametric models still perform better than the base specification of Zeta(2), as they work through optimization. We also highlighted how this distance increases as the number of nodes increases, allowing one to identify two groups of languages, based on network size, for which the estimated distribution would be more or less accurate. To fit the tail of the distribution, which is largely noisy, we have proposed a new model which does not allow the probability to reach infinitesimal values. Future work could involve the empirical implementation of such model.

5 Figures

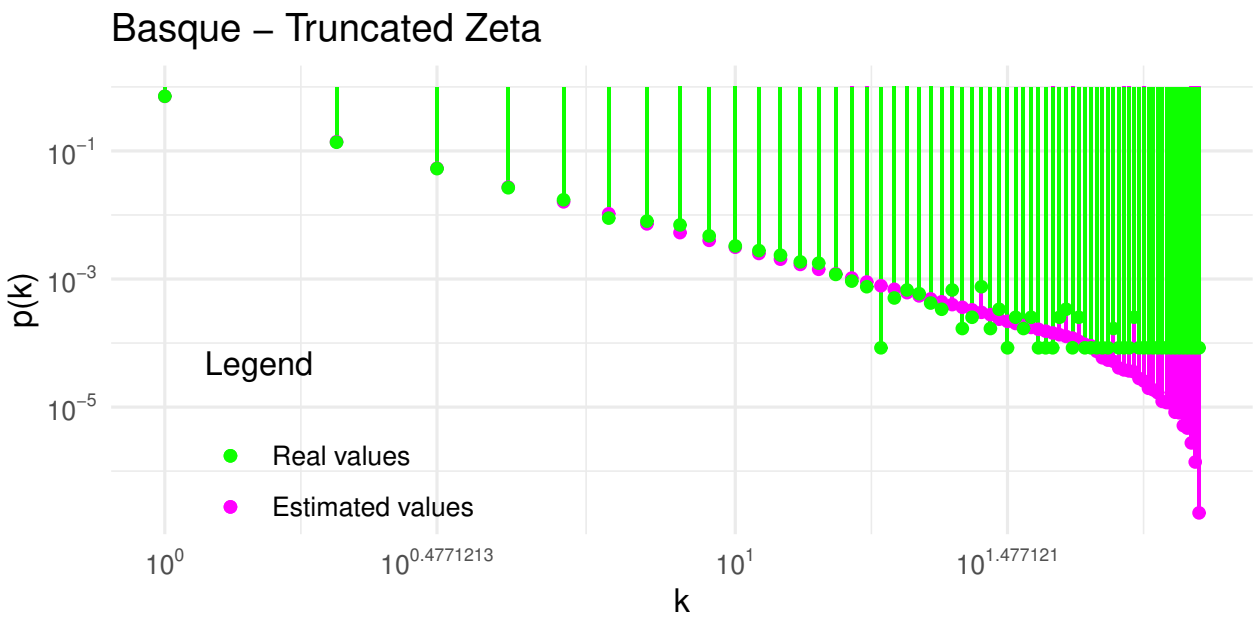

Figure 4: Real values and best model for Basque

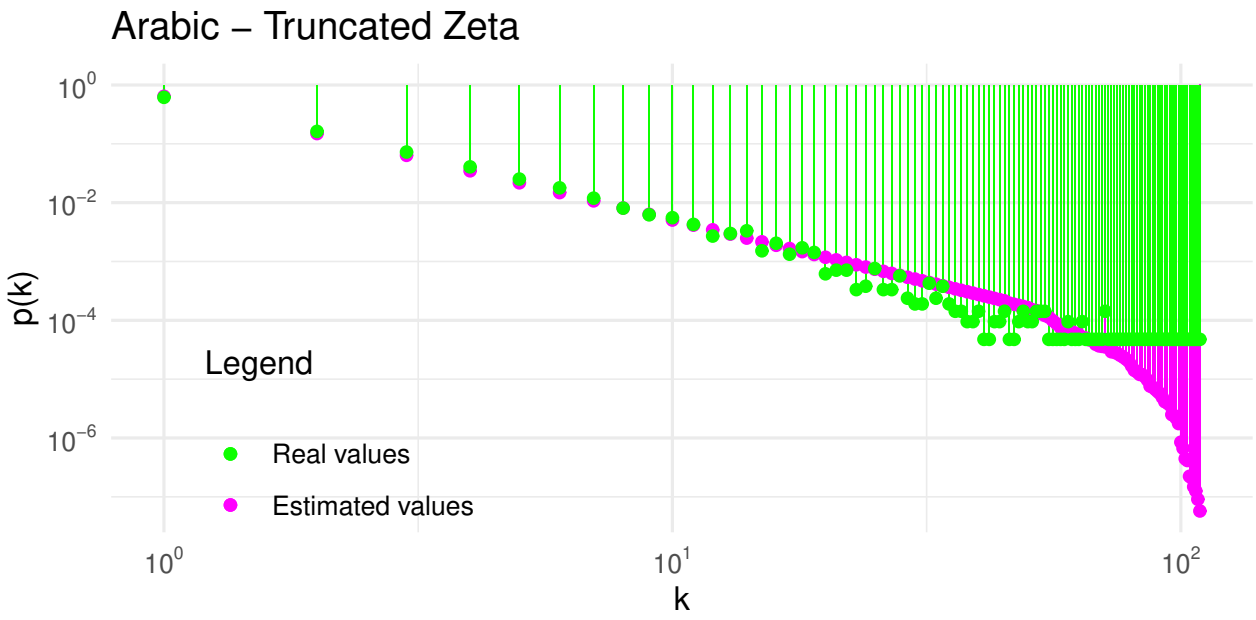

Figure 5: Real values and best model for Arabic

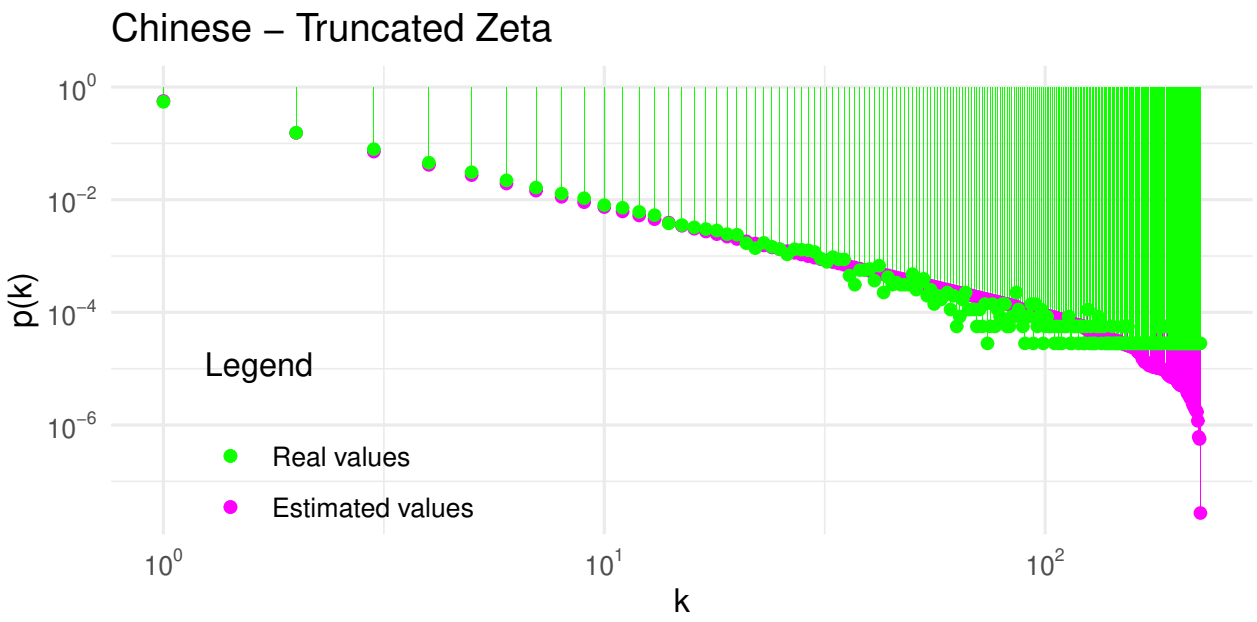

Figure 6: Real values and best model for Chinese

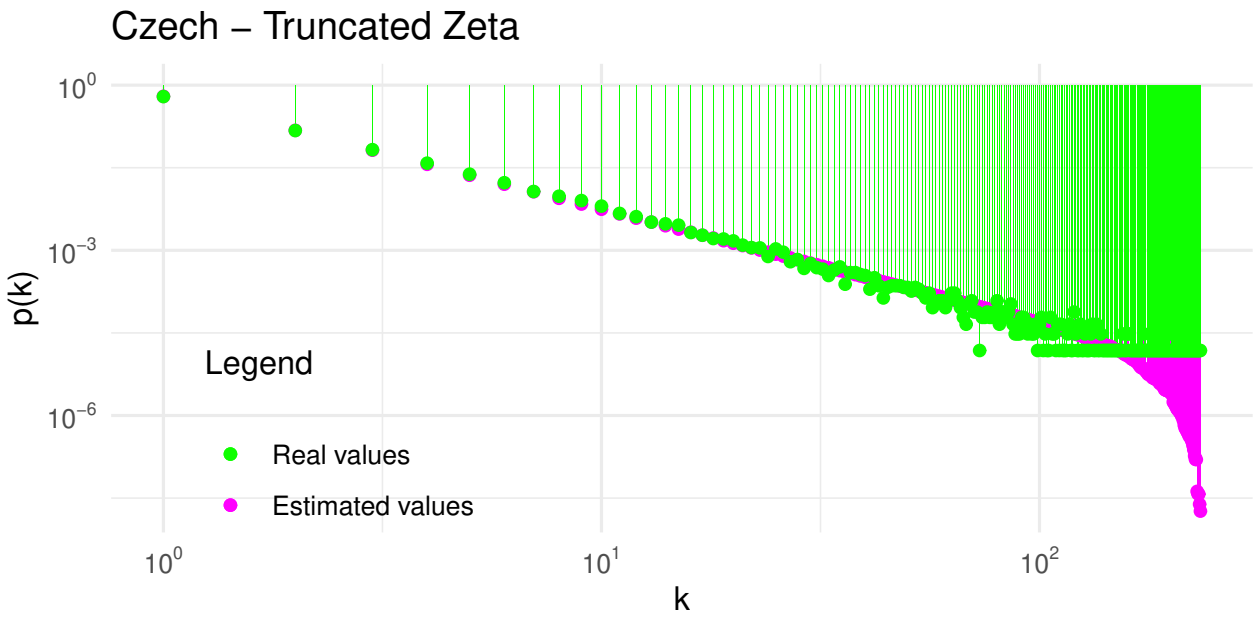

Figure 7: Real values and best model for Czech

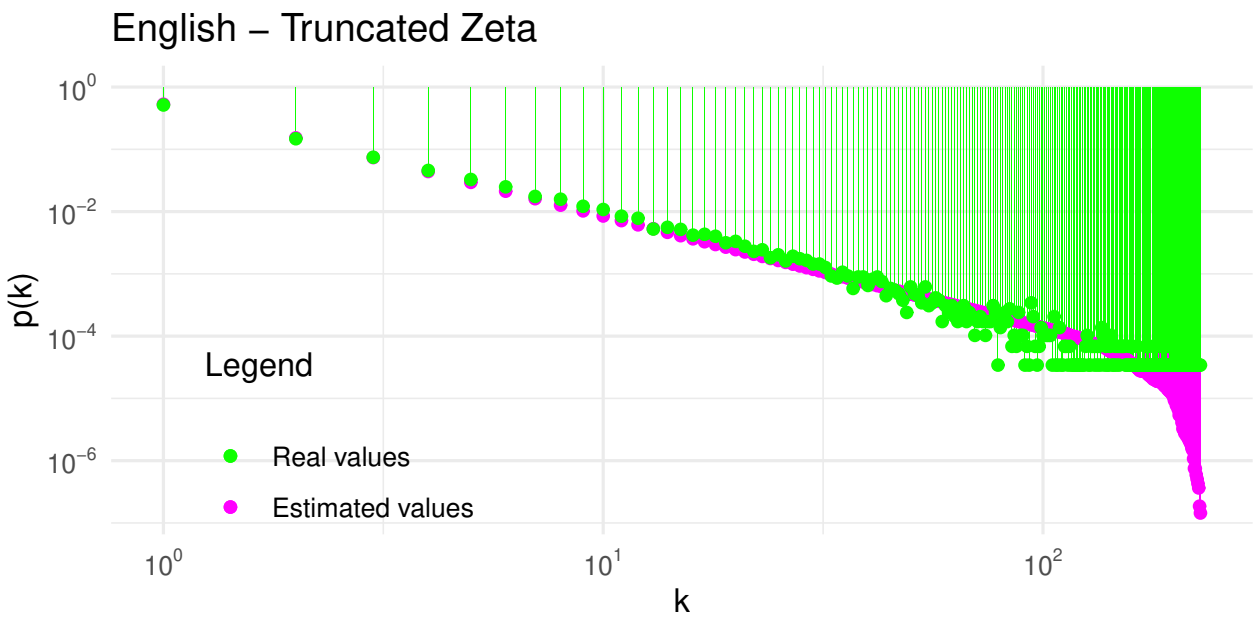

Figure 8: Real values and best model for English

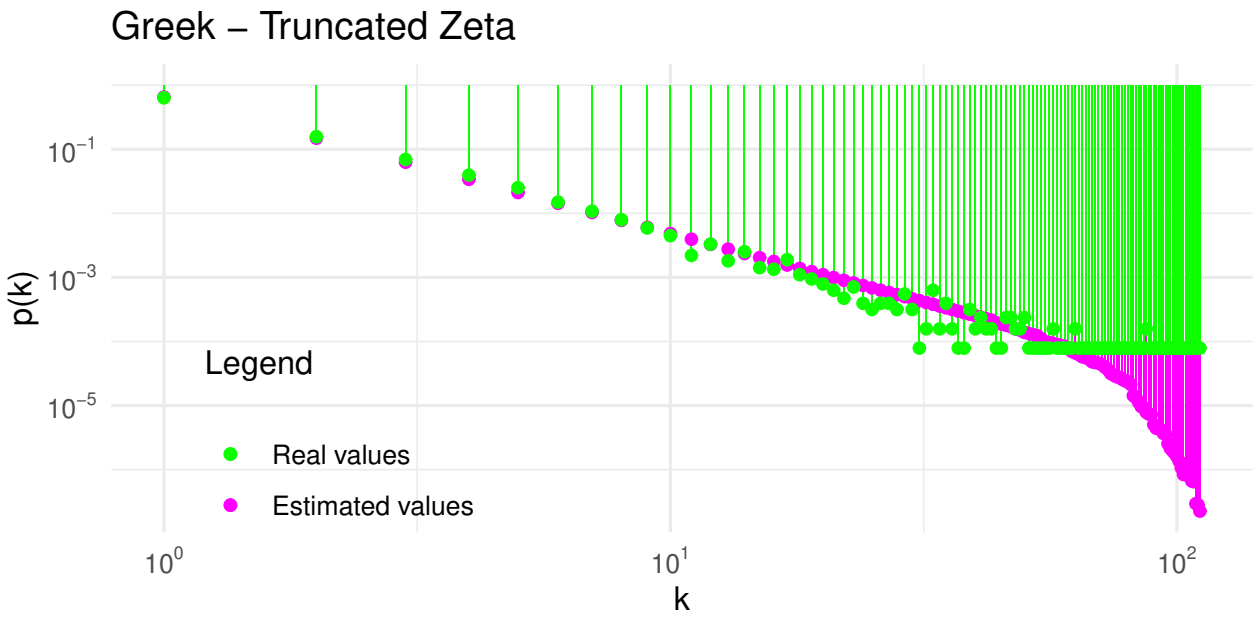

Figure 9: Real values and best model for Greek

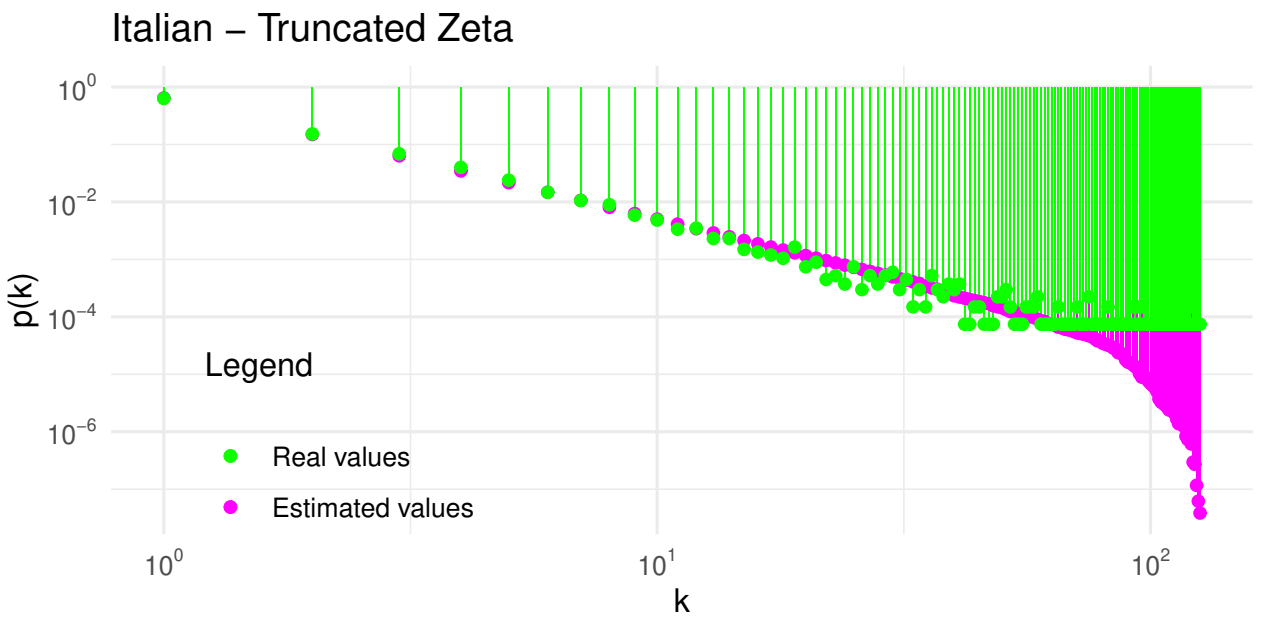

Figure 10: Real values and best model for Italian

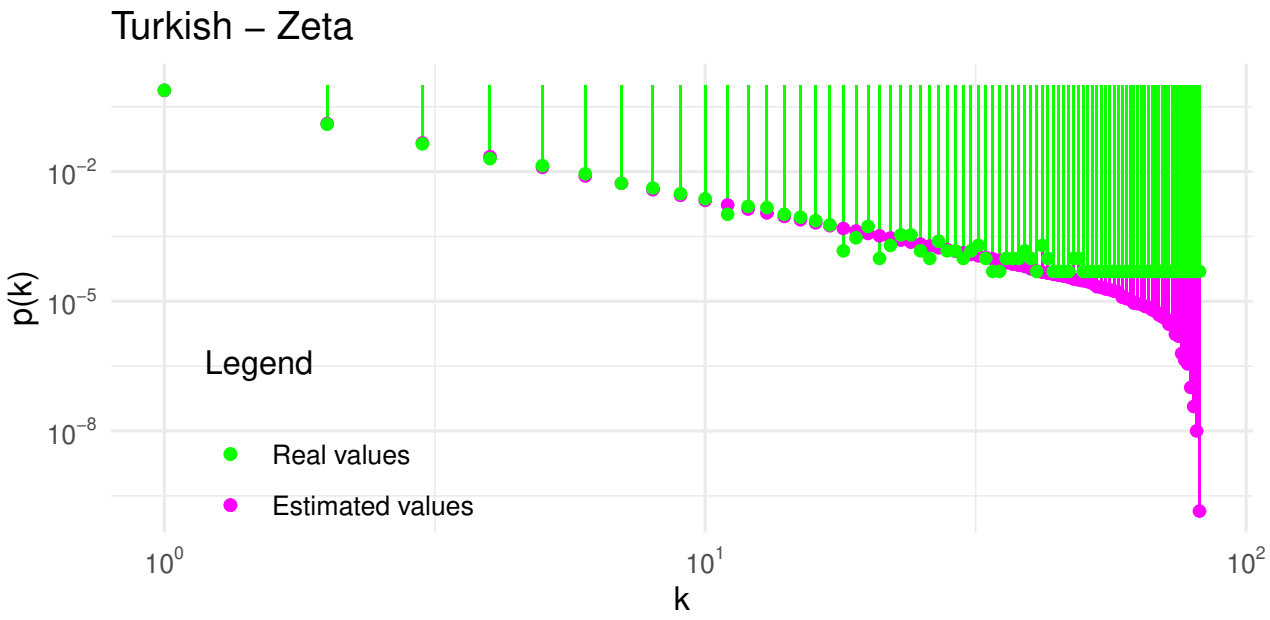

Figure 11: Real values and best model for Turkish

References

[1] Ferrer-i-Cancho, R., Sole, R.V., Kohler, R.: *Patterns in syntactic dependency networks*, Physical Review E 69, 051915 (2004).

[2] Corral, A., Boleda, G., Ferrer-i-Cancho, R.: *Zipf's law for word frequencies: word forms versus lemmas in long texts.* PLoS ONE, 10:e0129031 (2015).