Complex and Social Networks:
Lab Sessions Report 2

Sara Montese
Marius Behret

UPC - FIB
05/10/2023

1 Introduction

Syntactic dependency networks are representations of the grammatical structure of a sentence or text, where words are connected by labeled arcs that indicate syntactic relationships between them. Analyzing patterns in syntactic dependency networks can provide valuable insights into the structure and meaning of language. A property of the structure of a network that is typically investigated is the distribution of the network node degrees, where the degree of a node is the number of its neighbors. For any integer $k \geq 0$, the quantity p_k is the fraction of nodes having degree k. It can also be seen as the probability that a randomly chosen node in the network has degree k. The quantities p_k, for $k \geq 0$, represent the degree distribution of the network. Directed networks have two different degree distributions, the in-degree and the out-degree distributions. By using the joint distribution of the two, we can investigate the correlation of the in- and out-degrees of vertices. Many observed networks can be viewed as only samples of some true underlying network. In this work, we study global syntactic dependency networks from nine languages (Arabic, Basque, Catalan, Chinese, Czech, English, Greek, Hungarian and Italian) and estimate the out-degree distribution of the true underlying network from its sampled network. We do that by employing a maximum likelihood estimation (MLE) for six different probability mass functions, to identify the best-fit probability distribution for the degree sequences. MLE allows us to estimate the parameters of a distribution by minimizing the negative log-likelihood function. We invert the sign of the formulae used in MLE to ensure accurate parameter estimation. Additionally, we utilize the Akaike Information Criterion (AIC) with a correction for sample size to perform model selection and identify the most suitable distribution for our data. We show that the model of the right-truncated zeta fits the real distribution best, compared to null models. The paper is organized as follows. First, the results are shown, which include - for every language and model - a summary of the properties of the degree sequences, the maximum likelihood estimates of the parameters, the AIC and the plots of the degree distributions. In Section 3, we interpret and discuss the previously presented results. Lastly, Section 4 shows a detailed explanation of our applied methodology and the decisions we made in the process.

2 Results

2.1 Description of Data

General information of the nine text corpora used for the syntactic dependency analysis is shown in Table 1. The number of nodes is represented as N and the maximum degree indicates the most number of connections a node possesses. To understand the average degree, the ratio of total degrees M to N was computed while the inverse of this ratio, N/M, illustrates the spread of the network.

Table 1: Summary of the properties of the degree sequences

Language	N	Maximum degree	M/N	N/M
Arabic	15678	4896	4.50	0.22
Basque	6188	2097	4.18	0.24
Catalan	24727	6622	8.25	0.12
Chinese	23946	7537	7.73	0.13
Czech	41912	12671	6.26	0.16
English	17775	7040	11.25	0.09
Greek	9280	2737	4.82	0.21
Hungarian	25534	1020	4.20	0.24
Italian	12285	1671	4.63	0.22

2.2 Maximum Likelihood Estimation

In Table 2 the estimated parameters across all languages and models are displayed, together with the standard error. The λ belongs to the displaced Poisson distribution, the q is the parameter of the displaced geometric, γ_1 refers to the zeta distribution, γ_2 and k_{max} refer to the right-truncated zeta distribution and γ and δ to the Altmann distribution. More information on the models used can be seen in Section 4.

Table 2: Summary of the parameters estimated with Maximum Likelihood

	Model						
	1	2	4		5		6
Language	λ	q	γ_1	γ_2	k_{max}	γ	δ
Arabic	4.45 ± 0.02	0.22 ± 0	1.8 ± 0.01	1.8 ± 0.01	15678 ± 0.13	1.55 ± 0.01	0.02 ± 0
Basque	4.11 ± 0.03	0.24 ± 0	1.89 ± 0.01	1.88 ± 0.01	6188 ± 0.13	1.76 ± 0.02	0.01 ± 0
Catalan	8.25 ± 0.02	0.12 ± 0	1.59 ± 0	1.58 ± 0	24727 ± 0.06	1.25 ± 0	0.02 ± 0
Chinese	7.72 ± 0.02	0.13 ± 0	1.66 ± 0	1.66 ± 0	23946 ± 0.08	1.47 ± 0	0.01 ± 0
Czech	6.24 ± 0.01	0.16 ± 0	1.69 ± 0	1.69 ± 0	41912 ± 0.11	1.44 ± 0	0.02 ± 0
English	11.25 ± 0.03	0.09 ± 0	1.55 ± 0	1.53 ± 0	17775 ± 0.04	1.25 ± 0	0.01 ± 0
Greek	4.78 ± 0.02	0.21 ± 0	1.7 ± 0.01	1.69 ± 0.01	9280 ± 0.07	1.19 ± 0.02	0.05 ± 0
Hungarian	4.13 ± 0.01	0.24 ± 0	1.77 ± 0.01	1.77 ± 0.01	25534 ± 0.14	1.35 ± 0.01	0.05 ± 0
Italian	4.58 ± 0.02	0.22 ± 0	1.7 ± 0.01	1.7 ± 0.01	12285 ± 0.08	1.16 ± 0.07	0.06 ± 0

2.3 Calculation of AIC

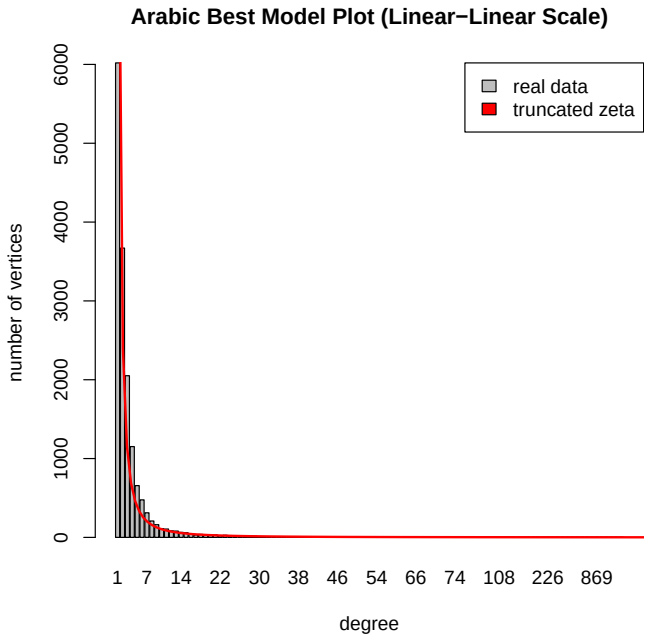
The AIC (1) of every model and language can be seen in Table 3. The model of each language with the lowest AIC has a 0 in the corresponding row of Δ. Models that with a value higher than 5000 of the lowest AIC show a >>0, while the other models that are higher than the lowest AIC but lower than the threshold of 5000 display a >0. It can be seen that the truncated zeta model has the lowest AIC across all languages.

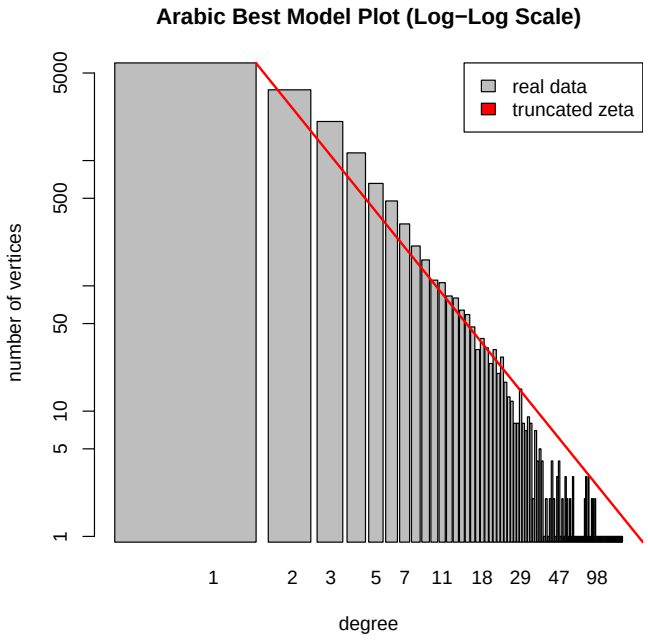
Table 3: AIC difference (Δ) of a model on a given source

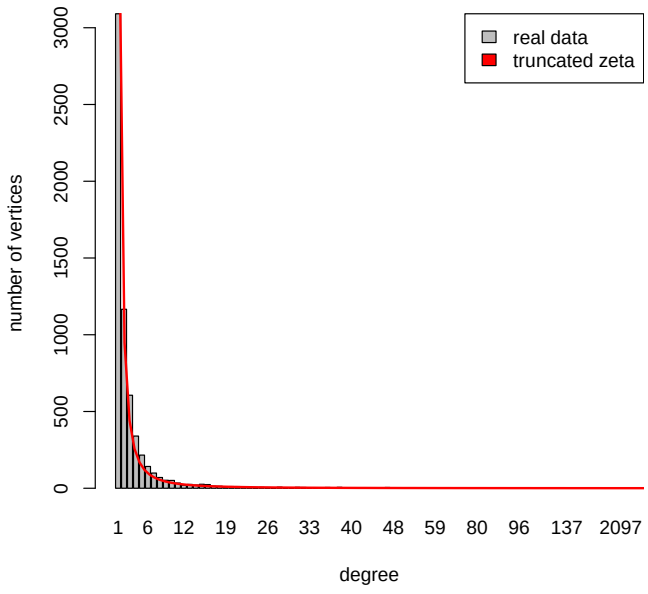
	Model				
	1	2	3	4	5
Arabic	268562.4	74763.7	65727.54	64942.68	64935.19
Δ	>>0	>>0	>0	>0	**0**
Basque	90108.35	28470.11	23085.07	23004.11	23002.65
Δ	>>0	>>0	>0	>0	**0**
Catalan	678249.04	150713.38	144317.52	136643.65	136550.02
Δ	>>0	>>0	>>0	>0	**0**
Chinese	723602.86	142572.77	123164.3	118841.21	118799.69
Δ	>>0	>>0	>0	>0	**0**
Czech	1024005.32	230438.45	205873.86	199874.02	199837.96
Δ	>>0	>>0	>>0	>0	**0**
English	752319.59	119980.52	113369.73	105772.02	105637.04
Δ	>>0	>>0	>>0	>0	**0**
Greek	134246.93	45696.26	44990.64	43754.21	43733.81
Δ	>>0	>0	>0	>0	**0**
Hungarian	274169.21	117692.4	111391.45	109641.22	109629.08
Δ	>>0	>>0	>0	>0	**0**
Italian	152883.69	59334.18	59035.91	57476.46	57455.51
Δ	>>0	>0	>0	>0	**0**

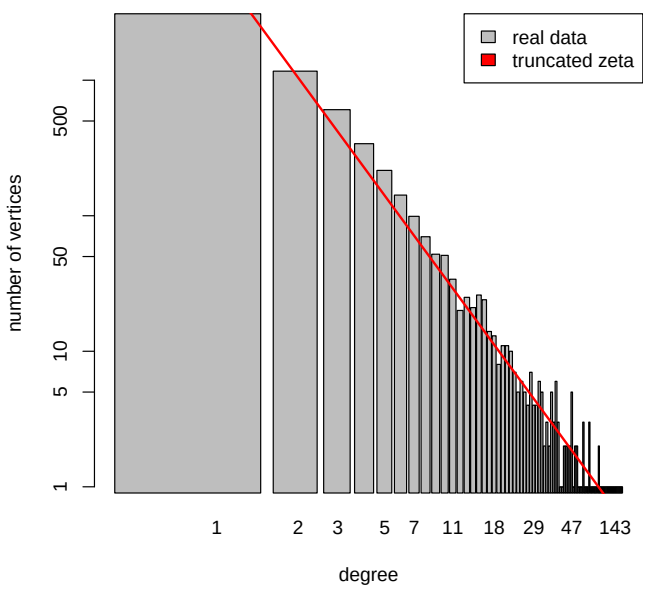
2.4 Best Model Plots

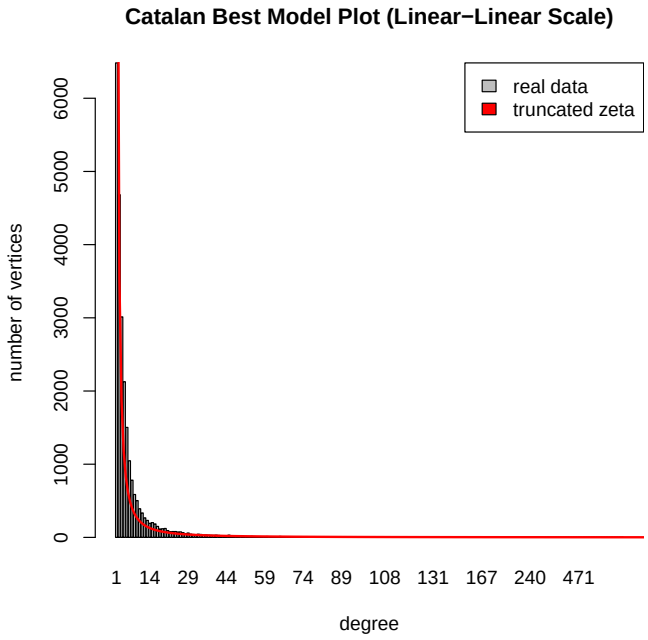
The following plots, displayed in linear-linear and log-log scales, show the best model, which is always the right truncated zeta, fitted to the real distribution of the data.

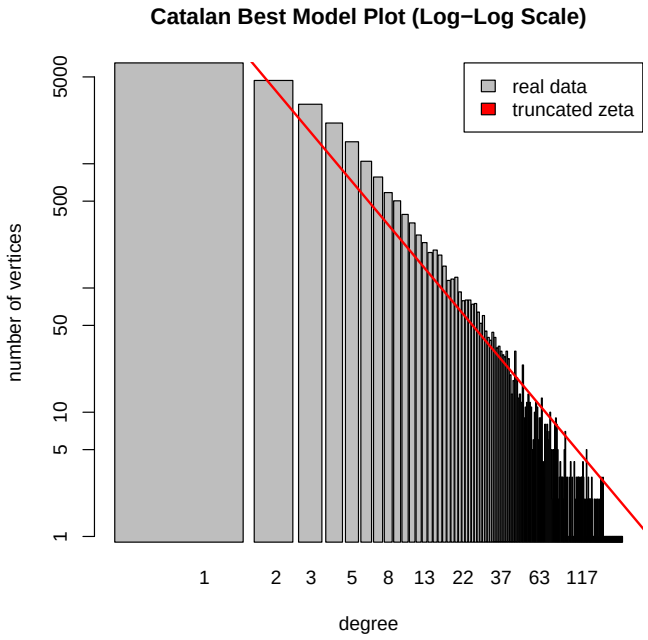

Arabic Best Model Plot (Linear−Linear Scale)
real data
truncated zeta
number of vertices
0
1000
2000
3000
4000
5000
6000
1
7
14
22
30
38
46
54
66
74
108
226
869
degree

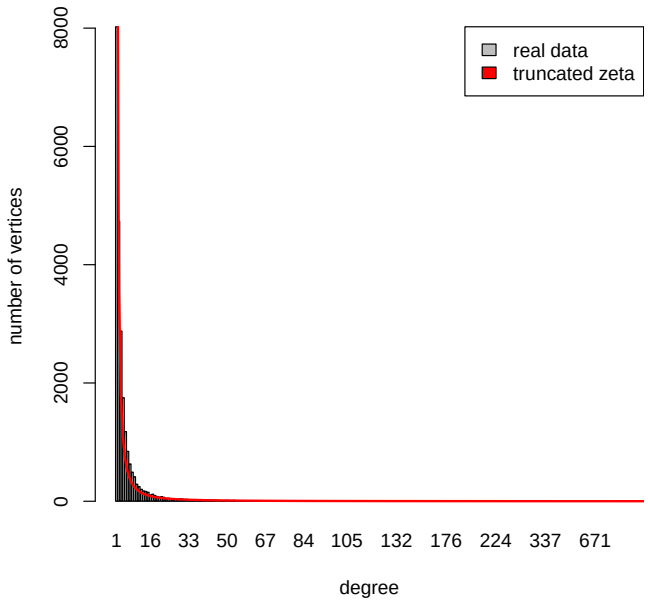

Arabic Best Model Plot (Log−Log Scale)
real data
truncated zeta
number of vertices
1
5
10
50
500
5000
1
2
3
5
7
11
18
29
47
98
degree

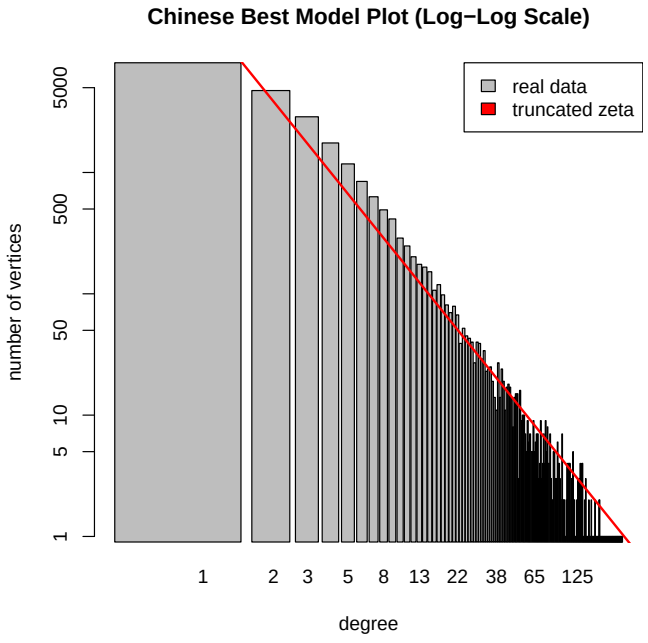

Basque Best Model Plot (Linear−Linear Scale)
real data
truncated zeta
number of vertices
0
500
1000
1500
2000
2500
3000
1
6
12
19
26
33
40
48
59
80
96
137
2097
degree

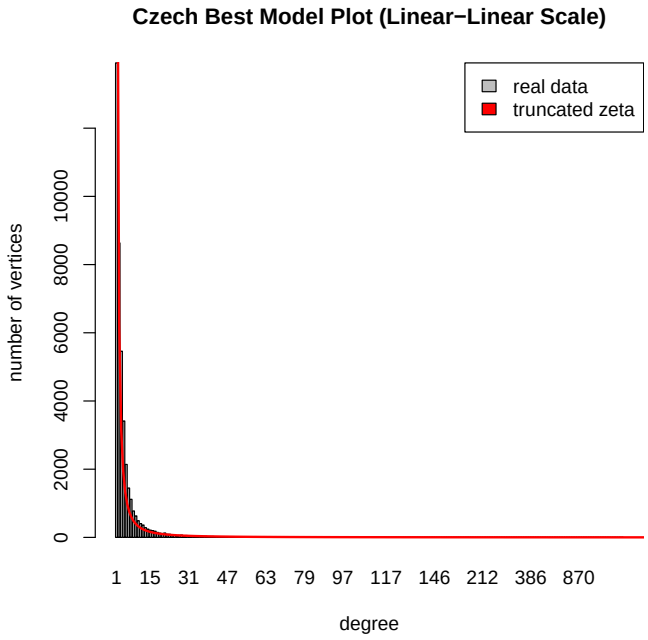

Basque Best Model Plot (Log−Log Scale)
real data
truncated zeta
number of vertices
1
5
10
50
500
1
2
3
5
7
11
18
29
47
143
degree

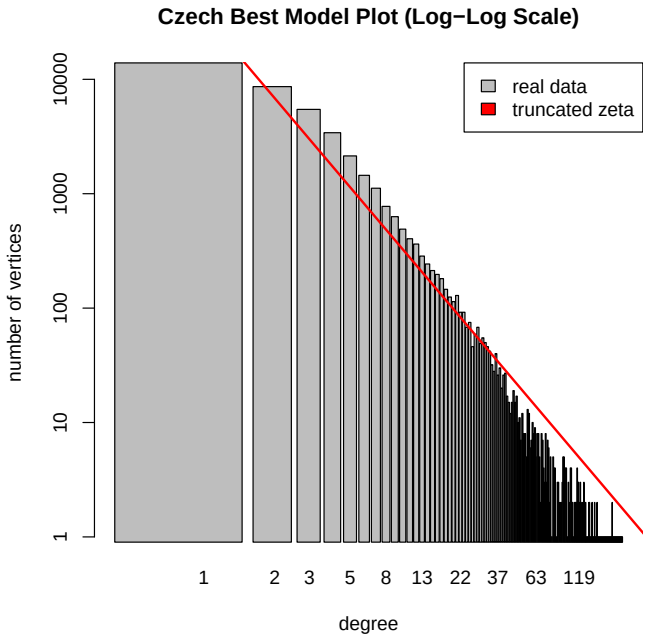

Catalan Best Model Plot (Linear−Linear Scale)
real data
truncated zeta
number of vertices
0
1000
2000
3000
4000
5000
6000
1
14
29
44
59
74
89
108
131
167
240
471
degree

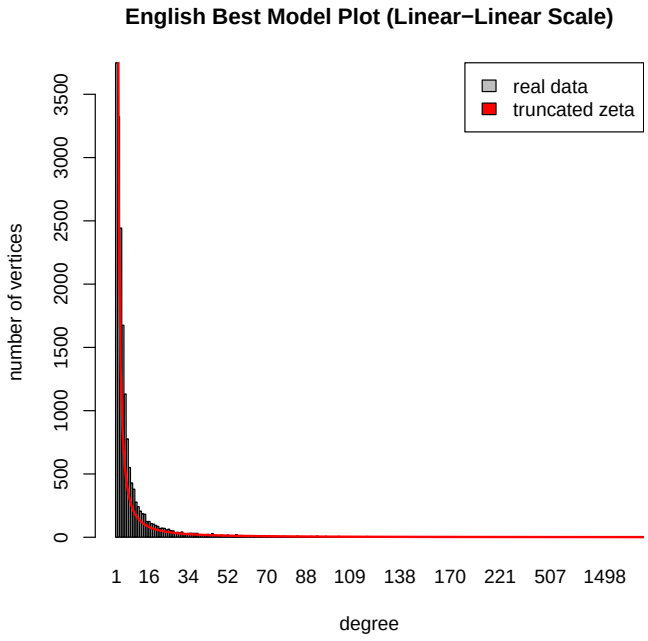

Catalan Best Model Plot (Log−Log Scale)
real data
truncated zeta
number of vertices
1
5
10
50
500
5000
1
2
3
5
8
13
22
37
63
117
degree

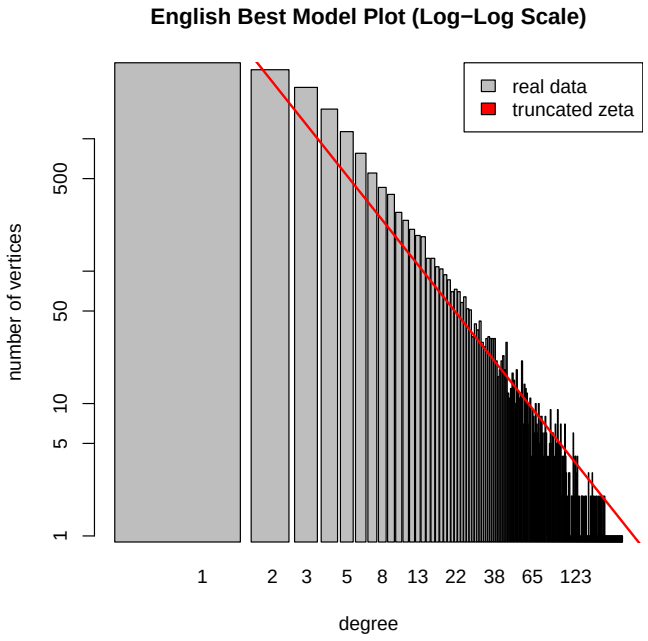

Chinese Best Model Plot (Linear−Linear Scale)
real data
truncated zeta
number of vertices
0
2000
4000
6000
8000
1
16
33
50
67
84
105
132
176
224
337
671
degree

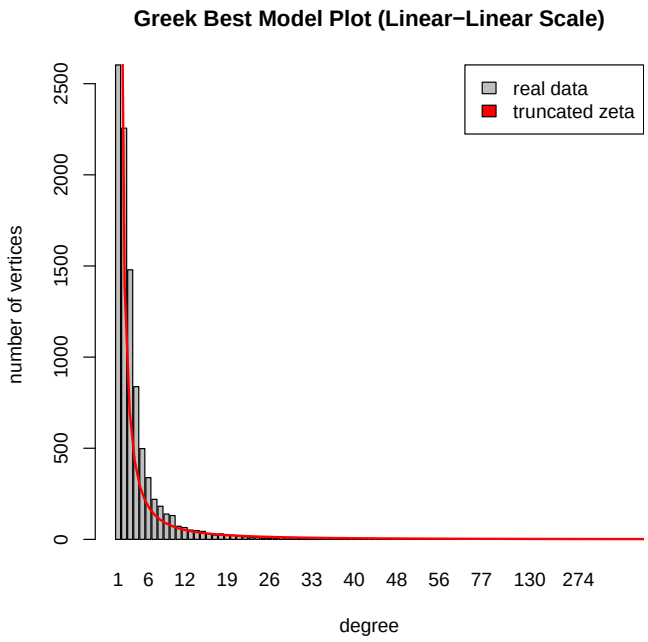

Chinese Best Model Plot (Log−Log Scale)
real data
truncated zeta
number of vertices
1
5
10
50
500
5000
1
2
3
5
8
13
22
38
65
125
degree

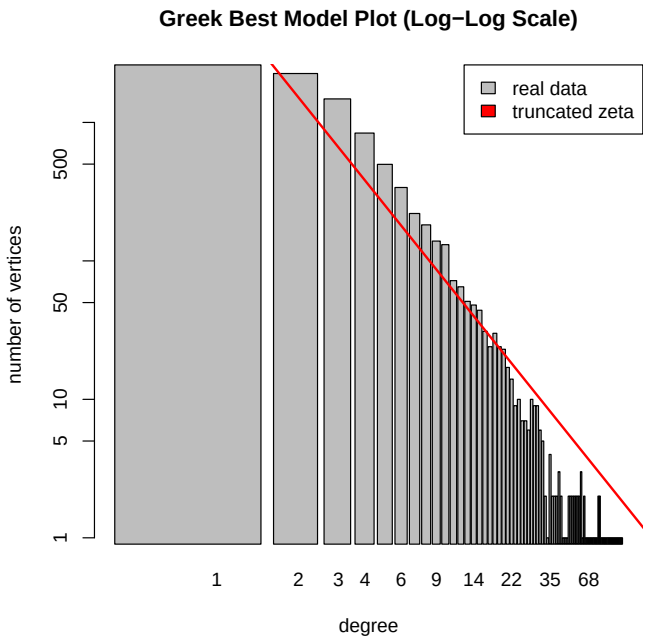

Czech Best Model Plot (Linear−Linear Scale)
real data
truncated zeta
number of vertices
0
2000
4000
6000
8000
10000
1
15
31
47
63
79
97
117
146
212
386
870
degree

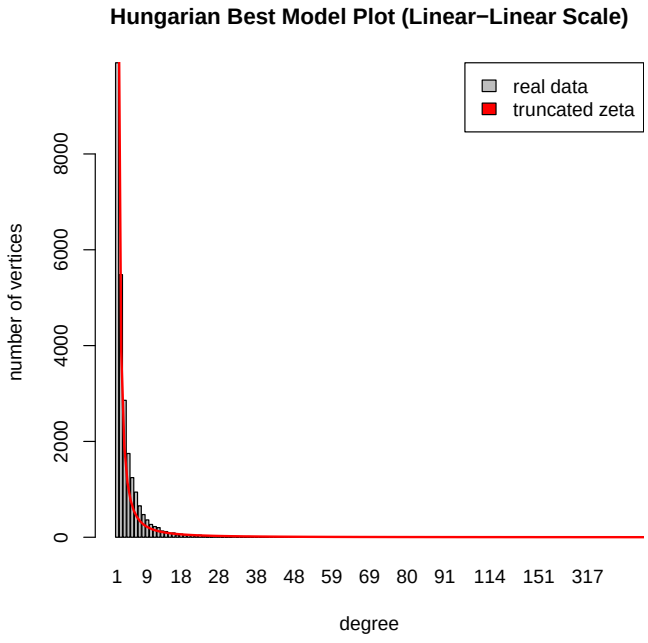

Czech Best Model Plot (Log−Log Scale)
real data
truncated zeta
number of vertices
1
10
100
1000
10000
1
2
3
5
8
13
22
37
63
119
degree

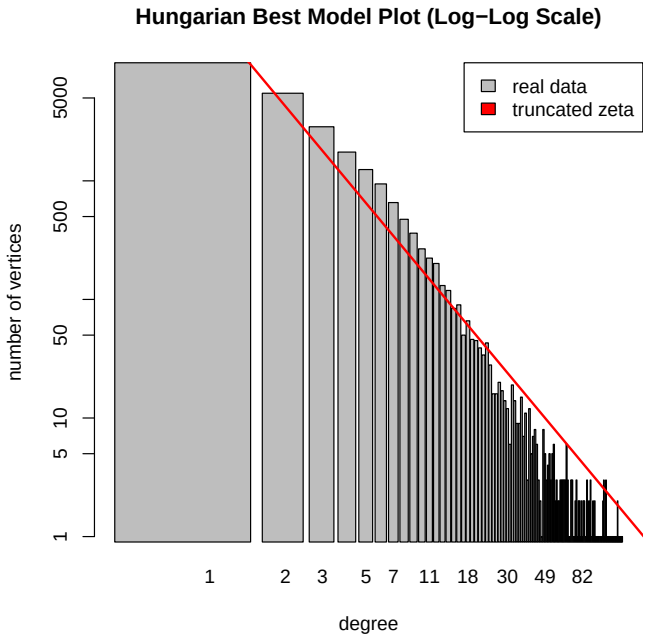

English Best Model Plot (Linear−Linear Scale)
real data
truncated zeta
number of vertices
0
500
1000
1500
2000
2500
3000
3500
1
16
34
52
70
88
109
138
170
221
507
1498
degree

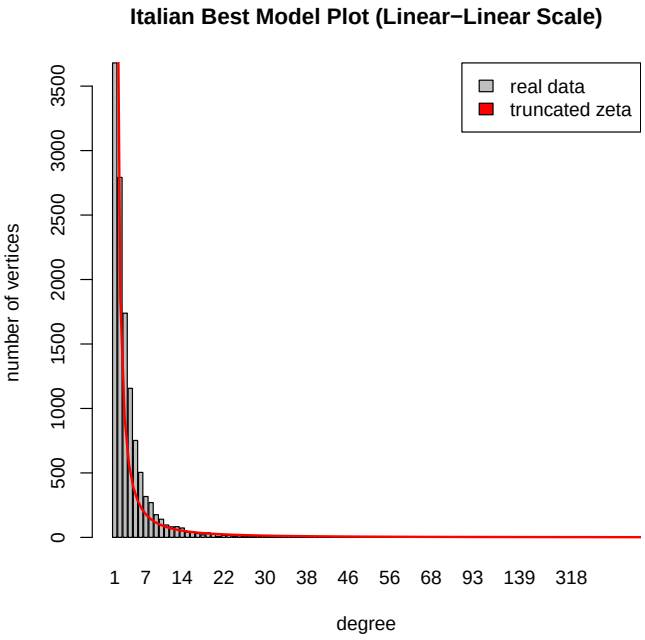

English Best Model Plot (Log−Log Scale)
real data
truncated zeta
number of vertices
1
5
10
50
500
1
2
3
5
8
13
22
38
65
123
degree

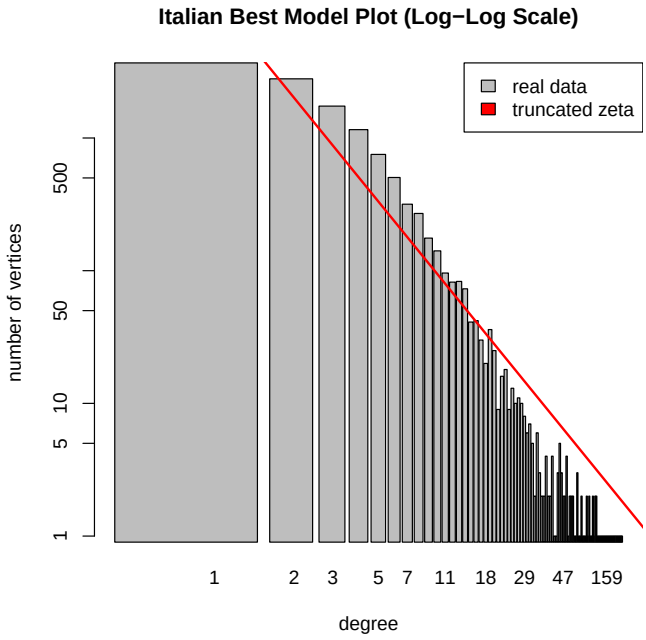

Greek Best Model Plot (Linear−Linear Scale)
real data
truncated zeta
number of vertices
0
500
1000
1500
2000
2500
1
6
12
19
26
33
40
48
56
77
130
274
degree


Greek Best Model Plot (Log−Log Scale)
real data
truncated zeta
number of vertices
1
5
10
50
500
1
2
3
4
6
9
14
22
35
68
degree


Hungarian Best Model Plot (Linear−Linear Scale)
real data
truncated zeta
number of vertices
0
2000
4000
6000
8000
1
9
18
28
38
48
59
69
80
91
114
151
317
degree


Hungarian Best Model Plot (Log−Log Scale)
real data
truncated zeta
number of vertices
1
5
10
50
500
5000
1
2
3
5
7
11
18
30
49
82
degree





3 Discussion

By analysing the AIC for each model, it can be seen that there is a significant difference between the fit of the distributions from null models (displaced Poisson and displaced geometric) compared to those of the zeta family. In general, null models perform poorly when trying to fit the networks under examination. Syntactic dependency networks can be highly complex due to the relationships between words in a sentence. Simple null models that assume randomness often oversimplify the complexity of linguistic dependencies, leading to inaccurate representations. In particular, the displaced Poisson is the model that is performing the worst. On the other hand, power-law models are a good fit for this kind of networks. For each language, the right-truncated zeta distribution is the best fit among the zeta family distributions. When we look at the real degree distributions of the languages, all of them follow a similar trend. In the visualizations of the right-truncated zeta distribution, we can see that it gives a reasonably good fit to the real distributed data. It shows a straight line in the log-log scaled plots and points to a dependency between the degree and the number of vertices. Furthermore, by analysing the Altmann distribution, it can be seen that it outperforms the best model for each language.

The differences in relative performance of other models and the absolute AIC values hint at the unique characteristics and complexities of individual languages. While statistical models can capture general patterns, the nuances and idiosyncrasies of each language can lead to variations in model performance.

4 Methods

4.1 Fitting empirical distributions to theoretical models

The goal is fitting an observed empirical data sample to a theoretical distribution model. The fitting problem can be split in three main tasks:

1. Selection of a suitable theoretical model;
2. Estimation of the model parameters;
3. Evaluation of the model.

To investigate the best model describing each degree sequence, we perform model selection considering an ensemble of degree distributions that is summarized in Table 4. The ensemble contains two distributions from null models of networks (Poisson and geometric distributions), three nested variants of the zeta distribution and the Altmann distribution. Here, N is number of nodes of the network, M is the sum of the degrees, and C is the sum of the logarithm of degree factorials. For simplicity, we assume that $k_i \geq 1$. This means that unlinked nodes, if any, are removed. By imposing this condition, we consider variants of the null models: the displaced Poisson and the displaced geometric distribution. Before applying standard model selection methods, the parameters giving

Table 4: The log-likelihood $\mathcal{L}$ for each of the probability mass functions

Model	Function	K	$\mathcal{L}$
1	Displaced Poisson	1	$M \log \lambda - N(\lambda + \log(1 - e^{-\lambda})) - C$
2	Displaced geometric	1	$(M - N)\log(1 - q) + N \log q$
3	Zeta with $\gamma = 2$	0	$-2M' - N \log \frac{\pi^2}{6}$
4	Zeta	1	$-\gamma M' - N \log \zeta(\gamma)$
5	Right-truncated zeta	2	$-\gamma M' - N \log H(k_{\max}, \gamma)$
6	Altmann (2)	2	$N \log c - \gamma \sum_{i=1}^{N} \log k_i - \delta \sum_{i=1}^{N} k_i$ (3)

the best fit must be obtained. To achieve this, for each model, we:

- compute the Maximum Likelihood estimator of the space of possible parameters;
- maximize the negative log-likelihood, by indicating the range of variation of the parameters.

The initial values and the bounds for the parameters to estimate are shown in Table 5. Given the set of models, we calculate the Akaike's Information Criterion (AIC) for each model and select the one with the lowest AIC value. The model with the lowest AIC is considered the best trade-off between goodness of fit and model complexity. In this analysis we make use of the corrected AIC, defined as follows:

$$\mathrm{AIC} = -2\mathcal{L} + \frac{2KN}{N - K - 1} \tag{1}$$

where K is the number of free parameters and N is the number of vertices of the network. Table 3 summarizes the AIC for each model. For each language, we highlight $\Delta = AIC$ - AIC_{best}, the so-called AIC difference. By comparing the AIC values and the estimates of $-2\mathcal{L}$ in Table 6, it can be seen that these values are very close. This implies that our metric is not heavily penalized

Table 5: Initial values and bounds for MLE

Model	Parameters	Initial values	Bounds
1	λ	M/N	$\lambda \geq 0$
2	q	N/M	$0 \leq q < 1$
3	-	-	-
4	γ_1	2	$\gamma_1 > 1$
5	γ_2	3.0000001	$\gamma_2 > 1$
	$k_{\max}$	N	$max(k) < k_{\max} \leq N$
6	γ	1	$\gamma > 0.01$
	δ	0.1	$\delta > 0.01$

by the number of parameters in the model. Finally, we investigate a new probability distribution that is able to give a better fit than the best model so far. In particular, we study the Altmann function, defined as follows:

$$p(k) = ck^{-\gamma}e^{-\delta k} \tag{2}$$

if $1 \leq k \leq$ N and $p(k) = 0$ otherwise, with

$$c = \frac{1}{\sum_{k=1}^{N} k^{-\gamma}e^{-\delta k}}$$

The log-likelihood $\mathcal{L}$ for the Altmann probability mass functions is

$$\begin{aligned}
\mathcal{L} &= \log\left(\prod_{i=1}^{N} p(k_i)\right) \\
&= \sum_{i=1}^{N} \log p(k_i) \\
&= \sum_{i=1}^{N} \log ck_i^{-\gamma}e^{-\delta k_i} \\
&= \sum_{i=1}^{N} \log[c + k_i^{-\gamma} + e^{-\delta k_i}] \\
&= \sum_{i=1}^{N} \log c + \sum_{i=1}^{N} \log k_i^{-\gamma} + \sum_{i=1}^{N} \log e^{-\delta k_i}] \\
&= N \log c - \gamma \sum_{i=1}^{N} \log k_i - \delta \sum_{i=1}^{N} k_i
\end{aligned} \tag{3}$$

From Eq.(1) and Eq.(3) we compute the AIC of the Altmann function. In Table 7 it can be seen that Altmann distribution outperforms the best model for each language.

Table 6: Estimates of maximum of likelihood function ($-2\mathcal{L}$)

		Model				
		1 Displaced Poisson	2 Displaced Geometric	3 Zeta with $\gamma = 2$	4 Zeta	5 Right-trucanted zeta
	Arabic	268560.4	74761.7	65727.54	64940.68	64931.19
	Basque	90106.35	28468.11	23085.07	23002.1	22998.65
	Catalan	678247	150711.4	144317.5	136641.6	136546
	Chinese	723600.9	142570.8	123164.3	118839.2	118795.7
$-2\mathcal{L}$	Czech	1024003	230436.5	205873.9	199872	199834
	English	752317.6	119978.5	113369.7	105770	105633
	Greek	134244.9	45694.26	44990.64	43752.21	43729.81
	Hungarian	274167.2	117690.4	111391.4	109639.2	109625.1
	Italian	152881.7	59332.18	59035.91	57474.46	57451.51

Table 7: AIC comparison between the best model and Altmann function

	Model	
	AIC_{best}	Altmann
Arabic	64935.19	64189.14
Δ	>0	0
Basque	23002.65	22901.82
Δ	>0	0
Catalan	136550.02	133329
Δ	>0	0
Chinese	118799.69	117474.3
Δ	>0	0
Czech	199837.96	196969
Δ	>0	0
English	105637.04	103449.1
Δ	>0	0
Greek	43733.81	42414.17
Δ	>0	0
Hungarian	109629.08	107193.4
Δ	>0	0
Italian	57455.51	55561.62
Δ	>0	0

4.2 Tests with artificial datasets

To be more certain about the correctness of the results on the ensemble of distributions, we run some tests on artificial datasets where the true distribution is known a priori. In particular, we consider three zeta distributions with γ in $\{2, 2.5, 3\}$ and two geometric distributions with q in $\{0.05, 0.4\}$. For both experiments, we proved that our methods are able to select the right distribution and

obtain the right parameters of the distribution. Results for artificial zeta datasets are showns in Tables 8 and 9.

Table 8: Models' AIC on artificial zeta datasets

z_{value}	Displaced Poisson	Displaced Geometric	*Zeta* (with $\gamma = 2$)	*Zeta* (without fixed γ)	Truncated *Zeta*	Altmann distribution
2	20284.24	5040.10	**3346.50**	3348.14	3348.71	3372.20
2.5	4610.11	2530.99	2222.16	**2122.12**	2124.07	2122.44
3	3084.22	1587.19	1657.44	**1363.12**	1365.12	1365.19

Table 9: Parameters estimates for artificial zeta datasets

Z_value	Model	Parameters Estimates
2.0	1	$\lambda = 5.04$
	2	q = 0.20
	3	$\gamma =$ **2**
	4	$\gamma = 1.98$
	5	$\gamma = 1.97$ $k\_max = 1000$
	6	$\gamma = 1.88$ $\delta = 0.01$
2.5	1	$\lambda = 1.37$
	2	q = 0.54
	3	$\gamma = 2$
	4	$\gamma =$ **2.45**
	5	$\gamma =$ **2.45** $k_{max} = 1000$
	6	$\gamma = 2.38$ $\delta = 0.01$
3.0	1	$\lambda = 0.66$
	2	q = 0.73
	3	$\gamma = 2$
	4	$\gamma =$ **3.0**
	5	$\gamma =$ **3.0** k_{max}= 1000
	6	$\gamma = 2.97$ $\delta = 0.01$