Master in Data Science

Complex and Social Networks Laboratory

Analysis of Degree Distributions in Language Dependency Networks

Raffaele D'Agostino
raffaele.d.agostino@estudiantat.upc.edu

Gabriele Villa
gabriele.villa@estudiantat.upc.edu

Project Report

Academic Year 2025/2026

October 15, 2025

Contents

1 Introduction

In this study, we analyzed the *in-degree distribution* of words within syntactic dependency networks constructed from a collection of 1,000 sentences for each language included in the *Parallel Universal Dependencies (PUD)* dataset (a multilingual corpus designed to provide consistent syntactic annotations across multiple languages). For each of the 21 PUD languages, we built a directed graph where each word is represented as a node, and each edge connects a word to its syntactic "head" (as indicated in the corresponding CoNLL-U file).

We then investigated which theoretical probability distributions best fit the empirical in-degree data. Specifically, we considered five candidate models:

Model	Distribution
1	Displaced Poisson
2	Displaced geometric
3	Zeta ($\gamma = 2$)
4	Zeta
5	Truncated Zeta

Table 1: Model numbering and corresponding probability distributions.

For each model, we estimated the optimal parameter values by maximizing the log-likelihood function. We then computed the *Akaike Information Criterion (AIC)* for all fitted models and identified, for each language, the model with the lowest AIC and RMSE values. Finally, we compared these best-fitting models to the *Altmann distribution* in order to assess whether the Altmann model could provide an even better explanation of the observed in-degree distributions.

2 Results

2.1 Basic Network Statistics

Once the dependency network was constructed for each language, we computed several fundamental statistical measures to characterize the network topology. For each language corpus, we calculated the number of nodes N (representing unique lexical items), the sum of all in-degrees M (corresponding to the total number of dependency relations in the network), and the mean degree $\langle k \rangle = M/N$. Additionally, we computed the inverse mean degree N/M.

Language	N	Maximum degree	M/N	N/M
Arabic	4578	564	3.588 ± 0.295	0.279 ± 0.023
Chinese	5222	881	3.268 ± 0.270	0.306 ± 0.025
Czech	5009	688	3.007 ± 0.258	0.333 ± 0.029
English	4478	857	3.800 ± 0.362	0.263 ± 0.025
Finnish	3839	574	3.038 ± 0.281	0.329 ± 0.030
French	4476	1394	4.094 ± 0.511	0.244 ± 0.031
Galician	4294	1363	4.091 ± 0.513	0.244 ± 0.031
German	5119	1527	3.398 ± 0.409	0.294 ± 0.035
Hindi	4373	809	3.570 ± 0.314	0.280 ± 0.025
Icelandic	4645	507	3.212 ± 0.276	0.311 ± 0.027
Indonesian	3646	579	4.154 ± 0.303	0.241 ± 0.018
Italian	4619	1600	3.937 ± 0.502	0.254 ± 0.032
Japanese	4715	1003	4.260 ± 0.466	0.235 ± 0.026
Korean	7765	520	1.831 ± 0.085	0.546 ± 0.025
Polish	4809	726	3.103 ± 0.257	0.322 ± 0.027
Portuguese	5711	1217	3.342 ± 0.374	0.299 ± 0.034
Russian	4891	793	3.168 ± 0.259	0.316 ± 0.026
Spanish	4343	1250	4.040 ± 0.491	0.248 ± 0.030
Swedish	4773	539	3.302 ± 0.281	0.303 ± 0.026
Thai	3963	380	4.120 ± 0.236	0.243 ± 0.014
Turkish	4428	577	3.073 ± 0.230	0.325 ± 0.024

Table 2: Network statistics for dependency networks across all 21 languages.

2.2 Parameter Estimation

Table 3 presents the estimated parameters for each probabilistic model fitted to the degree distributions of the dependency networks across all languages. These parameters were obtained through maximum likelihood estimation (MLE), which involves finding the parameter values that maximize the log-likelihood function for each model given the observed degree distribution data.

Each estimated parameter is accompanied by its associated uncertainty, derived from the standard deviation of the estimate. These standard deviations provide a measure of the precision of the parameter estimates and reflect the variability inherent in the fitting procedure. Model 1 is characterized by the parameter λ, while Model 2 is characterized by the parameter q, the probability parameter of the geometric distribution. Model 4 features the exponent γ_1, while Model 5 is defined by both the exponent γ_2 and the maximum degree cutoff k_{max} (we will delve into this later in the report).

	Model				
	1	2	4	5	
Language	λ	q	γ_1	γ_2	k_{max}
Arabic	3.4769 ± 0.0288	0.2787 ± 0.0035	2.0663 ± 0.0169	2.0600 ± 0.0171	564
Chinese	3.1246 ± 0.0258	0.3060 ± 0.0035	2.1690 ± 0.0175	2.1666 ± 0.0176	881
Czech	2.8292 ± 0.0254	0.3326 ± 0.0038	2.2173 ± 0.0187	2.2149 ± 0.0188	688
English	3.7068 ± 0.0298	0.2631 ± 0.0034	2.0962 ± 0.0176	2.0926 ± 0.0178	857
Finnish	2.8652 ± 0.0292	0.3291 ± 0.0044	2.1771 ± 0.0206	2.1735 ± 0.0208	574
French	4.0208 ± 0.0309	0.2442 ± 0.0032	2.1025 ± 0.0177	2.1003 ± 0.0178	1394
Galician	4.0177 ± 0.0315	0.2444 ± 0.0032	2.0892 ± 0.0178	2.0868 ± 0.0180	1363
German	3.2688 ± 0.0266	0.2943 ± 0.0035	2.2940 ± 0.0199	2.2934 ± 0.0199	1527
Hindi	3.4576 ± 0.0294	0.2801 ± 0.0036	2.1025 ± 0.0179	2.0988 ± 0.0181	809
Icelandic	3.0617 ± 0.0272	0.3113 ± 0.0038	2.3070 ± 0.0211	2.3049 ± 0.0212	507
Indonesian	4.0842 ± 0.0344	0.2407 ± 0.0035	1.9737 ± 0.0171	1.9641 ± 0.0175	579
Italian	3.8535 ± 0.0298	0.2540 ± 0.0032	2.1111 ± 0.0176	2.1093 ± 0.0177	1600
Japanese	4.1963 ± 0.0306	0.2347 ± 0.0030	2.1097 ± 0.0174	2.1069 ± 0.0175	1003
Korean	1.3624 ± 0.0157	0.5461 ± 0.0042	2.6631 ± 0.0217	2.6628 ± 0.0217	520
Polish	2.9382 ± 0.0263	0.3223 ± 0.0038	2.1745 ± 0.0183	2.1717 ± 0.0185	726
Portuguese	3.2066 ± 0.0250	0.2992 ± 0.0033	2.2856 ± 0.0187	2.2848 ± 0.0187	1217
Russian	3.0120 ± 0.0263	0.3157 ± 0.0037	2.1652 ± 0.0180	2.1625 ± 0.0182	793
Spanish	3.9631 ± 0.0311	0.2475 ± 0.0033	2.0960 ± 0.0179	2.0935 ± 0.0180	1250
Swedish	3.1619 ± 0.0272	0.3029 ± 0.0037	2.2649 ± 0.0200	2.2625 ± 0.0202	539
Thai	4.0482 ± 0.0329	0.2427 ± 0.0034	1.9635 ± 0.0162	1.9490 ± 0.0167	380
Turkish	2.9049 ± 0.0273	0.3254 ± 0.0040	2.1547 ± 0.0188	2.1507 ± 0.0189	577

Table 3: Fitted parameters with standard deviations for different network models across languages.

2.3 AIC-Based Model Comparison

To systematically compare the performance of the five candidate models, we computed the Akaike Information Criterion (AIC) for each model using the optimal parameters estimated via maximum likelihood.

For each language, we calculated the AIC difference (ΔAIC) by subtracting the minimum AIC value (i.e., the AIC of the best-performing model) from the AIC of each competing model.

Table 4 presents the ΔAIC values for all models across the 21 languages examined.

	Model				
Language	1	2	3	4	5
Arabic	46828.66	5465.62	18.62	4.43	**0.00**
Chinese	47932.85	6648.14	102.63	0.15	**0.00**
Czech	42466.12	6003.65	153.87	**0.00**	0.03
English	56001.73	6358.79	31.24	1.22	**0.00**
Finnish	31086.79	4297.55	81.95	0.47	**0.00**
French	72229.71	7208.98	34.00	**0.00**	0.21
Galician	67973.37	6735.74	24.63	**0.00**	0.04
German	67576.24	8578.50	263.55	**0.00**	1.59
Hindi	46939.37	5654.74	34.37	1.21	**0.00**
Icelandic	49148.57	7297.52	256.78	**0.00**	0.54
Indonesian	38862.64	4437.54	7.77	7.45	**0.00**
Italian	70655.70	7140.69	41.11	**0.00**	0.51
Japanese	79855.29	8124.77	41.62	0.58	**0.00**
Korean	22290.48	5842.65	1376.53	**0.00**	1.78
Polish	40593.62	5594.14	100.34	0.39	**0.00**
Portuguese	74072.96	9235.83	280.47	**0.00**	1.34
Russian	41840.03	5811.41	92.36	0.35	**0.00**
Spanish	67214.95	6775.94	28.95	0.06	**0.00**
Swedish	52176.20	7351.06	206.23	**0.00**	0.17
Thai	37789.31	4603.79	17.68	14.74	**0.00**
Turkish	32701.67	4810.52	74.96	1.31	**0.00**

Table 4: The AIC difference (Δ) of each model relative to the best model for each language. Models are numbered according to Table 1. Bold values indicate the best AIC value for each language.

2.4 Root Mean Squared Error Analysis

In addition to the Akaike Information Criterion, we computed the Root Mean Square Error (RMSE) for each model to provide an alternative performance metric. Unlike AIC, which balances model fit with complexity through penalization, RMSE focuses purely on the magnitude of prediction errors, making it a complementary tool for evaluating model performance. Table 5 presents the RMSE values for all five models across the 21 languages.

	Model				
Language	1	2	3	4	5
Arabic	0.068	0.043	0.0300	0.031	**0.00390**
Chinese	0.070	0.046	0.0210	0.024	**0.00120**
Czech	0.068	0.044	0.0230	0.027	**0.00270**
English	0.071	0.046	0.0260	0.028	**0.00190**
Finnish	0.072	0.047	0.0260	0.029	**0.00260**
French	0.076	0.050	0.0270	0.029	**0.00220**
Galician	0.075	0.049	0.0290	0.030	**0.00330**
German	0.077	0.053	0.0170	0.022	**0.00074**
Hindi	0.073	0.047	0.0290	0.031	**0.00320**
Icelandic	0.075	0.052	0.0150	0.019	**0.00270**
Indonesian	0.076	0.045	0.0300	0.030	**0.00120**
Italian	0.073	0.047	0.0270	0.029	**0.00360**
Japanese	0.074	0.049	0.0240	0.026	**0.00130**
Korean	0.057	0.042	0.0063	0.017	**0.00230**
Polish	0.070	0.045	0.0250	0.028	**0.00240**
Portuguese	0.074	0.050	0.0190	0.023	**0.00140**
Russian	0.070	0.045	0.0240	0.027	**0.00150**
Spanish	0.075	0.049	0.0280	0.030	**0.00310**
Swedish	0.075	0.051	0.0180	0.022	**0.00120**
Thai	0.064	0.041	0.0290	0.028	**0.00091**
Turkish	0.074	0.049	0.0250	0.027	**0.00120**

Table 5: Root Mean Square Error (RMSE) for all five probability distribution models fitted to the degree distributions across languages. Bold values indicate the best-performing model for each language.

2.5 Exploring Altmann distribution

After considering the Altmann model for the in-degree distribution, we estimated its two parameters: γ and $\delta \geq 0$. Table 6 presents the estimated values along with their standard errors for all 21 languages analyzed, Table 7 shows the difference of the AIC values between the best three models, while Table 8 analyzes the different RMSE values among these models.

Language	γ	δ
Arabic	2.0462 ± 0.0193	0.00163 ± 0.00080
Chinese	2.1686 ± 0.0176	0.00000 ± 0.00008
Czech	2.2170 ± 0.0189	0.00000 ± 0.00011
English	2.0955 ± 0.0178	0.00000 ± 0.00011
Finnish	2.1718 ± 0.0213	0.00034 ± 0.00039
French	2.1018 ± 0.0180	0.00000 ± 0.00011
Galician	2.0884 ± 0.0181	0.00000 ± 0.00012
German	2.2938 ± 0.0200	0.00000 ± 0.00014
Hindi	2.0982 ± 0.0183	0.00026 ± 0.00019
Icelandic	2.3069 ± 0.0213	0.00000 ± 0.00021
Indonesian	1.9329 ± 0.0215	0.00286 ± 0.00112
Italian	2.1104 ± 0.0178	0.00000 ± 0.00010
Japanese	2.1091 ± 0.0176	0.00000 ± 0.00010
Korean	2.6631 ± 0.0217	0.00000 ± 0.00007
Polish	2.1742 ± 0.0185	0.00000 ± 0.00011
Portuguese	2.2855 ± 0.0187	0.00000 ± 0.00008
Russian	2.1648 ± 0.0182	0.00000 ± 0.00010
Spanish	2.0952 ± 0.0181	0.00000 ± 0.00012
Swedish	2.2647 ± 0.0202	0.00000 ± 0.00016
Thai	1.9140 ± 0.0211	0.00362 ± 0.00123
Turkish	2.1434 ± 0.0203	0.00110 ± 0.00071

Table 6: Fitted parameters via maximum likelihood estimation with standard deviations for Altmann distribution

Language	ΔAIC Zeta	ΔAIC Zeta Trunc	ΔAIC Altmann
Arabic	4.43	**0.00**	1.64
Chinese	0.15	**0.00**	1.88
Czech	**0.00**	0.03	1.83
English	1.22	**0.00**	2.70
Finnish	0.47	**0.00**	1.82
French	**0.00**	0.21	1.51
Galician	**0.00**	0.04	1.44
German	**0.00**	1.59	1.92
Hindi	1.21	**0.00**	2.05
Icelandic	**0.00**	0.54	1.92
Indonesian	9.96	2.51	**0.00**
Italian	**0.00**	0.51	1.54
Japanese	0.58	**0.00**	2.13
Korean	**0.00**	1.78	2.00
Polish	0.39	**0.00**	2.13
Portuguese	**0.00**	1.34	1.91
Russian	0.35	**0.00**	2.07
Spanish	0.06	**0.00**	1.54
Swedish	**0.00**	0.17	1.89
Thai	15.25	0.51	**0.00**
Turkish	1.31	**0.00**	1.55

Table 7: AIC comparison for Zeta, Truncated Zeta, and Altmann distributions. Values represent ΔAIC relative to the best model for each language. Lower values indicate better fit.

Language	RMSE Zeta	RMSE Zeta Trunc	RMSE Altmann
Arabic	0.03100	0.00390	**0.00120**
Chinese	0.02400	0.00120	**0.00033**
Czech	0.02700	0.00270	**0.00084**
English	0.02800	0.00190	**0.00057**
Finnish	0.02900	0.00260	**0.00081**
French	0.02900	0.00220	**0.00049**
Galician	0.03000	0.00330	**0.00073**
German	0.02200	0.00074	**0.00016**
Hindi	0.03100	0.00320	**0.00088**
Icelandic	0.01900	0.00270	**0.00097**
Indonesian	0.03000	0.00120	**0.00034**
Italian	0.02900	0.00360	**0.00075**
Japanese	0.02600	0.00130	**0.00037**
Korean	0.01700	0.00230	**0.00071**
Polish	0.02800	0.00240	**0.00072**
Portuguese	0.02300	0.00140	**0.00035**
Russian	0.02700	0.00150	**0.00046**
Spanish	0.03000	0.00310	**0.00072**
Swedish	0.02200	0.00120	**0.00041**
Thai	0.02800	0.00091	**0.00026**
Turkish	0.02700	0.00120	**0.00004**

Table 8: RMSE comparison for Zeta, Truncated Zeta, and Altmann distributions

2.6 Comprehensive Model Comparison

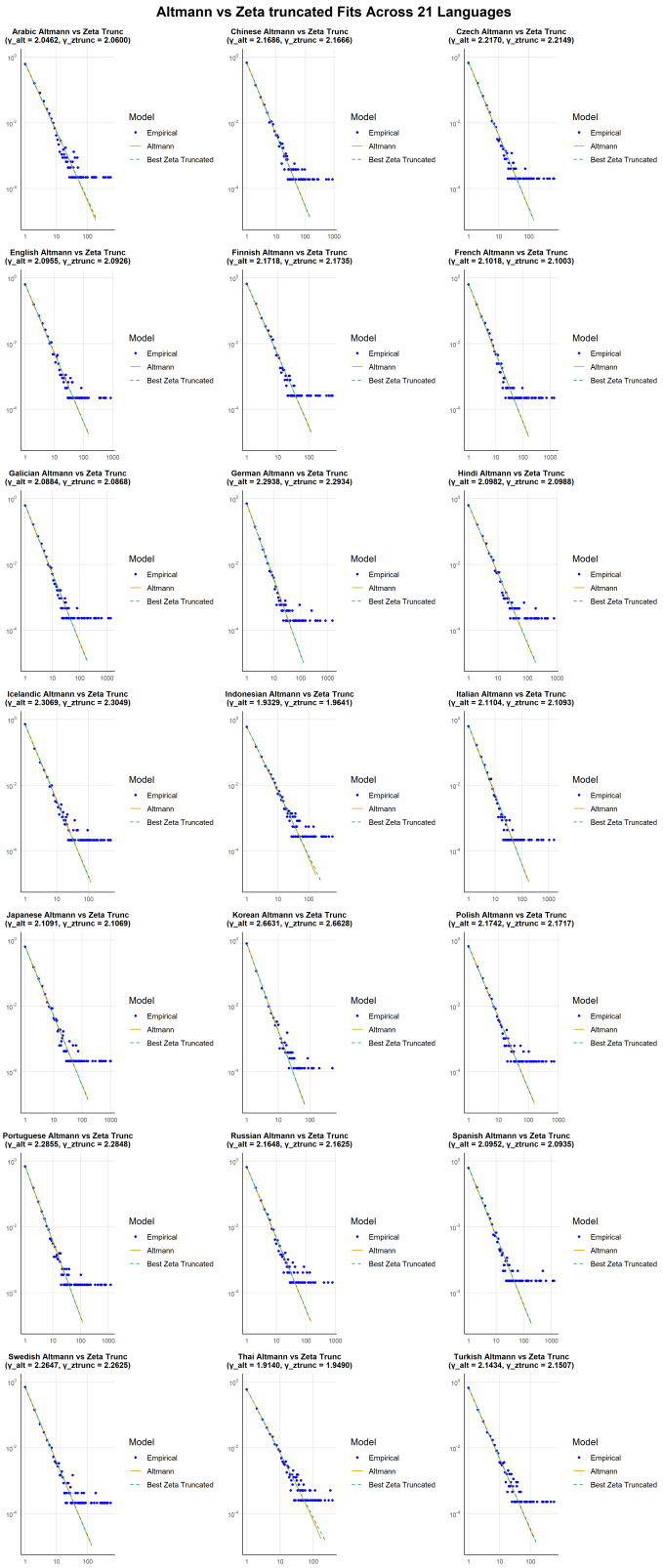


Figure 1: Fitted degree distributions for all languages in log-log scale. The right-truncated Zeta distribution (dashed line) and Altmann distribution (solid line) exhibit nearly identical behavior across all languages, with minor deviations observed only for Indonesian and Thai, where the Altmann model shows a small exponential correction in the tail region.

3 Discussion

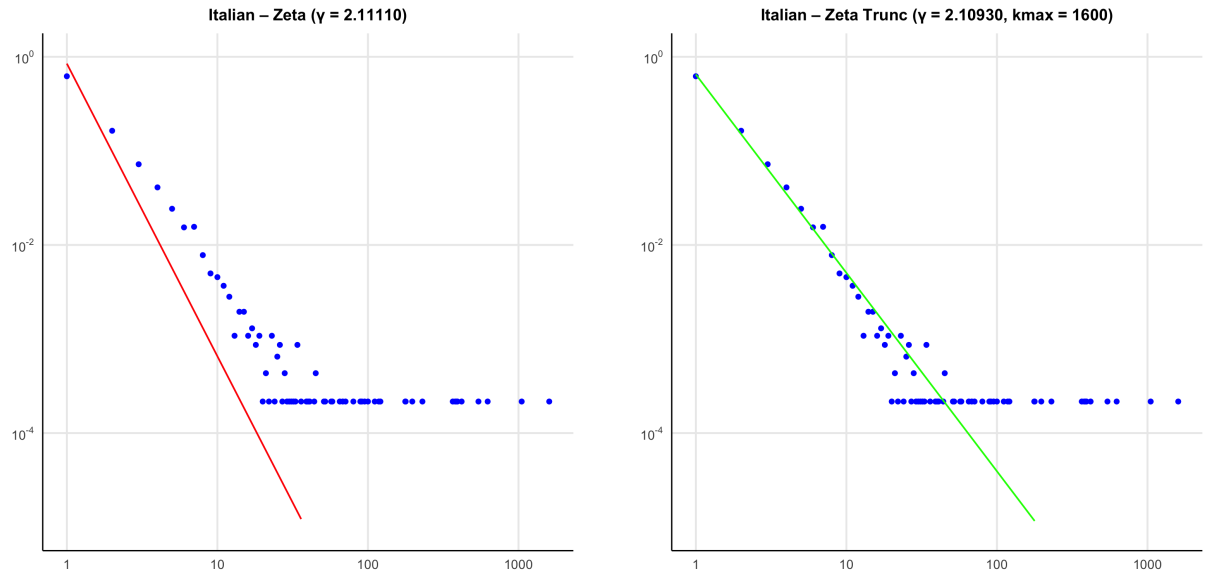
3.1 Best Model Selection among the first 5 models considered

As outlined in the methodology section (4), after analyzing the AIC and RMSE values of the five candidate models, a clear pattern emerges. While the Zeta and Truncated Zeta distributions show competitive AIC values (with one or the other achieving lower values depending on the language, as presented in Table 4) these differences are minimal, typically differing by only a few decimal digits.

However, a more comprehensive evaluation incorporating both visual inspection of the fitted distributions and quantitative RMSE analysis reveals a decisive advantage for the Truncated Zeta model. The visual comparison demonstrates that the Truncated Zeta distribution captures the in-degree sequence distribution with notably higher fidelity than the standard Zeta distribution. This observation is corroborated by the RMSE values presented in Table 5, which consistently show lower errors for the Truncated Zeta model across all of the analyzed languages.

As an illustrative example, Figure 2 shows the difference between the Zeta and Truncated Zeta distributions for the Italian language. Despite the Zeta distribution achieving a marginally lower AIC value for Italian (see Table 4), the visual comparison reveals a markedly inferior fit to the empirical data.

This pattern of superior performance by the Truncated Zeta distribution is consistently observed across all analyzed languages, as evidenced by the comprehensive visual comparison provided in Appendix A (Figure A.1 and A.2).



(a) Zeta distribution

(b) Truncated Zeta distribution

Figure 2: Comparison of Zeta and Truncated Zeta distributions for Italian in-degree sequence. It's clear that the Zeta Truncated, although having a slightly higher AIC, exibits better performances

This AIC preference for the simpler Zeta model likely stems from the criterion's inherent penalty for model complexity: the Zeta distribution has only one parameter, while the Truncated Zeta has two, and AIC favors more parsimonious models.

Therefore, considering the combination of competitive information criteria, superior predictive accuracy, and excellent visual agreement with the observed degree sequences, we select the Truncated Zeta distribution as the optimal model for characterizing the in-degree distributions in our dataset.

3.2 Comparison with the Altmann Distribution

After establishing the right-truncated Zeta distribution as the best-performing model among the initial five candidates, we extended our analysis to include the Altmann distribution. The maximum likelihood estimation reveals that for the vast majority of languages, the fitted exponential decay parameter δ is estimated to be approximately zero ($\delta \approx 0$). This result indicates that the exponential component of the Altmann distribution is effectively suppressed, causing the model to reduce to a pure power-law form.

Importantly, the joint optimization of both parameters (γ and δ) allows the Altmann distribution to implicitly adapt to the finite support of the empirical data. Specifically, by selecting a small δ, the model automatically implements a soft cutoff near $k_{\max}$, thereby assigning negligible probability mass to degrees beyond the observed maximum. Consequently, the fitted Altmann distribution exhibits behavior nearly identical to that of the right-truncated Zeta distribution across most languages.

Notable exceptions occur in Indonesian and Thai, where the estimated δ parameter is significantly different from zero, suggesting a genuine exponential correction to the power-law tail. For these two languages, both AIC values and RMSE computations favor the Altmann distribution as the optimal model. However, the magnitude of improvement is modest: the differences in AIC are small (typically $\Delta\text{AIC} < 2$), and visual inspection of the fitted distributions (Figure 1) confirms that the exponential correction primarily affects the extreme tail, with minimal impact on the bulk of the distribution.

In summary, while the Altmann distribution offers additional flexibility through its exponential decay parameter, this flexibility is rarely exploited by the data, resulting in convergence to a truncated power-law for the majority of languages analyzed.

4 Methods

4.1 Node Representation: Lexical Forms

In constructing the dependency network, the edges connect nodes from head to dependent following the syntactic structure of each sentence. A critical design choice concerned the lexical representation of nodes: whether to use surface forms (words as they appear in text) or lemmatized base forms.

We adopted lemmatized base forms for most languages, as this approach offers several analytical advantages. First, lemmatization reduces data sparsity by consolidating morphological variants (e.g., "running," "runs," "ran") into a single node ("run"), enabling more robust statistical analysis of network properties. Second, it captures semantic relationships more accurately, as different inflections of the same lexeme represent the same underlying concept and should contribute to a unified network position. Third, this normalization enhances the comparability of structural metrics across languages with varying morphological complexity, preventing highly inflected languages from artificially inflating network size and fragmenting connectivity patterns.

However, we made exceptions for Korean and Portuguese, where we retained surface forms. In these corpora, a substantial proportion of tokens lacked lemmatization data in the dependency annotations, likely due to preprocessing pipeline limitations or morphological complexity. Using incomplete lemmatization would introduce systematic bias, creating artificially isolated nodes for non-lemmatized tokens while merging others. Maintaining surface forms for these languages ensures internal consistency and prevents measurement artifacts in network topology metrics.

4.2 Preprocessing of In-Degree Sequences

Before analyzing the in-degree sequences, we excluded all nodes with in-degree zero. This step is necessary because the family of Zeta distributions considered in this study cannot produce degree zero (by definition, their support starts from one). Removing zero-degree nodes ensures the validity of the model fitting process and allows for meaningful parameter estimation on the observed degree distribution.

4.3 Observation on a closed form solution of the maximum likelihood estimation for the displaced geometric distribution

During the estimation of the displaced geometric distribution, the parameter q was initially computed via Maximum Likelihood Estimation (MLE). However, it can be noted that q actually admits a closed-form solution derived directly from the characteristics of the network. This means that the MLE procedure is not strictly necessary, since the analytical expression of q can be obtained as:

$$\hat{q} = \frac{N}{M}$$

where N is the size of the network and M is the sum of the in-degrees of all nodes. This ratio corresponds to the inverse of the *mean degree* of the network. Here, we show the proof:

$$\mathcal{L}(q) = (M - N)\,\log(1 - q) + N\,\log q$$

Taking the derivative of the log-likelihood with respect to q:

$$\frac{d\mathcal{L}}{dq} = -\frac{M-N}{1-q} + \frac{N}{q} = 0,$$

we obtain the maximum-likelihood estimate

$$\hat{q} = \frac{M}{N}.$$

This observation simplifies the estimation process and confirms the consistency between the computational and analytical results.

4.4 Observation on the monotonicity of the likelihood function of the right Truncated Zeta function with respect to the $k_{\max}$ parameter

In the context of the right-truncated Zeta distribution, the parameter $k_{\max}$ represents the upper bound of the domain. A crucial point in the fitting procedure is that $k_{\max}$ is not a free parameter to be estimated via numerical optimization in the same manner as the exponent γ. Instead, its Maximum Likelihood Estimate (MLE) is directly determined by the observed data.

Let our dataset be a sample of N positive integers $\{k_1, k_2, \ldots, k_N\}$, and let $k_{\text{obs max}} = \max_i k_i$ be the maximum value observed in the sample. The estimation of $k_{\max}$ follows from two main arguments: a validity condition and the behavior of the likelihood function.

1. Validity Condition The right-truncated Zeta distribution is defined over the support $k \in \{1, 2, \ldots, k_{\max}\}$. This implies that the probability mass function $P(k|\gamma, k_{\max})$ is zero for any $k > k_{\max}$. For the observed sample to be possible under the model, every observation k_i must have a non-zero probability. This imposes the necessary condition that every k_i must be supported by the distribution. Consequently, the truncation parameter $k_{\max}$ must be at least as large as the maximum observed value:

$$k_{\max} \geq k_{\text{obs max}}$$

If this condition is not met (i.e., if $k_{\max} < k_{\text{obs max}}$), the probability of observing at least one data point $k_i > k_{\max}$ would be zero. This would lead to a likelihood of zero for the entire sample, and a log-likelihood of $\ell = -\infty$. Such a parameter value is therefore inadmissible.

2. Likelihood Maximization The log-likelihood function for the sample is given by:

$$\ell(\gamma, k_{\max}) = \sum_{i=1}^{N} \log P(k_i|\gamma, k_{\max}) = \sum_{i=1}^{N} [-\gamma \log k_i - \log H(k_{\max}, \gamma)]$$

which simplifies to:

$$\ell(\gamma, k_{\max}) = -\gamma M' - N \log H(k_{\max}, \gamma)$$

To find the MLE for $k_{\max}$, we analyze the behavior of ℓ as a function of $k_{\max}$ for any fixed γ. The first term, $-\gamma \sum \log k_i$, does not depend on $k_{\max}$. Therefore, maximizing the log-likelihood with respect to $k_{\max}$ is equivalent to maximizing the term $-N \log H(k_{\max}, \gamma)$.

This, in turn, is equivalent to minimizing $\log H(k_{\max}, \gamma)$, and since the logarithm is a monotonically increasing function, this is equivalent to minimizing the normalization factor $H(k_{\max}, \gamma)$ itself. The normalization factor is the generalized harmonic number:

$$H(k_{\max}, \gamma) = \sum_{j=1}^{k_{\max}} j^{-\gamma}$$

Given that $j^{-\gamma} > 0$ for any $j \geq 1$, the sum $H(k_{\max}, \gamma)$ is a strictly monotonically increasing function of its upper limit, $k_{\max}$.

Conclusion To maximize the likelihood, we must choose the smallest possible value for $k_{\max}$. From the validity condition, we know that the smallest admissible value is $k_{\text{obs max}}$. Combining these two points, the value of $k_{\max}$ that satisfies the validity constraint and simultaneously maximizes the likelihood function is:

$$\hat{k}_{\max}^{\text{MLE}} = k_{\text{obs max}}$$

Therefore, $k_{\max}$ is not treated as a parameter to be fitted but is set to the maximum of the observed data. The estimation problem then simplifies to finding the optimal value of γ given $k_{\max} = k_{\text{obs max}}$.

4.5 Estimation of the uncertainties

Uncertainties on the language statistics were computed using standard statistical methods. For the mean degree $\langle k \rangle = M/N$, where M is the total number of edges and N is the number of nodes, the standard error of the mean was calculated as:

$$\sigma_{\langle k \rangle} = \frac{\sigma_k}{\sqrt{N}}$$

where σ_k is the sample standard deviation of the degree distribution.

For the inverse quantity $N/M = 1/\langle k \rangle$, error propagation was applied. Given the functional dependence $f(x) = 1/x$, the propagated uncertainty is:

$$\sigma_{N/M} = \left| \frac{\partial}{\partial(M/N)} \left(\frac{N}{M} \right) \right| \sigma_{M/N} = \frac{1}{(M/N)^2} \cdot \sigma_{M/N}$$

Equivalently, using the standard error propagation formula:

$$\sigma_{N/M} = \left(\frac{N}{M} \right)^2 \cdot \sigma_{M/N} = \frac{1}{\langle k \rangle^2} \cdot \frac{\sigma_k}{\sqrt{N}}$$

All reported uncertainties represent one standard deviation confidence intervals.

For all the models we have fitted (the displaced Geometric, the displaced Poisson, the Zeta, the Truncated Zeta, and the two-parameter Altmann distribution) the uncertainties of the estimated parameters were computed using the `mle()` function in `R`. The `mle()` function estimates the standard deviations of the parameters based on the curvature of the log-likelihood function at its maximum. Formally, for a parameter vector θ, the asymptotic covariance matrix is given by the inverse of the negative Hessian of the log-likelihood function evaluated at the maximum likelihood estimates $\hat{\theta}$:

$$\mathrm{Cov}(\hat{\theta}) \approx - \left. \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^\top} \right|^{-1}_{\theta = \hat{\theta}}.$$

For single-parameter models, such as the displaced geometric, Poisson, Zeta, or Truncated Zeta (with fixed $k_{\max}$), this reduces to the reciprocal of the negative second derivative of the log-likelihood with respect to the parameter. For multi-parameter models, like the Altmann distribution, the covariance matrix captures the correlations between parameters, and the standard deviations of each parameter correspond to the square roots of the diagonal elements of the covariance matrix. This approach reflects the intrinsic variability of the maximum likelihood estimates due to the finite sample size and the shape of the likelihood function around its maximum.

4.6 Model Selection Strategy

Model selection in statistical analysis benefits from a multi-criteria approach rather than reliance on a single metric. While individual measures such as the Akaike Information Criterion or Root Mean Square Error each provide valuable insights into model performance, they capture different aspects of model quality and may emphasize different properties of the fit.

The AIC balances goodness of fit against model complexity by penalizing additional parameters: this helps prevent overfitting and favors parsimonious models that capture essential patterns without unnecessary elaboration. On the contrary, RMSE provides a direct, intuitive measure of predictive accuracy by quantifying the average magnitude of residuals between model predictions and observed data. Unlike AIC, RMSE does not explicitly account for model complexity, instead focusing purely on deviation from empirical observations.

Given that these metrics embody distinct statistical philosophies and can potentially favor different models, we adopted a triangulated approach to model evaluation. For each candidate distribution, we computed both AIC and RMSE values across all languages in our corpus. Additionally, we conducted visual inspection of fitted distributions plotted against empirical degree frequencies. This multi-faceted strategy ensures that our final model selection is robust and logical, not an artifact of any single evaluation criterion. When multiple independent lines of evidence converge on the same model, confidence in that selection is substantially strengthened.

4.7 Extended Model Comparison: Altmann Family

Following the initial evaluation of the five canonical probability distributions, we extended our analysis to include an additional family of functions proposed by Altmann.

The Altmann distribution is defined as:

$$p(k) = c\, k^{-\gamma} e^{-\delta k}$$

where c is a normalization constant ensuring that the distribution sums to unity, γ controls the power-law exponent governing the initial decay rate, and δ modulates the exponential cutoff at higher degree values. The normalization constant is given by:

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

We compared the Altmann distribution against the best-performing models from the initial analysis using the same evaluation framework: AIC for information-theoretic model comparison, RMSE for quantifying prediction accuracy, and visual inspection of fitted distributions. This consistent methodology enables direct comparison of the Altmann family with previously evaluated models, allowing us to assess whether the additional functional flexibility justifies the increased model complexity.

A Figures

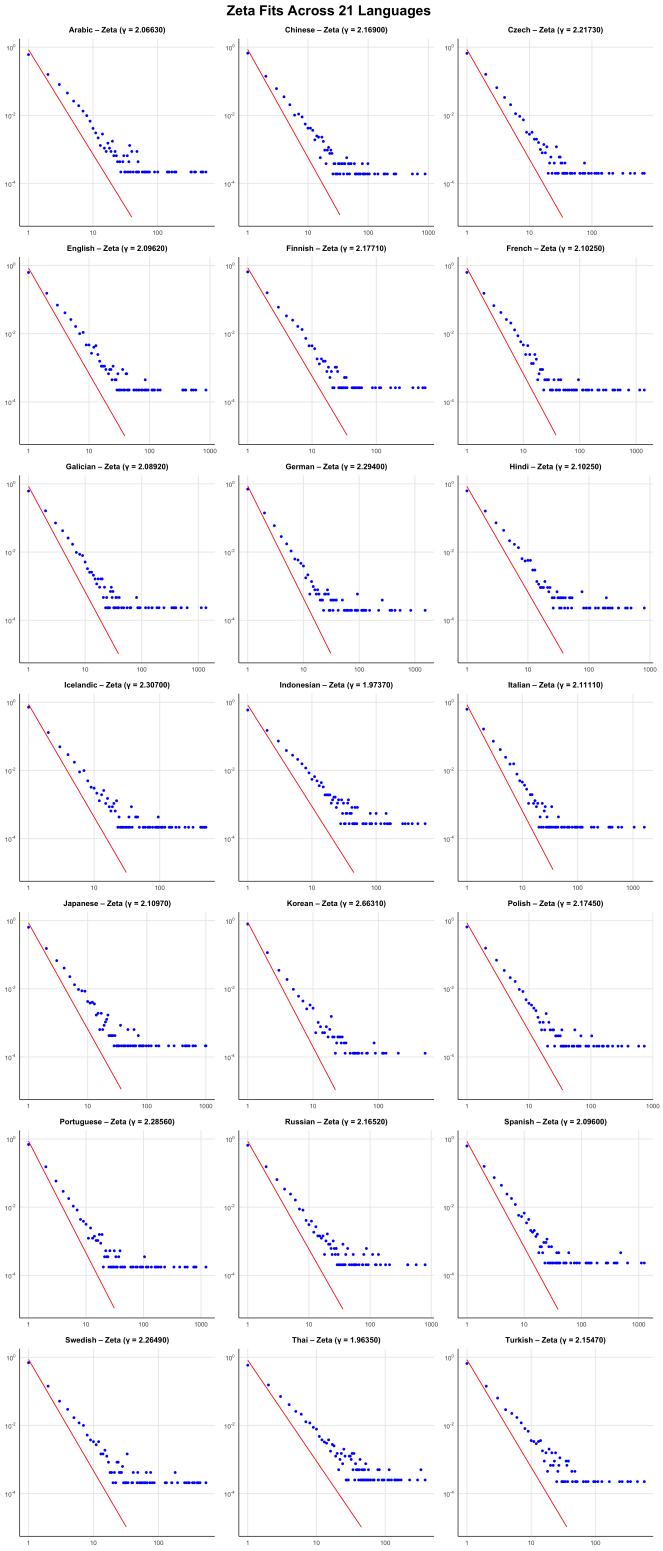


Figure A.1: Fitted Zeta distribution compared to empirical degree distributions across all languages (log-log scale). The model exhibits systematic deviations from the data due to its infinite support assumption, which fails to capture the finite cutoff at $k_{\max}$ observed in real networks.

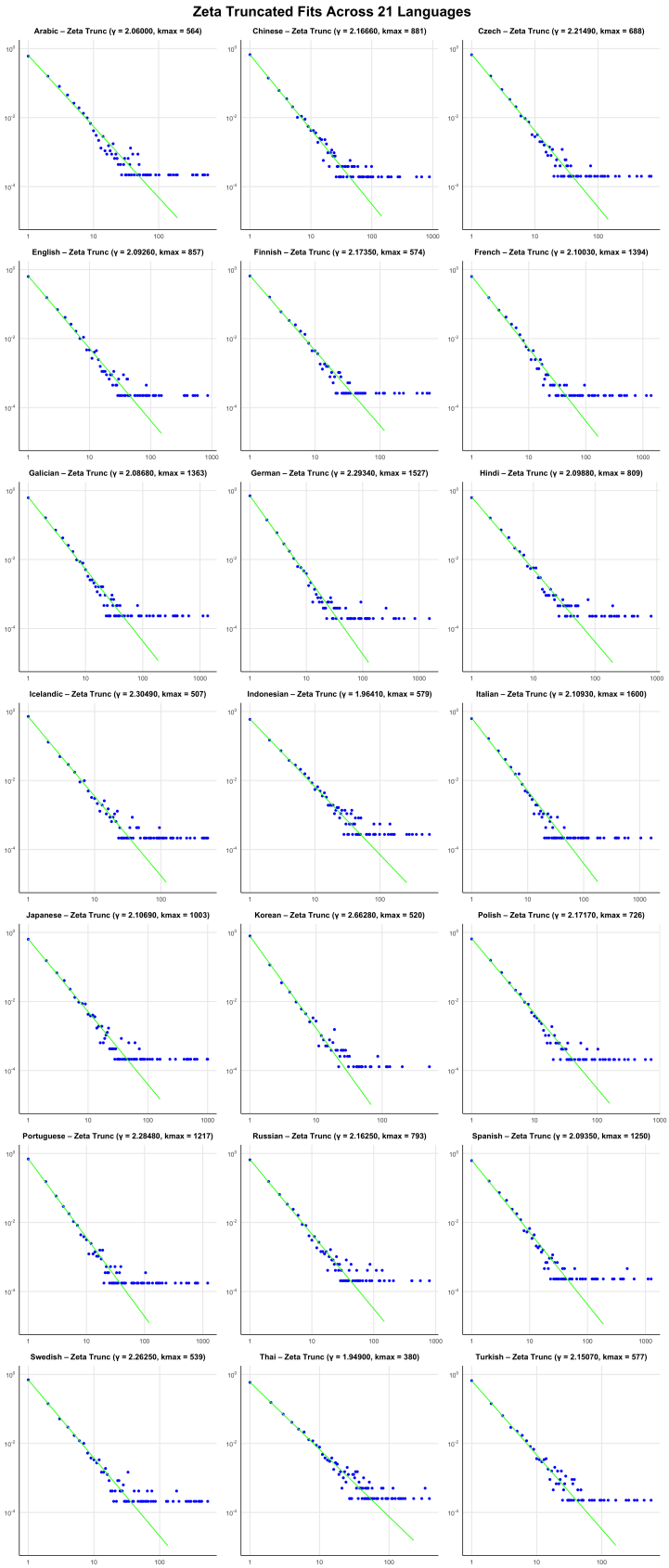


Figure A.2: Fitted right-truncated Zeta distribution compared to empirical degree distributions across all languages (log-log scale). The truncated model accurately captures the power-law behavior and natural cutoff at the maximum observed degree $k_{\max}$, resulting in superior fit quality compared to the infinite-support Zeta distribution.