Master in Data Science

Complex and Social Networks Laboratory

Non-linear regression on dependency trees

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

Ilaria Boschetto
ilaria.boschetto@estudiantat.upc.edu

Project Report

Academic Year 2025/2026

November 19, 2025

Contents

1 Introduction

In this laboratory session, we analyzed the statistical properties of syntactic dependency lengths across the Parallel Universal Dependencies (PUD) treebanks. The aim was to examine how dependency distances distribute within and across languages and to evaluate whether these distributions exhibit scaling behavior consistent with previously observed linguistic laws.

To achieve this, we implemented a series of functions in R designed to compute and analyze several distance-based metrics. Firstly, we extracted all head–dependent pairs from the annotated treebanks and calculated their linear distances. These raw distances are then aggregated and processed through logarithmic binning, which smooths frequency fluctuations and allows a clearer observation of potential non-linear patterns.

Subsequently, we fitted multiple statistical models to the empirical distributions. Model comparison was performed using the Akaike Information Criterion (AIC), allowing us to identify the model that provided the best balance between goodness of fit and model complexity.

Overall, this analysis allowed us to evaluate whether the observed patterns of dependency distances reflect universal principles of syntactic organization or arise from random structural constraints within specific treebanks.

2 Results

2.1 Preliminary results

Table 1: Summary of the properties of the degree sequences. N is the sample size (the number of sentences or dependency trees), μ_n and σ_n are, respectively, the mean and the standard deviation of n, the sentence length (n is the number of vertices of a tree), μ_d and σ_d are the mean and the standard deviation of $\langle d \rangle$.

Language	**N**	μ_n	σ_n	μ_d	σ_d
Arabic	1000	20.75	8.61	3.01	0.58
Chinese	1000	21.41	8.62	3.29	0.88
Czech	1000	18.61	7.76	3.01	0.73
English	1000	21.18	8.22	3.16	0.66
Finnish	1000	15.81	6.53	2.83	0.68
French	1000	24.73	10.02	3.09	0.58
Galician	1000	23.51	9.80	3.07	0.61
German	1000	21.33	8.54	3.70	0.93
Hindi	1000	23.83	9.56	3.53	0.95
Icelandic	1000	18.83	7.67	2.86	0.54
Indonesian	1000	19.45	7.69	2.84	0.60
Italian	1000	23.73	9.85	3.09	0.63
Japanese	1000	28.79	11.00	2.87	0.58
Korean	1000	16.58	6.68	2.54	0.71
Polish	1000	18.38	7.43	2.87	0.62
Portuguese	1000	23.41	9.49	3.07	0.59
Russian	1000	19.36	8.05	2.91	0.67
Spanish	1000	23.28	9.45	3.07	0.59
Swedish	1000	19.08	7.65	3.03	0.61
Thai	1000	22.32	8.99	2.37	0.51
Turkish	1000	16.88	6.63	2.70	0.76

Table 1 summarizes several statistical properties of sentence structures across the languages in our sample, focusing on the distribution of sentence length n and mean dependency length $\langle d \rangle$. Languages such as *French*, *Italian*, *Portuguese*, and *Spanish* display relatively long average sentence lengths ($\mu_n \approx$ 23–24 words), whereas languages like *Finnish*, *Korean*, and *Turkish* tend to have much shorter sentences on average ($\mu_n \approx$ 15–17 words). The remaining languages occupy intermediate positions, indicating substantial cross-linguistic variation in typical sentence length even within this restricted set of treebanks.

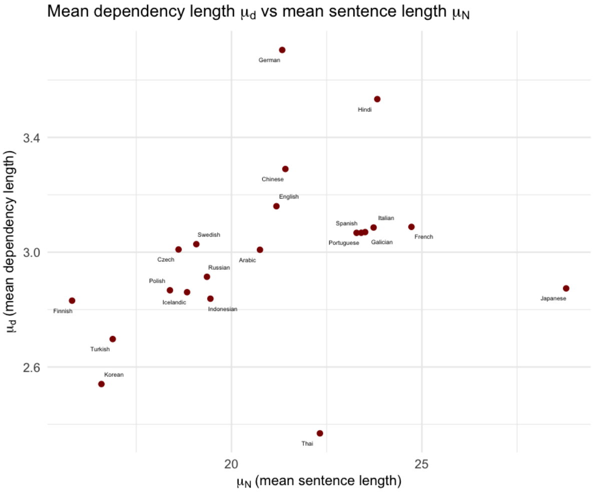


Figure 1: Mean dependency length μ_d as a function of mean sentence length μ_N for all languages in our sample.

Figure 1 complements this picture by plotting mean dependency length μ_d as a function of mean sentence length μ_N for each language. Overall, the points suggest a mild positive association: languages with longer sentences often have slightly larger mean dependency length, but there is considerable dispersion around any simple trend, and several languages depart clearly from the main cluster. Most languages lie in a relatively narrow band with μ_d close to 3, despite differences of almost a factor of two in μ_N, which indicates that mean dependency length varies less across languages than mean sentence length itself.

Some languages occupy more extreme regions of the plot. *German* and *Hindi*, for instance, combine medium-to-long sentences with comparatively high mean dependency length, while *Thai* and *Korean* show lower μ_d values for their sentence lengths. *Japanese* is notable in that it has the largest μ_N in the sample but only a moderate μ_d, showing that longer sentences do not necessarily correspond to proportionally higher dependency length. Taken together, the table and the scatter plot highlight both broad similarities, since most languages cluster around $\mu_d \approx 3$, and meaningful differences in how sentence length and dependency length combine across languages, without pointing to a single universal quantitative relationship between these two quantities.

Table 2 reports the residual standard error for each model. Almost all parametric models (that is, all models except the null model) have a small residual standard error, which means that they fit the data reasonably well. For many languages several different models reach very similar, very low error values, so their quality of fit is almost indistinguishable. In these cases, choosing a single "best" model is difficult, especially when the corresponding gains in AIC are only marginal.

Table 3 reports the AIC values for all candidate models across languages. In most cases the preferred specifications are relatively simple: the logarithmic models AIC$_4$ (one parameter) and AIC$_{4+}$ (two parameters) are very often among the best-performing options. For many languages these models obtain the lowest AIC, but the advantage over other reasonable candidates is usually small. In Section 3 we return to this point and examine in detail a specific issue of the extended models, which suggests that AIC rankings should be interpreted with some caution when comparing "+" and non-"+" models.

Table 4 summarizes the relative performance of the candidate models using ΔAIC values, obtained by subtracting, for each language, the smallest AIC across all specifications. By definition, the best model for a language has $\Delta\text{AIC} = 0$, and these entries are highlighted in bold in the table. The pattern of ΔAIC values confirms that no single model is best for all languages: in some cases a simple or intermediate specification is already competitive, while in others one of the more flexible models attains the minimum AIC, with the remaining options showing noticeably higher ΔAIC.

Table 5 reports the optimal parameter estimates for models 1–5, together with their standard errors, across all languages. Overall, the estimates of the exponent b in Models 1 and 2 are fairly stable and typically lie between about 0.3 and 0.7, suggesting a broadly similar scaling pattern of mean dependency length with sentence length across languages. The parameter a varies more strongly, especially in Models 2 and 3, indicating language-specific differences in the overall level of dependency length.

Tables 6 and 7 report the parameter estimates for the extended models including an intercept d. Overall, the exponents b in Models 1+ remain in a similar range to the corresponding models without d, indicating that adding an intercept does not substantially change the inferred scaling behaviour. The new parameter d is typically positive and of moderate size for many languages, but it can be negative or very large in magnitude for some cases (for example in Models 2+ and 3+), suggesting considerable language-specific shifts in baseline dependency length that the simpler models could not capture; we will return to the interpretation of these extreme values in more detail in the Section 3. Taken together, these results show that adding d mainly affects the intercept level rather than the shape of the functional relationship, and that for some languages the gain in fit comes at the cost of higher uncertainty in individual parameters.

Table 2: Residual standard error s for different models by language

Language	s_0	s_1	s_2	s_3	s_4	s_5	s_{1+}	s_{2+}	s_{3+}	s_{4+}	s_{5+}
Arabic	5.182	0.284	0.104	0.215	0.110	0.101	0.111	0.091	0.189	0.086	0.096
Chinese	5.120	0.192	0.146	0.318	0.169	0.090	0.153	0.086	0.263	0.079	0.080
Czech	4.843	0.216	0.120	0.303	0.092	0.096	0.128	0.087	0.244	0.083	0.091
English	4.921	0.245	0.125	0.209	0.121	0.132	0.128	0.128	0.183	0.125	0.133
Finnish	3.235	0.163	0.122	0.280	0.083	0.089	0.128	0.080	0.227	0.075	0.083
French	7.470	0.233	0.080	0.184	0.092	0.083	0.081	0.083	0.148	0.096	0.084
Galician	6.412	0.247	0.120	0.259	0.100	0.120	0.127	0.111	0.226	0.102	0.111
German	5.006	0.308	0.160	0.422	0.191	0.129	0.177	0.116	0.329	0.116	0.121
Hindi	5.601	0.136	0.135	0.371	0.313	0.090	0.138	0.079	0.274	0.081	0.083
Icelandic	4.978	0.279	0.104	0.226	0.120	0.079	0.113	0.082	0.199	0.077	0.080
Indonesian	5.231	0.182	0.108	0.241	0.087	0.099	0.113	0.093	0.200	0.089	0.098
Italian	6.177	0.275	0.109	0.242	0.112	0.111	0.115	0.103	0.207	0.101	0.105
Japanese	7.807	0.134	0.077	0.176	0.078	0.080	0.078	0.078	0.147	0.081	0.079
Korean	4.273	0.081	0.083	0.207	0.150	0.087	0.083	0.085	0.151	0.108	0.085
Polish	4.306	0.172	0.053	0.147	0.077	0.054	0.053	0.054	0.109	0.081	0.054
Portuguese	6.146	0.223	0.125	0.267	0.111	0.128	0.130	0.117	0.226	0.115	0.118
Russian	5.143	0.162	0.086	0.250	0.076	0.064	0.092	0.057	0.197	0.055	0.060
Spanish	6.421	0.254	0.079	0.201	0.080	0.079	0.084	0.076	0.167	0.076	0.079
Swedish	4.214	0.205	0.065	0.190	0.056	0.063	0.071	0.056	0.151	0.059	0.058
Thai	6.718	0.125	0.057	0.132	0.095	0.052	0.059	0.051	0.117	0.047	0.053
Turkish	3.991	0.096	0.081	0.294	0.199	0.055	0.077	0.056	0.198	0.090	0.057

Table 3: AIC values for different models by language

Language	**AIC$_0$**	**AIC$_1$**	**AIC$_2$**	**AIC$_3$**	**AIC$_4$**	**AIC$_5$**	**AIC$_{1+}$**	**AIC$_{2+}$**	**AIC$_{3+}$**	**AIC$_{4+}$**	**AIC$_{5+}$**
Arabic	81.67	7.09	−18.13	0.74	−17.59	−18.11	−16.50	−20.80	−1.82	**-22.99**	−18.73
Chinese	75.25	−2.58	−8.38	10.34	−5.61	−19.19	−7.22	−20.15	6.54	**-23.12**	−21.58
Czech	85.90	−0.24	−15.89	10.12	−24.16	−21.29	−13.91	−24.09	4.86	**-26.03**	−21.94
English	74.30	3.21	−12.03	0.33	**-13.60**	−10.05	−11.52	−10.67	−2.11	−11.97	−9.27
Finnish	69.41	−7.26	−14.02	7.63	−24.77	−21.42	−12.81	−24.24	2.89	**-26.56**	−22.48
French	91.18	1.93	**-25.05**	−3.24	−22.07	−23.13	−24.66	−23.07	−8.15	−20.08	−22.23
Galician	87.20	3.53	−14.50	5.56	**-20.12**	−13.70	−12.97	−16.88	2.78	−18.74	−15.26
German	92.88	10.19	−8.50	20.51	−4.09	−14.24	−5.52	−17.52	13.89	**-18.27**	−15.55
Hindi	83.69	−12.05	−11.34	14.94	9.62	−20.99	−10.83	−24.59	7.84	**-24.62**	−22.56
Icelandic	80.62	6.66	−18.05	2.07	−15.25	−24.66	−16.07	−23.50	−0.43	**-25.89**	−23.51
Indonesian	81.91	−4.40	−17.13	3.72	**-23.63**	−18.56	−16.00	−20.26	−0.33	−22.23	−18.16
Italian	92.71	6.50	−18.54	3.90	−18.53	−17.18	−16.96	−19.26	0.31	**-20.59**	−18.04
Japanese	92.32	−12.40	−25.98	−4.39	**-26.39**	−24.07	−25.46	−24.91	−8.41	−24.69	−23.91
Korean	76.65	**-25.55**	−23.89	−0.28	−9.54	−21.92	−23.95	−22.49	−7.75	−17.04	−21.96
Polish	71.09	−5.20	−32.73	−8.17	−24.39	−31.36	**-32.75**	−30.81	−14.61	−22.42	−30.70
Portuguese	92.57	0.70	−14.67	6.63	**-18.77**	−13.17	−13.49	−15.72	2.68	−16.99	−14.77
Russian	81.47	−7.51	−23.21	4.63	−26.99	−29.98	−21.31	−33.16	−0.70	**-34.62**	−30.99
Spanish	87.24	4.26	−25.19	−0.96	−25.89	−24.56	−23.53	−25.60	−5.02	**-26.42**	−23.85
Swedish	70.57	−1.03	−27.80	−1.95	**-32.21**	−27.89	−25.70	−30.43	−6.76	−30.28	−29.15
Thai	81.77	−12.99	−30.81	−10.72	−19.58	−32.18	−29.94	−33.22	−12.87	**-35.50**	−31.32
Turkish	80.48	−22.91	−26.89	9.32	−2.58	**-36.68**	−28.30	−36.49	−0.99	−23.75	−35.10

Table 4: ΔAIC values for different models by language

Language	**ΔAIC$_{0}$**	**ΔAIC$_{1}$**	**ΔAIC$_{2}$**	**ΔAIC$_{3}$**	**ΔAIC$_{4}$**	**ΔAIC$_{5}$**	**ΔAIC$_{1+}$**	**ΔAIC$_{2+}$**	**ΔAIC$_{3+}$**	**ΔAIC$_{4+}$**	**ΔAIC$_{5+}$**
Arabic	104.661	30.080	4.857	23.727	5.405	4.879	6.495	2.189	21.172	**0.000**	4.264
Chinese	98.374	20.547	14.746	33.462	17.512	3.937	15.905	2.977	29.664	**0.000**	1.547
Czech	111.930	25.787	10.143	36.152	1.866	4.745	12.121	1.942	30.886	**0.000**	4.086
English	87.898	16.816	1.571	13.935	**0.000**	3.549	2.077	2.934	11.495	1.636	4.329
Finnish	95.977	19.306	12.543	34.190	1.793	5.147	13.757	2.327	29.450	**0.000**	4.084
French	116.222	26.980	**0.000**	21.811	2.978	1.921	0.382	1.972	16.897	4.963	2.812
Galician	107.324	23.652	5.620	25.685	**0.000**	6.424	7.149	3.245	22.902	1.376	4.862
German	111.155	28.458	9.773	38.780	14.183	4.034	12.748	0.750	32.165	**0.000**	2.720
Hindi	108.307	12.572	13.276	39.559	34.241	3.629	13.784	0.028	32.455	**0.000**	2.063
Icelandic	106.508	32.543	7.840	27.951	10.638	1.228	9.813	2.381	25.453	**0.000**	2.376
Indonesian	105.541	19.229	6.502	27.351	**0.000**	5.072	7.626	3.370	23.302	1.400	5.466
Italian	113.298	27.090	2.043	24.484	2.059	3.405	3.628	1.324	20.898	**0.000**	2.543
Japanese	118.712	13.994	0.409	21.999	**0.000**	2.315	0.925	1.474	17.979	1.696	2.476
Korean	102.203	**0.000**	1.656	25.273	16.009	3.627	1.596	3.063	17.797	8.513	3.587
Polish	103.842	27.548	0.016	24.574	8.354	1.384	**0.000**	1.936	18.141	10.326	2.049
Portuguese	111.345	19.477	4.102	25.399	**0.000**	5.602	5.286	3.048	21.449	1.782	4.000
Russian	116.087	27.111	11.405	39.247	7.631	4.638	13.310	1.463	33.922	**0.000**	3.632
Spanish	113.664	30.681	1.236	25.460	0.536	1.869	2.898	0.826	21.400	**0.000**	2.578
Swedish	102.786	31.185	4.408	30.262	**0.000**	4.319	6.513	1.780	25.449	1.929	3.066
Thai	117.274	22.519	4.697	24.787	15.927	3.328	5.568	2.284	22.630	**0.000**	4.186
Turkish	117.159	13.770	9.787	45.998	34.095	**0.000**	8.375	0.183	35.685	12.924	1.575

Table 5: Model parameters (models 1-5) by language

Language	**Model 1**	**Model 2**		**Model 3**		**Model 4**	**Model 5**		
	b	a	b	a	c	a	a	b	c
Arabic	0.463 ± 0.011	1.165 ± 0.060	0.316 ± 0.017	2.11 ± 0.09	0.01485 ± 0.00162	1.023 ± 0.011	1.01 ± 0.13	0.395 ± 0.065	−0.0040 ± 0.0031
Chinese	0.495 ± 0.007	0.900 ± 0.070	0.423 ± 0.025	2.09 ± 0.14	0.0183 ± 0.0023	1.094 ± 0.017	0.54 ± 0.07	0.685 ± 0.066	−0.0122 ± 0.0030
Czech	0.483 ± 0.008	0.970 ± 0.052	0.389 ± 0.017	1.99 ± 0.12	0.0185 ± 0.0021	1.065 ± 0.009	0.76 ± 0.08	0.528 ± 0.053	−0.0072 ± 0.0026
English	0.487 ± 0.009	1.054 ± 0.070	0.366 ± 0.021	2.12 ± 0.09	0.0172 ± 0.0017	1.074 ± 0.012	1.03 ± 0.19	0.378 ± 0.096	−0.0006 ± 0.0046
Finnish	0.496 ± 0.008	0.871 ± 0.055	0.426 ± 0.022	1.72 ± 0.11	0.0269 ± 0.0032	1.049 ± 0.009	0.64 ± 0.07	0.632 ± 0.066	−0.0140 ± 0.0043
French	0.443 ± 0.008	1.095 ± 0.044	0.329 ± 0.012	2.23 ± 0.08	0.01173 ± 0.00097	1.001 ± 0.008	1.12 ± 0.12	0.318 ± 0.049	0.0004 ± 0.0018
Galician	0.450 ± 0.009	1.08 ± 0.06	0.336 ± 0.018	2.12 ± 0.11	0.0137 ± 0.0016	1.007 ± 0.009	0.96 ± 0.13	0.396 ± 0.066	−0.0027 ± 0.0028
German	0.536 ± 0.009	1.005 ± 0.059	0.425 ± 0.018	2.27 ± 0.15	0.0185 ± 0.0021	1.241 ± 0.017	0.77 ± 0.08	0.567 ± 0.053	−0.0068 ± 0.0024
Hindi	0.5073 ± 0.0042	0.754 ± 0.049	0.487 ± 0.020	2.06 ± 0.14	0.0193 ± 0.0021	1.137 ± 0.029	0.51 ± 0.06	0.683 ± 0.054	−0.0084 ± 0.0022
Icelandic	0.453 ± 0.012	1.171 ± 0.062	0.305 ± 0.017	2.06 ± 0.10	0.01474 ± 0.00186	0.997 ± 0.012	0.89 ± 0.09	0.459 ± 0.052	−0.0081 ± 0.0027
Indonesian	0.453 ± 0.007	0.965 ± 0.055	0.367 ± 0.018	1.94 ± 0.10	0.0169 ± 0.0019	0.988 ± 0.008	0.79 ± 0.10	0.478 ± 0.067	−0.0055 ± 0.0032
Italian	0.452 ± 0.010	1.125 ± 0.055	0.324 ± 0.015	2.12 ± 0.10	0.01367 ± 0.00142	1.014 ± 0.010	1.05 ± 0.12	0.362 ± 0.056	−0.0017 ± 0.0024
Japanese	0.3983 ± 0.0050	0.952 ± 0.042	0.333 ± 0.013	1.96 ± 0.07	0.01183 ± 0.00105	0.881 ± 0.007	0.92 ± 0.11	0.347 ± 0.054	−0.0005 ± 0.0020
Korean	0.4476 ± 0.0038	0.715 ± 0.034	0.456 ± 0.016	1.58 ± 0.08	0.0248 ± 0.0021	0.946 ± 0.015	0.70 ± 0.09	0.466 ± 0.067	−0.0006 ± 0.0037
Polish	0.474 ± 0.008	0.989 ± 0.029	0.371 ± 0.010	1.92 ± 0.07	0.01978 ± 0.00144	1.021 ± 0.008	1.04 ± 0.08	0.342 ± 0.043	0.0016 ± 0.0023
Portuguese	0.455 ± 0.008	1.007 ± 0.058	0.359 ± 0.018	2.04 ± 0.10	0.01504 ± 0.00156	1.014 ± 0.010	0.94 ± 0.13	0.398 ± 0.065	−0.0018 ± 0.0028
Russian	0.465 ± 0.006	0.933 ± 0.041	0.387 ± 0.014	1.95 ± 0.10	0.0178 ± 0.0019	1.017 ± 0.007	0.74 ± 0.06	0.513 ± 0.042	−0.0062 ± 0.0020
Spanish	0.450 ± 0.009	1.134 ± 0.045	0.321 ± 0.012	2.19 ± 0.09	0.01278 ± 0.00124	1.007 ± 0.007	1.03 ± 0.10	0.370 ± 0.048	−0.0021 ± 0.0020
Swedish	0.489 ± 0.009	1.024 ± 0.036	0.372 ± 0.012	2.00 ± 0.09	0.0196 ± 0.0018	1.060 ± 0.006	0.92 ± 0.08	0.434 ± 0.049	−0.0034 ± 0.0026
Thai	0.356 ± 0.006	1.019 ± 0.041	0.275 ± 0.013	1.79 ± 0.06	0.01132 ± 0.00121	0.781 ± 0.009	0.87 ± 0.09	0.356 ± 0.049	−0.0036 ± 0.0021
Turkish	0.4693 ± 0.0043	0.651 ± 0.028	0.503 ± 0.014	1.54 ± 0.10	0.0277 ± 0.0028	0.988 ± 0.020	0.52 ± 0.03	0.638 ± 0.037	−0.0079 ± 0.0021

Table 6: Model parameters with intercept d (models 1+–3+) by language

Language	**Model 1+**		**Model 2+**			**Model 3+**		
	b	d	a	b	d	a	c	d
Arabic	0.390 ± 0.012	0.536 ± 0.067	27.580 ± 160.400	0.030 ± 0.163	-27.160 ± 160.900	429.800 ± 72520.000	0.000 ± 0.018	-427.900 ± 72520.000
Chinese	0.465 ± 0.013	0.262 ± 0.098	42.810 ± 206.700	0.029 ± 0.129	-43.400 ± 207.400	910.200 ± 258300.000	0.000 ± 0.020	-908.400 ± 258300.000
Czech	0.442 ± 0.010	0.330 ± 0.068	30.300 ± 101.400	0.034 ± 0.104	-30.350 ± 101.800	738.700 ± 149500.000	0.000 ± 0.017	-737.000 ± 149500.000
English	0.431 ± 0.013	0.449 ± 0.083	3.471 ± 6.613	0.187 ± 0.232	-2.898 ± 7.328	297.200 ± 27010.000	0.000 ± 0.017	-295.300 ± 27010.000
Finnish	0.466 ± 0.013	0.200 ± 0.069	36.560 ± 161.600	0.028 ± 0.117	-36.630 ± 162.000	796.100 ± 203100.000	0.000 ± 0.025	-794.700 ± 203100.000
French	0.390 ± 0.007	0.474 ± 0.052	1.250 ± 1.093	0.309 ± 0.137	-0.208 ± 1.447	339.300 ± 27940.000	0.000 ± 0.010	-337.300 ± 27940.000
Galician	0.397 ± 0.011	0.441 ± 0.077	18.170 ± 58.080	0.047 ± 0.132	-17.910 ± 58.540	476.000 ± 78530.000	0.000 ± 0.015	-474.000 ± 78530.000
German	0.493 ± 0.011	0.447 ± 0.084	11.980 ± 14.730	0.093 ± 0.091	-12.140 ± 15.130	975.200 ± 179600.000	0.000 ± 0.014	-973.300 ± 179600.000
Hindi	0.501 ± 0.009	0.065 ± 0.078	11.410 ± 11.590	0.103 ± 0.081	-12.200 ± 12.020	723.600 ± 95850.000	0.000 ± 0.013	-722.000 ± 95850.000
Icelandic	0.378 ± 0.013	0.524 ± 0.069	41.760 ± 357.100	0.019 ± 0.157	-41.280 ± 357.500	817.900 ± 318100.000	0.000 ± 0.021	-816.000 ± 318100.000
Indonesian	0.413 ± 0.011	0.298 ± 0.067	23.290 ± 98.700	0.039 ± 0.149	-23.230 ± 99.230	512.100 ± 87520.000	0.000 ± 0.017	-510.300 ± 87520.000
Italian	0.392 ± 0.010	0.492 ± 0.064	6.685 ± 11.010	0.104 ± 0.132	-6.118 ± 11.400	498.500 ± 77880.000	0.000 ± 0.014	-496.500 ± 77880.000
Japanese	0.367 ± 0.007	0.247 ± 0.051	2.349 ± 2.676	0.204 ± 0.142	-1.722 ± 3.071	316.000 ± 30310.000	0.000 ± 0.011	-314.200 ± 30310.000
Korean	0.452 ± 0.008	-0.028 ± 0.047	1.145 ± 0.796	0.366 ± 0.127	-0.607 ± 1.058	286.100 ± 16500.000	0.000 ± 0.014	-284.800 ± 16500.000
Polish	0.424 ± 0.006	0.344 ± 0.034	0.850 ± 0.498	0.399 ± 0.110	0.193 ± 0.705	265.800 ± 15030.000	0.000 ± 0.013	-264.100 ± 15030.000
Portuguese	0.412 ± 0.010	0.359 ± 0.071	6.584 ± 10.460	0.114 ± 0.136	-6.242 ± 10.900	503.700 ± 70660.000	0.000 ± 0.014	-501.900 ± 70660.000
Russian	0.430 ± 0.008	0.272 ± 0.054	16.350 ± 27.810	0.057 ± 0.084	-16.380 ± 28.120	549.200 ± 84540.000	0.000 ± 0.016	-547.500 ± 84540.000
Spanish	0.388 ± 0.008	0.509 ± 0.053	5.006 ± 6.934	0.130 ± 0.129	-4.391 ± 7.343	421.700 ± 55120.000	0.000 ± 0.013	-419.700 ± 55120.000
Swedish	0.434 ± 0.008	0.400 ± 0.045	4.926 ± 5.000	0.146 ± 0.106	-4.466 ± 5.310	386.200 ± 40900.000	0.000 ± 0.017	-384.500 ± 40900.000
Thai	0.312 ± 0.008	0.269 ± 0.044	16.480 ± 80.480	0.035 ± 0.156	-15.960 ± 80.880	333.500 ± 66070.000	0.000 ± 0.017	-331.800 ± 66070.000
Turkish	0.485 ± 0.006	-0.111 ± 0.038	2.229 ± 0.859	0.274 ± 0.061	-2.081 ± 1.013	805.800 ± 116100.000	0.000 ± 0.014	-804.600 ± 116100.000

Table 7: Model parameters with intercept d (models 4+–5+) by language

Language	**Model 4+**		**Model 5+**		
	a	d	a	b	c
Arabic	0.913 ± 0.039	0.320 ± 0.110	19.020 ± 454.400	0.043 ± 0.963	-0.000 ± 0.003
Chinese	1.344 ± 0.040	-0.744 ± 0.116	11.080 ± 65.430	0.113 ± 0.544	-0.001 ± 0.007
Czech	1.128 ± 0.033	-0.181 ± 0.093	12.230 ± 100.400	0.078 ± 0.560	-0.000 ± 0.003
English	1.109 ± 0.063	-0.101 ± 0.183	14.880 ± 836.600	0.051 ± 2.668	0.001 ± 0.021
Finnish	1.112 ± 0.034	-0.164 ± 0.085	10.780 ± 79.630	0.093 ± 0.599	-0.001 ± 0.008
French	1.006 ± 0.043	-0.015 ± 0.135	10.650 ± 407.100	0.058 ± 2.008	0.001 ± 0.018
Galician	0.977 ± 0.042	0.094 ± 0.127	23.640 ± 564.300	0.034 ± 0.776	0.000 ± 0.002
German	1.445 ± 0.042	-0.595 ± 0.118	13.150 ± 87.560	0.085 ± 0.495	0.000 ± 0.002
Hindi	1.596 ± 0.036	-1.397 ± 0.108	9.058 ± 43.810	0.125 ± 0.485	-0.000 ± 0.003
Icelandic	0.853 ± 0.035	0.413 ± 0.097	8.347 ± 65.260	0.099 ± 0.654	-0.001 ± 0.009
Indonesian	1.016 ± 0.040	-0.082 ± 0.114	11.260 ± 153.300	0.075 ± 0.894	-0.000 ± 0.004
Italian	0.937 ± 0.039	0.232 ± 0.116	20.910 ± 550.000	0.035 ± 0.878	0.000 ± 0.004
Japanese	0.899 ± 0.036	-0.058 ± 0.113	11.210 ± 310.400	0.055 ± 1.379	0.000 ± 0.007
Korean	1.108 ± 0.049	-0.450 ± 0.131	8.946 ± 321.000	0.071 ± 2.351	0.002 ± 0.034
Polish	1.015 ± 0.041	0.017 ± 0.113	6.879 ± 234.000	0.080 ± 2.475	0.002 ± 0.034
Portuguese	1.033 ± 0.045	-0.057 ± 0.131	23.400 ± 696.900	0.034 ± 0.961	0.000 ± 0.006
Russian	1.101 ± 0.025	-0.245 ± 0.070	8.751 ± 52.240	0.098 ± 0.497	-0.000 ± 0.003
Spanish	0.956 ± 0.034	0.157 ± 0.102	9.432 ± 132.600	0.075 ± 0.920	0.000 ± 0.001
Swedish	1.067 ± 0.030	-0.020 ± 0.082	10.930 ± 145.900	0.072 ± 0.858	0.000 ± 0.001
Thai	0.644 ± 0.024	0.421 ± 0.072	7.547 ± 103.600	0.075 ± 0.896	-0.000 ± 0.006
Turkish	1.237 ± 0.036	-0.676 ± 0.095	0.916 ± 1.038	0.479 ± 0.298	-0.004 ± 0.007

2.2 Preliminary plots

Figure 2 shows the raw data together with the null-model prediction, and illustrates how the empirical mean dependency lengths stay systematically below the random baseline. This pattern is consistent with the dependency length minimization principle, according to which human languages tend to organize word order in such a way that syntactic dependencies are, on average, as short as possible.

2.3 Best Models through AIC selection

In Figure 3 we show the fit of the best-performing model to the observed data for each language. These models were selected using AIC, which we employed to rank the competing specifications in terms of their balance between goodness of fit and model complexity. Although this specification achieves the lowest AIC among the candidates and therefore represents our preferred model from a theoretic perspective, we recall the caveats discussed in Section 3, where we highlighted the limitations of this choice and the potential impact of model misspecification on the interpretation of the results.

3 Discussion

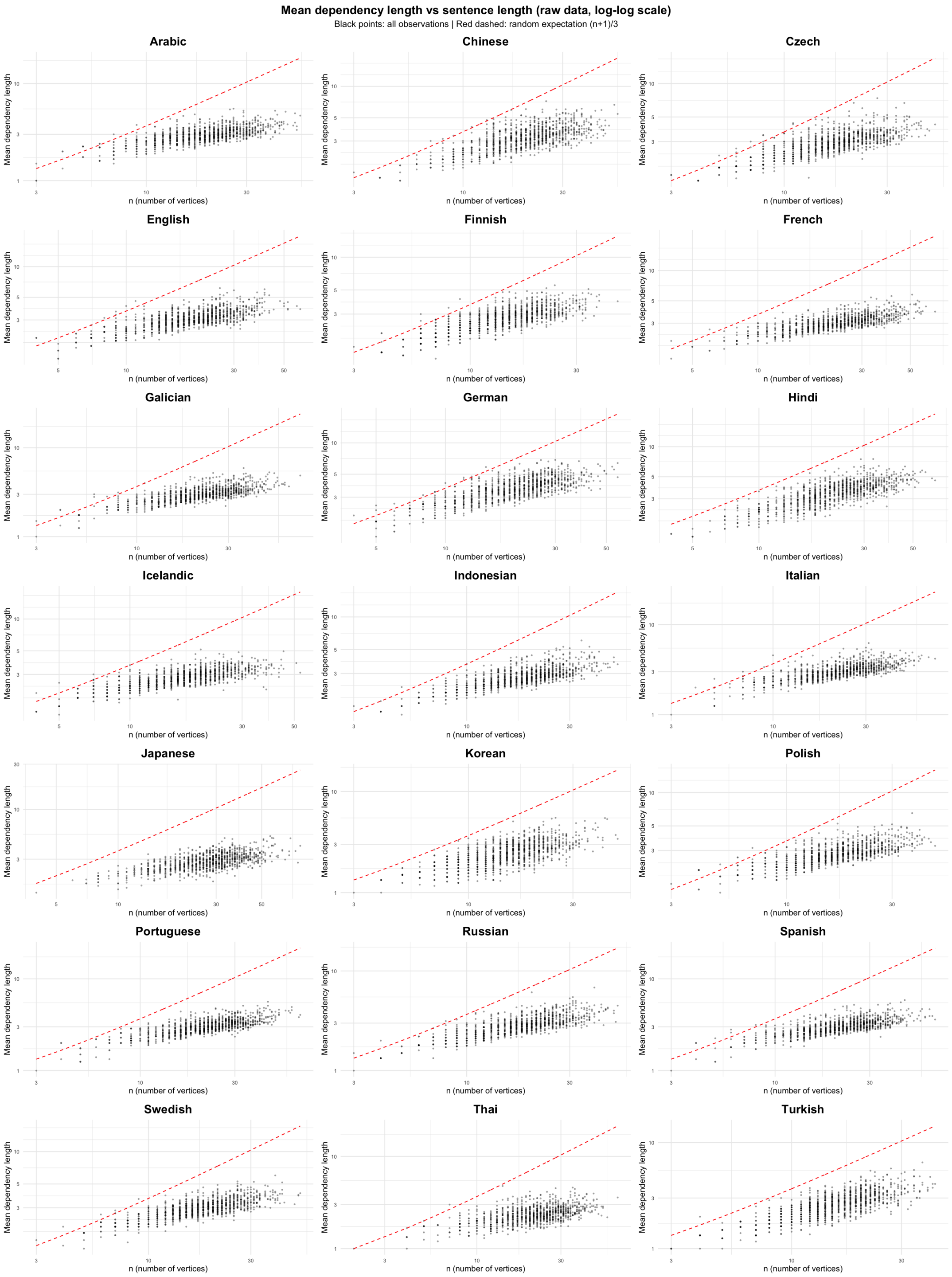
3.1 Model selection and parameter stability

When inspecting the nonlinear fits, an unexpected pattern emerged in the estimated parameters for the + models, except for 1+ and 4+ , as you can easily see in Table 6 and 7: for almost all languages, the standard errors associated with the scale parameter a and with the intercept d in the extended models were extremely large, often even larger than the parameter estimates themselves. This was extremely unexpected, because the fitted curves look visually reasonable and the residual sums of squares are generally small across a wide range of specifications. In other words, the models manage to track the empirical trends quite well, but some of their parameters are statistically very poorly determined. This discrepancy between "good-looking fits" and "strange standard errors" is one of the central observations in our results, and we try to motivate it by taking a closer look at the geometry of the fitting problem.

To interpret these unstable standard errors, we used the geometric perspective on nonlinear regression developed by Bates and Watts [1]. In this framework, the least-squares criterion defines a surface over the parameter space, and the covariance matrix of the estimators is (locally) proportional to the inverse of the curvature of this surface. When two parameters, such as a and d, can compensate each other in producing almost the same fitted curve, the surface becomes very flat along certain directions: a wide "valley" of nearly equivalent solutions appears. In this situation the information matrix is close to singular and the corresponding standard errors inflate, even though the fitted values themselves remain stable. This is exactly what we may be observing in our case: a and d are highly correlated and their individual estimates vary a lot, while the predicted mean dependency lengths barely change. Our best reading of the results is therefore that some of the extended models with d suffer from a form of local non-identifiability in (a, d), rather than from a fundamental mismatch between model family and data.

These issues in parameter stability have a direct impact on model selection. Across languages, the residual sums of squares are relatively low and very similar for several competing models within each language, so that information criteria such as AIC are often driven more by the number of parameters than by large differences in fit quality. Model 4+ (the logarithmic specification with offset) is frequently among the best-scoring models, and in contrast to other extended specifications it does not show the extreme standard errors for a and d that we observe elsewhere. Its AIC advantage over the corresponding model without d is usually small, but at least it is not clearly undermined by severe instability in the parameters. By contrast, some of the richer "plus" models with more parameters combine very modest AIC gains with highly uncertain estimates, so any preference for them based on AIC alone should be interpreted with caution.

Another aspect of the results is their cross-linguistic heterogeneity. The descriptive statistics and model-based analyses reveal a mixed picture regarding how much languages resemble or differ from each other. On the one hand, several properties are remarkably stable across the sample. Most languages exhibit mean distances μ_d close to 3 with relatively small standard deviations, and the exponents b in Models 1–2 (and their plus-variants) typically lie between about 0.3 and 0.7, indicating a broadly similar



Mean dependency length vs sentence length (raw data, log-log scale)

Black points: all observations | Red dashed: random expectation (n+1)/3

Arabic, Chinese, Czech, English, Finnish, French, Galician, German, Hindi, Icelandic, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, Turkish

Figure 2: Mean dependency length as a function of sentence length for all languages in the sample, on a log–log scale. Each panel shows one language, with black points for individual sentences and a red dashed line for the random baseline.

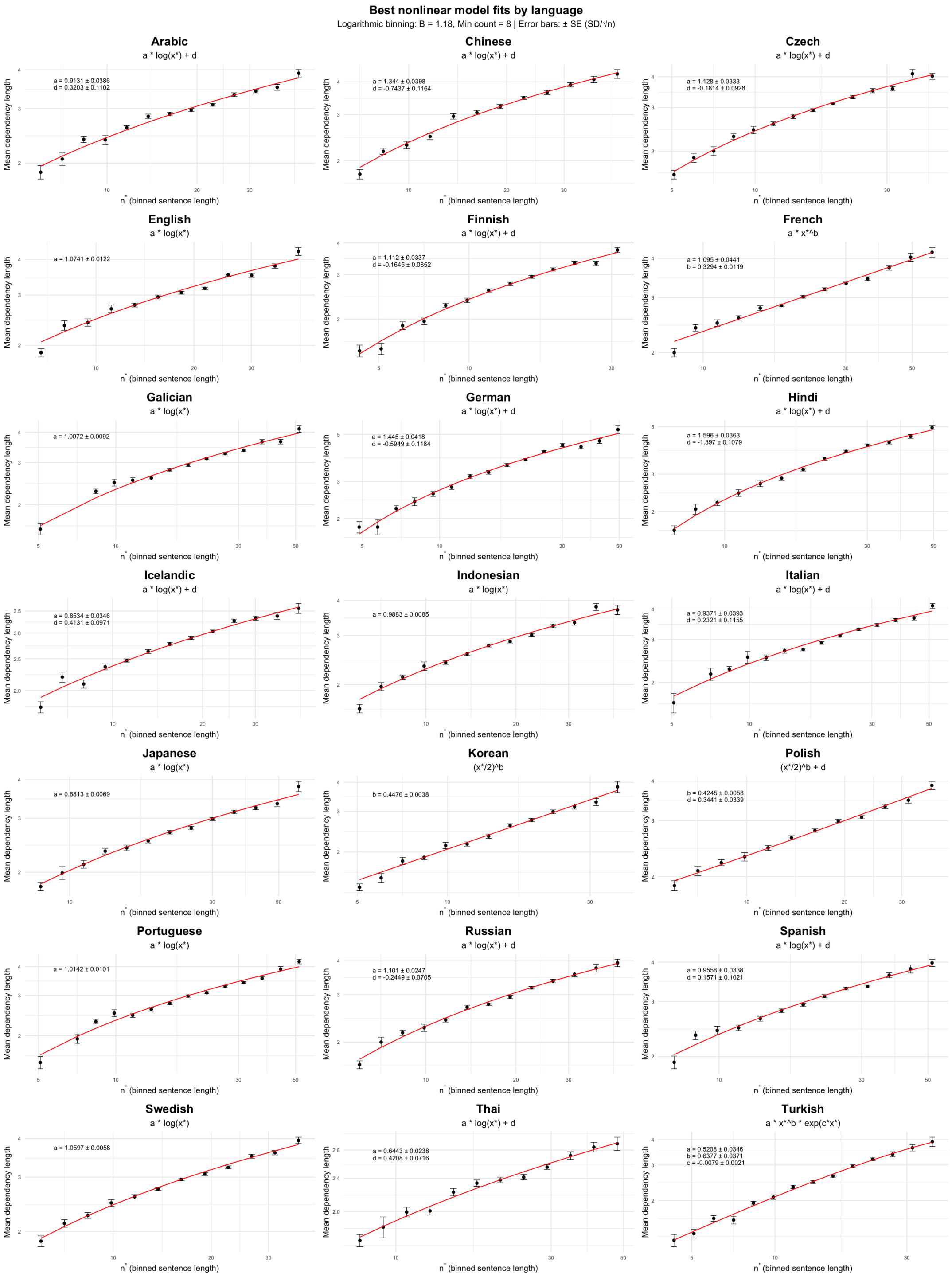


Figure 3: Best-fitting nonlinear models for mean dependency length as a function of sentence length in each language (logarithmic binning). Each panel shows the empirical binned means with error bars and the corresponding AIC-selected model (red curve).

scaling regime in which mean dependency length grows sublinearly with sentence length.

On the other hand, there are clear and systematic cross-linguistic contrasts. Romance languages such as French, Italian, Portuguese, and Spanish cluster together in having relatively long sentences ($\mu_n \approx 23$–24 words) but mean dependency lengths that are close to the overall average. By contrast, languages like Finnish, Korean, Turkish, and Icelandic show much shorter sentences ($\mu_n \approx 15$–19), and some languages in this group have lower mean degrees (e.g. Thai and Korean), while others such as German and Hindi have noticeably higher values. The fitted scale parameter a in Models 2–3 also varies substantially across languages, indicating differences in the overall level of dependency length even when the scaling exponent is similar. Taken together, these observations suggest that the dependency-length patterns encoded in the different treebanks are not captured by a single, language-independent law within the set of models considered here.

At the same time, many different models achieve very similar residual sums of squares, which means that under our current preprocessing and unweighted fitting strategy the data do not seem to provide enough information to clearly separate closely related functional forms. The strong heteroscedasticity documented earlier, which we deliberately ignore by treating all bins as equally weighted, is likely to contribute both to the apparent similarity in fit quality across models and to the poor identifiability of parameters such as a and d. From the point of view of model selection, no single functional form is always preferred across all languages: some are best described by simple power laws, others by logarithmic or exponential curves, and in many cases several competitors lie within a very small ΔAIC range. Overall, this pattern is compatible with the idea that a similar general tendency towards short and slowly growing dependency lengths can be realized through different quantitative profiles, with languages differing in typical sentence length, syntactic density, and in which specific non-linear function provides the most economical summary of the data.

In summary, our analyses present a coherent picture: across all the languages examined, mean dependency length grows non-linearly with sentence length, suggesting a shared pressure to keep dependencies relatively short. At the same time, the data do not seem to point to a single functional law or a universally preferred parametrization. Several models account for the observed patterns similarly well, and languages differ both in the specifics of their best-fitting parameters and in how clearly any one model stands out.

4 Methods

4.1 Data preprocessing

In this step, we analyse the syntactic annotation files provided in the Universal Dependencies format for each language. For every sentence, the program reads the corresponding dependency tree, that is, the information about which words depend on which other words in the sentence. From these trees, we compute three summary measures that are stored as plain text. These three columns are written to simple text files, one per language, so that we can easily compare the structural properties of dependency trees across the different corpora.

4.2 Data validation

Following the preprocessing phase, we performed a systematic validation to ensure the quality and theoretical consistency of the extracted metrics. The dependency metrics obtained for the PUD treebanks were validated against the theoretical bounds for the second moment of the degree distribution $\langle k^2 \rangle$ and the mean dependency length $\langle d \rangle$. Tables 8 and 9 report, for each language, the empirical estimates together with the corresponding lower and upper bounds derived from these results. As can be seen, all observed values of $\langle k^2 \rangle$ and $\langle d \rangle$ fall within their respective theoretical intervals, indicating that the empirical dependency structures are fully consistent with the graph-theoretic constraints and supporting the reliability of the metrics used in our analysis.

Table 8: Validation of the second moment k^2 for dependency tree degree distributions.

Language	k^2	**Lower Bound**	**Upper Bound**	**Valid**
Arabic	5.216	3.644	19.747	Yes

Language	k^2	**Lower Bound**	**Upper Bound**	**Valid**
Chinese	5.667	3.661	20.415	Yes
Czech	5.392	3.605	17.609	Yes
English	5.814	3.662	20.180	Yes
Finnish	5.361	3.540	14.813	Yes
French	5.756	3.706	23.726	Yes
Galician	5.582	3.684	22.510	Yes
German	5.788	3.660	20.332	Yes
Hindi	5.777	3.696	22.829	Yes
Icelandic	5.431	3.615	17.833	Yes
Indonesian	5.330	3.629	18.446	Yes
Italian	5.710	3.689	22.732	Yes
Japanese	5.997	3.753	27.788	Yes
Korean	4.729	3.569	15.584	Yes
Polish	5.233	3.606	17.384	Yes
Portuguese	5.599	3.685	22.407	Yes
Russian	5.274	3.620	18.355	Yes
Spanish	5.535	3.686	22.284	Yes
Swedish	5.584	3.622	18.076	Yes
Thai	5.213	3.670	21.322	Yes
Turkish	4.967	3.574	15.881	Yes

Table 9: Validation of the mean distance $\langle d \rangle$ for dependency trees.

Language	$\langle d \rangle$	**Lower Bound**	**Upper Bound**	**Valid**
Arabic	3.008	1.584	14.818	Yes
Chinese	3.290	1.700	15.319	Yes
Czech	3.010	1.645	13.216	Yes
English	3.160	1.750	15.143	Yes
Finnish	2.831	1.649	11.121	Yes
French	3.088	1.722	17.801	Yes
Galician	3.070	1.681	16.889	Yes
German	3.705	1.743	15.256	Yes
Hindi	3.533	1.720	17.129	Yes
Icelandic	2.861	1.653	13.383	Yes
Indonesian	2.838	1.622	13.843	Yes
Italian	3.086	1.716	17.056	Yes
Japanese	2.874	1.770	20.847	Yes
Korean	2.541	1.447	11.698	Yes
Polish	2.867	1.599	13.047	Yes
Portuguese	3.068	1.685	16.812	Yes
Russian	2.914	1.607	13.775	Yes
Spanish	3.067	1.666	16.720	Yes
Swedish	3.028	1.698	13.565	Yes
Thai	2.369	1.565	15.999	Yes
Turkish	2.697	1.527	11.921	Yes

4.3 Logarithmic Binning

To obtain smooth and comparable degree distributions across languages, we represent the empirical data using logarithmic binning, following the guidelines in [2]. Starting from a lower bound a_0 slightly smaller than the minimum observed value, we construct bins of the form $[a_0 B^{k-1}, a_0 B^k)$ with a fixed multiplicative factor $B = 1.18$. In this way bin widths grow geometrically in n while remaining approximately uniform on a $\log n$ scale, so that medium and small values are represented with narrow bins and large values with wider bins. This construction increases the signal-to-noise ratio in the tail without distorting the overall shape of the distribution in log–log plots.

In our case, however, the data are not only sparse in the tail but also at the lower end of the range: very short sentences (or structures with very few vertices) are relatively infrequent in several treebanks, and naive logarithmic binning would still produce highly unstable estimates in the first bins. For each language we therefore apply an additional merging step: after assigning all observations to logarithmic bins, we iteratively merge adjacent bins until each contains at least eight observations. This merging is applied symmetrically at both ends of the range, so that very small and very large values are pooled with their neighbours until the corresponding bins reach a minimally robust sample size. This does not significantly affect the shape of the distribution, but helps smoothen the curve.

Within each merged bin we compute the mean of the variable of interest and place the point at an effective bin centre (given by the geometric mean of the bin boundaries, which we will call from now on n^*), obtaining a smoothed curve that tracks the underlying trend while suppressing fluctuations driven by few data points. The parameters $B = 1.18$ and a minimum of eight points per bin are chosen by visual inspection as a compromise between variance reduction and resolution: smaller bins or lower thresholds yield noisy, highly irregular curves, whereas larger bins or higher thresholds oversmooth the distributions and may hide genuine cross-linguistic differences. In Figure 4 we show a representative comparison between standard (linearly binned) and logarithmically binned representations,[1] where the overall shape of the distribution is clearly preserved but random fluctuations, especially at the extremes, are substantially reduced.

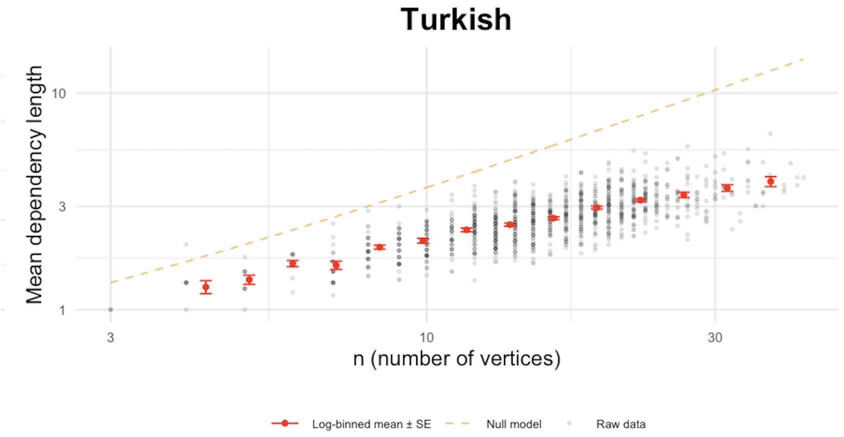


Figure 4: Comparison between standard binned raw data and logarithmically binned data (averaged on the same bin). On the y-axis the mean distances for each n, on the x-axis n and its equivalent n^* after the log-transformation. The plot has a log scale in both teh x and y-axis.

4.4 Heteroscedasticity and variance structure

Logarithmic binning partially regularizes the data but does not remove the strong heteroscedasticity present in our measurements. As shown in Figure 5, where we report the empirical variance of the mean dependency length in each merged bin n^* for all languages, the variance exhibits a clear dependence on n^* and is far from constant across the range. This behaviour indicates that the residuals of any regression model fitted to these data cannot reasonably be assumed to have homogeneous variance.

The non-linear interpolation model that we employ is estimated using the `nls()` routine, which by default assumes homoscedastic errors. In principle, a statistically coherent way to account for heteroscedasticity in non-linear regression would be to introduce observation weights based on the inverse of the variance, so that bins with higher uncertainty contribute less to the fit. Implementing such a weighted non-linear least squares, however, would require specifying or estimating a functional dependence of the variance on the bin position n (or on its logarithmic counterpart n^*), introducing additional parameters and model complexity. This, in turn, would affect information criteria such as AIC and complicate fair comparisons both across competing functional forms and across languages.

In our case, the empirical variance curves do not follow a simple, common functional form across languages, and they are not easily interpolated in a stable and interpretable way. For this reason we decided not to introduce an explicit variance model into the fitting procedure. Instead, we proceeded by fitting the non-linear model to the binned means of the mean dependency length in each bin n^*, treating

[1] Analogous plots for the other languages are produced by the accompanying analysis code.

all bins as equally weighted and thus effectively ignoring the heteroscedasticity in the formal estimation step. The results of these unweighted non-linear fits should therefore be interpreted with caution: they are useful to capture overall trends and to compare languages at a qualitative level, but they do not represent fully efficient or variance-optimal estimators.

4.5 Initial parameter estimation for nonlinear fits

Nonlinear least squares fitting via `nls()` requires initial guesses for the model parameters. To obtain these starting values we performed a preliminary linear regression on the log-transformed data (i.e. fitting $\log y$ versus $\log n^*$) as illustrated in Figure 6. For models that represent pure power-law behaviour, such as Model 1 ($f(n^*) = (n^*/2)^b$) and Model 2 ($f(n^*) = a\,(n^*)^b$), this strategy is theoretically well-motivated: the slope of the linear fit in the log–log plane directly estimates the exponent b, while the intercept provides a natural estimate for the prefactor a.

For models that do not correspond to power laws, such as the exponential Model 3 ($f(n^*) = a\ \exp(c\,n^*)$) or the logarithmic Model 4 ($f(n^*) = a\ \log(n^*)$), the log–log linear fit is not formally justified, since these functional forms are not linear on a doubly logarithmic scale. Nevertheless, the resulting initial parameters (a, b) provide a reasonable first approximation to the overall scale and rate of growth in the data and, in practice, we found that `nls()` converged without difficulty for these models across all languages, without being necessary to perform other kinds of initial parameters estimation for those models.

For the mixed model Model 5 ($f(n^*) = a\,(n^*)^b\ \exp(c\,n^*)$), which combines power-law and exponential behaviour, we initialized the power-law parameters a and b using the same log–log linear fit, and set the exponential coefficient $c = 0$ as a starting value. This choice reflects the assumption that the exponential term acts as a small correction to an underlying power-law trend, so that c is expected to be small in magnitude. Table 10 provides a summary of the estimated initial parameters.

Table 10: Initial parameter values used for nonlinear model fitting, based on log–log linear regression on binned means.

Language	a_0	b_0	c_0	d_0
Arabic	1.122	0.329	0.0	0.0
Chinese	0.815	0.455	0.0	0.0
Czech	0.910	0.411	0.0	0.0
English	1.025	0.375	0.0	0.0
Finnish	0.805	0.456	0.0	0.0
French	1.086	0.332	0.0	0.0
Galician	1.016	0.355	0.0	0.0
German	0.925	0.452	0.0	0.0
Hindi	0.682	0.519	0.0	0.0
Icelandic	1.119	0.321	0.0	0.0
Indonesian	0.920	0.383	0.0	0.0
Italian	1.082	0.337	0.0	0.0
Japanese	0.936	0.338	0.0	0.0
Korean	0.698	0.464	0.0	0.0
Polish	0.989	0.371	0.0	0.0
Portuguese	0.952	0.378	0.0	0.0
Russian	0.886	0.405	0.0	0.0
Spanish	1.107	0.328	0.0	0.0
Swedish	0.998	0.381	0.0	0.0

The same strategy was adopted for the extended versions of these models that include an additional additive constant d, i.e. functions of the form $f(n^*) + d$. In these cases the parameters a, b (and c, when present) were initialized exactly as described above from the log–log linear fit, while the offset d was always initialized to zero, corresponding to the assumption that any constant shift around the main trend should be relatively small.

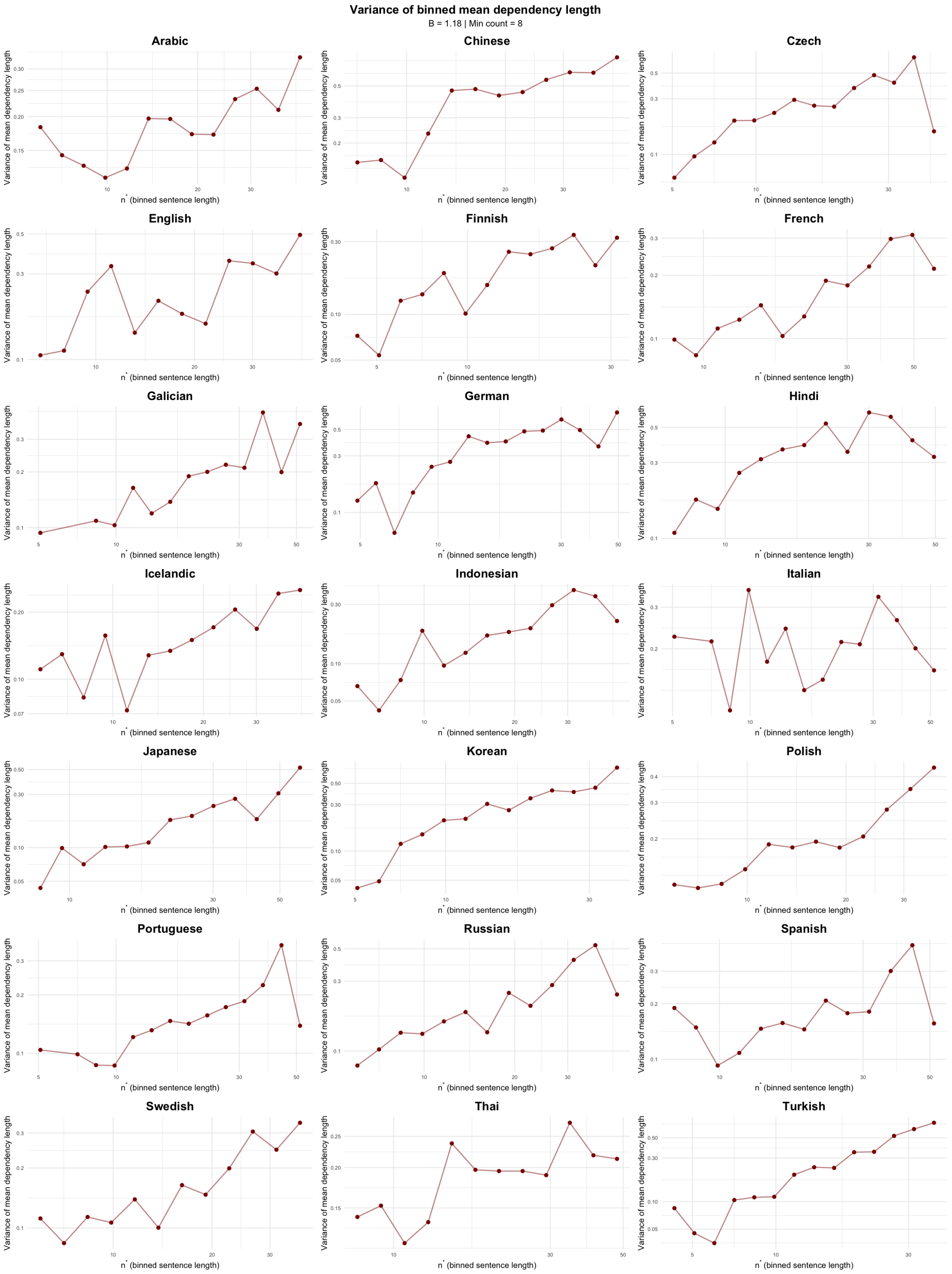


Figure 5: Empirical variance of mean dependency length as a function of the logarithmic bin centre n^* for all languages. Each panel corresponds to one language and shows the variance computed over merged bins obtained from the logarithmic binning procedure. The clear non-constant behaviour of the variance across n^* illustrates the strong heteroscedasticity present in the data, even after bin merging.

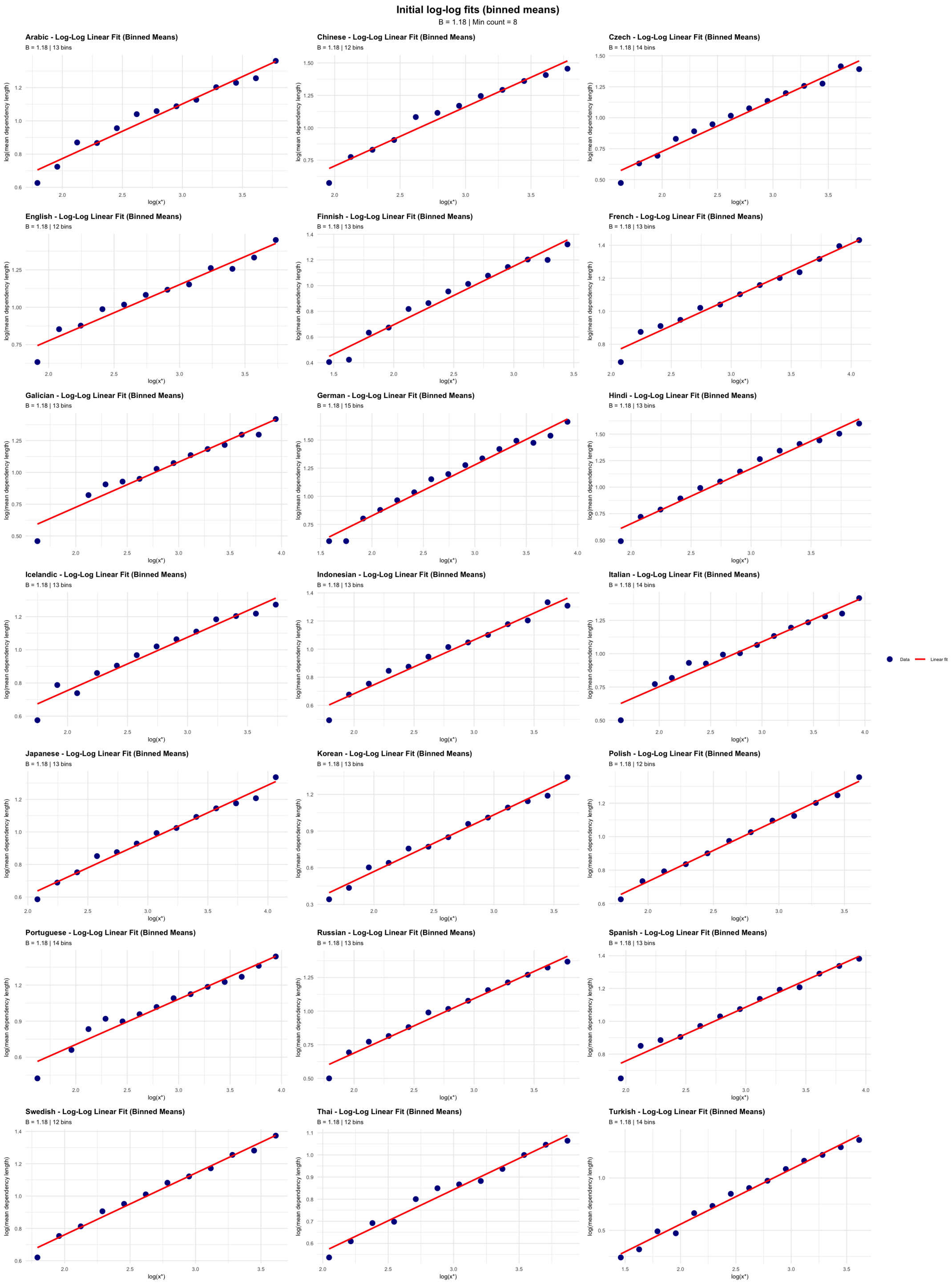


Figure 6: Linear regression fits on log-transformed data for all languages. Each panel shows log(mean dependency length) versus $\log(n^*)$, where n^* denotes the logarithmic bin centre. The slope and intercept of these linear fits provide initial parameter estimates for the nonlinear models.

4.6 Nonlinear model fitting

Once the initial parameters were obtained, we fitted all candidate functions to the binned means of the mean dependency length in each bin n^* using nonlinear least squares. Specifically, for every language and each functional form (Models 0–5 and their variants with an additional constant term d, Models 0+–5+) we minimized the unweighted sum of squared residuals between the observed binned means and the model predictions, treating all bins as equally informative despite the heteroscedasticity discussed above. The quality of these fits, and the differences between models, are illustrated in section 2.3, where we display the best-fitting curves for each language.

As a default solver we used the `nls()` function in R, which implements a Gauss–Newton–type algorithm for nonlinear least squares and is efficient when the objective surface is well behaved and the starting values are close to the optimum. However, for more complex models—most notably the combined power-law–exponential model (Model 5) and all the extensions with an additive constant d—`nls()` frequently encountered convergence issues, such as failure to decrease the residual sum of squares. In these cases we resorted to `nlsLM()` from the `minpack.lm` package, which employs a Levenberg–Marquardt algorithm that interpolates between Gauss–Newton and gradient descent, that makes the optimization substantially more robust to poor conditioning and suboptimal initial values. This combination of `nls()` for simpler models and `nlsLM()` for the more parameter-rich specifications ensured stable convergence of the nonlinear fits across all languages.

References

[1] Douglas M Bates and Donald G Watts. *Nonlinear regression analysis and its applications*, volume 2. Wiley New York, 1988.

[2] Anna Deluca and Álvaro Corral. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophysica*, 61(6):1351–1394, 2013.