LAB 04

Non-linear regression on dependency trees

By
Johannes Weinert & Gabriele Villa

November 19, 2025

Contents

List of Figures

List of Tables

1 Introduction

In this laboratory session, we investigated how to fit different non-linear functions to model the distribution of the mean edge length ($\langle d \rangle$) as a function of n of graphs obtained from the PUD dataset. After obtaining introductory statistical results for the graphs and metrics, we visually analyzed the relationship between the number of nodes in a sentence and the mean edge length $\langle d \rangle$ to assess whether the assumption of homoscedasticity was satisfied.

The model ensemble consisted of a null hypothesis (a linear model) and five different non-linear models. To obtain a more complete picture and consider more complex models, we also evaluated the addition of a y-axis shift parameter ($+d$) to the models.

Here, we report a summary table of the models we tried:

Table 1: Ensemble of candidate functions $f(n)$ used to model the scaling of mean edge length with sentence length, including the null model, basic models (1–5), and their + variants with an additive constant.

Model	**Formula**
Model 0	$f(n) = \frac{n}{3} + \frac{1}{3}$
Model 1	$f(n) = \left(\frac{n}{2}\right)^b$
Model 2	$f(n) = an^b$
Model 3	$f(n) = ae^{cn}$
Model 4	$f(n) = a \log n$
Model 5	$f(n) = an^b e^{cn}$
Model 1+	$f(n) = \left(\frac{n}{2}\right)^b + d$
Model 2+	$f(n) = an^b + d$
Model 3+	$f(n) = ae^{cn} + d$
Model 4+	$f(n) = a \log n + d$
Model 5+	$f(n) = an^b e^{cn} + d$

After carefully choosing the most accurate initial parameters, we use built-in R algorithms to fit the parameters for each model, and, through the evaluation of both AIC and BIC, as well as visual validation, we chose the best one and plotted it.

2 Results

2.1 Preliminary Results

To gain an initial intuition about how the mean edge length $\langle d \rangle$ scales with sentence size n, we begin with a set of exploratory plots and summary statistics based on the raw and aggregated data.
First, Table 2 displays preliminary statistics for all 21 languages. Second, Figure 1 presents preliminary visualizations of the relationship between sentence length and mean edge length, including the empirical mean and the random linear arrangement baseline.

Table 2: 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), $\mu_{\langle d \rangle}$ and $\sigma_{\langle d \rangle}$ are the mean and the standard deviation of the mean edge length $\langle d \rangle$.

Language	N	μ_n	σ_n	$\mu_{\langle d \rangle}$	$\sigma_{\langle d \rangle}$
Arabic	1000	20.747	8.614	3.008	0.581
Chinese	1000	21.415	8.622	3.290	0.876
Czech	1000	18.622	7.764	3.010	0.734
English	1000	21.187	8.213	3.160	0.664
Finnish	1000	15.817	6.527	2.831	0.678
French	1000	24.726	10.018	3.088	0.578
Galician	1000	23.510	9.804	3.070	0.614
German	1000	21.332	8.541	3.705	0.927
Hindi	1000	23.829	9.558	3.533	0.954
Icelandic	1000	18.833	7.668	2.861	0.543
Indonesian	1000	19.446	7.695	2.838	0.604
Italian	1000	23.732	9.850	3.086	0.626
Japanese	1000	28.788	10.997	2.874	0.575
Korean	1000	16.584	6.676	2.541	0.710
Polish	1000	18.384	7.428	2.867	0.621
Portuguese	1000	23.407	9.489	3.068	0.594
Russian	1000	19.355	8.048	2.914	0.674
Spanish	1000	23.284	9.446	3.067	0.586
Swedish	1000	19.085	7.652	3.028	0.612
Thai	1000	22.322	8.992	2.369	0.514
Turkish	1000	16.881	6.632	2.697	0.757

Figure 1: Preliminary plots of mean edge length $\langle d \rangle$ as a function of sentence length n. For each language, the figure shows the raw data (scatter), the empirical mean $\langle d \rangle$ aggregated by number of vertices (solid green line), and the expected mean under a random linear arrangement, $f(n) = (n+1)/3$ (dashed red line), in different axis scales.

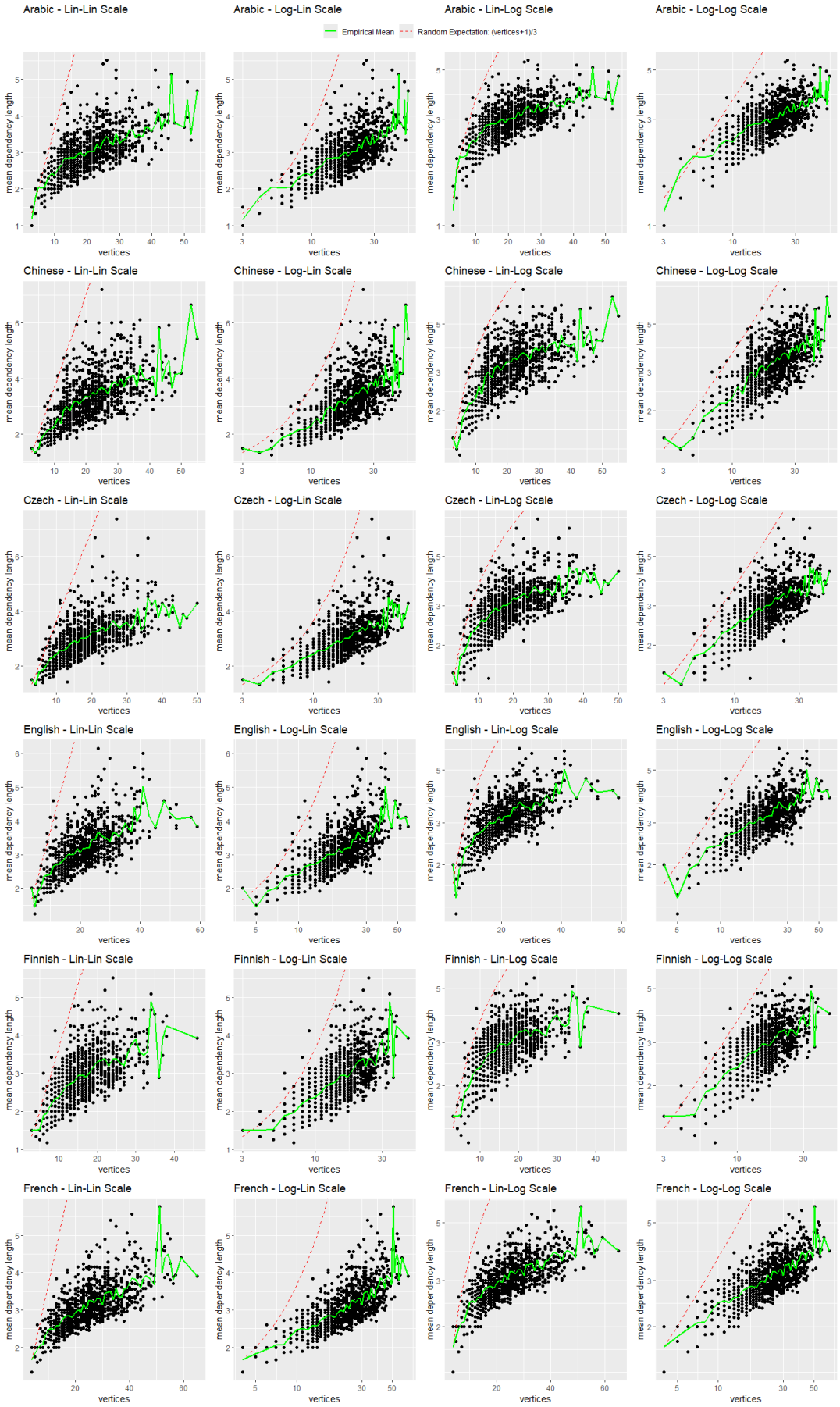


Figure 1: Preliminary plots of mean edge length $\langle d \rangle$ as a function of sentence length n (cont.).

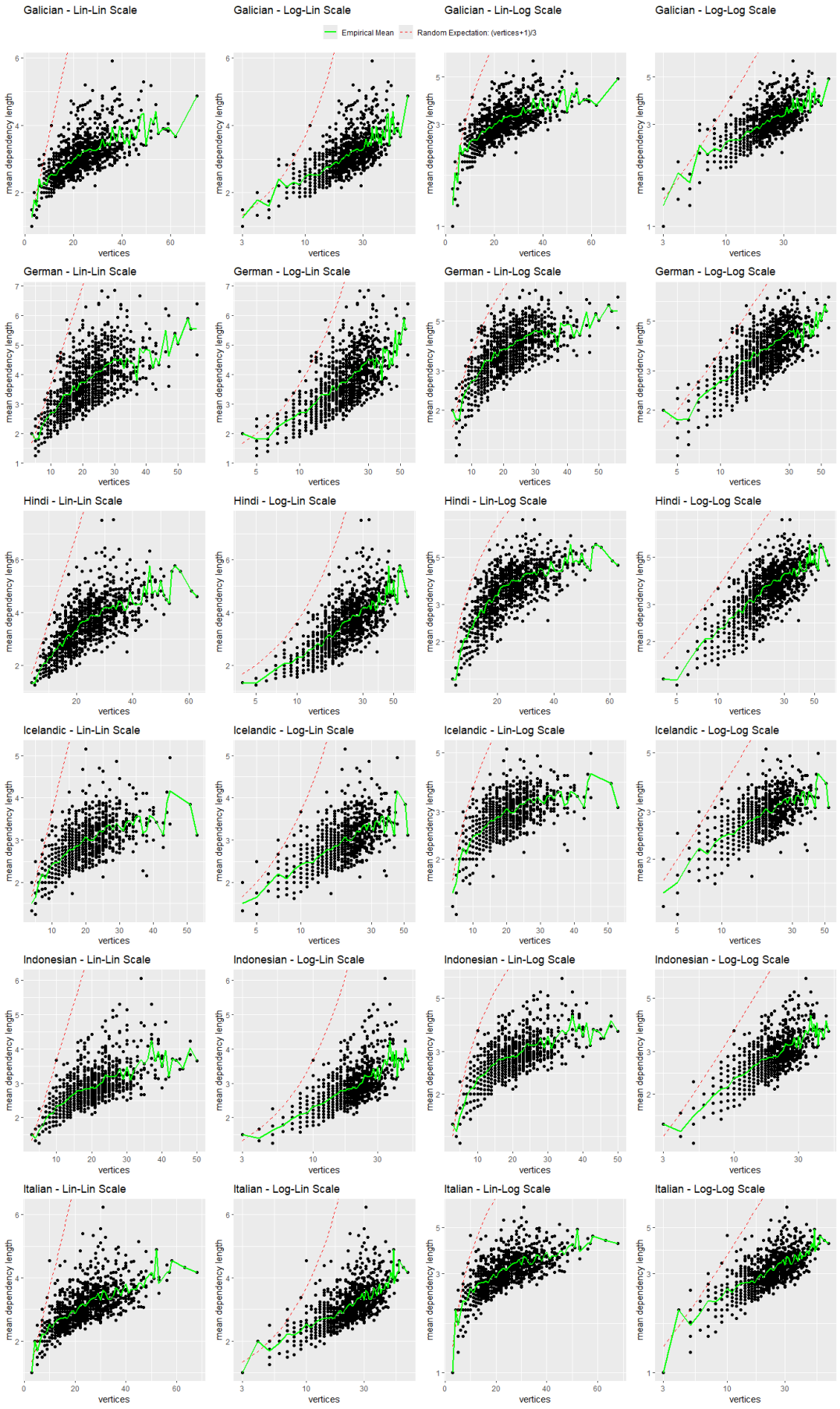


Figure 1: Preliminary plots of mean edge length $\langle d \rangle$ as a function of sentence length n (cont.).

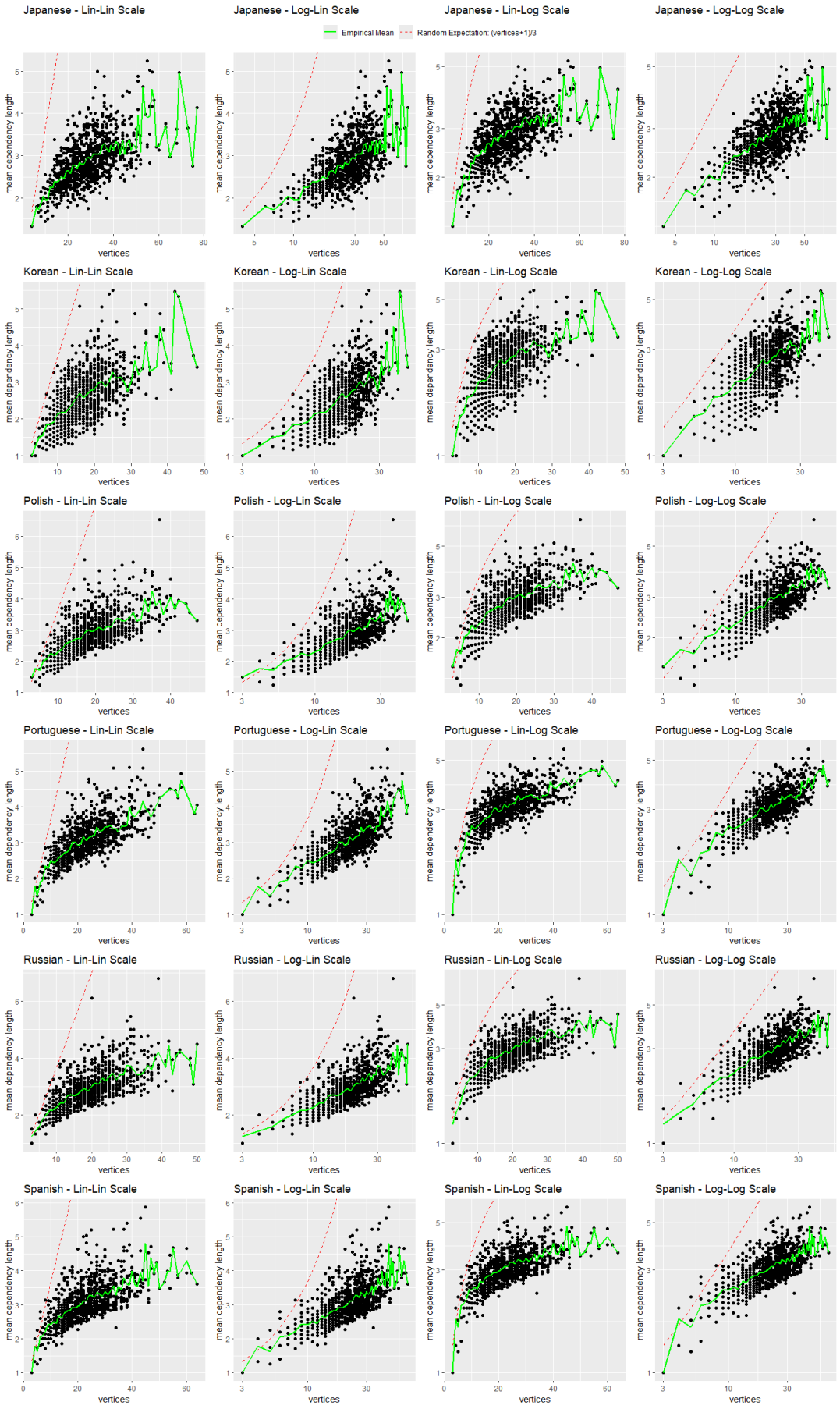
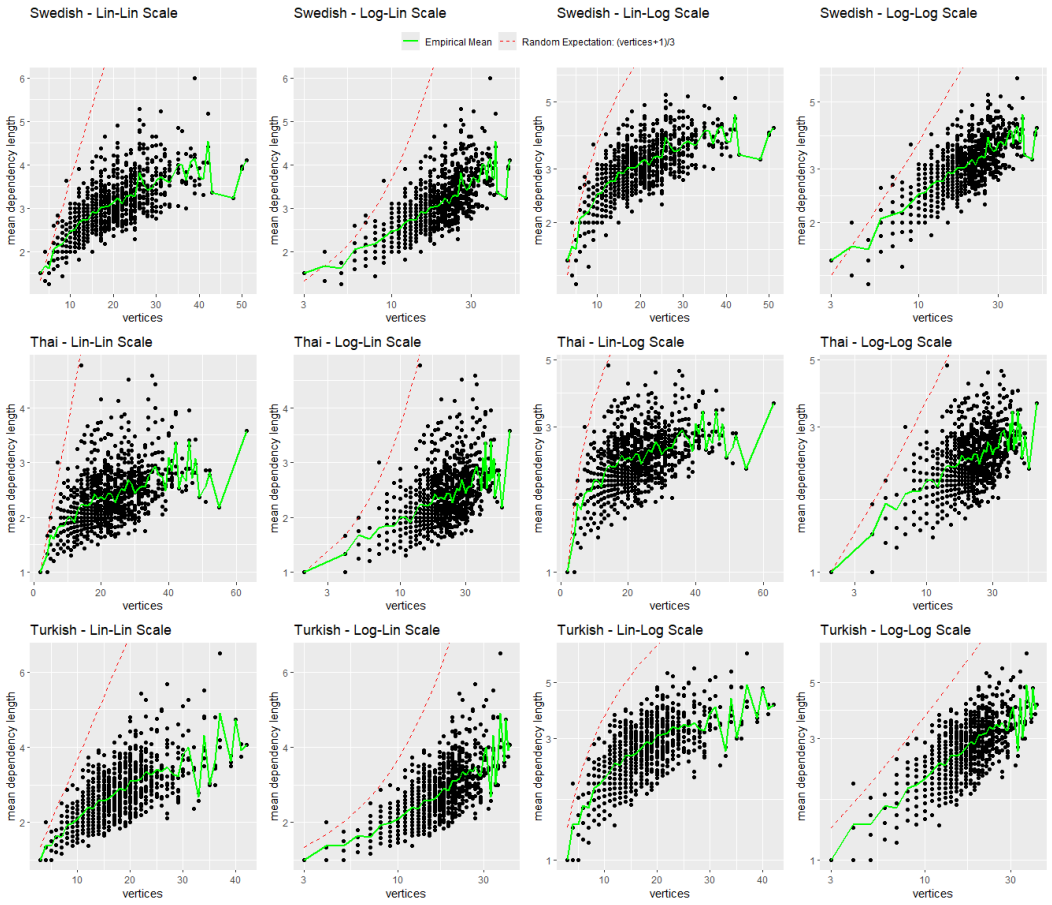


Figure 1: Preliminary plots of mean edge length $\langle d \rangle$ as a function of sentence length n (cont.).



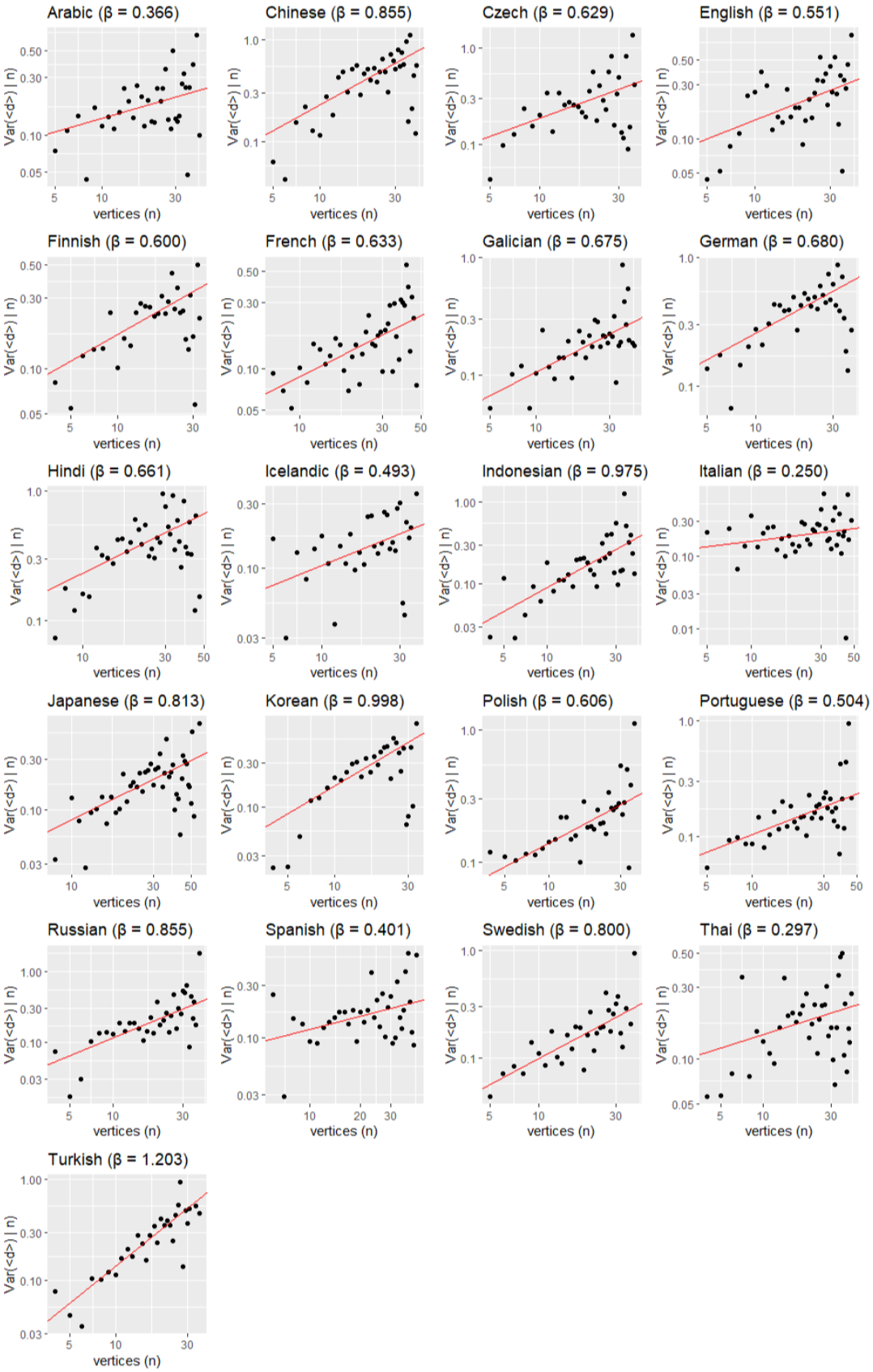
2.2 Homoscedasticity results

For each language, we estimated the scaling exponent β from the log–log regression of $\mathrm{Var}(\langle d \rangle \mid n)$ on n as explained in Section 4.1. Table 3 reports the fitted parameters for each language, sorted by slope β (ascending), including the intercept α, standard error, t-statistic, p-value for the test $H_0 : \beta = 0$, and the corresponding 95% confidence interval. All slopes were positive, ranging from $\beta = 0.25$ (Italian) to $\beta = 1.20$ (Turkish), with a median around $\beta = 0.65$, while intercepts α ranged approximately between -1.0 and -2.0. The results are visualized in Figure 2, where each point represents the empirical variance $\mathrm{Var}(\langle d \rangle \mid n)$ and the red line shows the fitted power-law relationship $\log_{10} \mathrm{Var}(\langle d \rangle \mid n) = \alpha + \beta \log_{10} n$ for the corresponding language.

Table 3: Fitted parameters of the log–log regression of $\mathrm{Var}(\langle d \rangle \mid n)$ on sentence length n for each language, sorted by slope β (ascending). For each language, the table reports the estimated slope β, intercept α, standard error of β, t-statistic and p-value for the hypothesis $H_0 : \beta = 0$, and the corresponding 95% confidence interval

Language	β	α	SE(β)	t-value	p-value	$\mathrm{CI}_{0.95}$
Italian	0.2495	-1.0438	0.179	1.39	1.71e-01	[-0.112, 0.611]
Thai	0.2973	-1.1394	0.158	1.88	6.78e-02	[-0.023, 0.617]
Arabic	0.3659	-1.2261	0.161	2.28	2.93e-02	[0.039, 0.693]
Spanish	0.4006	-1.3238	0.174	2.30	2.75e-02	[0.047, 0.754]
Icelandic	0.4935	-1.4747	0.203	2.44	2.10e-02	[0.080, 0.907]
Portuguese	0.5039	-1.4913	0.114	4.42	8.32e-05	[0.273, 0.735]
English	0.5508	-1.3885	0.193	2.86	7.21e-03	[0.159, 0.942]
Finnish	0.5999	-1.3688	0.135	4.43	1.32e-04	[0.322, 0.877]
Polish	0.6061	-1.4585	0.129	4.71	4.86e-05	[0.344, 0.868]
Czech	0.6285	-1.3522	0.187	3.35	2.12e-03	[0.246, 1.011]
French	0.6331	-1.6960	0.146	4.34	1.01e-04	[0.338, 0.928]
Hindi	0.6613	-1.2945	0.155	4.27	1.30e-04	[0.348, 0.975]
Galician	0.6746	-1.6444	0.145	4.66	4.45e-05	[0.381, 0.968]
German	0.6802	-1.2680	0.134	5.08	1.60e-05	[0.407, 0.953]
Swedish	0.7998	-1.8009	0.156	5.14	1.59e-05	[0.482, 1.118]
Japanese	0.8129	-1.9111	0.158	5.14	6.40e-06	[0.494, 1.132]
Russian	0.8550	-1.7862	0.162	5.28	8.18e-06	[0.525, 1.185]
Chinese	0.8554	-1.4937	0.158	5.40	4.76e-06	[0.534, 1.177]
Indonesian	0.9748	-2.0185	0.174	5.59	2.92e-06	[0.620, 1.329]
Korean	0.9982	-1.7679	0.182	5.47	7.63e-06	[0.625, 1.372]
Turkish	1.2028	-2.0579	0.126	9.54	2.68e-10	[0.945, 1.461]

Figure 2: Log–log plots of the conditional variance of $\langle d \rangle$ as a function of sentence length n.



2.3 Model Fitting Results

All nonlinear regressions were performed on aggregated data, as discussed in Section 3.1. Table 4 reports the residual standard error s for each model and language, while Table 5 presents the corresponding AIC values. Table 6 shows the AIC differences ΔAIC, computed for each language as the difference between the AIC of every model and that of the best model for that language. Analogously, Table 7 lists the BIC values and Table 8 their associated differences ΔBIC. Finally, Table 9 and Table 10 summarize all fitted parameters for all models across all languages.

Table 4: Residual standard error for each model and language

Languages	s_0	s_1	s_2	s_3	s_4	s_5	s_{1+}	s_{2+}	s_{3+}	s_{4+}	s_{5+}
Arabic	7.470	0.343	0.265	0.339	0.274	0.268	0.267	0.267	0.315	0.270	0.267
Chinese	7.078	0.459	0.449	0.521	0.477	0.453	0.449	0.453	0.491	0.466	0.455
Czech	6.627	0.328	0.261	0.393	0.246	0.247	0.269	0.250	0.349	0.249	0.250
English	7.274	0.358	0.262	0.395	0.251	0.247	0.268	0.255	0.348	0.253	0.245
Finnish	5.260	0.348	0.315	0.421	0.309	0.312	0.319	0.314	0.379	0.314	0.317
French	8.946	0.336	0.292	0.350	0.307	0.295	0.292	0.294	0.318	0.310	0.297
Galician	9.114	0.323	0.222	0.333	0.231	0.224	0.225	0.220	0.297	0.223	0.220
German	6.882	0.374	0.272	0.415	0.299	0.272	0.278	0.269	0.357	0.278	0.271
Hindi	8.189	0.339	0.326	0.528	0.388	0.290	0.331	0.297	0.440	0.298	0.293
Icelandic	6.956	0.324	0.210	0.319	0.210	0.182	0.218	0.191	0.293	0.189	0.184
Indonesian	6.989	0.256	0.198	0.323	0.192	0.186	0.202	0.191	0.281	0.191	0.187
Italian	8.776	0.285	0.194	0.320	0.201	0.196	0.198	0.193	0.276	0.199	0.192
Japanese	11.191	0.379	0.363	0.432	0.360	0.363	0.364	0.363	0.408	0.363	0.364
Korean	6.253	0.428	0.432	0.498	0.471	0.438	0.432	0.438	0.461	0.456	0.443
Polish	6.394	0.286	0.201	0.316	0.210	0.194	0.205	0.199	0.275	0.201	0.188
Portuguese	8.486	0.262	0.197	0.346	0.192	0.194	0.201	0.190	0.294	0.194	0.191
Russian	6.816	0.291	0.253	0.385	0.240	0.241	0.257	0.243	0.292	0.195	0.191
Spanish	8.882	0.341	0.260	0.389	0.246	0.249	0.265	0.247	0.338	0.244	0.249
Swedish	6.590	0.349	0.247	0.381	0.233	0.219	0.255	0.231	0.345	0.228	0.219
Thai	8.712	0.267	0.235	0.301	0.269	0.213	0.237	0.230	0.288	0.227	0.232
Turkish	5.626	0.341	0.346	0.436	0.379	0.348	0.345	0.347	0.392	0.356	0.352

Table 5: AIC values for each model and language

Languages	AIC_0	AIC_1	AIC_2	AIC_3	AIC_4	AIC_5	AIC_{1+}	AIC_{2+}	AIC_{3+}	AIC_{4+}	AIC_{5+}
Arabic	338.113	37.358	13.040	36.979	15.251	15.040	13.818	14.514	30.597	14.430	15.460
Chinese	332.843	65.552	64.477	79.179	69.434	66.436	64.643	66.357	74.088	68.189	67.719
Czech	299.909	30.364	11.094	47.720	4.361	6.640	13.179	8.011	37.912	6.274	8.623
English	308.301	38.117	11.022	47.871	5.984	6.593	13.043	9.489	37.490	7.828	6.858
Finnish	229.852	29.978	23.567	45.016	21.069	23.680	24.277	24.148	37.959	23.030	25.559
French	406.336	39.975	25.029	45.207	29.414	26.834	24.866	26.862	36.835	31.414	28.807
Galician	415.681	35.951	-6.002	40.115	-2.406	-4.716	-3.956	-5.989	28.308	-5.661	-4.631
German	323.397	44.863	15.265	55.858	23.372	16.097	17.458	14.985	42.325	17.118	16.719
Hindi	389.388	40.068	36.592	89.781	54.905	24.878	38.259	27.510	70.786	26.604	26.869
Icelandic	290.833	28.211	-8.093	27.800	-9.257	-19.715	-5.149	-15.306	21.226	-17.623	-22.318
Indonesian	311.421	8.147	-14.302	30.767	-18.232	-19.366	-12.230	-16.896	18.963	-17.752	-18.021
Italian	397.005	20.029	-20.129	34.657	-17.654	-18.508	-17.986	-20.128	19.380	-17.743	-19.710
Japanese	485.095	59.542	55.100	77.058	52.755	55.735	50.505	56.164	70.792	54.384	57.596
Korean	268.659	49.849	51.566	63.084	57.672	53.557	51.567	53.566	57.753	55.857	55.413
Polish	283.594	17.362	-11.821	26.911	-9.332	-14.310	-10.117	-11.810	16.023	-11.816	-15.758
Portuguese	379.084	11.315	-18.220	41.877	-21.254	-18.400	-15.590	-20.847	25.029	-19.267	-19.738
Russian	302.436	19.691	7.982	45.721	2.328	4.620	9.542	5.427	34.889	4.145	6.620
Spanish	405.529	41.272	11.701	57.107	5.055	8.012	14.403	7.181	46.143	5.207	9.201
Swedish	286.191	34.428	5.561	44.727	-0.090	-3.621	8.396	0.738	35.496	-1.316	-2.634
Thai	367.534	13.021	0.935	26.384	13.845	0.289	1.840	-0.303	22.631	-2.366	1.593
Turkish	247.408	29.715	31.655	49.956	37.988	33.137	31.601	32.933	42.458	37.337	34.913

Table 6: AIC differences for each model and 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	325.072	24.318	**0.000**	23.939	2.211	2.000	0.778	1.474	17.557	1.390	2.420
Chinese	268.366	1.075	**0.000**	14.702	4.958	1.959	0.166	1.880	9.611	3.712	3.260
Czech	295.548	26.003	6.733	43.359	**0.000**	2.281	8.818	3.650	33.516	1.913	4.262
English	302.317	32.133	5.039	41.887	**0.000**	0.609	7.059	3.505	31.505	1.844	0.874
Finnish	208.783	8.903	2.498	23.948	**0.000**	2.611	3.208	3.079	16.890	1.960	4.587
French	381.471	15.107	0.163	20.342	4.548	1.968	**0.000**	1.996	11.969	6.548	3.942
Galician	421.683	41.953	**0.000**	46.117	3.596	1.286	2.046	0.013	34.310	0.341	1.372
German	308.412	29.878	0.279	40.872	8.387	1.112	2.473	**0.000**	27.340	2.132	1.733
Hindi	364.510	15.190	11.713	64.903	30.027	**0.000**	13.381	2.632	45.908	1.726	1.991
Icelandic	310.548	47.926	11.622	47.515	10.455	**0.000**	14.566	4.409	40.941	2.093	1.763
Indonesian	330.787	27.513	5.064	50.133	1.134	**0.000**	7.136	2.470	38.329	1.614	1.345
Italian	417.134	40.159	**0.000**	54.786	2.475	1.621	2.143	0.001	39.509	2.745	0.418
Japanese	432.341	6.788	2.347	24.305	**0.000**	2.982	2.752	3.411	18.039	1.990	4.843
Korean	218.810	**0.000**	1.717	13.235	7.823	3.708	1.718	3.717	7.904	6.008	5.564
Polish	299.352	33.120	3.937	42.669	6.426	1.449	5.641	3.948	31.781	3.943	**0.000**
Portuguese	400.338	32.568	3.034	63.131	**0.000**	2.854	5.664	0.407	46.505	1.987	1.518
Russian	300.108	17.363	5.654	43.393	**0.000**	2.292	7.214	3.099	32.561	1.817	4.292
Spanish	400.474	36.216	6.645	52.051	**0.000**	2.956	9.348	2.126	41.087	0.151	4.145
Swedish	289.813	38.049	9.183	48.195	3.532	**0.000**	12.018	4.360	39.119	2.306	0.969
Thai	369.900	15.386	3.330	28.754	16.210	2.655	4.197	2.063	24.969	**0.000**	3.965
Turkish	217.692	**0.000**	1.940	20.241	8.273	3.421	1.886	3.218	12.743	4.258	5.198

Table 7: BIC values for each model and language

Languages	BIC_0	BIC_1	BIC_2	BIC_3	BIC_4	BIC_5	BIC_{1+}	BIC_{2+}	BIC_{3+}	BIC_{4+}	BIC_{5+}
Arabic	340.004	41.141	18.716	42.654	19.034	22.607	19.494	22.081	38.164	20.106	24.919
Chinese	334.735	69.336	70.152	84.855	73.218	74.003	70.319	73.924	81.656	73.865	77.179
Czech	301.715	33.978	16.514	53.140	7.974	13.869	18.600	15.237	45.139	11.694	17.657
English	310.108	41.730	16.443	53.291	9.597	13.820	18.463	16.716	44.716	13.248	15.809
Finnish	231.463	33.194	28.399	49.849	24.291	30.124	29.109	30.591	44.403	27.862	33.710
French	408.362	44.023	31.105	51.283	33.464	34.935	30.942	34.963	44.962	37.490	38.934
Galician	417.724	40.038	0.127	46.244	1.680	3.456	2.173	2.183	36.480	0.468	5.585
German	325.269	48.606	20.878	61.471	27.114	23.582	23.072	22.470	49.810	22.731	26.075
Hindi	391.396	44.083	42.614	95.803	58.919	32.907	44.281	35.539	78.816	32.626	36.905
Icelandic	292.594	31.733	-2.810	33.084	-5.737	-12.670	0.135	-8.262	28.271	-12.339	-9.146
Indonesian	313.249	11.805	-8.816	36.253	-14.575	-12.051	-6.744	-9.581	26.277	-12.267	-8.877
Italian	399.012	24.044	-14.107	40.679	-13.064	-10.479	-11.964	-12.098	27.409	-11.362	-9.673
Japanese	487.238	63.828	61.529	83.487	57.040	64.305	61.934	64.614	79.365	61.173	68.312
Korean	270.373	53.276	56.707	68.225	61.099	60.412	56.708	60.420	64.607	61.000	63.981
Polish	285.355	20.885	-6.538	32.195	-5.809	-7.265	4.834	-4.765	23.068	-6.532	-6.952
Portuguese	381.054	15.255	-12.309	47.787	-17.313	-10.519	-9.679	-12.097	33.132	-13.356	-9.850
Russian	304.243	23.304	13.402	51.141	5.941	11.847	14.962	12.653	42.116	9.565	15.653
Spanish	407.554	45.323	17.777	63.183	9.106	16.113	20.480	15.283	54.244	11.283	19.327
Swedish	287.952	37.950	10.844	49.856	3.432	3.423	13.680	7.782	42.542	3.968	6.153
Thai	369.466	16.885	6.731	32.184	17.708	8.016	7.626	7.425	30.359	3.430	11.258
Turkish	249.071	33.042	36.646	54.947	41.315	39.791	36.592	39.587	49.113	38.964	43.231

Table 8: BIC differences for each model and language

Language	ΔBIC_0	ΔBIC_1	ΔBIC_2	ΔBIC_3	ΔBIC_4	ΔBIC_5	ΔBIC_{1+}	ΔBIC_{2+}	ΔBIC_{3+}	ΔBIC_{4+}	ΔBIC_{5+}
Arabic	321.289	22.426	**0.000**	23.939	0.319	3.892	0.778	3.366	19.449	1.390	6.204
Chinese	265.399	**0.000**	0.817	15.519	3.882	4.668	0.983	4.588	12.319	4.529	7.843
Czech	293.741	26.003	8.540	45.165	**0.000**	5.895	10.625	7.263	37.165	3.720	9.683
English	300.510	32.133	6.845	43.693	**0.000**	4.222	8.866	7.118	35.118	3.651	6.296
Finnish	207.173	8.903	4.109	25.558	**0.000**	5.833	4.819	6.300	20.112	3.571	9.419
French	377.420	13.082	0.163	20.342	2.526	3.934	**0.000**	4.021	13.995	6.548	7.923
Galician	417.597	39.910	**0.000**	46.117	1.553	3.328	2.046	2.056	36.353	0.341	5.458
German	304.391	27.727	**0.000**	40.593	6.236	2.704	2.194	1.592	28.931	1.853	5.196
Hindi	358.770	11.457	9.988	63.177	26.924	0.282	11.653	2.913	46.190	**0.000**	4.280
Icelandic	305.264	44.404	9.861	45.754	6.933	**0.000**	12.805	4.508	40.941	0.335	3.524
Indonesian	327.824	26.380	5.759	50.829	**0.000**	2.524	7.831	4.993	40.852	2.309	5.698
Italian	413.119	38.151	**0.000**	54.786	0.467	3.627	2.143	2.009	41.517	2.745	4.336
Japanese	430.198	6.788	4.490	26.448	**0.000**	7.268	4.895	7.639	22.325	4.133	11.272
Korean	217.097	**0.000**	3.431	14.949	7.823	7.136	3.432	7.144	11.331	7.721	10.705
Polish	292.619	28.149	0.727	39.457	1.455	**0.000**	2.431	2.499	30.333	0.733	0.313
Portuguese	398.367	32.568	5.005	65.101	**0.000**	6.794	7.634	4.348	50.455	3.958	7.429
Russian	298.302	17.363	7.461	45.199	**0.000**	5.906	9.021	6.712	36.174	3.624	9.712
Spanish	398.448	36.216	8.671	54.076	**0.000**	7.007	11.373	6.176	45.138	2.176	10.221
Swedish	284.530	34.527	7.421	46.434	0.010	**0.000**	10.257	4.359	39.119	0.545	2.783
Thai	366.036	13.455	3.330	28.754	14.278	4.587	4.197	3.995	26.929	**0.000**	7.828
Turkish	216.029	**0.000**	3.603	21.905	8.273	6.749	3.549	6.545	16.070	5.922	10.188

Table 9: Fitted parameters for basic models: 1 to 5

	Model								
	1	2		3		4	5		
Language	*b*	*a*	*b*	*a*	*c*	*a*	*a*	*b*	*c*
Arabic	0.45	1.13	0.33	2.17	1.35e-02	1.03	1.13	0.33	1.74e-05
Chinese	0.49	0.88	0.43	2.12	1.73e-02	1.14	0.92	0.41	8.88e-04
Czech	0.47	1.07	0.36	2.16	1.52e-02	1.06	0.78	0.52	-7.74e-03
English	0.47	1.15	0.34	2.36	1.22e-02	1.07	0.84	0.49	-6.22e-03
Finnish	0.48	0.98	0.39	2.02	1.89e-02	1.07	0.77	0.52	-6.94e-03
French	0.44	1.05	0.34	2.22	1.21e-02	1.02	1.11	0.32	1.07e-03
Galician	0.44	1.17	0.31	2.29	1.08e-02	1.01	1.09	0.35	-1.34e-03
German	0.53	1.08	0.40	2.54	1.52e-02	1.27	0.95	0.47	-2.58e-03
Hindi	0.50	0.85	0.45	2.35	1.48e-02	1.19	0.49	0.70	-9.03e-03
Icelandic	0.44	1.23	0.29	2.22	1.16e-02	0.98	0.83	0.49	-9.36e-03
Indonesian	0.44	1.03	0.34	2.04	1.46e-02	0.99	0.80	0.48	-6.42e-03
Italian	0.44	1.10	0.33	2.24	1.17e-02	1.01	1.05	0.35	-8.84e-04
Japanese	0.39	0.99	0.32	2.14	9.14e-03	0.89	0.80	0.41	-2.93e-03
Korean	0.45	0.68	0.48	1.72	2.08e-02	0.99	0.69	0.46	5.77e-04
Polish	0.46	1.10	0.33	2.10	1.50e-02	1.02	0.89	0.45	-5.66e-03
Portuguese	0.44	1.03	0.35	2.20	1.23e-02	1.02	0.92	0.41	-2.22e-03
Russian	0.46	0.98	0.37	2.06	1.55e-02	1.03	0.74	0.52	-6.87e-03
Spanish	0.44	1.14	0.32	2.27	1.11e-02	1.01	0.89	0.44	-4.64e-03
Swedish	0.46	1.16	0.33	2.23	1.36e-02	1.05	0.81	0.52	-9.08e-03
Thai	0.34	1.07	0.26	1.81	9.94e-03	0.77	0.90	0.35	-4.25e-03
Turkish	0.47	0.70	0.48	1.72	2.29e-02	1.02	0.60	0.56	-4.39e-03

Table 10: Fitted parameters for plus-models: 1+ to 5+

	Model													
	1+		2+			3+			4+		5+			
Language	*b*	*d*	*a*	*b*	*d*	*a*	*c*	*d*	*a*	*d*	*a*	*b*	*c*	*d*
Arabic	0.40	0.50	2.44	0.21	-1.57	356.41	1.27e-04	-354.44	0.94	2.85e-01	2.12e+01	0.03	4.94e-04	-20.49
Chinese	0.47	0.24	1.28	0.36	-0.57	387.66	1.64e-04	-385.90	1.30	-5.30e-01	9.28e+00	0.08	1.72e-03	-8.84
Czech	0.42	0.44	14.06	0.06	-13.80	593.70	8.90e-05	-591.81	1.05	4.70e-02	6.26e-01	0.58	-9.06e-03	0.22
English	0.41	0.55	10.84	0.08	-10.48	738.43	6.21e-05	-736.33	1.05	6.89e-02	9.67e-02	1.11	-1.93e-02	1.37
Finnish	0.44	0.33	4.15	0.17	-3.70	561.89	1.14e-04	-560.17	1.09	-4.34e-02	5.48e-01	0.62	-9.27e-03	0.32
French	0.40	0.42	0.72	0.41	0.48	350.78	1.20e-04	-348.79	1.02	-8.52e-05	4.54e+00	0.11	1.78e-03	-3.63
Galician	0.38	0.54	3.63	0.16	-2.85	483.09	7.82e-05	-481.01	0.92	3.03e-01	1.65e+01	0.04	3.25e-04	-15.80
German	0.48	0.57	3.26	0.23	-2.78	539.18	1.19e-04	-537.02	1.44	-5.67e-01	1.06e+01	0.09	7.23e-04	-10.38
Hindi	0.48	0.20	11.43	0.10	-12.07	775.25	8.03e-05	-773.37	1.55	-1.22e+00	5.55e-01	0.67	-8.43e-03	-0.12
Icelandic	0.36	0.58	62.80	0.01	-62.29	560.69	6.77e-05	-558.65	0.83	4.62e-01	2.24e+00	0.28	-4.79e-03	-1.71
Indonesian	0.40	0.37	6.30	0.11	-5.83	504.11	9.34e-05	-502.30	0.94	1.52e-01	2.74e-01	0.78	-1.33e-02	0.79
Italian	0.39	0.47	2.74	0.20	-1.97	506.10	8.10e-05	-504.10	0.97	1.55e-01	1.84e+01	0.04	4.25e-04	-17.75
Japanese	0.36	0.29	4.17	0.14	-3.70	511.30	5.89e-05	-509.38	0.89	-2.31e-02	2.68e-01	0.67	-6.58e-03	0.83
Korean	0.46	-0.08	0.69	0.47	-0.03	389.83	1.68e-04	-388.47	1.18	-6.04e-01	7.13e+00	0.08	2.62e-03	-6.79
Polish	0.40	0.46	3.61	0.16	-2.89	519.59	9.48e-05	-517.72	0.93	2.88e-01	5.94e-02	1.29	-2.77e-02	1.41
Portuguese	0.41	0.38	4.01	0.16	-3.49	570.81	7.66e-05	-568.88	1.02	-1.30e-02	1.67e+01	0.05	3.51e-04	-16.30
Russian	0.42	0.32	7.84	0.10	-7.58	582.06	8.99e-05	-580.27	1.05	-6.65e-02	7.54e-01	0.51	-6.74e-03	-0.02
Spanish	0.39	0.50	33.59	0.03	-33.31	601.72	6.50e-05	-599.68	0.95	1.97e-01	9.12e+00	0.09	-2.50e-04	-8.75
Swedish	0.40	0.54	39.14	0.02	-38.80	688.11	6.92e-05	-686.11	0.97	2.70e-01	2.10e-01	0.91	-1.90e-02	0.98
Thai	0.30	0.31	34.49	0.02	-33.83	424.89	6.10e-05	-423.19	0.58	6.27e-01	1.10e+01	0.05	-2.27e-04	-10.38
Turkish	0.47	-0.04	1.66	0.32	-1.33	413.66	1.73e-04	-412.30	1.23	-6.20e-01	3.04e+00	0.21	1.38e-03	-2.82

2.4 Best Models

Table 11 reports for every language, the model selected by AIC together with the model that minimizes BIC. Those models where the best model according to AIC differs from the one chosen by BIC are highlighted in green.

Table 11: Best model for every language according to AIC and BIC (highlighted in green where the best model differs)

Language	AIC-best model	BIC-best model
Arabic	Model 2	Model 2
Chinese	**Model 2**	**Model 1**
Czech	Model 4	Model 4
English	Model 4	Model 4
Finnish	Model 4	Model 4
French	Model 1+	Model 1+
Galician	Model 2	Model 2
German	**Model 2+**	**Model 2**
Hindi	**Model 5**	**Model 4+**
Icelandic	Model 5	Model 5
Indonesian	**Model 5**	**Model 4**
Italian	Model 2	Model 2
Japanese	Model 4	Model 4
Korean	Model 1	Model 1
Polish	**Model 5+**	**Model 5**
Portuguese	Model 4	Model 4
Russian	Model 4	Model 4
Spanish	Model 4	Model 4
Swedish	Model 5	Model 5
Thai	Model 4+	Model 4+
Turkish	Model 1	Model 1

2.4.1 Visual AIC-best versus BIC-best Comparison

As shown in Table 11, for five languages the model with the lowest AIC and the one with the lowest BIC differed. Therefore, we plotted them to see whether we could visually detect a model being better than the other one:

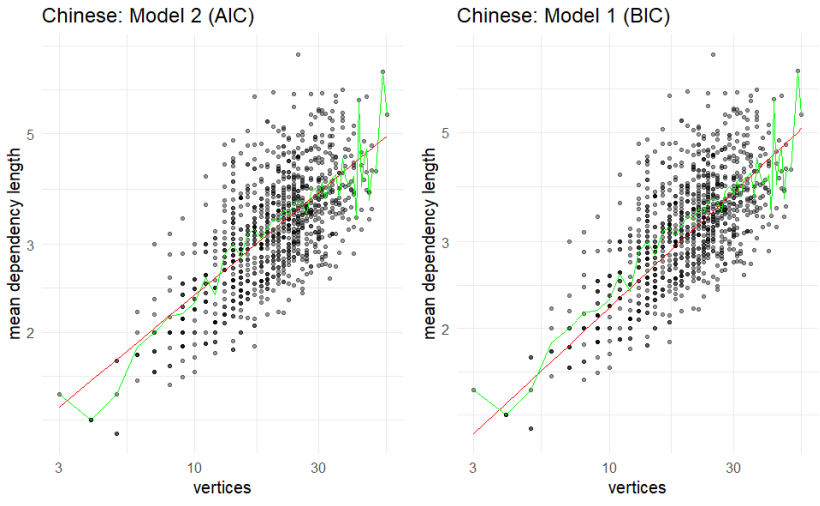


Figure 3: Comparison of best model according to AIC (model 2) and BIC (model 1) for Chinese

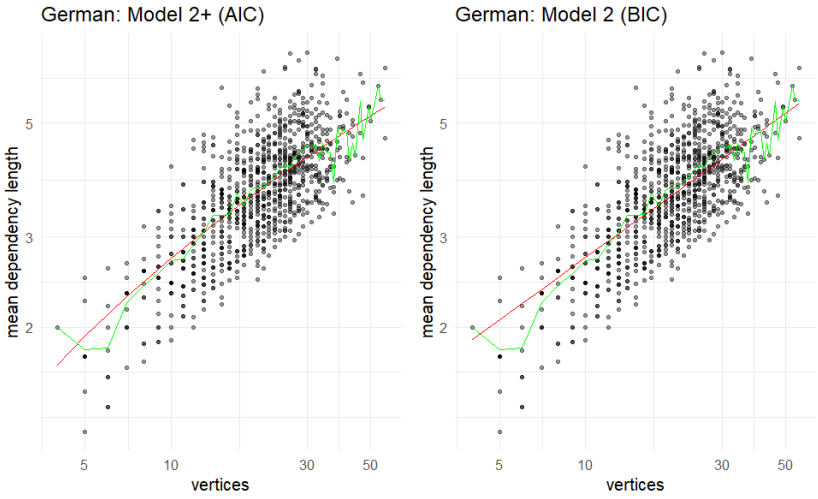


Figure 4: Comparison of best model according to AIC (model 2+) and BIC (model 2) for German

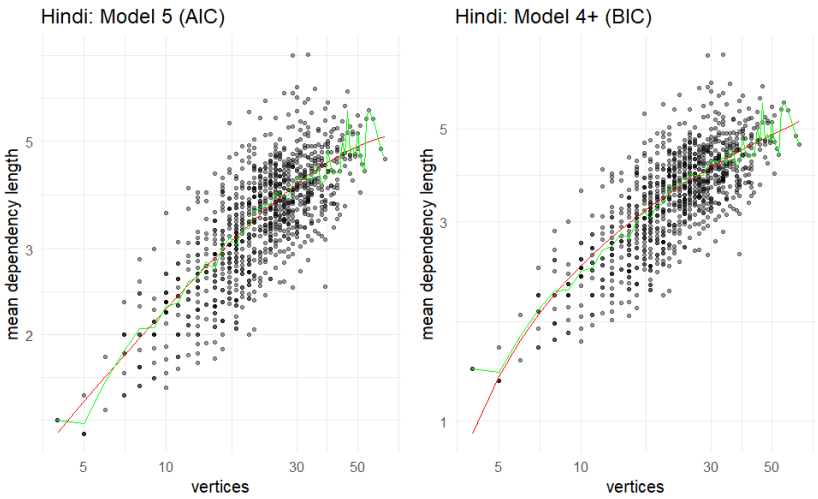


Figure 5: Comparison of best model according to AIC (model 5) and BIC (model 4+) for Hindi

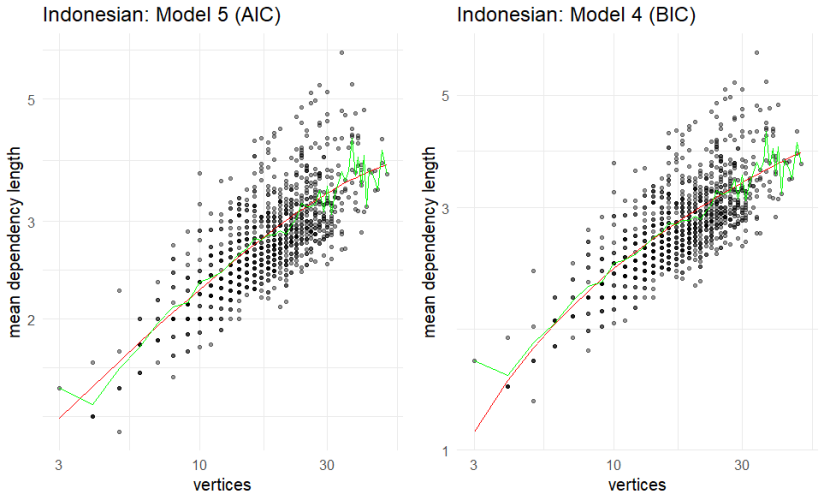


Figure 6: Comparison of best model according to AIC (model 5) and BIC (model 4) for Indonesian

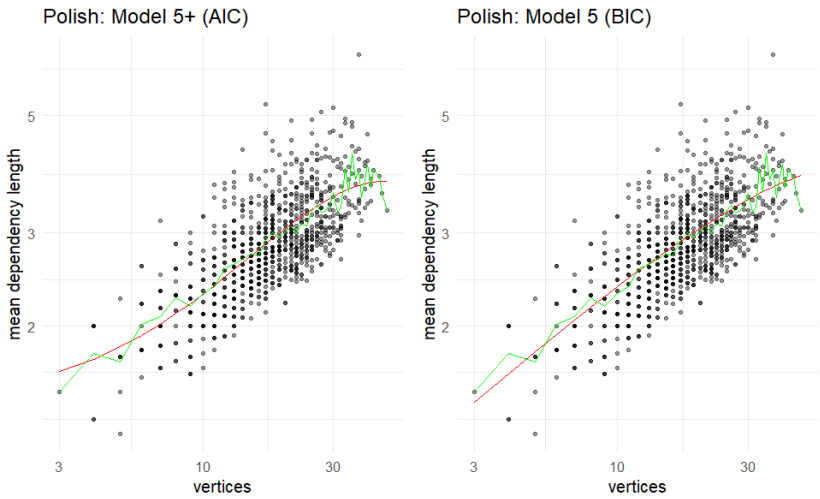
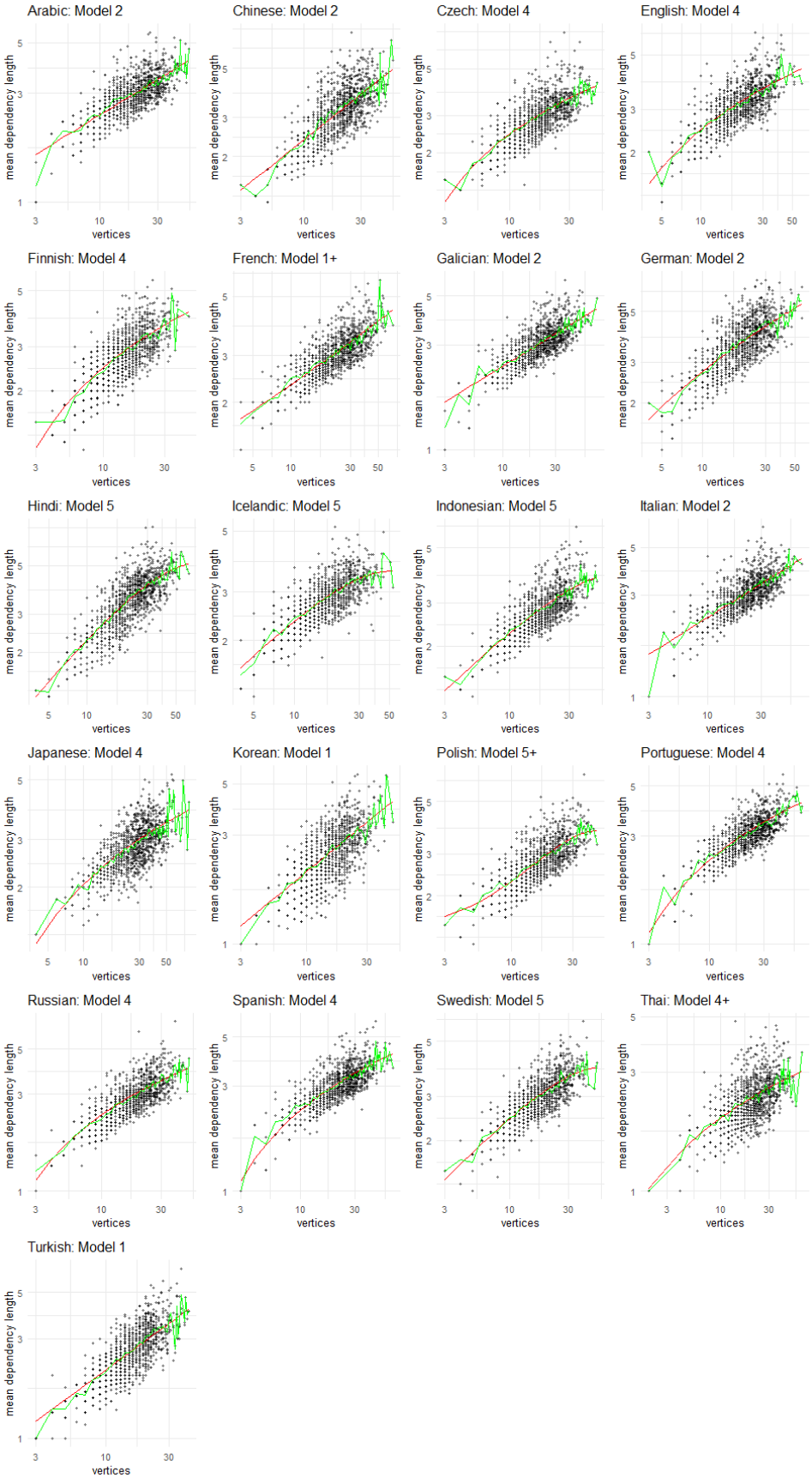


Figure 7: Comparison of best model according to AIC (model 5+) and BIC (model 5) for Polish

2.4.2 All Fitted Best-AIC Models

Figure 8 illustrates, for each language, the best-fitting model selected by AIC: the figure shows the raw sentence-level data cloud, the aggregated mean edge length $\langle d \rangle$ for each sentence length n, and the corresponding fitted curve of the chosen model.

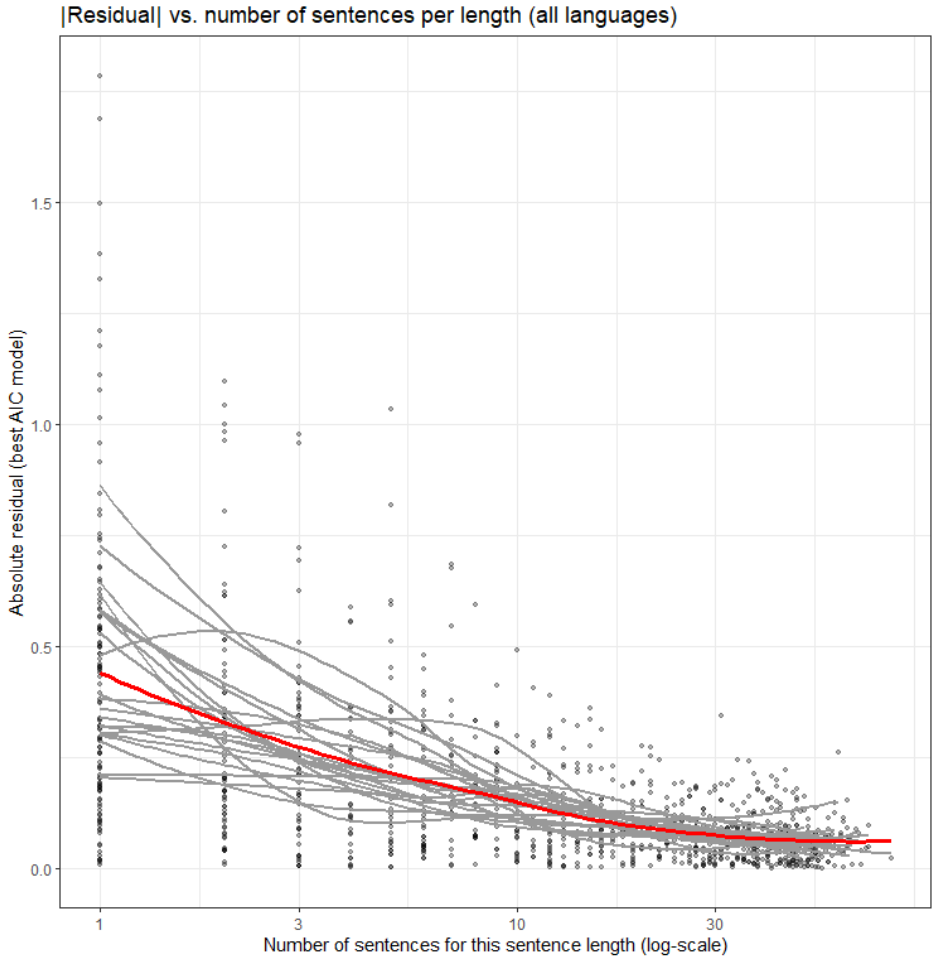
Figure 8: Best fit model (red line) according to AIC for every language together with the raw datapoints and the mean-aggregated data (green line)



2.4.3 Residuals of Fitted Final Models

Figure 9 displays, for all languages jointly, the absolute residuals of the finally selected models. For each language, the absolute residuals at each sentence length n were computed from the aggregated data point (mean $\langle d \rangle$ at that n) and plotted against the number of sentences available for that sentence length. All languages' points are overlaid in a single figure. The light grey curves represent a LOESS smooth fitted separately for each language, while the red curve shows a global LOESS fit computed over the entire combined dataset across all languages.

Figure 9: Absolute residuals of the best-fitting models for all languages. Each point represents the absolute residual at a given sentence length n, plotted against the number of sentences observed for that n. Light grey lines show per-language LOESS smooths, and the red line shows a global LOESS smooth fitted to all languages combined.



2.4.4 Language–Model Selection Patterns

Figure 10 summarizes how the selected AIC-best models are distributed across languages by showing for each model, the languages that selected it, together with the corresponding counts. Figure 11 visualizes how these choices relate to basic summary statistics. It displays standardized summary statistics (μ_n, σ_n, $\mu_{\langle d \rangle}$, $\sigma_{\langle d \rangle}$) plotted against the chosen model, with each point representing one language. This provides a compact view of how sentence length and dependency-length characteristics vary across the set of models selected by AIC.

Figure 10: Number of languages selecting each AIC-best model, with language names displayed inside the bars.

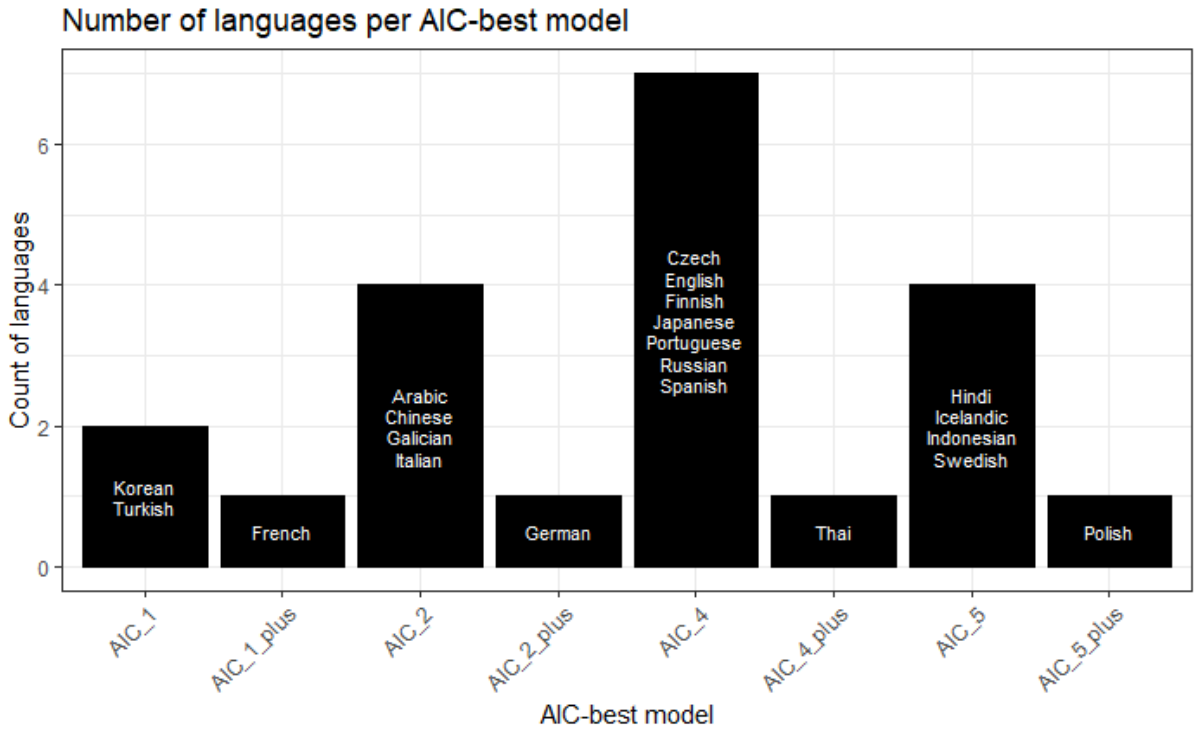
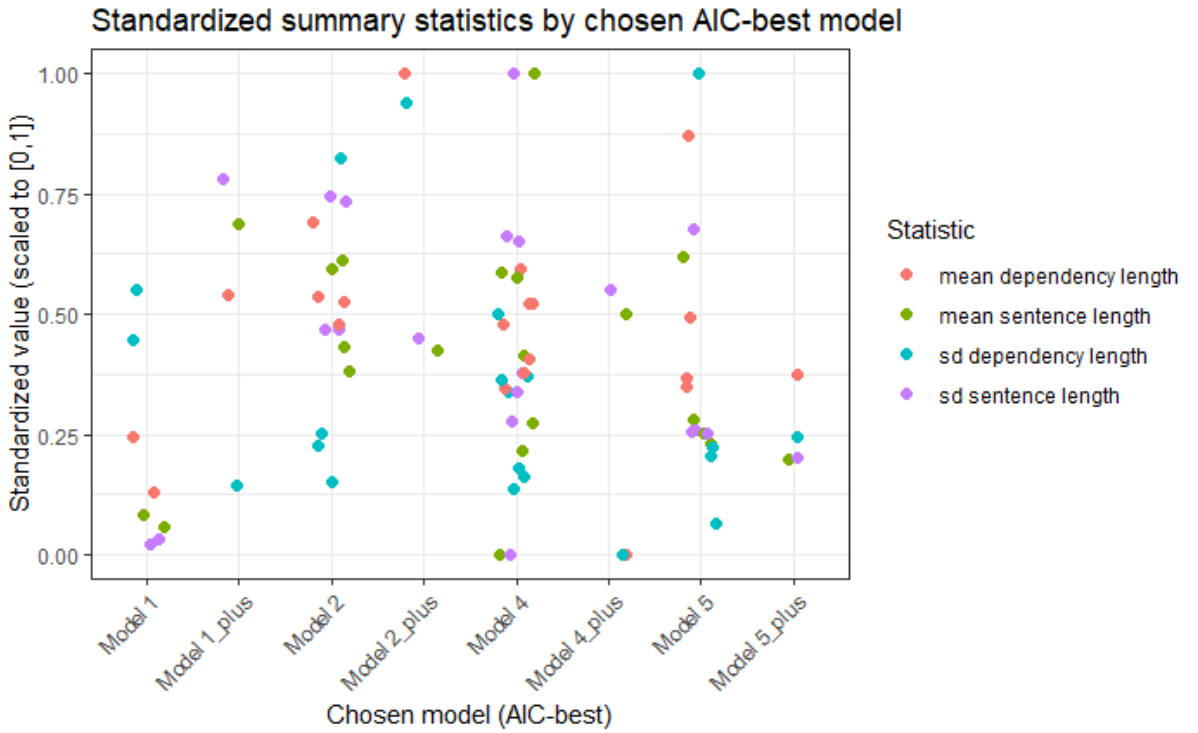


Figure 11: Standardised summary statistics for all languages plotted against the selected AIC-best model. Each point represents one language.



3 Discussion

3.1 Homoscedasticity Assumption

For each language, we assessed the homoscedasticity assumption by regressing $\log_{10} \mathrm{Var}(\langle d \rangle \mid n)$ on $\log_{10} n$, as described in Section 4.1. The estimated slopes β and their standard errors, t-statistics, p-values and 95% confidence intervals are reported in Table 3. All point estimates were positive, ranging from $\beta = 0.25$ (Italian) to $\beta = 1.20$ (Turkish), with a median around $\beta \approx 0.65$. For 19 out of 21 languages, the 95% confidence interval for β does not include zero, providing clear evidence that the variance of $\langle d \rangle$ increases systematically with sentence length. Only Italian and Thai have confidence intervals that intersect zero, indicating weaker statistical evidence for a deviation from homoscedasticity in these two cases.

These results imply that, for most languages in the corpus, the strict homoscedasticity assumption of ordinary nonlinear least-squares regression on the raw sentence-level data is not satisfied. Following the lab instructions, we therefore fitted all nonlinear models to the aggregated data, where each observation corresponds to the mean edge length $\langle d \rangle$ for a given sentence length n. Aggregating at the level of n reduces within-length variability (each point is an average over multiple sentences) and mitigates the dependence of the residual variance on n. In other words, the nonlinear regressions are fitted to a much smoother mean curve, for which the residuals are expected to be more stable across sentence lengths.

At the same time, the values of β in Table 3 reveal clear cross-linguistic differences in how strongly the dispersion of $\langle d \rangle$ grows with sentence length. Some languages exhibit relatively steep scaling exponents (larger β), while others have noticeably flatter slopes. We do not pursue a detailed linguistic interpretation here, but these results show that although the direction of the effect is consistent (variance increasing with n), the strength of heteroscedasticity is not uniform across languages.

3.2 Model Selection

A first and important step in the model selection process was comparing the null hypothesis (Model 0), which represents a simple linear relationship, with the set of nonlinear alternatives. Across all languages, Model 0 produced substantially higher values of AIC, BIC, and residual standard error s than any nonlinear model. This consistent gap makes it straightforward to reject the linear null model and proceed under the assumption that the relationship between mean edge length $\langle d \rangle$ and sentence length n is nonlinear.

Table 11 summarizes, for each language, the best models according to AIC and BIC. In most cases the two criteria agree, but five languages differ in their preferred model. Based on a visual comparison (Plots in Section 2.4.1), we find that the fitted curves of the best AIC models seem to select models that match the observed behavior of short sentence lengths more closely and otherwise behave very similar, we decided to adopt the AIC-best model as the final choice for all languages. Similarly, the preliminary plots in Figure 1 indicate that $\langle d \rangle$ grows more slowly than the random-reference expectation $(n+1)/3$, instead suggesting a sublinear, nearly power-law–like increase, fully consistent with the need for nonlinear models.

The resulting AIC-best models, plotted in Figure 8, show a good qualitative agreement with the aggregated

observations. Each curve provides a reasonable estimate of the expected mean edge length across the full range of sentence lengths. It should be emphasized, however, that these models are fitted only to aggregated means. They are therefore intended as estimators of the expected value, not as descriptions of the full variability of the data. The underlying data clouds are noticeably dispersed, which is expected, and the fitted curves cannot account for this spread.

To evaluate the behaviour of residuals, we examined the fitted curves against the aggregated means (the green curves in Figure 8). Even though aggregation was applied specifically to reduce heteroscedasticity, clear patterns remain, especially for large sentence lengths where the residuals tend to fan out. To understand these patterns more directly, Figure 9 plots the absolute residuals against the number of available sentences for each sentence length (across all languages). The plot shows a clear trend: sentence lengths with very few observations tend to have larger absolute residuals on average. This suggests that much of the remaining heteroscedasticity is driven by underrepresented, extreme sentence lengths. With larger datasets these fluctuations would likely diminish.

One possible strategy would have been to filter out sparsely represented sentence lengths to stabilize the residuals further. However, such filtering would remove information exactly in the regions where sentence lengths naturally become rare. Since the fitted curves already provide a satisfactory representation of the mean behavior, we retained all available aggregated data. It should nevertheless be noted that uncertainty is higher in the tails of the distribution, and conclusions in that region should be interpreted with this in mind.

3.3 Conclusions

The plots in Section 2.4.4 show that different models are preferred for different languages. Model 4 is the most frequently selected AIC-best model, being chosen for seven languages (Czech, English, Finnish, Japanese, Portuguese, Russian, and Spanish). Models 2 and 5 are each selected for four languages: Model 2 for Arabic, Chinese, Galician, and Italian, and Model 5 for Hindi, Icelandic, Indonesian, and Swedish. Model 1 is chosen for two languages (Korean and Turkish), while Model 1+ (French), Model 2+ (German), Model 4+ (Thai), and Model 5+ (Polish) each occur only once.

Some specific groupings can be related to the summary statistics using Figure 11. Korean and Turkish have very similar profiles, with relatively low mean sentence length, low mean edge length, and comparatively small sentence-length variability. Both are best described by Model 1. Thai stands out with the lowest overall mean edge length and the lowest standard deviation of dependency length, and it is the only language for which Model 4+ is selected. German, conversely, has the highest mean edge length and the second highest standard deviation of edge length, and it is uniquely assigned to Model 2+. Beyond these cases, however, we do not observe simple rules that map summary statistics directly to model choice. For example, Finnish has the smallest mean sentence length and smallest sentence-length standard deviation, whereas Japanese has the largest values for both, yet both are best modelled by Model 4 according to AIC. This suggests that the preferred model is not determined by any single summary statistic in isolation, but rather by more complex interactions between sentence length, dependency length, and their distributional shapes. Exploring these interactions in more detail would be a natural direction for future work on the relationship between model choice and corpus-level summary statistics.

4 Methods

4.1 Visual Test for Homoscedasticity

To verify whether the variance of the mean edge length $\langle d \rangle$ changes systematically with sentence length n, we computed the empirical variance $\mathrm{Var}(\langle d \rangle \mid n)$ for all sentences sharing the same n. If the dispersion of $\langle d \rangle$ remains constant across different sentence sizes, the data can be considered approximately homoscedastic. Otherwise, variance increasing or decreasing with n indicates heteroscedasticity.

Both variables were log-transformed and fitted using the model

$$\log_{10} \mathrm{Var}(\langle d \rangle \mid n) = \alpha + \beta \log_{10} n + \varepsilon,$$

weighted by the number of sentences per length to reduce the influence of rare sentence sizes. Only lengths represented by at least five sentences were retained to ensure stable variance estimates. A slope $\beta \approx 0$ supports the homoscedasticity assumption, while $\beta > 0$ indicates increasing variance with sentence length.

4.2 Initial values of the parameters

Reliable initial values are crucial for the convergence of non-linear least-squares estimation with the `nls()` function in R. To obtain consistent and data-driven starting points for the parameters, we exploited log- or semi-log transformations that linearize the functional forms of the non-plus models. All fittings were performed on the aggregated datasets (one observation per sentence length n).

Model 1: $f(n) = (n/2)^b$. Taking logarithms yields

$$\log\langle d \rangle = b \log(n/2),$$

which is a linear relationship without intercept. Hence, the slope of the regression `lm(log(mean_length) ~ 0 + log(vertices/2))` provides the initial value of b.

Model 2: $f(n) = an^b$. After taking logarithms,

$$\log\langle d \rangle = \log a + b \log n,$$

so the intercept and slope of the regression `lm(log(mean_length) ~ log(vertices))` yield the initial values $a_0 = e^{\text{intercept}}$ and $b_0 =$ slope.

Model 3: $f(n) = ae^{cn}$. The semi-logarithmic transformation

$$\log\langle d \rangle = \log a + cn$$

leads to a simple linear fit `lm(log(mean_length) ~ vertices)`, from which the initial values $a_0 = e^{\text{intercept}}$ and $c_0 =$ slope are obtained.

Model 4: $f(n) = a \log n$. This model is already linear in a and requires no logarithmic transformation: the slope of `lm(mean_length ~ log(vertices))` serves as the initial value a_0.

Model 5: $f(n) = an^b e^{cn}$. Taking logarithms gives

$$\log\langle d \rangle = \log a + b \log n + cn.$$

The parameters are initialized from the multiple regression `lm(log(mean_length) ~ log(vertices) + vertices)`, yielding $a_0 = e^{\text{intercept}}$, b_0 = coefficient of `log(vertices)`, and c_0 = coefficient of `vertices`.

"Plus" models. For the additive variants (Model 1+, 2+, 3+, 4+, and 5+), the parameters of the corresponding non-plus model were first estimated as described above. Their fitted values were then used as initial guesses for the same parameters in the "+" models, while the additive constant d was initialized to zero, i.e. `start = list(a = a_fit, b = b_fit, ..., d = 0)`. This procedure ensures that all models start from consistent and theoretically grounded points while allowing the optimizer to adjust d freely during fitting.

4.3 Nonlinear Model Fitting Algorithm

Before selecting the final models, we encountered practical difficulties when estimating the nonlinear parameters for some of the more flexible model variants. In particular, the plus models (Model 2+ and Model 3+) proved unstable in initial experiments. Despite using the same initial values that worked for the base models, both Model 2+, Model 3+ and Model 5+ consistently failed for several languages. We therefore experimented with different strategies for choosing the starting value of the offset parameter d. This included both setting d to 0 and computing a more informed initial estimate by first fitting the corresponding base model, then using the residuals to derive a plausible starting value for d. While this improved the situation for most languages, convergence problems remained, for example, Model 2+ for Swedish and Model 5+ in general continued to fail under `nls`, even with substantial variation in the initial values during manual grid searches.

Because these issues arose from the sensitivity of the Gauss–Newton algorithm used by `nls`, we switched to `nlsLM` from the `minpack.lm` package. The Levenberg–Marquardt algorithm implemented in `nlsLM` is known to be more robust, particularly in the presence of complicated curvature or when parameters may initially lie far from a good solution. Importantly, `nlsLM` performs the same nonlinear least-squares estimation as `nls`, the improvement stems solely from increased numerical stability. After switching to `nlsLM` and using $d = 0$ as a unified and simple initial value for all plus models, all remaining convergence problems were resolved. This allowed us to estimate the full set of model parameters consistently across all languages, ensuring that the subsequent model comparison via AIC and BIC was based on reliable fits.

4.4 Choice of the best model

After fitting the eleven models, the selection of the best-performing model was primarily based on the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) values, with the

objective of identifying the model that yielded the lowest values for both metrics.

- **Consistent Results:** If the lowest AIC value corresponded to the lowest BIC value, that model was unequivocally selected as the best one for the language.
- **Conflicting Results:** In cases where AIC and BIC suggested contrasting best models (i.e., Model A had the lowest AIC, but Model B had the lowest BIC), we made a decision based on a visual assessment