CSN - Second Lab

Kymry Burwell, Laura Cebollero

October 22nd, 2018

Introduction

The aim of this lab project is to analyze a degree distribution and select a theoretic model that best fits it. There are three sequences of which we can work on:

1. Undirected degree sequence.
2. In-degree sequence.
3. Out-degree sequence.

In our case, we have chosen to work with the out-degree one for 10 different languages. The distributions we will be testing are the following:

- Poisson distribution (λ parameter)
- Geometric distribution (q parameter)
- Zeta distribution (γ parameter)
- Zeta distribution ($\gamma = 2$ parameter)
- Right truncated distribution (γ and k_{max} parameters)
- Altmann distribution (δ, and γ parameters)

The first step on the analysis is to compute different metrics for each language. Additionally, to make our computations easier we have added a couple of metrics that we were not required, namely MP and C. The metrics of the English out-degree sequence can be seen in the following table.

Table 1: Summary Table of English Out-degree Metrics

Language	N	M	Maximum Degree	M/N	N/M	MP	C
Arabic	15678	70589	4896	4.502424	0.2221026	12530.413	165907.83
Basque	6188	25876	2097	4.181642	0.2391405	4231.383	54154.09
Catalan	24727	204095	6622	8.253933	0.1211544	29926.062	561322.53
Chinese	23946	185013	7537	7.726259	0.1294287	24832.108	549519.06
Czech	41912	262218	12671	6.256394	0.1598365	41038.656	721024.15
English	17775	200041	7040	11.254065	0.0888568	23919.120	657764.54
Greek	9280	44768	2737	4.824138	0.2072909	8938.332	91074.93
Hungarian	25534	107178	1020	4.197462	0.2382392	21493.722	177186.08
Italian	12285	56829	1671	4.625885	0.2161748	11701.853	104228.03
Turkish	15287	47186	4488	3.086675	0.3239732	8162.505	108443.77

In the table 1 above, N represents the number of nodes in the network, M is the sum of degrees of all nodes, Maximum Degree is the largest out-degree, M/N is the average degree, N/M is the inverse of the average degree, MP is the sum of the log of degrees, and C is the the following $\sum_{i=1}^{N} \sum_{j=2}^{k_i} log(j)$.

Next, we will look at a few bar plots for the English langauge to get a visual idea of the degree distribution. We can see from figures 1 and 2 below that nodes with small out-degree are more common than nodes with high out-degree.

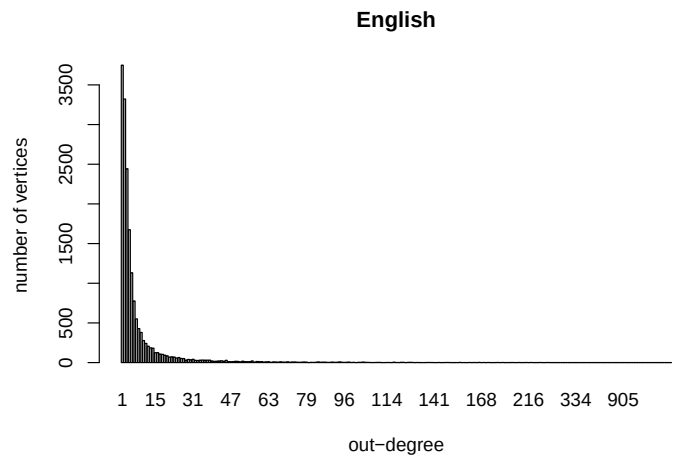

Figure 1: English out-degree distribution

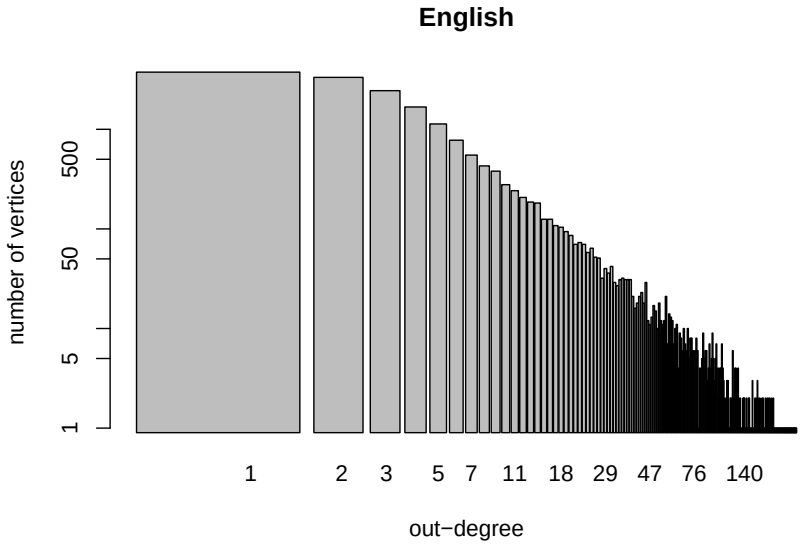

Figure 2: English out-degree distribution (loglog scale)

Results (Without Altmann)

Having computed the basic metrics, we now proceed to compute the most likely parameters for the various distributions. To do this, we are trying to find the parameters that minimize the minus log-likelihood function. To help expedite the process, we begin with default parameters, which act as our best initial guess. These consist of the following:

- $\lambda_0 = \mathrm{M}/\mathrm{N}$
- $q_0 = \mathrm{N}/\mathrm{M}$
- $\gamma_0 = 2$
- $k_{max,0} = \mathrm{N}$

Using the out-degree sequences for each different language, we have obtained the following coeffcients (Note that we have not included the coefficient for Zeta distribution with $\gamma = 2$ in any of the coefficient table in this report):

Table 2: Summary of the most likely parameters

Language	λ	q	γ	γ_2	k_{max}
Arabic	4.450	0.222	1.798	1.795	15678
Basque	4.113	0.239	1.887	1.885	6188
Catalan	8.252	0.121	1.591	1.583	24727
Chinese	7.723	0.129	1.663	1.658	23946
Czech	6.244	0.160	1.691	1.688	41912
English	11.254	0.089	1.545	1.532	17775
Greek	4.784	0.207	1.699	1.693	9280
Hungarian	4.130	0.238	1.769	1.767	25534
Italian	4.578	0.216	1.705	1.700	12285
Turkish	2.920	0.324	2.043	2.042	15287

Before proceeding onto computing the delta AIC and selecting the best model, we first check that our methods have been correctly addressed and implented. In order to verify the methods we have used to compute the minus log-likelihood parameters, we will test them on contrived data sets where the distribution is known a priori. We have 8 different data sets that were created using geometric and zeta distributions of varying parameters.

Geometric test

For the different instances of the geometric test data, we have computed the coefficients and delta AIC tables:

Table 3: Summary of the most likely parameters for geometric test sequence

Test Data	λ	q	γ	γ_2	k_{max}
probability_0.05	19.242	0.052	1.332	1.155	1000
probability_0.1	10.298	0.097	1.419	1.312	1000
probability_0.2	4.984	0.199	1.572	1.524	1000
probability_0.4	2.213	0.402	1.891	1.882	1000
probability_0.8	1.000	0.790	2.999	2.999	1000

Table 4: ΔAIC for Geometric test model

Test Data	Poisson	Geometric	Zeta	Zeta(γ=2)	RT Zeta
probability_0.05	11537.515	0	1334.446	3082.141	1036.041
probability_0.1	5180.882	0	957.304	1939.303	831.689
probability_0.2	1376.431	0	686.564	1040.700	654.660
probability_0.4	254.848	0	349.163	359.490	348.379
probability_0.8	209.342	0	56.200	351.582	58.208

As we can see, in this delta AIC table, the preferred and selected method is the geometric one. So in this case, the test is passed.

Zeta test

Now we proceed to test with data sets consisting of Zeta distributions.

Table 5: Summary of the most likely parameters for Zeta test sequence

Test Data	λ	q	γ	γ_2	k_{max}
exponent_2	5.044	0.197	1.980	1.975	1000
exponent_2.5	1.368	0.545	2.451	2.450	1000
exponent_3	1.000	0.733	2.996	2.996	1000
exponent_3.5	1.000	0.786	3.354	3.354	1000

Table 6: ΔAIC for the Zeta test model

Test Data	Poisson	Geometric	Zeta	Zeta(γ=2)	RT Zeta
exponent_2	16110.133	1693.604	1.651	0.000	2.224
exponent_2.5	1481.541	408.878	0.000	100.043	1.963
exponent_3	652.243	224.065	0.000	294.312	2.007
exponent_3.5	800.014	275.808	0.000	410.666	2.008

Again, it appears that our methods and implementation are correct as the appropriate Zeta distributions were chosen.

Delta AIC of our models

Having obtained the parameters and checked our methods, we can now proceed to obtain the -2 log Likelihood for each model and compute the AIC for the real cases. Once computed, we can produce the delta AIC table by substracting the best AIC of each Language from the other methods' AIC. The resulting table is the following:

Table 7: ΔAIC for various models

Test Data	Poisson	Geometric	Zeta	Zeta(γ=2)	RT Zeta
Arabic	195283.11	9828.508	7.494	792.354	0.000
Basque	62820.67	5467.463	1.456	82.421	0.000
Catalan	532711.68	14163.356	93.626	7767.499	0.000
Chinese	593680.93	23773.076	41.521	4364.603	0.000
Czech	804878.45	30600.485	36.051	6035.890	0.000
English	641485.32	14343.475	134.968	7732.678	0.000
Greek	86904.59	1962.453	20.398	1256.834	0.000
Hungarian	150811.66	8063.318	12.138	1762.370	0.000
Italian	90326.61	1878.663	20.941	1580.394	0.000
Turkish	155497.43	11594.540	0.000	21.107	1.207

We can see that, in the case of the out-degree sequence, the method that fits best is almost exclusevily the Right-Truncated Zeta, except for the Turkish language, which seems to fit better a Zeta distribution.

Plots of real data vs best-fit distribution

To visually verify this model selection, we are now going to check how the real data aligns with the distribution itself. To do so, we are going to work with three very distinct languages:

- **English**, which uses many Greek and Latin roots.
- **Chinese**, whose procedence is totally unrelated with no roots to Greek and Latin whatsoever.

- **Basque**, whose procedence is unknown and is also unrelated to Greek and Latin.

The blue dots are the actual data and the green line is the distribution (RT-Zeta in this case).

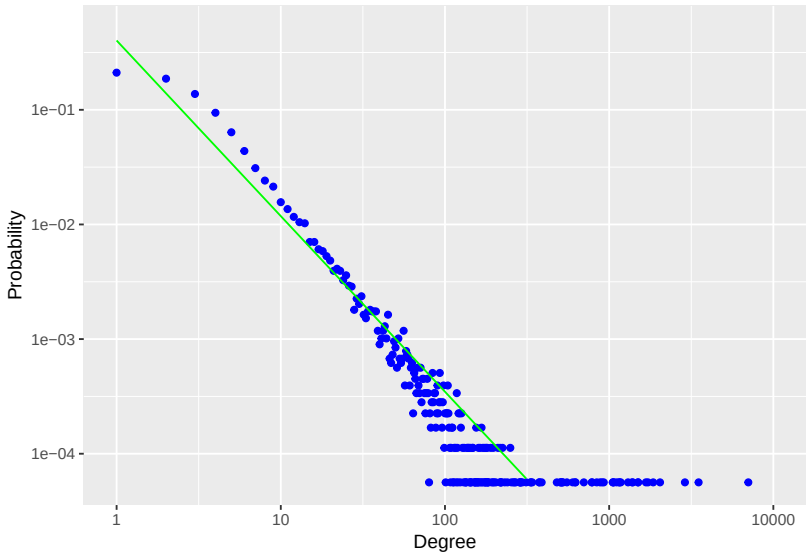

Figure 3: RT Zeta Distribution Comparison with English

As we can see, the Right truncated Zeta distribution fits well with the data's observations for the coefficient of gamma obtained in the case of English.

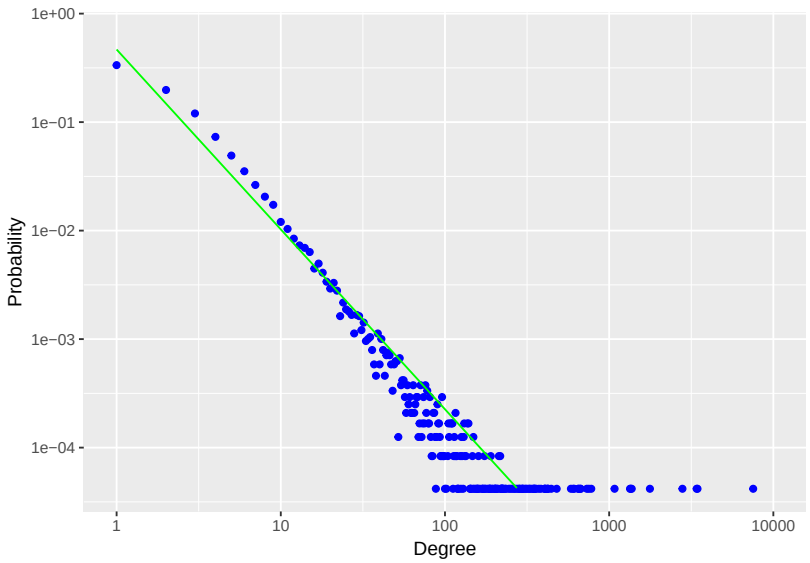

Figure 4: RT Zeta Distribution Comparison with Chinese

As we can see, the Right truncated Zeta distribution fits well with the data's observations for the coefficient of gamma obtained in the case of Chinese as well.

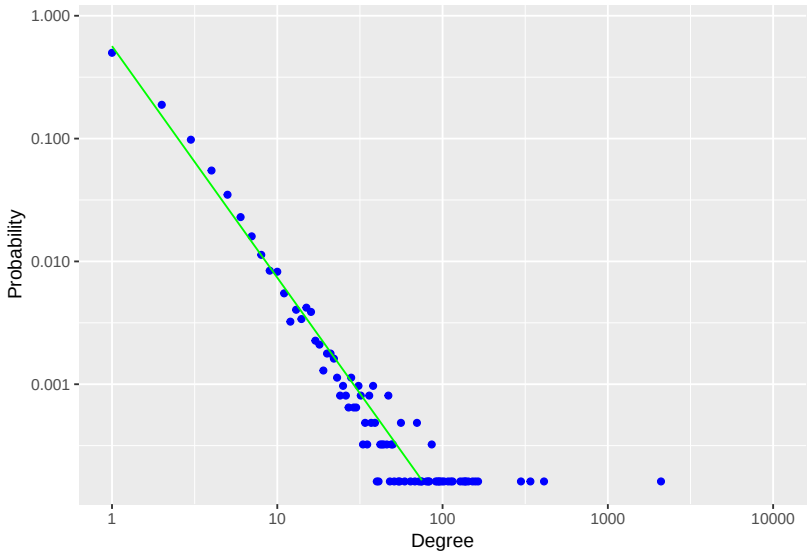

Figure 5: RT Zeta Distribution Comparison with Basque

It also seems to fit the Basque language.

So, as we can see, the Right-Truncated Zeta distribution fits quite well, although not perfectly, with the three languages.

Below, we plot the Basque language against the Geometric distribution, in order to exemplify a bad case and to show how poor of a fit it is:

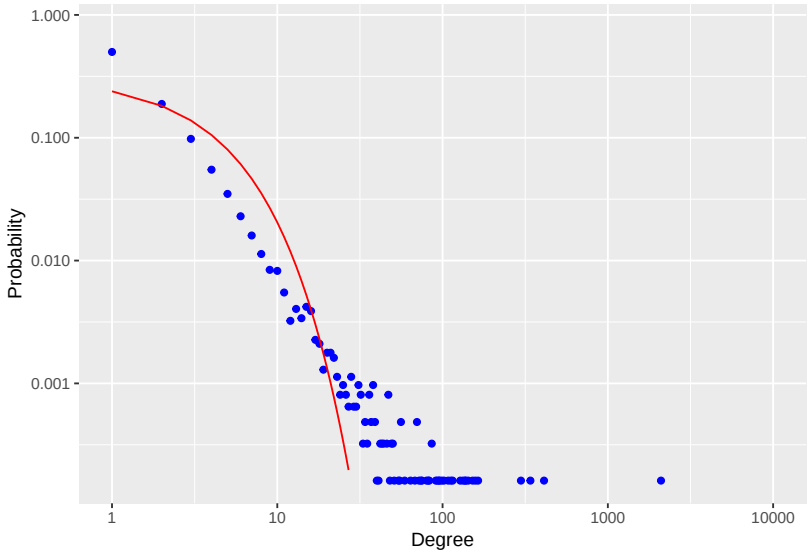

Figure 6: Geometric Distribution Comparison with English

Clearly, the data does not fit the geometric distribution.

Results (With Altmann)

Next, we add the Altmann function to our set of distributions, to see if it will provide a better fit of our data. The two added parameters, $\gamma 3$ and δ are for the Altman funcion.

Table 8: Summary of the most likely parameters

Test Data	λ	q	γ	γ_2	k_{max}	$\gamma 3$	δ
Arabic	4.450	0.222	1.798	1.795	15678	1.555	0.022
Basque	4.113	0.239	1.887	1.885	6188	1.763	0.011
Catalan	8.252	0.121	1.591	1.583	24727	1.249	0.019
Chinese	7.723	0.129	1.663	1.658	23946	1.467	0.010
Czech	6.244	0.160	1.691	1.688	41912	1.439	0.017
English	11.254	0.089	1.545	1.532	17775	1.256	0.011
Greek	4.784	0.207	1.699	1.693	9280	1.196	0.053
Hungarian	4.130	0.238	1.769	1.767	25534	1.353	0.048
Italian	4.578	0.216	1.705	1.700	12285	1.156	0.061
Turkish	2.920	0.324	2.043	2.042	15287	1.950	0.011

Table 9: ΔAIC for various models

Test Data	Poisson	Geometric	Zeta	Zeta(γ=2)	RT Zeta	Altmann
Arabic	196029.16	10574.560	753.546	1538.406	746.052	0
Basque	62921.60	5568.399	102.392	183.357	100.936	0
Catalan	535932.72	17384.395	3314.665	10988.538	3221.039	0
Chinese	595006.42	25098.563	1367.008	5690.090	1325.487	0
Czech	807747.40	33469.430	2904.996	8904.835	2868.945	0
English	643673.24	16531.387	2322.880	9920.590	2187.912	0
Greek	88224.23	3282.089	1340.034	2576.470	1319.636	0
Hungarian	153247.32	10498.974	2447.794	4198.026	2435.656	0
Italian	92220.50	3772.553	1914.831	3474.284	1893.890	0
Turkish	155611.67	11708.776	114.236	135.343	115.443	0

We can see that the Altmann function is now chosen as the best model for our data. Finally, to visually confirm these, we plot the Altmann function against the English language, using the same type of plot as before.

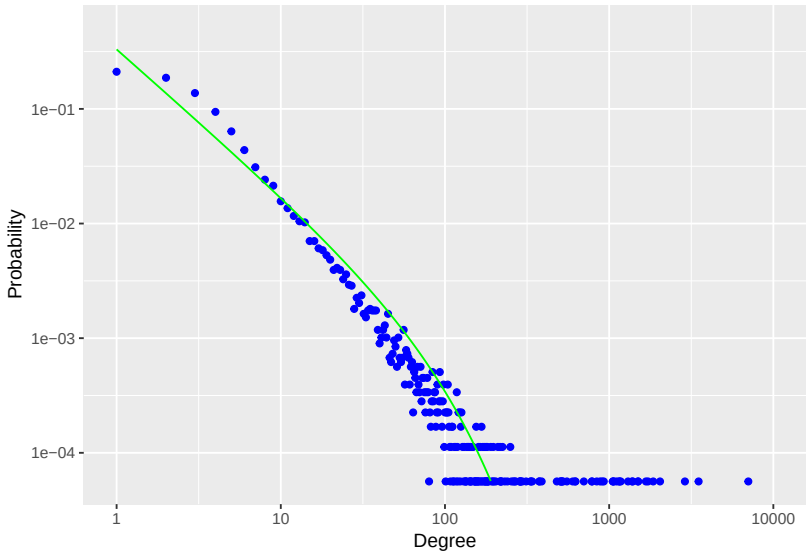

Figure 7: Altmann Distribution Comparison with English

We can also plot it with the Basque language:

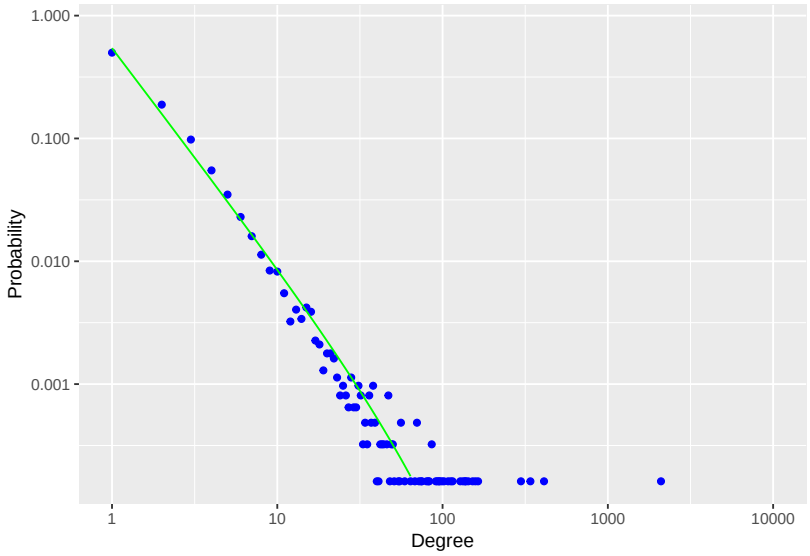

Figure 8: Altmann Distribution Comparison with Basques

Again, we can see that it fits well the data.

Discussion

Table 1 displays some useful metrics of the out-degree distributios of 10 different languages. A few items stand out: Czech has a maximum degree that nearly doubles every other language, Greek and Basque have far fewer nodes than the rest, and the average degree of English is quite high relative to the other languages. This table gives us some good insight into how the lanugages differ form a high level perspective.

Figures 2 and 3 give a visual representation of the English language out-degree distribution. Figure one really shows that the node degree is concentrated around the smallers numbers, with a very long tail. This long tail makes sense, when we see that the maximum degree of English is over 7,000 and the average degree is 11.

Table 2 shows the best fit parameters we found from minimizing the minus log-likelihood. By comparing Tables 1 and 2, we can see that theses parameters are quite close to our initial guess. In the English language for example, our initial guess for the parameters were $\lambda = 11.25$, $q = 0.088$, $\gamma = 2$ and $kmax = 17,775$ and the best fit parameters were $\lambda = 11.254$, $q = 0.089$, $\lambda = 1.545$, and $kmax = 17,775$.

Tables 3 and 4 are the results of our Geometric test. Table 3 displays the best fit parameters of 5 data sets consting of geometric distributions with a varying q parameter. We can see that the q parameters our methods chose, were nearly identical to the actual parameters. Table 4 shows delta AIC for the same data sets. We can see that the geometric distribution was chosen correctly for each.

Tables 5 and 6 are the results of our Zeta distribution test. Table 5 displays the best fit parameters of 5 data sets consisting of Zeta distributions. It's notable that our methods correctly chose the Zeta distribution with the fixed $\gamma = 2$ for the data set with the same parameter. Table 6 is the delta AIC table confirming that Zeta distribution was chosen as best.

Table 7 is the delta AIC table for the out-degree sequence of the 10 languages. We can see that RT Zeta was chosen in all cases except for Turkish. The Zeta distribution came in a very close second for Arabic and Basque. It's also worth noting that, based on the large values present in this table, Poisson and Geometric distributions should not be used to model out-degree distributions (At least with the langauges studied here). Looking at the large values for most of the languages in the Zeta model with $\gamma = 2$, we can see that the parameter selection is extermely important.

Figures 3, 4, and 5 provide a visual comparison of the RT Zeta distribution (green line) to the actual out-degree data (blue dots) of the English, Chinese, and Basque languages, respectively. We can see that the fit is quite good in all cases, but particulary good for the Basque language. Please note that the y-axis was limited to minimum probability in the actual data. These three plots allow us to visually confirm that the distribution of each network differs slightly, as is the case with many real-world networks, meaning it a best fit distribution for one network isn't necessarily best for all networks. Additionally, we can see that the tails of the three languages differs, with the Chinese and English having slightly longer tails than Basque.

Figure 6 compares the English out-degree network to a geometric distribution. We did this to show just how poor of a fit it provides and to emphasize the importance of determing the correct (or best fit) distribution of the network.

Tables 8 and 9 are the coefficient and delta AIC tables with the Altmann function included. $\gamma 3$ and δ are the new parameters for the Altmann fnction. Looking at table 9, we can see that the Altman function was chosen as the best fit for every language. It's worth noting that Zeta was a fairly close second for both Turkish and Basque, but not the other languages. Figures 7 and 8 are another visual comparison of the Altmann function (green line) with the English and Basque. Comparing with figures 3 and 5, we can see that the Altmann function provides a slightly better fit.

Methods

Our work included, among other things:

- Computing basic metrics as well as C, for example, in order to use it aftwerwards.
- Compute the log likelihood of choosing a method minimising each distribution function with the `mle` function.
- On the same line, we used the suggested `L-BFGS-B` method as the solver of the `mle` function. This method gave us lots of trouble for if there was a slight error on the passed function to the method, it would complain about trying to solve a singular system, which was not possible.
- To compute the log likelihood of the Altmann function, we had to apply a logarithmic scale to the whole function. Otherwise, the `mle` method could not solve it due to the reason explained in the previous point.
- Graphics were used to check visually the how the data fitted the theoretical distributions.