Lab 4: Non-linear regression on dependency trees

Rodrigo Arias Mallo

November 17, 2017

1 Introduction

A list of datasets with information of syntactic dependency trees are used in this report to derive conclusions about the relation of a metric and the size of the sentence n.

1.1 Selection of the metric

For selecting the metric, I built a small python program, that receives as input a list of elements, and a number. The program concatenate the list and the number by using `;` as separator, and a MD5 hash is computed over the resulting string. Finally, an element of the list is chosen by using the first byte of the hexadecimal hash, converted to a index modulus the size of the list. Then the item with the corresponding index is selected.

The last digit of my identification number is used as the input number, so the output is unique for me.

```
% python choice.py 64718
degree 2nd moment
```

Finally the chosen metric is the degree 2nd moment $\langle k^2 \rangle$.

2 Results

All the analysis has been made in python, by using numeric and statistical packages. The datasets were read and the test

$$4 - 6/n \leq \langle k^2 \rangle \leq n - 1$$

was failed in some elements, because the rounding errors. After allowing a small error $\epsilon = 5 \times 10^{-6}$ the test

$$4 - 6/n - \epsilon \leq \langle k^2 \rangle \leq n - 1 + \epsilon$$

was succesfully passed in all languages.

2.1 Summary

In the table 1 a summary of the properties of the degree sequences is shown. The sample mean and standard deviation of the metric x are represented by $\overline{x}$ and s_x respectively.

Language	N	$\overline{n}$	s_n	$\overline{x}$	s_x
Arabic	4108	26.958	20.647	4.160	1.275
Basque	2933	11.335	6.527	4.143	1.089
Catalan	15053	25.572	13.618	4.962	0.824
Chinese	54238	6.249	3.310	3.218	1.066
Czech	25037	16.428	10.721	4.293	1.299
English	18779	24.046	11.223	5.170	0.802
Greek	2951	22.820	14.379	4.600	1.070
Hungarian	6424	21.660	12.565	5.956	1.706
Italian	4144	18.407	13.344	4.340	1.170
Turkish	6030	11.102	8.281	3.759	0.934

Table 1: The measures of the datasets.

2.2 Models

The models tested are presented in the table 2. The model 0 is used as reference, and has no parameters.

Model	Function	Parameters
0	$f(n) = (1 - 1/n)(5 - 6/n)$	
1	$f(n) = (n/2)^b$	b
2	$f(n) = an^b$	a, b
3	$f(n) = ae^{cn}$	a, c
4	$f(n) = a \log n$	a
5	$f(n) = an^b e^{cn}$	a, b, c
1+	$f(n) = (n/2)^b + d$	b, d
2+	$f(n) = an^b + d$	a, b, d
3+	$f(n) = ae^{cn} + d$	a, c, d
4+	$f(n) = a \log n + d$	a, d
5+	$f(n) = an^b e^{cn} + d$	a, b, c, d

Table 2: The list of models to test.

2.3 Non-linear regression

In the table 3 the difference AIC metric is shown, with respect to the best model in each case. Note that the metric can be affected by the value of the outliers if we use the aggregate mean, so the full dataset is used to get the best parameters.

Language	0	1	2	3	4	5
Arabic	210.1	6276.8	2446.3	5436.0	2290.0	1426.9
Basque	2003.9	5476.5	1626.6	3417.0	1143.0	600.7
Catalan	12 644.6	26 268.4	3326.8	8266.6	7741.9	1329.2
Chinese	11 896.5	85 699.6	30 496.7	74 717.7	10 359.2	8129.4
Czech	6340.2	25 546.9	5039.1	15 949.9	3661.5	4014.9
English	18 717.1	28 508.2	2309.5	5793.3	7775.1	1207.6
Greek	1673.5	5279.4	1450.1	3211.2	1596.4	713.5
Hungarian	6888.6	4987.7	258.3	1271.2	138.3	155.4
Italian	1736.1	7250.8	1995.7	4642.6	1860.2	936.9
Turkish	920.2	14 390.5	4259.5	7868.1	5634.7	1773.2

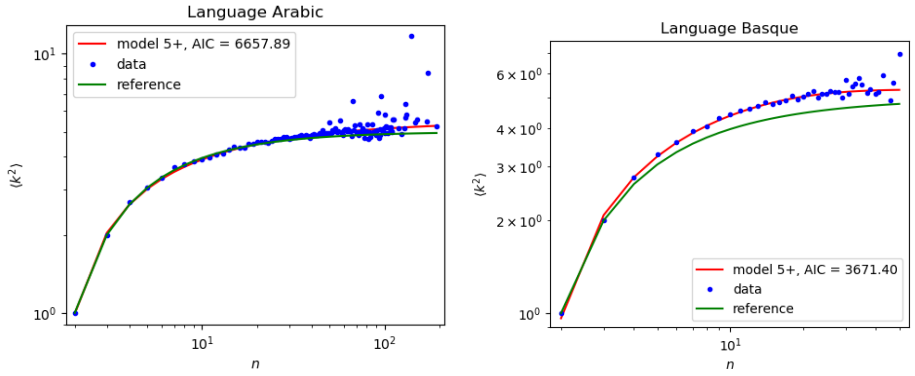
Language	1+	2+	3+	4+	5+
Arabic	3053.3	1.0	392.9	1353.7	0.0
Basque	2095.3	14.2	1 457 178.6	804.9	0.0
Catalan	4427.3	0.0	737.0	2134.3	1.0
Chinese	38 903.3	816.2	26 797 895.7	10 347.4	0.0
Czech	7082.3	13.1	1349.5	1859.7	0.0
English	3075.1	1.2	619.7	1590.0	0.0
Greek	1878.8	0.4	278.1	813.4	0.0
Hungarian	467.4	4.1	54.2	52.8	0.0
Italian	2602.5	0.0	239.9	1022.1	1.9
Turkish	5139.7	38.8	3 717 244.6	2851.8	0.0

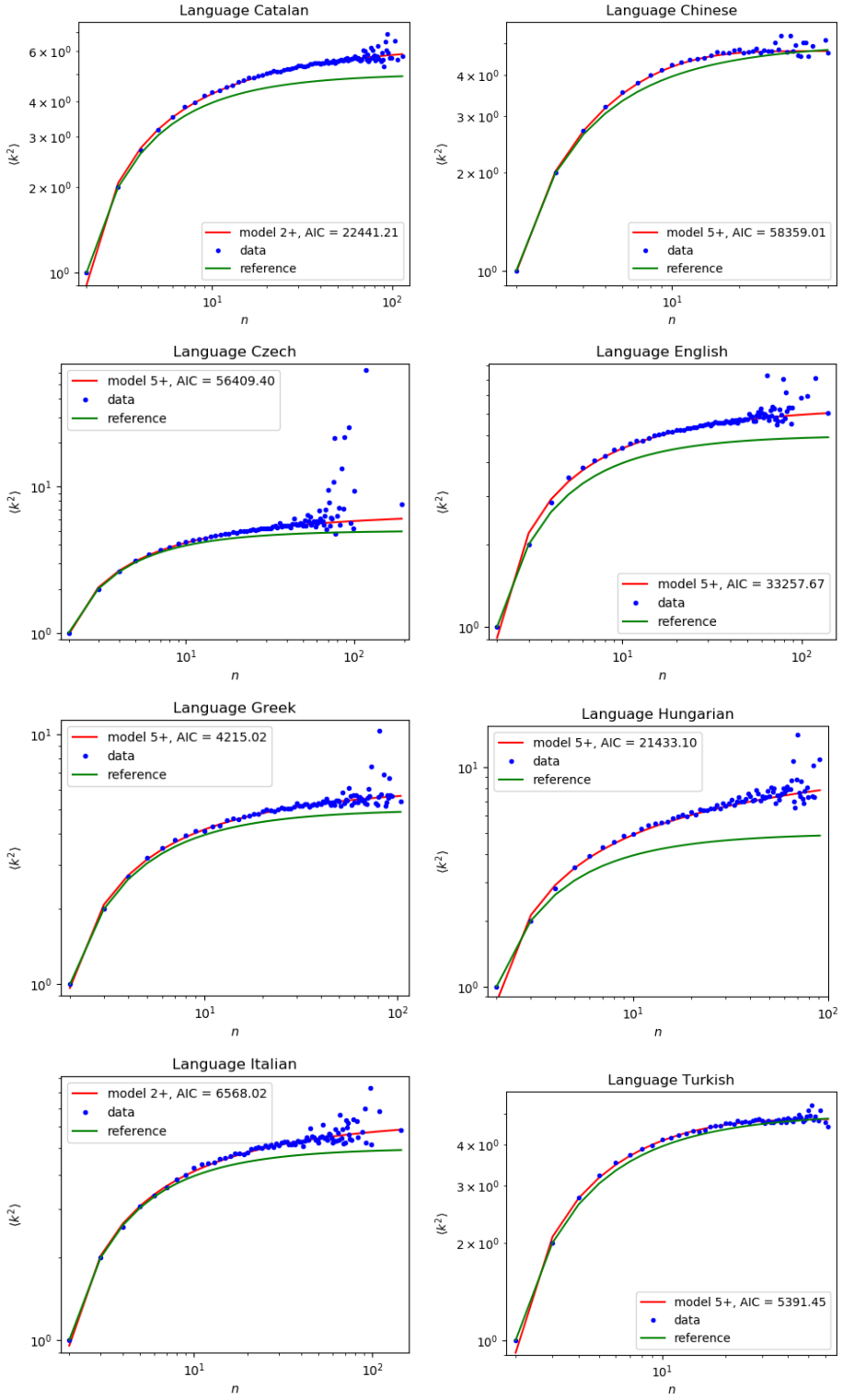
Table 3: The ΔAIC of the models.

We see that the model 5+ seems to better fit the data, followed by the 2+ model.

3 Plots of the models

The best model has been selected to be plotted along with the reference model 0, and the aggregate mean data points. The logarithmic scale is enabled in both axis.




Language Catalan
model 2+, AIC = 22441.21
data
reference
Language Chinese
model 5+, AIC = 58359.01
data
reference
Language Czech
model 5+, AIC = 56409.40
data
reference
Language English
model 5+, AIC = 33257.67
data
reference
Language Greek
model 5+, AIC = 4215.02
data
reference
Language Hungarian
model 5+, AIC = 21433.10
data
reference
Language Italian
model 2+, AIC = 6568.02
data
reference
Language Turkish
model 5+, AIC = 5391.45
data
reference
$\langle k^2 \rangle$
n

3.1 Model parameters

The best model parameters are shown in the table 4. Each column is one parameter of the model, named $m\,p$ where m is the model name and p the parameter name.

Language	1 b	2 a	2 b	3 a	3 c	4 a	5 a	5 b	5 c
Arabic	0.535	2.003	0.243	3.605	0.006	1.385	1.326	0.464	−0.010
Basque	0.767	1.887	0.340	3.236	0.022	1.813	1.106	0.741	−0.034
Catalan	0.604	2.730	0.191	4.229	0.006	1.581	1.875	0.376	−0.008
Chinese	0.932	1.267	0.526	2.208	0.059	1.876	0.734	1.159	−0.088
Czech	0.659	1.950	0.298	3.439	0.014	1.641	1.280	0.564	−0.017
English	0.635	2.954	0.179	4.396	0.006	1.658	2.132	0.345	−0.008
Greek	0.603	2.316	0.233	3.893	0.008	1.558	1.517	0.459	−0.011
Hungarian	0.712	2.503	0.285	4.582	0.011	1.996	1.782	0.464	−0.008
Italian	0.627	2.116	0.264	3.646	0.010	1.607	1.389	0.509	−0.013
Turkish	0.695	2.044	0.273	3.167	0.016	1.690	1.320	0.597	−0.025

Language	1+ b	1+ d	2+ a	2+ b	2+ d	3+ a	3+ c	3+ d
Arabic	0.394	1.609	−7.046	−0.647	5.493	−5.193	−0.179	4.839
Basque	0.550	1.656	−8.393	−0.616	6.386	5.000	5.000	5.000
Catalan	0.372	2.456	−8.227	−0.591	6.367	−4.571	−0.130	5.434
Chinese	0.737	0.952	−8.086	−0.572	6.383	5.000	5.000	5.000
Czech	0.483	1.666	−7.956	−0.525	6.506	−5.389	−0.169	5.191
English	0.372	2.685	−8.594	−0.664	6.342	−4.741	−0.147	5.536
Greek	0.409	2.042	−7.831	−0.594	6.157	−5.291	−0.162	5.210
Hungarian	0.510	2.581	−11.970	−0.346	10.312	−5.949	−0.083	7.418
Italian	0.442	1.832	−7.905	−0.546	6.370	−5.462	−0.170	5.187
Turkish	0.463	1.680	−7.846	−0.831	5.251	5.000	5.000	5.000

Language	4+ a	4+ d	5+ a	5+ b	5+ c	5+ d
Arabic	1.075	1.002	−7.012	−0.656	−0.002	5.429
Basque	1.535	0.663	−7.427	−0.515	−0.087	5.325
Catalan	0.979	1.921	−7.874	−0.621	−0.015	5.879
Chinese	1.870	0.010	−6.861	−0.344	−0.183	4.743
Czech	1.343	0.816	−7.859	−0.538	−0.002	6.370
English	0.945	2.242	−8.543	−0.676	−0.003	6.231
Greek	1.141	1.280	−7.767	−0.606	−0.002	6.049
Hungarian	1.689	0.910	−10.962	−0.403	−0.004	9.074
Italian	1.246	1.019	−7.355	−0.581	−0.022	5.664
Turkish	1.170	1.207	−7.338	−0.688	−0.079	4.801

Table 4: The models parameters.

4 Methods

The proposed R function `nls` is the abreviature of Nonlinear Least-Squares. The basic procedure is to modify the parameters of the model in order to reduce the sum of the square distance between the data points and the predicted points of the model. This function is available in the python package `scipy.optimize` as `least_squares`. However the wrap function `curve_fit` let us use it more easily, and is what has been used in order to fit the models.

The initial parameters are set by default to 1. Except in models 2+ and 3+ which had been modified a bit, but with no luck. The default parameters are defined in the model function:

```
def m2d(n, a=0.06, b=0.8, d=1):   return a*n**b + d
def m3d(n, a=-1, c=-0.05, d=13):  return a*np.exp(c*n) + d
```

Some models can't find a optimum value and a exception is returned. In this case a second attempt is made, in order to find better parameters for the model. The function `differential_evolution` let us use an evolutionary algorithm to find values where the least-squares fails. It takes a bit more, but almost always improves the parameters. For more details see `fit.py`.

5 Discussion

We see that the reference model 0 is not a very good predictor of the data. As n increases the distance from the mean and the predicted value is bigger. The best model seems to be the model 5+, which is the one with the best AIC overall. Homoscedasticity was not tested in any way, only visually looked. By looking at all the models the selected one matches with the one that graphically appears to better fit the data. We conclude that the mean metric can be successfully predicted by the nonlinear regression model 5+.