Vowel-pair rank–frequency distributions are polylogarithmic

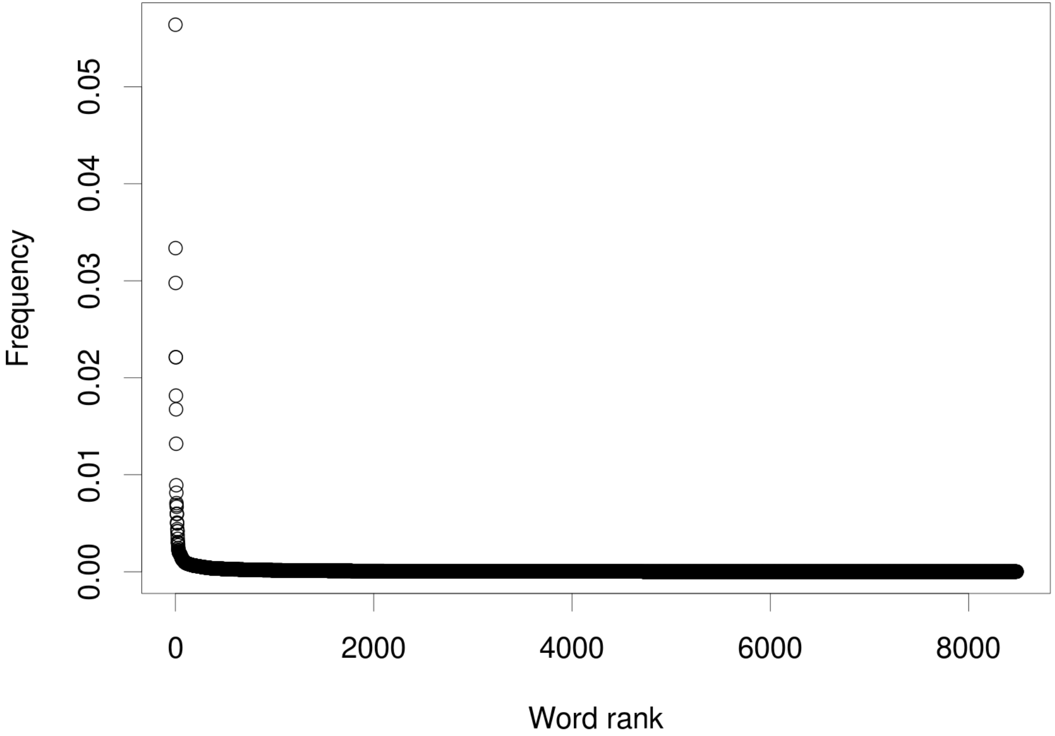
/vaʊ̯əl pɛə̯ ɹænk fɹiːkwənsi dɪstɹɪbjuːʃənz ɑː pɒlilɒgəɹɪðmɪk/

Stephen Nichols & Henri Kauhanen

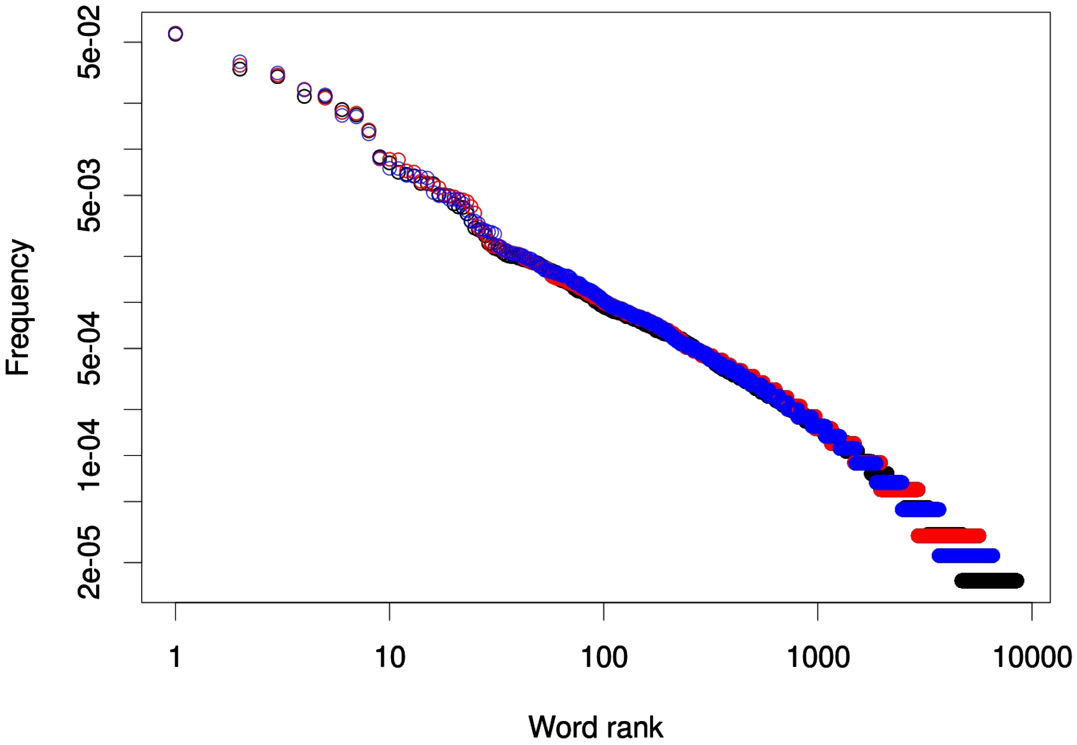
University of Manchester

LAGB, 11 September 2019
Queen Mary University of London

Introduction: Rank-frequency distributions in natural language



Introduction: Rank-frequency distributions in natural language



Introduction: Rank–frequency distributions in natural language

Mathematically, Zipf's Law is (Zipf 1949):

(1) $$f(r) = ar^{-b}$$

r – word's rank

$f(r)$ – relative frequency in corpus

a – normalisation constant

b – scaling parameter

This is a ***power law*** – one of many ***long-tailed*** distributions (see e.g. Newman 2005).

Introduction: Rank-frequency distributions in natural language

Phonemes follow a similar curve (Martindale et al. 1996, Tambovtsev & Martindale 2007):

(2) $$f(r) = ar^{-b}c^{r}$$

r – phoneme's rank

f(*r*) – relative frequency in the lexicon

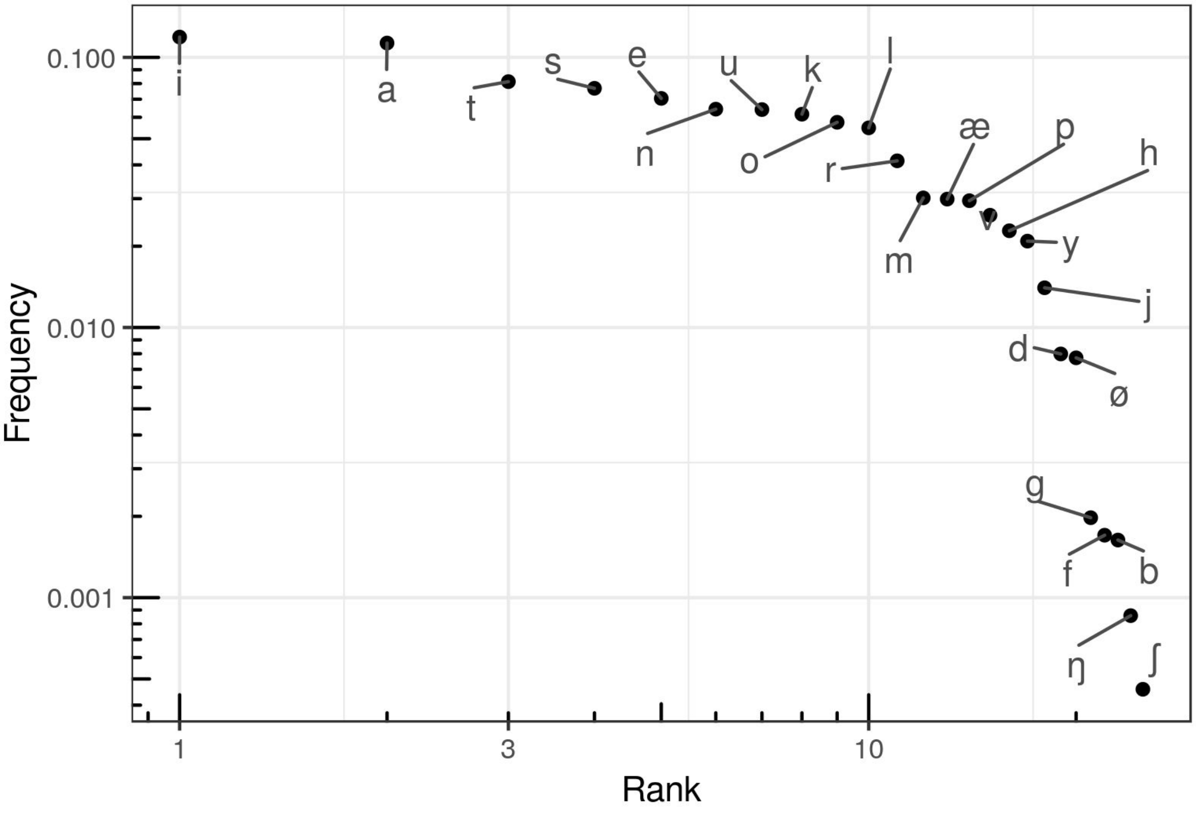
a – normalisation constant

b and *c* – scaling and shape parameters

This is a ***polylogarithmic distribution*** (Kemp 1995: 110).

Note that this directly generalises Zipf's Law (1) by the addition of the $\boldsymbol{c^r}$ factor, which introduces an ***exponential cut-off***.

Introduction: Rank-frequency distributions in natural language



‹tangent›

Martindale et al. (1996), Martindale & Konopka (1996), Tambovtsev & Martindale (2007) and, hence, many subsequent papers call this a ***Yule distribution***.

We follow Kemp (1995: 110) and use ***polylogarithmic distribution*** in order to avoid confusion with the ***Yule–Simon distribution***.

This is different to (2) but also often referred to as the ***Yule distribution*** (e.g. Yule 1924, Simon 1955, Chung & Cox 1994, Newman 2005).

Yet others (e.g. Zörnig & Altmann 1995, Eeg-Olofsson 2008, Klar et al. 2010) dub this the ***Good distribution*** after Good (1953).

‹/tangent›

Introduction: Our questions

If the rank–frequency distribution of phonemes in language is described by (2):

- Does this hold for ***dependencies***, i.e. combinations of phonemes?
- If not, can the deviations be explained?
- Would other theoretical distributions fit better?

In this talk, we limit ourselves to a consideration of ***vowel pairs***:

- 1σ, 0p: *dog* – /dɒg/
- 2σ, 1p: *spanner* – /sp**æ**n**ə**/
- 3σ, 2p: *Manchester* – /m**æ**ntʃ**ɛ**stə/, /mæntʃ**ɛ**st**ə**/

Introduction: Four distributions

	Zipf (Zipf 1949)	***Polylogarithmic*** (Simon 1955)	***Sigurd*** (Sigurd 1968)	***Borodovsky–Gusein-Zade*** (Borodovsky & Gusein-Zade 1989)
Formula	$f(r) = ar^{-b}$	$f(r) = ar^{-b}c^{-r}$	$f(r) = a(1-b)b^{r-1}/(1-b^{n})$	$f(r) = (a/n)\log[(n+1)/r]$
Parameters	1	2	1	0
Remarks	A plain power law; linear in log-log space	Power law with exponential cutoff	A geometric series: the ratio $f(r)/f(r+1)$ is constant (the parameter b)	Linear in semi-log space

N.B. (1) n = number of elements. (2) Normalisation constant a does not count as a parameter, as its value is determined as soon as the values of the other parameters are known from the requirement $\sum f(r) = 1$.

Methodology: Data sources

Language	Genus (family)	Macro-area	Source	Lemma count
Breton	Indo-European (Celtic)	Eurasia	Wiktionary	10,259
Finnish	Uralic (Finnic)	Eurasia	Kotus	93,087
Georgian	Kartvelian (Kartvelian)	Eurasia	Wiktionary	10,084
Italian	Indo-European (Italic)	Eurasia	phonItalia	42,127
Lozi	Niger–Congo (Bantu)	Africa	CBOLD	14,863
Malagasy	Austronesian (Barito)	Africa	Wiktionary	24,220
Northern Sami	Uralic (Saami)	Eurasia	Wiktionary	35,970
Serbo-Croatian	Indo-European (Slavic)	Eurasia	Wiktionary	23,624
Tagalog	Austronesian (Gr. Central Philippine)	Papunesia	Ispell	18,202

Methodology: Data extraction & processing

Pre-processing differed slightly according to the source of each data set.

- All data from Wiktionary were culled from XML data dumps.
- Kotus, phonItalia, CBOLD and Ispell came in the form of text files.

An R script was then used to find the ***observed frequency*** of each vowel pair.

- For a five-vowel language such as Serbo-Croatian, this yields 25 possible pairs.

Methodology: Data analysis

The polylogarithmic distribution (2) was fit to the data using non-linear least squares and normalisation to unity.

This procedure was repeated for the Zipf, Sigurd and BGZ distributions.

Model selection based on:

a. R^2 (regression on observed v. predicted frequencies)
b. **RSS** (residual sum of squares, i.e. goodness of fit)
c. **BIC** (Bayesian information criterion)

In order to estimate noise resulting from potential sampling biases, each lexicon was randomly sampled 100 times (bootstrapping).

Word of warning: fitting long-tailed distributions is fraught with difficulty (Clauset, Shalizi & Newman 2009); our approach may not necessarily be the best one, we leave refinements for future research.

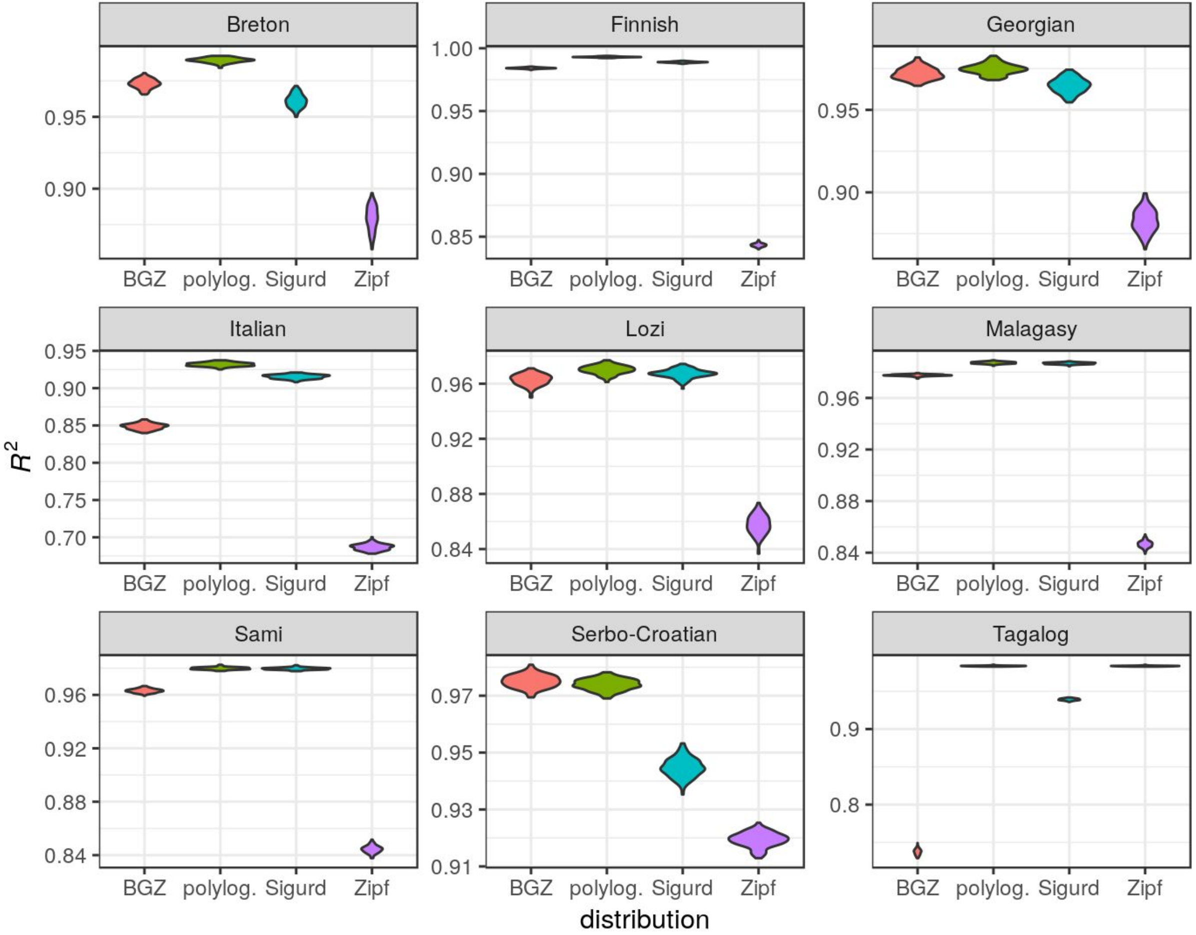
Results: R^2

Coefficient of determination for regression of observed v. predicted frequencies

Higher is better

Standard measure in previous literature…

… ***but***: does not penalise model complexity!



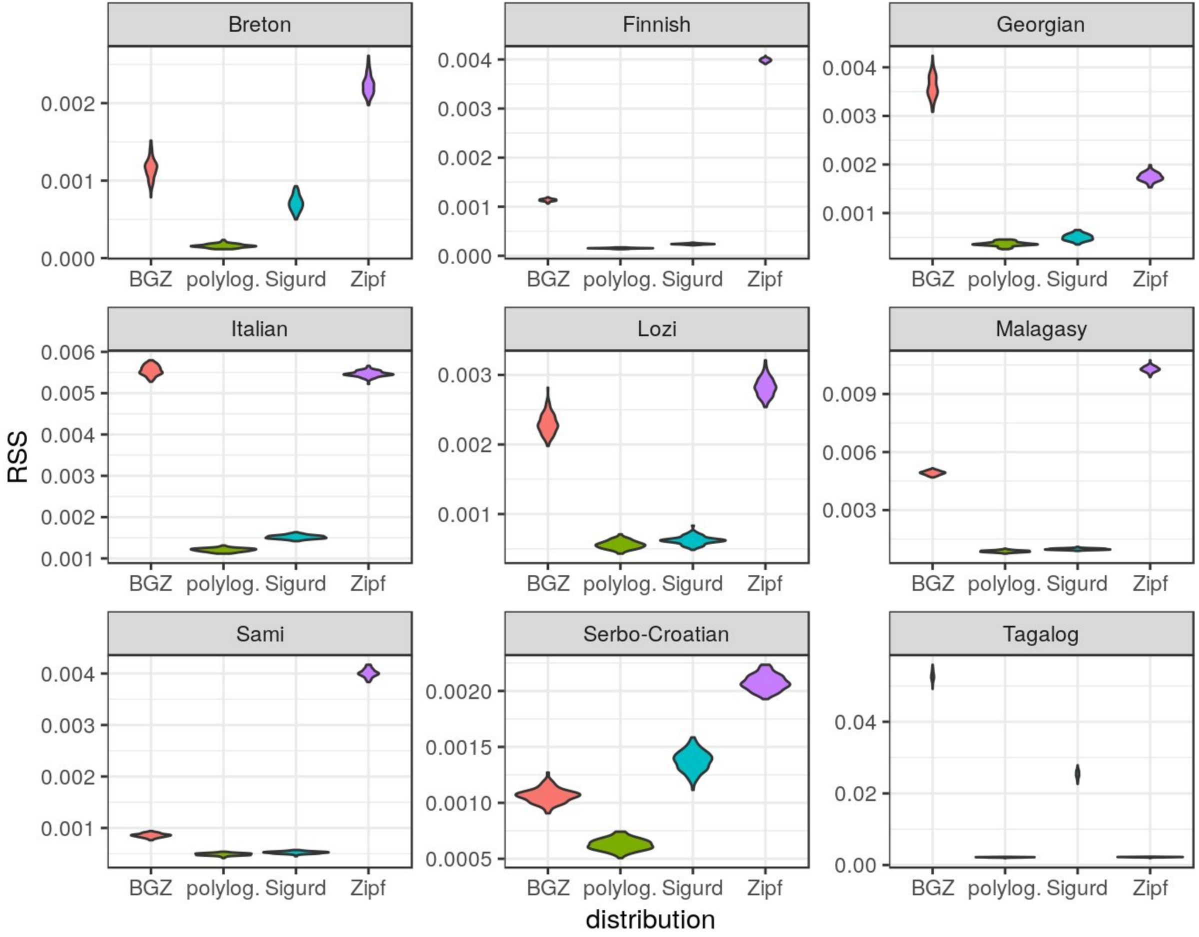
Results: RSS

Goodness of fit between empirical frequencies and theoretical distribution.

Lower is better.

Does not penalise model complexity…

… ***but***: serves as an intermediate step towards better measures.



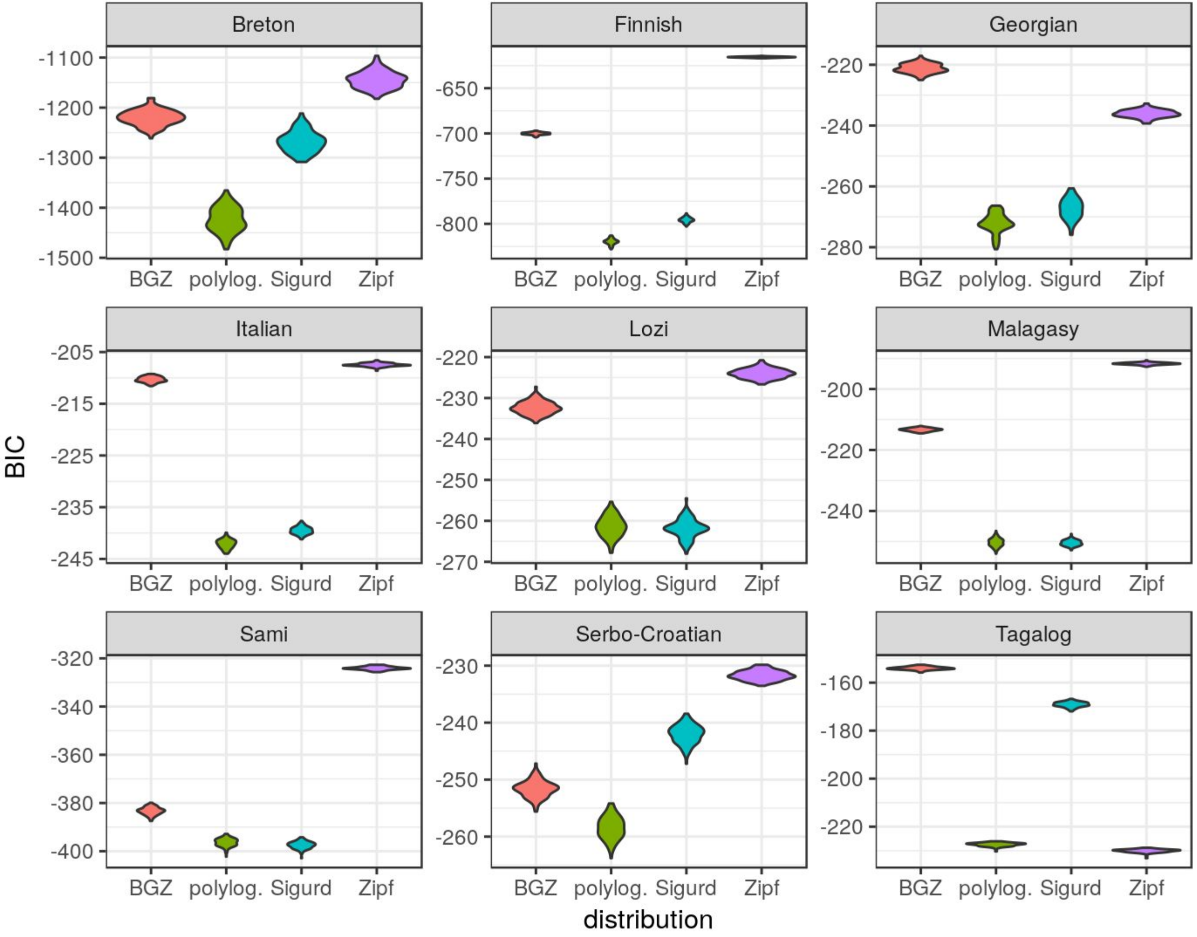
Results: BIC

Bayesian information criterion

Calculated from RSS and the number of parameters in the distribution

Distributions with more parameters incur a penalty

Lower is better



Results: Median BIC scores across bootstrap (to 1 significant decimal)

	Breton	**Finnish**	**Georgian**	**Italian**	**Lozi**	**Malagasy**	**Northern Sami**	**Serbo-Croatian**	**Tagalog**
BGZ	−1219.9	−700.1	−221.4	−210.4	−232.5	−213.3	−383.3	−251.6	−154.0
polylog.	***−1427.0***	***−819.5***	***−272.1***	***−242.0***	−261.4	−250.3	−396.3	***−258.4***	−227.3
Sigurd	−1267.9	−795.8	−267.0	−239.5	***−261.9***	***−250.5***	***−397.4***	−241.9	−169.1
Zipf	−1145.3	−615.6	−236.2	−207.5	−224.1	−191.7	−324.2	−231.8	***−230.0***

The polylogarithmic distribution often wins. When it doesn't, the BIC difference to the winning distribution is < 2 i.e. 'not worth more than a bare mention' (Kass & Raftery 1993: 777).

Tagalog stands out as the exception, with Zipf fitting the best.

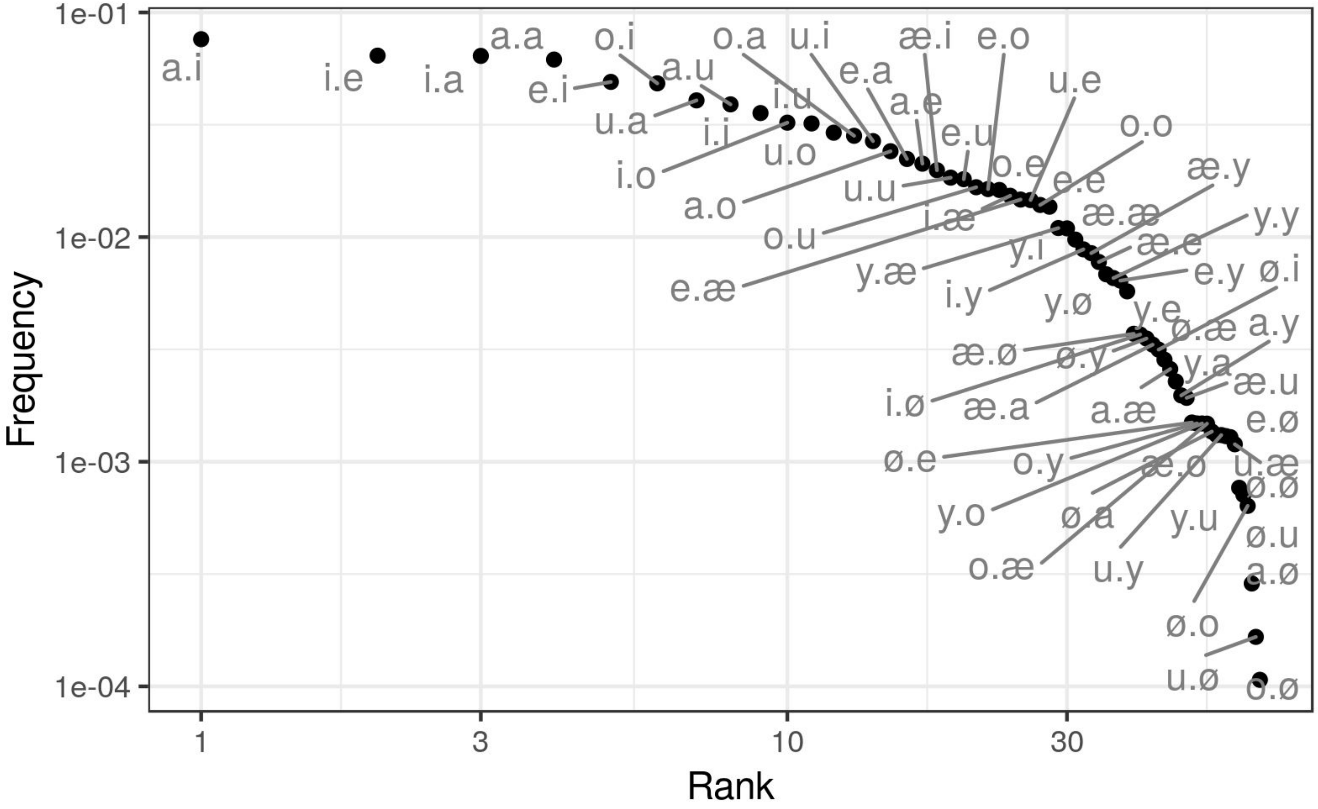
Discussion: Initial remarks

Our results show that vowel-pair frequencies do indeed conform closely to the ***polylogarithmic*** distribution.

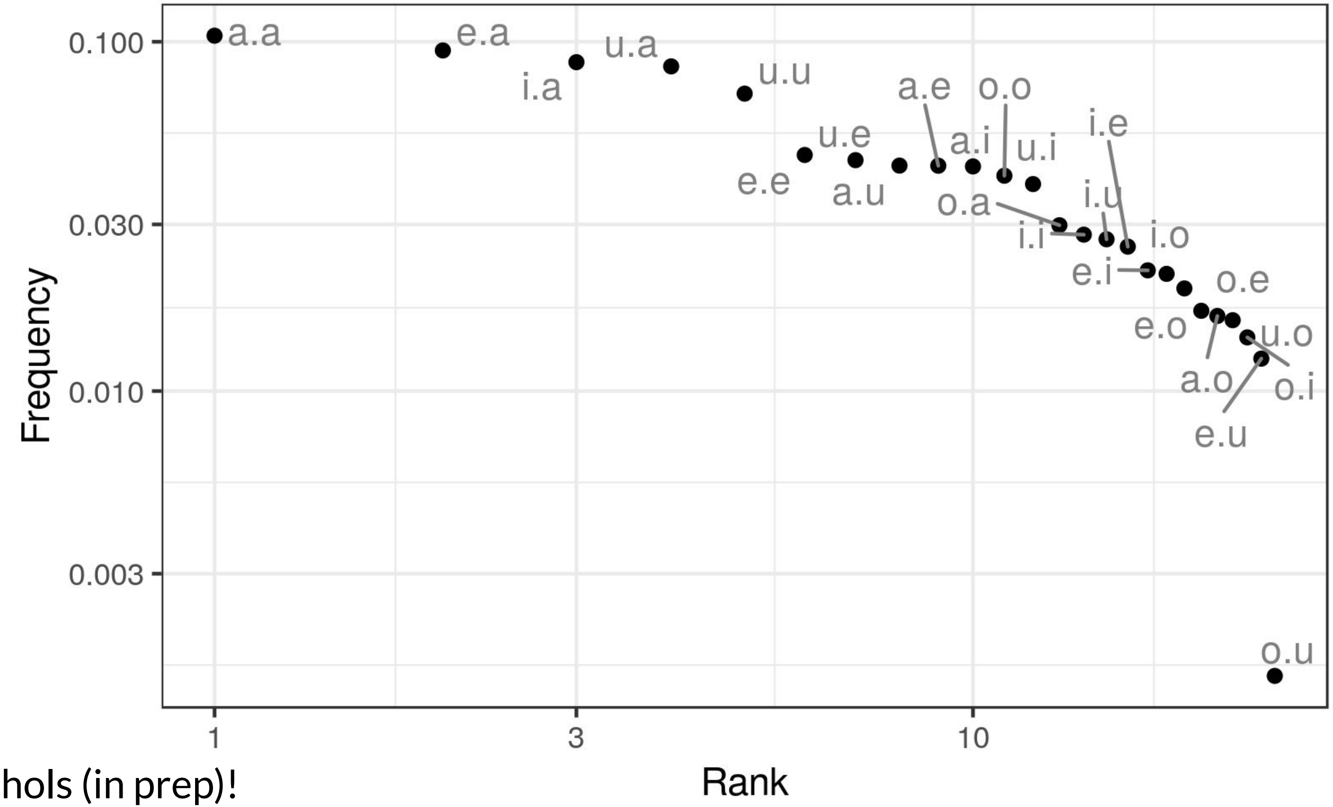
However, when individual fits are examined in detail, it is possible to discern slight ***deviations*** from the theoretical distributions.

These are mostly due to ***phonological*** or ***morphological*** effects that skew the distribution away from what would be expected under purely random combination.

Discussion: Individual deviations (Finnish)



Discussion: Individual deviations (Lozi)



See Nichols (in prep)!

Discussion: General remarks

Why do vowel pairs seemingly follow a polylogarithmic distribution?

Skewed long-tailed distributions like this can arise from a ***preferential attachment*** process:

- Items are chosen with a probability proportional to their frequency, so that ***"the rich get richer"*** (e.g. Yule 1924, Champernowne 1953, Simon 1955, Price 1976, Chung & Cox 1994, Martindale & Konopka 1996, Newman 2005).

Long-tailed distributions are found in various non-linguistic areas (e.g. genetics, ecology, economics, sociology, among others).

Discussion: General remarks

Unsure as to what the linguistic equivalent of this could be – modelling work is needed...

However, Ceolin & Sayeed (2019) and Ceolin (2019) show that the long-tailed distribution of singletons can be derived from a “null” model of sound change incorporating mergers and splits only.

Can something similar be devised to predict the distribution of pairs?

For now, we leave this question open...

Summary & conclusions

The rank–frequency distribution of phonemes is polylogarithmic.

And it seems that this is also the case for the dependent distribution of vowel pairs.

However, languages do exhibit deviations from this, but this appears to be due to language-specific phonotactic or morphological reasons.

Future work

Redux of Tambovtsev & Martindale (2007) study of phonemes.

As for vowel pairs, continue with more languages and bigger and better data sets.

Explore the effect that source type has, e.g. dictionaries/lexica v. corpora.

Investigate the implications for modelling sound change, esp. null/neutral models?

Examine not just the goodness of fit, but also the distribution parameters:

> What makes a language conform to a certain shape of the distribution? Is there a meaningful relation between the number of phonemes and the distribution parameters, for example?

References

Borodovsky, M. Yu. & S. M. Gusein-Zade (1989) A general rule for ranged series of codon frequencies in different genomes. *Journal of Biomolecular Structure and Dynamics* 6: 1000–12.

Ceolin, A. (2019) A *Null Model of Sound Change*. Talk given at RUSE, Manchester UK, 21st August. [`https://www.ling.upenn.edu/~ceolin/ruse2019.pdf`]

Ceolin, A. & O. Sayeed (2019) Modeling markedness with a split-and-merger model of sound change. In N. Tahmasebi, L. Borin, A. Jatowt & Y. Xu (eds.), *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. ACL.

Clauset, A., C. R. Shalizi & M. E. J. Newman (2009) Power-law distributions in empirical data. *SIAM Review* 51: 661–703.

Champernowne, D. G. (1953) A Model of Income Distribution. *The Economic Journal* 63(250): 318–51.

Chung, K. H. & R. A. K. Cox (1994) A stochastic model of superstardom: an application of the Yule distribution. *The Review of Economics and Statistics* 76: 771–5.

Eeg-Olofsson, M. (2008) Why is the Good distribution so good? Towards an explanation of word length regularity. *Lund University Department of Linguistics and Phonetics Working Papers* 53: 15–21.

Good, I. J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika* 40: 237–64.

Kass, R. E. & A. E. Raftery (1995) Bayes factors. *Journal of the American Statistical Association* 90: 773–795.

Kemp, A. W. (1995) Splitters, lumpers and species per genus. *Mathematical Scientist* 20: 107–18.

Klar, B., P. R. Parthasarathy & N. Henze (2010) Zipf and Lerch limit of birth and death processes. *Probability in the Engineering and Informational Sciences* 24: 129–44.

Martindale, C., S. M. Gusein-Zade, D. McKenzie & M. Yu. Borodovsky (1996) Comparison of equations describing the ranked frequency distributions of graphemes and phonemes. *Journal of Quantitative Linguistics* 3(2): 106–12.

Martindale, C. & A. K. Konopka (1996) Oligonucleotide frequencies in DNA follow a Yule distribution. *Computers & Chemistry* 20(1): 35–8.

Nichols, S. (in prep) Vowel-pair frequencies and phonotactic restrictions in Lozi. Manuscript, University of Manchester.

Newman, M. E. J. (2005) Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46: 323–51.

Price, D. (1976) A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* 27: 292–306.

Sigurd, B. (1968) Rank frequency distributions for phonemes. *Phonetica* 18: 1–15.

Simon, H. A. (1955) On a class of skew distribution functions. *Biometrika* 42: 425–40.

Tambovtsev, Yu. A. & C. Martindale (2007) Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics* 4: 1–11.

Yule, G. U. (1924) A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S. *Philosophical Transactions B* 213: 21–87.

Zipf, G. K. (1949) *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

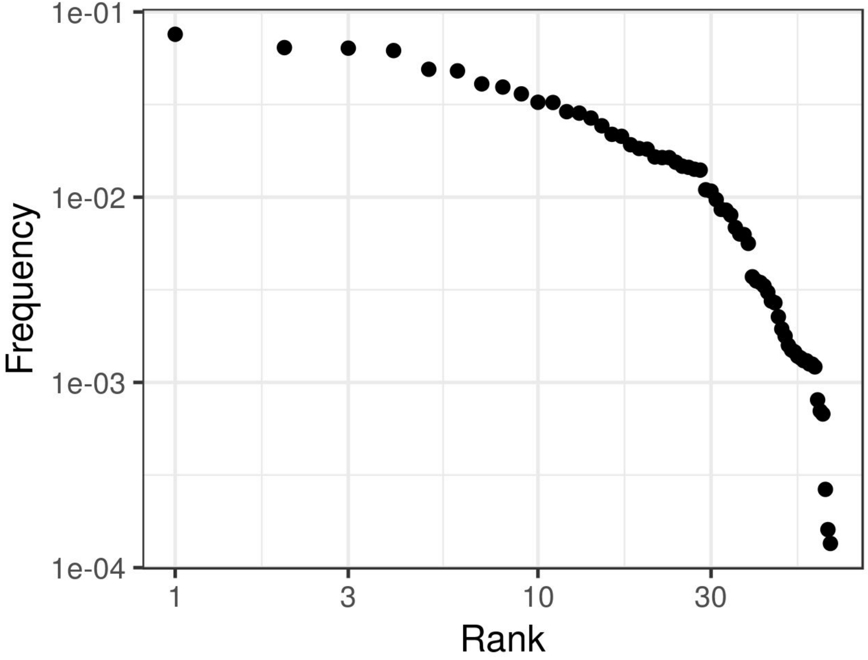
Zörnig, P. & G. Altmann (1995) Unified representation of Zipf distributions. *Computational Statistics & Data Analysis* 19: 461–73.

Thank you!

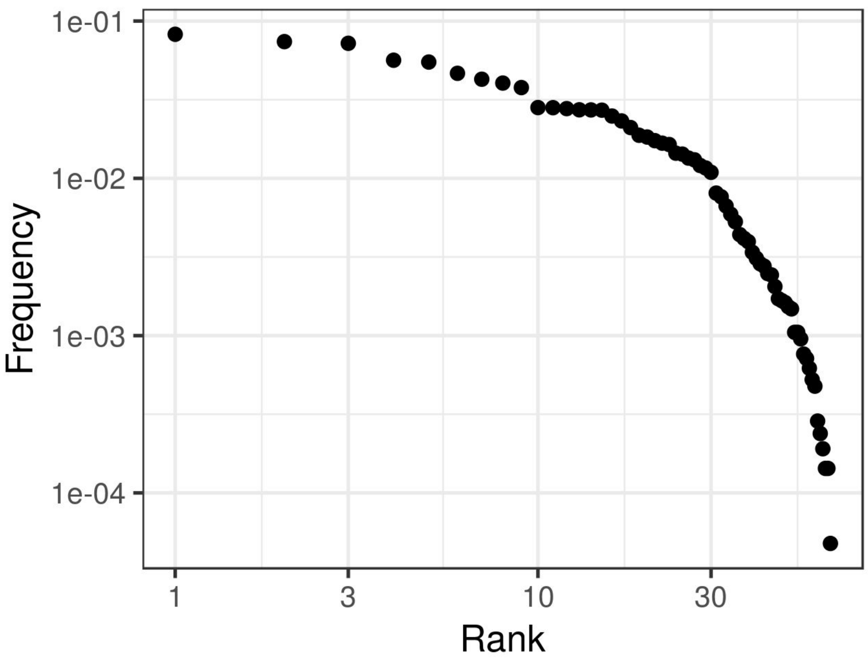
Plus thanks to ***Andrea Ceolin***, ***Ollie Sayeed*** and ***Christopher Quarles*** for discussions on phonology, typology, sound change, power laws and complex systems.

HK received funding from the ***Economic and Social Research Council***.

Appendix: Kotus v. Wiktionary comparison (Finnish)



Kotus



Wiktionary