The zipfR library:
Words and other rare events in R

Stefan Evert & Marco Baroni

University of Osnabrück, Germany
stefan.evert@uos.de

University of Bologna, Forlì, Italy
baroni@sslmit.unibo.it

useR! 2006, Vienna, 15 June 2006

Outline

(Computational) linguistics

Statistical inference in (computational) linguistics

Zipf's law and the LNRE problem

LNRE models for linguistic populations

Model estimation: The frequency spectrum

The zipfR library

Extrapolation of VGCs

Further work

Availability

What is (computational) linguistics?

The science of **linguistics** is concerned with ...

- natural language as a formal system (phonology, morphology, syntax, semantics, etc.)
- human language production and understanding, including the acquisition of language competence

Computational linguistics ...

- applies computers and electronic resources to linguistic research questions
- makes use of linguistic insights to build automatic natural language processing (NLP) systems

Corpora in (computational) linguistics

- increasing focus on language use and empirical evidence in recent years
- based on **corpora** = (usually large) machine-readable samples of naturally ocurring language
- some applications of corpus data
  - test hypotheses about formal system of language
  - validation of linguists' introspective judgements
  - observable result of human language production
  - model for linguistic experience of human speaker
  - training data for statistical NLP applications
- corpus = sample ➜ need for **statistical analysis**
  - standard methodologies are being established
  - random sample assumption is controversial for most corpora ➜ statistical inference may be unreliable
  - ☞ ongoing research into appropriate statistical models

Statistical inference from corpus data

- only observable data are **corpus frequencies**
- commonly used terminology: **types** vs. **tokens**
  - **tokens** can be running words, sentences in a text, instances of syntactic constructions, documents, etc.
  - categorization into fixed or open-ended set of **types**: distinct word forms or lemmas, parts of speech, etc.
  - of central interest are **type frequencies** $f(\omega)$
  - corpus is interpreted as a **random sample** of tokens
    ➜ inferences about type probabilities π_ω from $f(\omega)$
- **linguistic populations** are characterized by ...
  1. finite or countably infinite set of types ω
  2. type probabilities π_ω

➥ **multinomial distribution** of observed frequencies
  - confidence intervals or Bayesian estimates
  - comparison of type probabilities ($H_0 : \pi_1 = \pi_2$)
  - statistical associations

A characteristic problem: Zipf's law

- linguistic population is usually characterized by a very large or even infinite number of type probabilities
- in addition, substantial portion of probability mass is distributed over very infrequent types ($\neq$ normal dist.)
- referred to as the **LNRE** property (Khmaladze 1987) (*large number of rare events*)
- popularly known as **Zipf's law**, based on the **Zipf-Mandelbrot law** for type probabilities $\pi_k = \pi_{w_k}$:

$$\pi_k \approx \frac{C}{(k+b)^a}$$

where $b > 0$ and $a > 1$ is usually close to 1
  - Zipf ranking: $\pi_1 \geq \pi_2 \geq \pi_3 \geq \ldots$
  - see e.g. Baayen (2001, 101) for Zipf-Mandelbrot law
  - can be derived from Markov process (Rouault 1978)

Consequences of Zipf's law

- most types occur just once in a sample (**hapax legomena**) or not at all (**out-of-vocabulary**, OOV)
- hypothesis tests, confidence intervals and Bayesian estimates (for uniform or beta priors) will be inaccurate

> Imagine a population with 500 highly frequent types ($\pi = 10^{-3}$) and 500,000 rare types ($\pi = 10^{-6}$). In a sample of size $N = 1000$ there will be approx. 500 of the rare types among the hapax legomena, but the p-value for each individual occurrence is $p < .001$ (binomial test).

- estimators can also be highly biased if unseen types (OOV) are not taken into account

LNRE models

- we need a population model for the distribution of type probabilities ➜ **LNRE model** (Baayen 2001)
- such LNRE models have a wide range of applications
  - analyze accuracy of hypothesis tests and confidence interval estimates (Evert 2004b, Ch. 4)
  - better prior distributions for Bayesian estimates
  - estimate population vocabulary size (number of types), e.g. in authorship attribution (Thisted and Efron 1987), stylometry, or early diagnosis of Alzheimer's disease (Garrard *et al.* 2005)
  - extrapolate vocabulary growth, e.g. to estimate proportion of OOV types in large amounts of text, or the proportion of typos on the Web
  - extrapolate proportion of hapaxes for measuring morphological productivity in word formation (Baayen 2003; Lüdeling and Evert 2003)

LNRE models based on the Zipf-Mandelbrot law

- most widely-used LNRE models are based on the Zipf-Mandelbrot law
- rewrite Zipf-Mandelbrot equation as **distribution function** for type probabilities (as r.v.)

$$F(\rho) \coloneqq \sum_{\pi_k \le \rho} \pi_k$$

  - F is an increasing step function with range $[0, 1]$
- **type distribution** function G is more useful:

$$G(\rho) \coloneqq \left| \{ \omega_k \mid \pi_k \ge \rho \} \right|$$

  - G is a decreasing step function
  - for $\rho \to 0$, we have $G(\rho) \to S$
    (S = population vocabulary size, which may be infinite)
  - can easily be specified for $\rho = \pi_k$

The Zipf-Mandelbrot LNRE model

Some simplifications . . .

- use Poisson sampling instead of multinomial distribution (not conditioned on sample size N)
- approximate step function $G(\rho)$ by continuous function with **type density** $g(\pi)$:

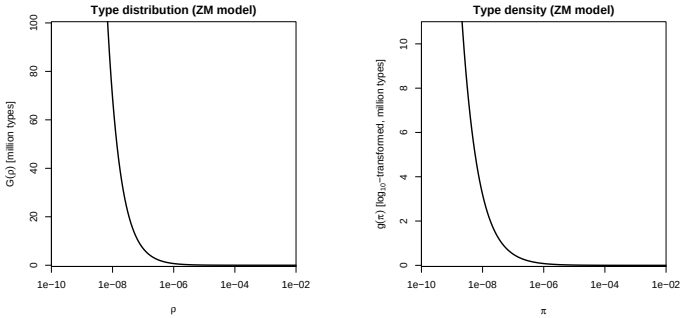
$$G(\rho) \approx \int_{\rho}^{\infty} g(\pi)\, d\pi$$

➥ the **Zipf-Mandelbrot** (ZM) model (Evert 2004a)

$$g(\pi) \coloneqq \begin{cases} C \cdot \pi^{-\alpha-1} & 0 \le \pi \le B \\ 0 & \text{otherwise} \end{cases}$$

  - free parameters are $0 < \alpha < 1$ and $0 < B \le 1$
  - relation to Zipf-Mandelbrot law: $\alpha = a^{-1}$

The Zipf-Mandelbrot LNRE model

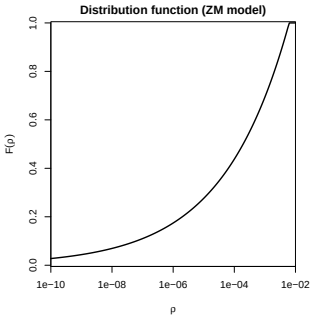


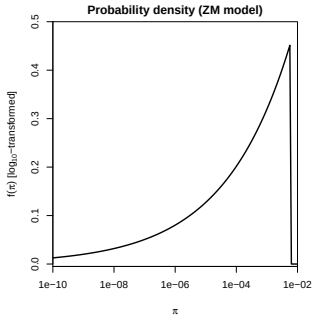
- type density function of Zipf-Mandelbrot LNRE model

$$g(\pi) = C \cdot \pi^{-\alpha-1} \quad (0 \le \pi \le B)$$

(densities in the images are $\log_{10}$-transformed)

The Zipf-Mandelbrot LNRE model





- corresponding **p.d.f.** for type probabilities

$$f(\pi) = C \cdot \pi^{-\alpha} \quad (0 \le \pi \le B)$$

(densities in the images are $\log_{10}$-transformed)

Extensions of the Zipf-Mandelbrot model

- **finite ZM** model adds lower threshold A for the type probabilities, i.e. $g(\pi) = 0$ for $\pi < A$ (Evert 2004a)
- **GIGP** model (Sichel 1971, 1975) with exponential attenuation instead of abrupt cutoff points, originally suggested by Good (1953, 249)

➥ allow better approximation of true population distribution, but mathematically less elegant and numerically more complex

Estimation of the model parameters

- estimate parameters of model from observed sample
  - type probabilities cannot be observed directly
  - many low-frequency types ➜ estimates unreliable
  - Zipf ranking of observed frequencies f_r may be different from Zipf ranking of type probabilities π_k
- individual hapaxes ($f_k = 1$) provide no useful information, but the number V_1 of such types does

➥ observed **frequency spectrum**

$$V_m := \left|\{\omega_k \mid f_k = m\}\right|$$

with **vocabulary size**

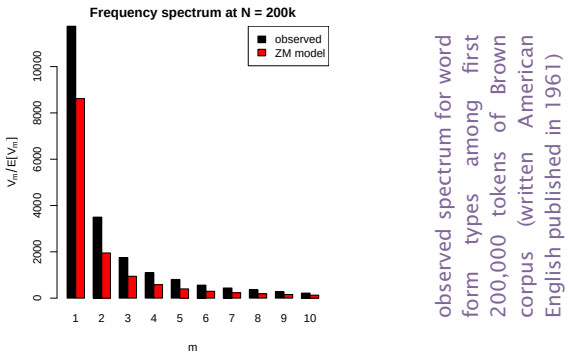
$$V := \left|\{\omega_k \mid f_k > 0\}\right| = \sum_{m=1}^{\infty} V_m$$

Observed and expected frequency spectrum

- expected spectrum can be calculated from $g(\pi)$:

$$\mathrm{E}[V_m] = \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi)\, d\pi$$

➥ leads to (incomplete) Gamma functions for ZM model



observed spectrum for word form types among first 200,000 tokens of Brown corpus (written American English published in 1961)

The zipfR library

The **zipfR** library for **R** implements:

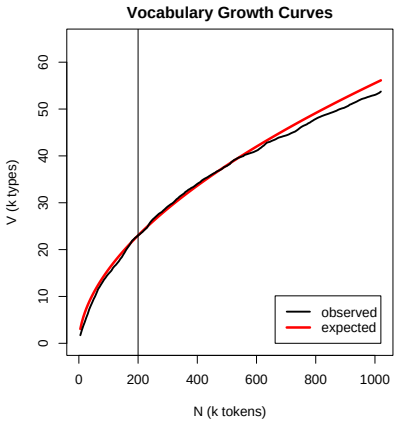
- LNRE models: ZM, finite ZM, GIGP
- parameter estimation from observed spectrum
- goodness-of-fit testing (Baayen 2001, 118–122)
- plots (spectrum, type & probability density)
- many utility functions for type frequency data
- fast subsampling & interpolation of observed spectrum

A zipfR example

- `spc <- read.spc("brown.200k.spc")`
  - ☞ load observed frequency spectrum from file
- `model <- lnre("zm", spc)`
  - ☞ estimate parameters of ZM model from spectrum
- `summary(model)`
  - ☞ displays model parameters & goodness-of-fit
- `spc.exp <- lnre.spc(model, N(spc))`
  - ☞ expected spectrum at this sample size
- `plot.spc(spc, spc.exp, m.max=10)`
  - ☞ plot expected vs. observed spectrum (as seen before)

Extrapolation of vocabulary growth

- LNRE models are often used for extrapolation of vocabulary growth beyond observed sample size
  - ☞ fully supported by **zipfR** library



extrapolation of vocabulary growth in Brown corpus from first 200,000 tokens to full size of 1 million word tokens, using the ZM model

Further work on the zipfR library

- more accurate and robust implementation of models
- better parameter estimation (plain `nlm()` for now)
- extended functionality for automation of experiments, e.g. extrapolation experiments with multiple randomizations (Baroni and Evert 2005)
- more advanced LNRE models for better goodness-of-fit
- corrections for non-randomness ➜ better extrapolation

☞ what do *you* want?

Availability

Availability of the **zipfR** library . . .

ahem

- but we can promise that it will be up on CRAN by end of July (in time for our ESSLLI course on *Counting Words*)
- some functionality (e.g. ZM and fZM models) already available in the **UCS** toolkit (`www.collocations.de`)
- we're also working on the **corpora** library for **R**, with basic statistical inference from corpus frequency data

Thank you!

Questions?

Fragen?

References I

Baayen, R. Harald (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Baayen, R. Harald (2003). Probabilistic approaches to morphology. In R. Bod, J. Hay, and S. Jannedy (eds.), *Probabilistic Linguistics*, chapter 7, pages 229–287. MIT Press, Cambridge.

Baroni, Marco and Evert, Stefan (2005). Testing the extrapolation quality of word frequency models. In P. Danielsson and M. Wagenmakers (eds.), *Proceedings of Corpus Linguistics 2005*, volume 1 of *The Corpus Linguistics Conference Series*. ISSN 1747-9398.

Evert, Stefan (2004a). A simple LNRE model for random character sequences. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, pages 411–422, Louvain-la-Neuve, Belgium.

Evert, Stefan (2004b). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from http://www.collocations.de/phd.html.

Garrard, Peter; Maloney, Lisa M.; Hodges, John R.; Patterson, Karalyn (2005). The effects of very early alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, **128**(2), 250–260.

References II

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**(3/4), 237–264.

Khmaladze, E. V. (1987). The statistical analysis of large number of rare events. Technical Report MS-R8804, Department of Mathematical Statistics, CWI, Amsterdam, Netherlands.

Lüdeling, Anke and Evert, Stefan (2003). Linguistic experience and productivity: corpus evidence for fine-grained distinctions. In D. Archer, P. Rayson, A. Wilson, and T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*, pages 475–483. UCREL.

Rouault, Alain (1978). Lois de Zipf et sources markoviennes. *Annales de l'Institut H. Poincaré (B)*, **14**, 169–188.

Sichel, H. S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In N. F. Laubscher (ed.), *Proceedings of the Third Symposium on Mathematical Statistics*, pages 51–97, Pretoria, South Africa. C.S.I.R.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, **70**, 542–547.

Thisted, Ronald and Efron, Bradley (1987). Did Shakespeare write a newly-discovered poem? *Biometrika*, **74**(3), 445–455.