ZipfR: Working with words and other rare events in R

Stefan Evert, University of Osnabrück, Germany (stefan.evert@uos.de)
Marco Baroni, University of Bologna, Italy (baroni@sslmit.unibo.it)

The field of linguistics has recently undergone a methodological revolution. Whereas earlier on most linguists had relied solely on introspection, recent years have seen the rise to prominence of *corpora*, i.e. large samples of texts, as the main source of linguistic data [5]. Because of this shift, statistical analysis plays an increasingly central role in the field. However, as has been known since the seminal work of George Kingsley Zipf (e.g. [6]), standard statistical models (in particular all those based on normality assumptions) are not suitable for analyzing the frequency distributions of words and other linguistic units. Even in the largest corpora currently available (containing above one billion running words of text), word frequency distributions are characterized by a high proportion of word types that occur only once or twice. When the sample size is increased further, a non-negligible number of new types will be encountered about which the original sample did not contain any information at all. Because of these properties, often referred to as the "Zipfianness" of language data, estimation of occurrence probabilities is unreliable (even when confidence interval estimates are used, cf. [3, Ch. 4]), the central limit theorem no longer guarantees the normality of sample averages for large samples, and the number of types in the population (which has an important linguistic interpretation as the overall vocabulary size of a certain language or sub-language) cannot easily be estimated from the observed data.

In the technical literature, various equations have been proposed for modelling the probability distribution of a Zipfian population. Baayen has summarized much of this work in [1], accompanied by a software package (`lexstats`) that can be used to estimate the parameters of different population models from an observed sample, and then calculate the expected values and variances of certain sample statistics (in particular, the number of distinct types in a sample of given size, as well as the number of types occurring once, twice, etc.). However, the `lexstats` package has only found limited use among linguists, for a number of reasons: `lexstats` is only supported under Linux, its ad-hoc Tk user interface has minimal functionality for graphing and data analysis, it has extremely restrictive input options (which make its use with languages other than English very cumbersome), and it works reliably only on rather small data sets, well below the sizes now routinely encountered in linguistic research (cf. the problems reported in [4]).

Following our positive experience implementing an R library of frequency comparison tests geared towards linguists with limited mathematical background (the `corpora` library available from CRAN), we decided to develop an R-based solution as an alternative to the `lexstats` suite. Our `ZipfR` library, which integrates code from the first author's UCS project (see `http://www.collocations.de/software.html`) currently provides implementations of three population models: Zipf-Mandelbrot, finite Zipf-Mandelbrot [2] and Generalized Inverse Gauss-Poisson (cf. [1]), although we expect to add other models in the future. The `ZipfR` library features an object-oriented design, in which units that should be intuitive to linguists (such as word frequency lists and vocabulary extrapolation experiments) are treated as objects. It relies on the R standard library for special functions (e.g. `besselI`) and statistical distributions, as well as general numerical utilities (e.g. `solve` for matrix inversion in a multivariate chi-squared test and `nlm` for parameter estimation). In our current tests, `ZipfR` model estimation has proven to be robust and efficient even for the largest data sets encountered in our experiments. A set of pre-designed plots are provided for the most common types of graphs (e.g. vocabulary growth curves and frequency spectra, see [1]), while allowing experienced users to take advantage of the full range of R graphics facilities. Furthermore, we have developed a package of auxiliary Perl scripts to extract word frequency data from corpora (including sophisticated randomization options in order to test the randomness assumption underlying the statistical models, which is

often problematic for language data), and to import existing data from `lexstats` as well as a range of other formats commonly used by linguists.

We hope that the availability of a powerful, user-friendly and flexible tool such as `ZipfR` will encourage more linguists to use advanced statistical models in their work that are suitable for the Zipfian distribution of word frequencies. As a welcome side-effect, this will also familiarize them with R itself and thus make the software more widespread in the linguistic community. A first public release of our library will be available from CRAN by the time of the conference, with supplementary information to be found on its homepage `http://purl.org/stefan.evert/ZipfR`.

References

[1] Baayen, Harald (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

[2] 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.

[3] Evert, Stefan (2004b). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

[4] Evert, Stefan and Baroni, Marco (2005). Testing the extrapolation quality of word frequency models. In: *Proceedings of Corpus Linguistics 2005*.

[5] McEnery, Tony and Andrew Wilson (2001). *Corpus Linguistics*. 2nd edition, Edinburgh: Edinburgh University Press.

[6] Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Cambridge MA: Addison-Wesley.