

Stefan Evert – Research – Teaching – CV – Publications – **Software** – Private Life

Software and Resources

zipfR – UCS – Web1T5–Easy – GOPHER – PerlTjes

zipfR – LNRE models of word frequency distributions in R

zipfR is a library for the statistical computing environment **R** (<<http://www.R-project.org/>>) that implements several statistical models for the distribution of word frequencies in natural language. These models can also be applied to many other linguistic and non-linguistic quantities and are often called **LNRE models** (for *large number of rare events*). Since its first public release in summer 2006, **zipfR** has attracted an international user community. Currently, work is under way on a new release that will feature more sophisticated models and other improvements. More information can be found on the project homepage: <<http://zipfR.R-Forge.R-project.org/>>

UCS – Association measures for collocation extraction

The **UCS toolkit** is a collection of libraries and scripts for the analysis of cooccurrence data, based on **statistical association measures**. Written in **Perl** and **R**, it provides reference implementations for a wide range of association measures, as well as various analysis and visualization techniques that can be used to achieve a better understanding of these measures.

The UCS toolkit accompanies my PhD thesis *The Statistics of Word Cooccurrences*, and a new release including code and data for (almost) all results presented in the thesis is currently in preparation. The software is open source can be obtained here: <<http://www.collocations.de/software.html#UCS>>.

Indexing Google's Web 1T 5–grams database with Perl and SQLite

Web1T5–Easy is a collection of Perl scripts for indexing and querying the **Google Web 1T 5–Grams** database [<<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13>>] with the open-source **SQLite** database engine [<<http://www.sqlite.org/>>]. This package offers a convenient way to build an interactively searchable version of the Web1T5 database, including a full collocation analysis and a simple, but powerful Web interface. It is not designed as a high-performance implementation and requires considerable amounts of disk space (approx. 220 GiB) as well as patience (indexing may take up to 2 weeks on a state-of-the-art server).

You can try an online demo of the Web1T5–Easy Web interface at <<http://cogsci.uos.de/~korpora/ws/Web1T5/>> or download the latest release **version 1.1** from <http://webascopus.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_10_Software&subpage=FILES_50_Goog>

GOPHER styles and the GopherPHP framework for simple Web sites

GOPHER is a set of CSS stylesheets for basic single-page or multi-page Web sites (e.g. the homepage of a software project, a conference Web site, etc.). They have been designed for a clear layout, good readability and compatibility with all major Web browsers. The acronym GOPHER stands for **Good Old Plain HTML** for the Occasional Writ**ER**. The associated **GopherPHP** framework automatically generates dynamic navigation bars for small Web sites and offers some other useful features (such as allowing users to change the font size of Web pages).

If you want to set up a simple Web site without worrying about CSS specifications and the details of page navigation, GOPHER is just what you need! More information and downloads can be found on the GOPHER homepage: <<http://purl.org/stefan.evert/GOPHER/>>

PerlTjes – Small but useful Perl scripts and modules

HotSync – Hoover

HotSync

HotSync is a convenient wrapper around the **rsync**, **ssh** and GNU **tar** programs for synchronizing files between two computers (over a network connection) as well as creating backup archives of directory trees (and storing them on a remote host). Synchronization and backup archive tasks are defined in a simple

configuration file and can be listed and invoked from the command line or through an interactive console.

- Download **HotSync version 0.8**: <<http://purl.org/stefan.evert/DOWNLOAD/HotSync-0.8.tgz>> (release date: 7 Oct 2008)
 - `mirror` option overwrites and deletes newer files on remote host without confirmation (useful for mirroring directory trees, as the name implies)
 - `download-mirror` option overwrites files on local host (useful for creating local backups of remote servers, e.g. a SourceForge project Web site)

Older versions: <<http://purl.org/stefan.evert/DOWNLOAD/HotSync-0.7.7.tgz>> (13 Nov 2007) – <<http://purl.org/stefan.evert/DOWNLOAD/HotSync-0.7.2.tgz>> (6 June 2005)

Hoover

When TeX/LaTeX is used to format a document, many temporary files are created in the same directory as the source file. Unfortunately, TeX does not offer an easy way to remove these temporary files when they are no longer needed. The **Hoover** program was designed to take over this job. It also deletes Emacs backup, recovery and font lock cache files.

- Download **Hoover version 0.99**: <<http://purl.org/stefan.evert/DOWNLOAD/Hoover-0.99.tgz>> (release date: 12 June 2005)