

KoGraR: standardized statistical analyses of corpus counts

Sascha Wolfer
IDS Mannheim
R5, 6-13
D-68161 Mannheim
wolfer@ids-
mannheim.de

Sandra Hansen-Morath
IDS Mannheim
R5, 6-13
D-68161 Mannheim
hansen@ids-
mannheim.de

Hans-Christian Schmitz*
Fraunhofer FKIE
Fraunhoferstr. 20
D-53353 Wachtberg
hans-
christian.schmitz@
fkie.fraunhofer.de

Within the project “Corpus grammar” (Korpusgrammatik) at the Institute for the German Language (Institut für Deutsche Sprache, IDS) in Mannheim, techniques and tools are developed for the description of grammatical phenomena based on analyses of very large morpho-syntactically annotated corpora. The goal of the project is a corpus-based grammar that captures variations of grammatical structure in present-day German. In the first project phase, pilot studies were conducted (cf. Bubenhofer et al., 2014; Fuß, 2014; Konopka, 2014) to exploit and evaluate various methodological approaches to variation phenomena. For each research question, statistical analyses were chosen and customized. From these analyses, a subset was extracted as the methodological core of the project, with the aim of supporting methodological coherence, interoperability of sub-projects and, finally, the descriptive coherence of the project result, that is, the grammar. The methodological core has been made available to project members via an easy-to-use web front-end: the results of corpus queries and other, user-defined data tables can be uploaded and analyzed automatically. The web front-end is called KoGraR.

A tool like KoGraR has to meet several requirements: (1) The statistical analyses that are conducted have to be general enough to study a wide range of variation (lexical, morpho-syntactic and syntactic) phenomena. (2) The tool has to incorporate tests of statistical significance but also effect size. This is necessary because analyses based on very large corpora tend to show significant results while the size of the effects may be very small or even negligible. (3) The tool has to be easy to use, and documentations on the im-

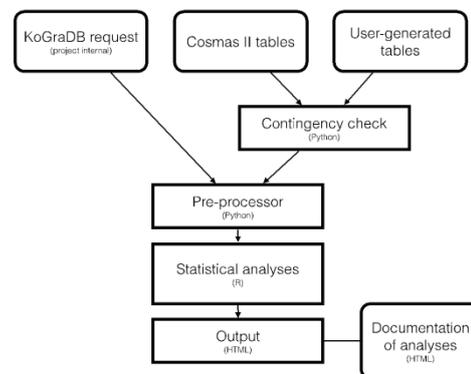


Figure 1: Schematic overview of KoGraR.

plemented statistical tests have to be accessible at a glance.

KoGraR is depicted schematically in Figure 1. As a basis for analyses, KoGraR takes frequency or contingency tables from various sources: firstly, KoGraR provides a direct interface to the KoGra database (KoGraDB) for processing the results of database queries (this resource is not open to the public). Secondly, arbitrary frequency tables can be entered manually (“User-generated tables” in Figure 1). Thirdly, frequency tables generated with Cosmas II can be uploaded.¹ It is possible to upload several tables at once and combine them in a multi-column table. The input is pre-processed, checked for contingency (which is most necessary for the user-generated tables), and transferred to a server-side installa-

* Hans-Christian Schmitz worked on KoGraR while he was a member of the IDS.

¹ Cosmas II is the Corpus Search, Management and Analysis System which makes huge portions of the IDS corpora available to the public (<http://www.ids-mannheim.de/cosmas2/> [last access: May 7th, 2015]).

tion of R, an open environment for statistical computing and graphics (R Core Team, 2015). Currently, the following statistical procedures are applied on the tables: (1) output of tables and diagrams for raw data, normed and relative values, (2) a Chi-Square test as well as expected frequencies and standardized cell residuals, (3) Phi / Cramér's V association coefficient, (4) association and mosaic plots, (5) tables and diagrams for confidence intervals and (6) dispersion measures and plots with a focus on DP(norm) (Lijffijt & Gries, 2012). The set of procedures is open, thus, further analyses can be added easily on demand. For each test implemented in KoGraR, a short documentation with further information and help for the interpretation of the test is made available. The R code used to conduct the analyses on the server can be accessed directly and copied into a local installation of R (via copy & paste). The code to create the table objects in R is also included so that the user is able to retrace every step of the analyses and adapt the code where appropriate.

KoGraR has a standardizing influence on the collaborative work within the project "Corpus grammar". The implemented statistical analyses are used as the "standard catalogue" necessary for the conception of a monograph containing a corpus-based description of present-day German variation phenomena. All researchers currently working on the monograph are supposed to consult KoGraR with their empirical questions in order to assure methodological coherence of the grammar monograph.

The set of statistical analyses and the associated documentations can be useful for a variety of other linguistic projects that are working with large corpora. The only restriction (for KoGraR in its current state) is that the data has to be arranged in a frequency table in a meaningful way. Of course, this does not restrict the scope of KoGraR to linguistics. In principle, every researcher interested in the statistical analyses of contingency tables can use KoGraR.

Elementary knowledge of statistical methodology is expected from the potential users of KoGraR. To meet needs that exceed this basic statistical knowledge, the specific tests carried out are thoroughly documented and explained.

References

- Noah Bubenhofer, Sandra Hansen-Morath, and Marek Konopka (2014). Korpusbasierte Exploration der Variation der nominalen Genitivmarkierung. *Zeitschrift für germanistische Linguistik* 42 (3), S. 379-419.
- Fred D. Davis (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13 (3), 319-340.
- Eric Fuß (2014). Endungslose Genitive. In: *grammis 2.0. – Korpusgrammatik*. Electronical resource - Mannheim: Institut für Deutsche Sprache.
- Stefan Th. Gries (2008): Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403-437.
- Marek Konopka (2014). Endungsvariation. In: *grammis 2.0 – Korpusgrammatik*. Variation der starken Genitivmarkierung. Electronical resource - Mannheim: Institut für Deutsche Sprache.
- R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>.