Abteilung Grammatik

INSTITUT FÜR DEUTSCHE SPRACHE

KOGRA-R: STANDARDISIERTE STATISTISCHE AUSWERTUNGEN VON KORPUSRECHERCHEN

Hans-Christian Schmitz, Sandra Hansen-Morath, Roman Schneider, Sascha Wolfer

PROJEKT „KORPUSGRAMMATIK – GRAMMATISCHE VARIATION IM STANDARDSPRACHLICHEN UND STANDARDNAHEN DEUTSCH“

Das Projekt zielt methodologisch darauf ab, Techniken und Werkzeuge zu entwickeln, um erstens grammatische Phänomene mit Bezug auf große Korpora zu beschreiben – **Deskription** –, zweitens Korpusdaten explorativ, mit dem Ziel der Mustererkennung zu untersuchen – **Exploration** – und drittens eine transparente quantitativ-statistische Basis für die Validierung von, neu aufgestellten oder bereits in der einschlägigen Literatur vertretenen, Hypothesen bereitzustellen – **Inferenz**.

Die korpuslinguistischen Methoden werden in Pilotstudien, die sich relevanten linguistischen Fragestellungen widmen, ausgesucht, angepasst, evaluiert und ggf. optimiert. Zu diesen Methoden gehören u.a. kanonische statistische Verfahren der Berechnung von Signifikanzen und Effektstärken, aber auch Techniken des maschinellen Lernens, wie sie sonst im Text- und Data Mining zur Anwendung kommen. Zusammengenommen werden die Methoden und Instrumente einen Werkzeugkasten für die korpusorientierte Grammatikforschung bilden.

Die Endphase des Projekts wird durch die Ausarbeitung der Konzeption und der Gliederung einer „Korpusgrammatik“ bestimmt, in der auf den Pilotstudien aufbauend die Erfassung grammatischer Variation innerhalb der systematischen Teilgebiete vervollständigt wird und diese miteinander in Bezug gesetzt werden.

KOGRA-DB UND KOGRA-R

Die Datengrundlage für die grammatischen Untersuchungen im Projekt ist das mit Metadaten (Medium, Register, Region, Zeit, ...) angereicherte, in einer relationalen Datenbank **(Kogra-DB)** bereitgestellte und durchsuchbar gemachte **Deutsche Referenzkorpus (DeReKo).**

Kogra-R ist eine Web-basierte Schnittstelle, über die vordefinierte, in R programmierte statistische Auswertungen durchgeführt werden. Mittels Kogra-R können Ergebnisse von Kogra-DB Recherchen, freie Tabellen und durch das Recherchesystem COSMAS erzeugte Frequenzlisten statistisch ausgewertet und miteinander verglichen werden.

Erzeugt werden bislang:

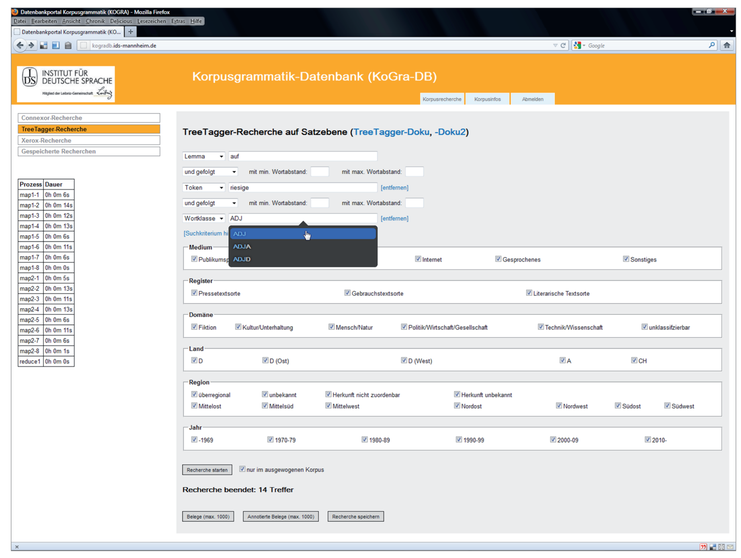
- **Tabellen und Diagramme für Rohdaten, normierte und relative Werte**
- **Chi²-Test, erwartete Häufigkeiten und Residuen**
- **Phi Assoziationsstärke/Cramérs V Assoziationsstärke**
- **Assoziationsplots, Mosaikplots**
- **Tabellen und Diagramme für Konfidenzintervalle**
- **Dispersionsmaße, insbesondere DP Norm (Gries 2008)**

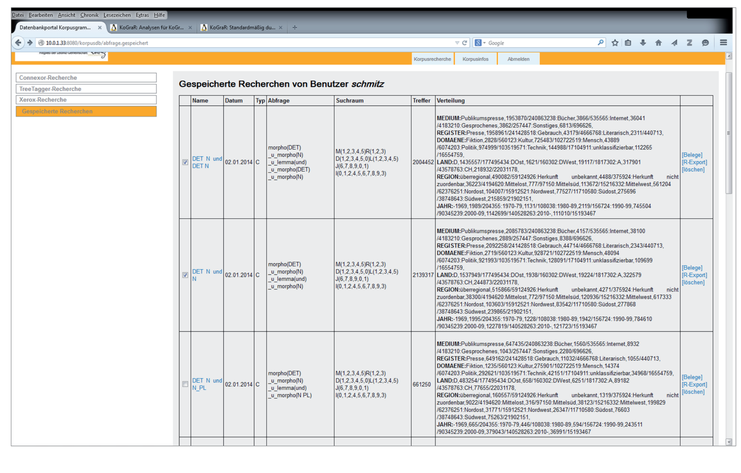
Kogra-R wirkt standardisierend auf die Projektarbeit, insofern durch die Schnittstelle eine Menge von statistischen Analysen definiert wird, die stets, bei allen Fragestellungen durchgeführt werden können und sollen.

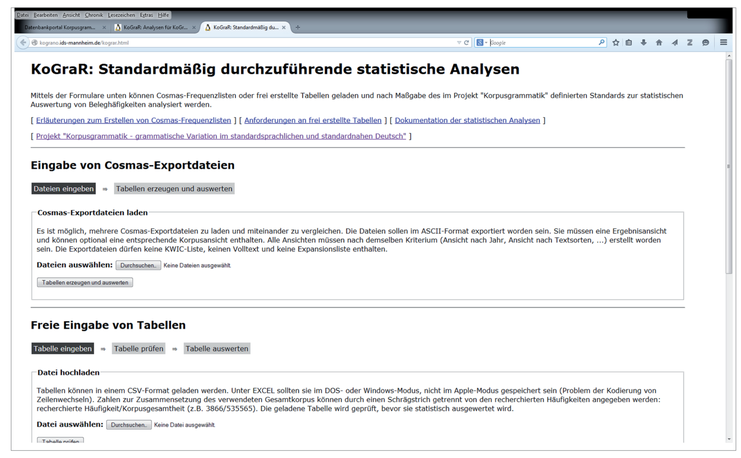
EVALUATION (*PERCEIVED USEFULNESS AND EASE OF USE*)

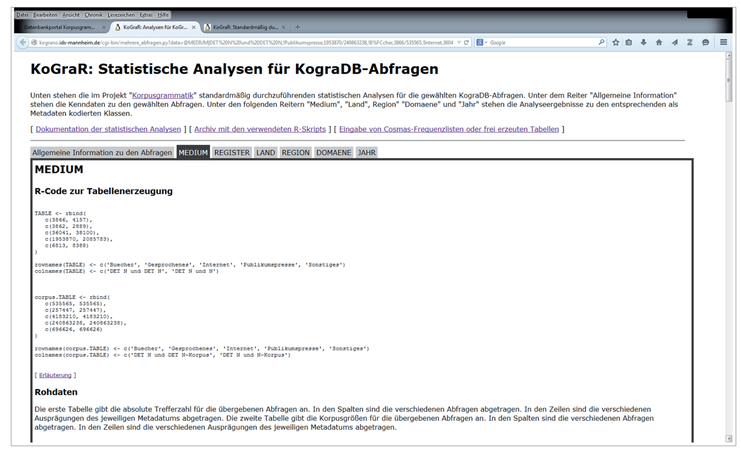
„Sie haben mit Kogra-R ein System zur automatischen Durchführung vorgegebener statistischer Analysen von Korpusrecherchen benutzt. Inwieweit stimmen Sie den folgenden Aussagen zu? (Skala 1-7, stimme gar nicht zu - stimme vollkommen zu)“

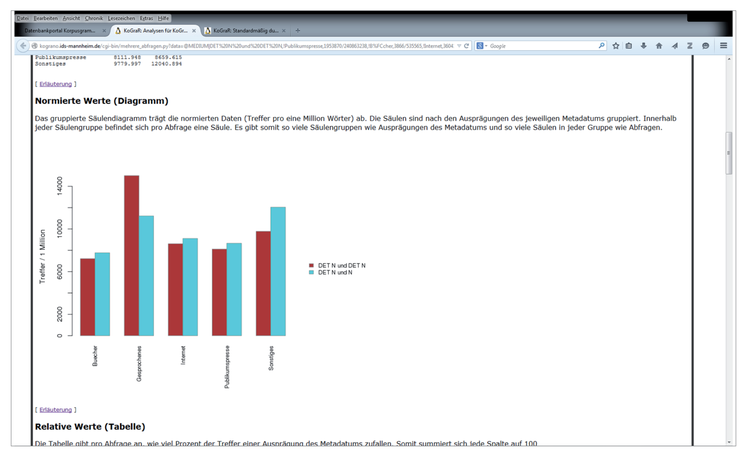
- Das System ist nützlich für mich.
  AM = 6.0, SD = 1.53, Median = 7.0, Modus = 7
- Ich kann meine Arbeitsziele aufgrund des Systems besser erreichen.
  AM = 6.0, SD = 1.53, Median = 7.0, Modus = 7
- Die Bedienung des Systems war einfach.
  AM = 4.67, SD = 1.80, Median = 5.0, Modus = 6
- Die Benutzung des Systems war frustrierend.
  AM = 2.5, SD = 1.26, Median = 2.5, Modi = 1, 4
- Die Benutzung des Systems war mühsam.
  AM = 2.67, SD = 1.11, Median = 2.5, Modus = 2, 4
- Die Benutzung des Systems war motivierend.
  AM = 4.00, SD = 1.60, Median = 5.0, Modi = 3, 5
- Durch das System konnte ich meine Aufgaben besser eigenständig bewältigen.
  AM = 5.5, SD = 1.90, Median = 6.5, Modus = 7
- Ich werde das System, sofern ich Zugang zu ihm habe, öfter benutzen.
  AM = 5.84, SD = 1.86, Median = 7.0, Modus = 7











Kontakt
Postadresse:
Dr. Hans-Christian Schmitz
Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim
schmitz@ids-mannheim.de

Tel.: +49 621 1581-234
Fax: +49 621 1581-200

Hausadresse:
Institut für Deutsche Sprache
R 5, 6-13
D-68161 Mannheim
Deutschland
Tel.: +49 621 1581-0
Fax: +49 621 1581-200
info@ids-mannheim.de
www.ids-mannheim.de

Mitglied der Leibniz-Gemeinschaft