

SANDRA HANSEN-MORATH / HANS-CHRISTIAN SCHMITZ / ROMAN SCHNEIDER / SASCHA WOLFER

KOGRA-R: STANDARDISIERTE STATISTISCHE AUSWERTUNG VON KORPUSRECHERCHEN

1. Einleitung

Das Projekt Korpusgrammatik hat zusätzlich zu seiner linguistischen Zielsetzung die Aufgabe, Techniken und Werkzeuge zu entwickeln, um grammatische Phänomene mit Bezug auf große Korpora geschriebener Sprache zu beschreiben, Korpusdaten explorativ zu untersuchen und eine transparente quantitativ-statistische Basis für die Validierung linguistischer Hypothesen bereitzustellen. Damit schafft es Grundlagen für differenzierte Untersuchungen, wie sie der geplanten Grammatik (vgl. das Vorwort zu diesem Band) zugrunde liegen sollen, aber auch anderen korpusorientierten Grammatikprojekten zugutekommen können.

Im Rahmen der im Projekt durchgeführten Pilotstudien (siehe v.a. die Studien in diesem Band sowie Konopka/Waßner 2013; Bubenhofer/Hansen-Morath/Konopka 2014; Bubenhofer/Konopka/Schneider 2014, S. 125ff.; Konopka/Fuß 2016) wurden u.a. statistische Analysearten bestimmt, die für die erste Exploration der Daten bei der Erstellung der Grammatik als sinnvoll und nützlich erscheinen. Sie bilden einen „Werkzeugkasten“ für korpusgestützte grammatische Untersuchungen. Die Menge der Analysearten ist prinzipiell offen und kann erweitert werden. Bislang umfasst sie

- die Darstellung von Tabellen und Diagrammen für Rohdaten, normierte und relative Werte,
- den Chi-Quadrat-Test, Berechnung von erwarteten Häufigkeiten und Residuen,
- die Berechnung der Assoziationsstärke Phi bzw. Cramérs V ,
- Assoziationsplot, Mosaikplot,
- die Darstellung von Tabellen und Diagrammen für Konfidenzintervalle und
- die Berechnung von Dispersionsmaßen, insbesondere die DP_{norm} (Gries 2008, 2009; Lijffijt/Gries 2012).

Um die Analysearten möglichst einfach zur Verfügung zu stellen, wurde KoGra-R entwickelt. KoGra-R ist ein webbasiertes Analysetool, über das die genannten, in der Programmiersprache R (R Core Team 2016) implementierten,

statistischen Auswertungen durchgeführt werden können. R ist eine freie und flexible Entwicklungsumgebung zur Umsetzung von statistischen Analysen, die u.a. zahlreiche Optionen zur Datenvisualisierung bereithält und sehr gut für große Datensätze geeignet ist (vgl. Wolfer/Hansen-Morath 2017). Die Programmiersprache ist mittlerweile in vielen Wissenschaftsbereichen zum Standard für quantitative Analysen geworden. Eine große Stärke von R liegt in der weltweiten Gemeinschaft der Nutzer/innen, die über Zusatzpakete immer neue Funktionen zur Verfügung stellen. KoGra-R ist über die Homepage unter www.ids-mannheim.de/kogra-r (letzter Zugriff: 27.11.2018) öffentlich verfügbar.¹

Die wichtigste Datengrundlage der geplanten Grammatik ist das Deutsche Referenzkorpus (DEReKo)² (Institut für Deutsche Sprache 2014), das projektintern mittels einer zu Testzwecken erstellten Datenbank (KoGra-DB)³ sowie des Recherchesystems COSMAS II⁴ durchsucht werden kann. Durch KoGra-R können KoGra-DB-Rechercheergebnisse und von COSMAS II erzeugte Frequenzlisten statistisch ausgewertet und miteinander verglichen werden. Darüber hinaus ist es möglich, beliebige, anderweitig erzeugte Kontingenztabellen einzugeben und auswerten zu lassen.

KoGra-R ist ein Werkzeug, dessen primärer Zweck es ist, die Arbeit der am Projekt beteiligten Linguist(inn)en zu vereinfachen. Die Bedienung von KoGra-R verlangt keine Programmierkenntnisse, statistische Auswertungen werden mithilfe der Dateneingabe über eine Benutzeroberfläche durchgeführt. Darüber hinaus soll KoGra-R standardisierend auf die Projektarbeit wirken, insofern durch das Analysetool die Menge derjenigen Analysen definiert wird, die möglichst bei allen Fragestellungen der geplanten Grammatik durchgeführt werden sollen. Wir erwarten, dass KoGra-R auch für Aufgaben jenseits der „Korpusgrammatik“ nützlich sein kann, und haben das Tool daher online für die Öffentlichkeit zugänglich gemacht.

Das vorliegende, methodische Kapitel ist wie folgt aufgebaut: Zunächst beschreiben wir in Abschnitt 2 verschiedene Nutzungsszenarien von KoGra-R. In Abschnitt 3 erläutern wir die in KoGra-R bislang realisierten statistischen

¹ In Hansen-Morath/Wolfer (2017) wird gezeigt, inwiefern das Tool dazu genutzt werden kann, Variationsphänomene auf verschiedenen linguistischen Ebenen zu untersuchen.

² Vgl. www1.ids-mannheim.de/kl/projekte/korpora (letzter Zugriff: 23.11.2017). Datengrundlage ist das DEReKo-Release 2014-II.

³ Siehe Abschnitt 4.1 und Bubenhofer/Konopka/Schneider (2014, S. 79ff.).

⁴ COSMAS II verfügt über eine Web-Schnittstelle: <https://cosmas2.ids-mannheim.de/cosmas2-web> (letzter Zugriff: 23.11.2017).

Analysen. In Abschnitt 4 beschreiben wir knapp die technische Implementierung. KoGra-R wurde zweimal evaluiert. In Abschnitt 5 stellen wir die Evaluationsergebnisse vor. Schließlich machen wir im letzten Abschnitt 6 eine kurze Bestandsaufnahme.

2. Nutzungsszenarien

Input für KoGra-R sind Kontingenztabellen mit Häufigkeitsangaben.⁵ Die Häufigkeitsangaben sind, so die Intention, Angaben sprachlicher Phänomenrealisierungen, die sich als Ergebnisse von Korpusrecherchen ergeben. Es gibt drei verschiedene Möglichkeiten, Tabellen an KoGra-R zu übergeben: Erstens können Ergebnisse von KoGra-DB-Recherchen direkt übergeben werden; die Ergebnisdarstellungen von KoGra-DB enthalten entsprechende Links. Zweitens können mit COSMAS II Frequenzlisten für Rechercheergebnisse exportiert und hochgeladen werden. Drittens können Tabellen als CSV-Dateien übergeben oder händisch in ein Formular eingetragen werden. Alle Modi der Dateneingabe sind auf der KoGra-R-Webseite ausführlich dokumentiert.

2.1 Eingabe von Daten via KoGra-DB

Eine in der Pilotphase genutzte Datengrundlage für grammatische Untersuchungen im Projekt Korpusgrammatik besteht aus mit Metadaten angereicherten, in einer relationalen Datenbank (KoGra-DB) durchsuchbar gemachten Teilkorpora des Deutschen Referenzkorpus (DEReKo).⁶ Ergebnisse von KoGra-DB-Recherchen können gespeichert und analysiert werden. Beides, Speicherung und Analyse, geschieht über entsprechende Links. Sobald die Rechercheergebnisse gespeichert sind, können sie über den Link „[KoGra-R]“ (vgl. Abb. 1) statistisch ausgewertet werden.

⁵ Vgl. z.B. Tabelle 1 in Abschnitt 3.1 unten.

⁶ Vgl. diesbezüglich Abschnitt 4.1.

Korpusgrammatik-Datenbank KoGra-DB (Testversion)

- Korpusrecherche
- Korpusinfos
- Abmelden

Gespeicherte Recherchen von Benutzer konopka

Name	Datum	Typ	Abfrage	Suchraum	Treffer	Verteilung	
<input checked="" type="checkbox"/> parken	23.01.2019	C	token(parken)	M(1,2,3,4,5)R(1,2,3) D(1,2,3,4,5,0)L(1,2,3,4,5,6) J(6,7,8,9,0,1) I(0,1,2,4,5,6,7,8,9,3)	23248	<p>MEDIUM:Publikumspresse,22591/380060468:Bücher,60/1472363:Internet,241/76306651:Gesprochenes,35028410491:Sonstiges,6/625010,</p> <p>REGISTER:Presse,22025/374110184:Gebrauch,1168/11647373:Literarisch,55/1117426,</p> <p>DOMAENE:Fiktion,58/1256989:Kultur,7311/201315611:Mensch,104/9269947:Politik,12773/216028418:Technik,22772/7531431:unklassifizierbar,725/21472587,</p> <p>LAND:D,19907/397804684:Dost,3/153469:DWest,71/1851459:A,297452453492:CH,150/32952890:LU,143/1658989,</p> <p>REGION:überregional,5021/62414992:Herkunft unbekannt,13/408257:Herkunft nicht zuordenbar,242/76358231:Mittelost,113/3722220:Mittelsüd,1207/2385679:Mittelwest,9079/102275873:Nordost,2999/55180337:Nordwest,1736/39773399:Südost,466/85319985:Südwest,273/37564960,</p> <p>JAHR:-1969,15/513899:1970-79,2/111804:1980-89,5/147334:1990-99,2503/98225315:2000-09,10034/184167499:2010,-7689/203709132</p>	(Beleg) (KoGra-DB) (loschen)
<input checked="" type="checkbox"/> parkieren	23.01.2019	C	token(parkieren)	M(1,2,3,4,5)R(1,2,3) D(1,2,3,4,5,0)L(1,2,3,4,5,6) J(6,7,8,9,0,1) I(0,1,2,4,5,6,7,8,9,3)	1446	<p>MEDIUM:Publikumspresse,1427/380060468:Bücher,0/1472363:Interne,5/76306651:Gesprochenes,14/28410491:Sonstiges,0/625010,</p> <p>REGISTER:Presse,1441/374110184:Gebrauch,5/111647373:Literarisch,0/1117426,</p> <p>DOMAENE:Fiktion,0/1256989:Kultur,515/201315611:Mensch,8/9269947:Politik,845/219928418:Technik,79372/1431:unklassifizierbar,2/21472587,</p> <p>LAND:D,16/397804684:Dost,0/153469:DWest,0/1851459:A,10/52453402:CH,1418/32952890:LU,2/1658989,</p> <p>REGION:überregional,6/62414992:Herkunft unbekannt,0/408257:Herkunft nicht zuordenbar,5/76358231:Mittelost,0/3722220:Mittelsüd,0/2385679:Mittelwest,3/102275873:Nordost,1/55180337:Nordwest,1/39773399:Südost,5/85319985:Südwest,1425/37564960,</p> <p>JAHR:-1969,0/513899:1970-79,0/111804:1980-89,0/147334:1990-99,2/4/98225315:2000-09,486/184167499:2010,-686/203709132</p>	(Beleg) (KoGra-DB) (loschen)

Abb. 1: Rechercheergebnisse in der KoGra-DB

Es können auch mehrere Abfragen miteinander verglichen werden. Dazu müssen sie in der Ergebnisliste markiert werden.

Die Daten der KoGra-DB sind mit Metadaten versehen, nämlich über das jeweilige Medium („Bücher“, „Gesprochenes“, „Internet“, „Publikumspresse“, „Sonstiges“), das Register („Presstexte“, „Gebrauchstexte“, „Literarische Texte“), das Land („Deutschland“, „Dtld. Ost“, „Dtld. West“, „Österreich“, „Schweiz“), die Region („überregional“, „Mittelost“, „Mittelsüd“, „Mittelwest“, „Nordost“, „Nordwest“, „Südost“, „Südwest“, „Herkunft nicht zuordenbar“, „Herkunft unbekannt“), die Domäne („Fiktion“, „Kultur/Unterhaltung“, „Mensch/Natur“, „Politik/Wirtschaft/Gesellschaft“, „Technik/Wissenschaft“, „unklassifizierbar“) und das Jahrzehnt (1960er-, 1970er-, 1980er-, 1990er-, 2000er-, 2010er-Jahre).⁷ Die Metadaten induzieren unterschiedliche Klassifizierungen der Rechercheergebnisse, die jeweils separat statistisch ausgewertet werden. Es wird geprüft, wie sich ein bestimmtes Phänomen über die verschiedenen Medien, Register, Länder etc. verteilt. Entsprechend werden die Verteilungen über die durch die Metadaten definierten Klassen separat berechnet und angezeigt.⁸ Die verschiedenen Auswertungen können, wie im entsprechenden Screenshot (siehe Abb. 2) zu sehen, über Reiter angesteuert werden.

⁷ Zur Auswahl und Beschreibung der Metadaten siehe Bubenhofer/Konopka/Schneider (2014, S. 84ff.).

⁸ Vgl. Abschnitt 3.

KoGra-R: Statistische Analysen für KograDB-Abfragen

Unten stehen die im Projekt "Korpusgrammatik" standardmäßig durchzuführenden statistischen Analysen für die gewählten KograDB-Abfragen. Unter dem Reiter "Allgemeine Information" stehen die Kenndaten zu den gewählten Abfragen. Unter der folgenden Reitern "Medium", "Land", "Region", "Domaene" und "Jahr" stehen die Analyseergebnisse zu den entsprechenden als Metadaten kodierten Klassen. "Aggregierte Daten" sind die Summen der Spalten der MEDIUM-Tabellen. Die Analyse der "aggregierten_Daten" werten die Rechercheergebnisse ohne Berücksichtigung einer Metadaten-basierten Klassifizierung aus.

[[Dokumentation der statistischen Analysen](#)] [[Archiv mit den verwendeten R-Skripts](#)] [[Eingabe von Cosmas-Frequenzlisten oder frei erzeugten Tabellen](#)]

Allgemeine Information zu den Abfragen
MEDIUM
REGISTER
LAND
REGION
DOMAENE
JAHR
aggregierte Daten

MEDIUM

R-Code zur Tabellenerzeugung

```
TABLE <- rbind(
  c(0),
  c(14),
  c(5),
  c(1427),
  c(0)
)

rownames(TABLE) <- c('Buecher', 'Gesprochenes', 'Internet', 'Publikumspresse', 'Sonstiges')
colnames(TABLE) <- c('parkieren')
```

```
corpus.TABLE <- rbind(
  c(1472363),
  c(28410491),
  c(76306651),
  c(380066460),
  c(625010)
)

rownames(corpus.TABLE) <- c('Buecher', 'Gesprochenes', 'Internet', 'Publikumspresse', 'Sonstiges')
colnames(corpus.TABLE) <- c('parkieren-Korpus')
```

[Erläuterung]

Abb. 2: Ergebnisse einer KoGra-DB-Auswertung

KoGra-DB-Rechercheergebnisse enthalten stets die Informationen über die Teilkorpusgrößen der verschiedenen Metadatenkategorien als Bezugsgrößen (= Korpusbezugsgrößen). Diese werden zur Berechnung u.a. von relativen Häufigkeiten verwendet.⁹

2.2 Eingabe von COSMAS-II-Frequenzlisten

COSMAS II ist ein im Web öffentlich zugängliches Recherche- und -analyse-system für die Korpora des Instituts für Deutsche Sprache (www.ids-mannheim.de/cosmas2, letzter Zugriff: 27.11.2018). Ergebnisse von COSMAS-II-Recherchen können exportiert und als Dateien lokal gespeichert werden. Die Dateien enthalten sogenannte Ergebnisansichten (Ansicht nach Quellen, Korpora, Dokumenten, Ländern, Textsorten, Themen, Jahrzehnten, Jahren, Monaten ...) und optional entsprechende Korpusansichten. Eine Ergebnisansicht enthält die Verteilung eines Rechercheergebnisses über den gewählten Bereich (die Quellen, Dokumente, Korpora ...). Diese Verteilung wird von KoGra-R ausgewertet, was analog zur oben beschriebenen Auswertung der Metadaten-induzierten Verteilungen von KoGra-DB-Ergebnissen geschieht. Eine Korpusansicht enthält die entsprechende Häufigkeitsverteilung für die einzelnen Teile des Gesamtkorpus. Diese können als Bezugsgröße zur Berechnung u.a. von normierten Häufigkeiten verwendet werden. Während bei KoGra-DB-

⁹ Vgl. hierzu genauer die Beschreibung in Abschnitt 3.

Auswertungen die Korpusbezugsgrößen stets berücksichtigt werden, ist ihre Angabe bei COSMAS-II-Auswertungen optional.

Es ist möglich, mehrere Exportdateien zugleich in KoGra-R zu laden und dadurch Ergebnisse verschiedener COSMAS-II-Recherchen miteinander zu vergleichen. Voraussetzung dafür ist, dass alle Dateien Ergebnis- und Korpusansichten des gleichen Typs enthalten.¹⁰

2.3 Freie Eingabe von Nutzer-definierten Tabellen

Mit KoGra-R besteht die Möglichkeit, selbst definierte Tabellen auszuwerten (vgl. hierzu das Eingabefeld in Abb. 3). Diese Tabellen müssen aus einer Kopfzeile, einer Kopfspalte und Datenfeldern bestehen. Die Datenfelder dürfen nur Zahlen enthalten. Zusätzlich zum eigentlichen Datenwert kann ein Datenfeld eine zweite Zahl als Bezugsgröße enthalten. Diese Zahl ist durch einen Schrägstrich von der ersten Zahl zu trennen. Beispiel: 3079/6914444 – 3079 dürfte im Regelfall die festgestellte Häufigkeit einer untersuchten Realisierung sein (z.B. eine bestimmte Anzahl von Tokens), 6914444 ist die entsprechende Größe der Korpusgesamtheit (z.B. die Anzahl der in Betracht gezogenen Tokens). Die Angabe der Bezugsgröße ist optional, muss aber, wenn sie in einer Zelle der Tabelle erfolgt, in allen anderen Zellen ebenfalls vorhanden sein.

Eine Tabelle kann als lokal gespeicherte CSV-Datei hochgeladen werden. Anders als bei der Übergabe von COSMAS-II-Exportdateien kann nur *eine* Datei (mit *einer* Tabelle) geladen werden; es ist nicht möglich, mehrere Dateien zugleich auszuwerten.¹¹ Alternativ kann die Tabelle in einem Textfeld erstellt oder in das Textfeld kopiert und von dort aus übergeben werden. Auf diese Weise können z.B. Tabellen aus Word-Dokumenten übergeben werden.

¹⁰ Es ist beispielsweise nicht möglich, eine Ansicht nach Themen mit einer Ansicht nach Jahren zu vergleichen.

¹¹ Dieses Vorgehen wäre auch nicht sinnvoll, da die Informationen über die Häufigkeiten verschiedener Realisierungen direkt in einer Kontingenztabelle angegeben werden können.

Freie Eingabe von Nutzer-definierten Tabellen

⇒ ⇒

Datei hochladen

Tabellen können in einem CSV-Format geladen werden. Zahlen zur Zusammensetzung des verwendeten Gesamtkorpus können durch einen Schrägstrich getrennt von den recherchierten Häufigkeiten angegeben werden: recherchierte Häufigkeit/Korpusgesamtheit (z.B. 3866/535565). Die geladene Tabelle wird geprüft, bevor sie statistisch ausgewertet wird. [[ausführlichere Erläuterungen zur Eingabe Nutzer-definierter Tabellen](#)]

Datei auswählen: keine Datei ausgewählt.

oder

Tabelle direkt eingeben

Tabellen können in einem CSV-Format eingegeben werden. Die Verwendung des Tabulators als Feldtrenner kann bei der Eingabe zu Problemen führen, weil die Tab-Taste bei HTML-Formularen wie diesem standardmäßig mit einer Sonderfunktion belegt ist. Zahlen zur Zusammensetzung des verwendeten Gesamtkorpus können durch einen Schrägstrich getrennt von den recherchierten Häufigkeiten angegeben werden: recherchierte Häufigkeit/Korpusgesamtheit (z.B. 3866/535565). Die eingegebene Tabelle wird geprüft, bevor sie statistisch ausgewertet wird. [[ausführlichere Erläuterungen zur Eingabe Nutzer-definierter Tabellen](#)]

```
eigentlich;tatsächlich
Gebrauch;3079/6914444;1754/6914444
Literarisch;884/182097;433/182097
Presse;1338/291657;834/291657
```

Abb. 3: Eingabe einer Nutzer-definierten Tabelle

Die geladenen oder eingegebenen Tabellen werden nicht direkt ausgewertet, sondern zuerst auf ihre Wohlgeformtheit geprüft. Das Ergebnis der Tabellenprüfung wird angezeigt. Falls die Tabelle nicht wohlgeformt ist, werden Fehlermeldungen gegeben und potenzielle Fehler markiert. Die in Abbildung 3 eingegebene Tabelle enthält einen Fehler. Dieser wird nach der Prüfung angezeigt (Abb. 4).

KoGra-R: Prüfung einer frei erzeugten Tabelle

[Eingabe von Cosmas-Frequenzlisten oder frei erzeugten Tabellen]

Prüfen und Auswerten der eingegebenen Tabelle

⇒ ⇒

Tabelle auswerten

eigentlich	tatsächlich	
Gebrauch	3079/6914444	1754/6914444
Literarisch	884/182097	433/182097
Presse	1338/291657	834/291657

- Anzahl der Spalten gemäß Kopfzeile: 1 (ohne Kopfspalte)
- Anzahl der Zeilen: 3 (ohne Kopfzeile)
- Es sind Bezugsdaten für die Auswertung vorhanden (Korpusgesamtheiten).
- Das erste Feld der Kopfzeile/-spalte ist nicht leer.
- Mindestens eine Zeile enthält eine andere Anzahl von Feldern als die Kopfzeile.

Die Tabelle kann noch nicht ausgewertet werden.

Tabelle ändern

Tabellen werden in einem CSV-Format eingegeben. Die Verwendung des Tabulators als Feldtrenner kann bei der Eingabe zu Problemen führen, weil die Tab-Taste bei HTML-Formularen wie diesem standardmäßig mit einer Sonderfunktion belegt ist. Zahlen zur Zusammensetzung des verwendeten Gesamtkorpus können durch einen Schrägstrich getrennt von den recherchierten Häufigkeiten angegeben werden: recherchierte Häufigkeit/Korpusgesamtheit (z.B. 3866/535565). Die eingegebene Tabelle muss geprüft werden, bevor sie statistisch ausgewertet werden kann.

```
eigentlich;tatsächlich
Gebrauch;3079/6914444;1754/6914444
Literarisch;884/182097;433/182097
Presse;1338/291657;834/291657
```

Abb. 4: Ergebnis der Prüfung einer nicht wohlgeformten Tabelle

Nicht wohlgeformte Tabellen wie die obige können im Textfeld korrigiert und erneut geprüft werden, bis sie wohlgeformt sind. Sobald die Tabelle wohlgeformt ist, d.h., sobald alle Fehler korrigiert sind, kann sie ausgewertet werden (Abb. 5).

KoGra-R: Prüfung einer frei erzeugten Tabelle

[Eingabe von Cosmas-Frequenzlisten oder frei erzeugten Tabellen]

Prüfen und Auswerten der eingegebenen Tabelle

Tabelle eingeben → Tabelle prüfen → Tabelle auswerten

Tabelle auswerten

	eigentlich	tatsächlich
Gebrauch	3079/6914444	1754/691444
Literarisch	884/182097	433/182097
Presse	1338/291657	834/291657

- Anzahl der Spalten gemäß Kopfzeile: 2 (ohne Kopfspalte)
- Anzahl der Zeilen: 3 (ohne Kopfzeile)
- Es sind Bezugsdaten für die Auswertung vorhanden (Korpusgesamtheiten).

Tabelle auswerten.

Tabelle ändern

Tabellen werden in einem CSV-Format eingegeben. Die Verwendung des Tabulators als Feldtrenner kann bei der Eingabe zu Problemen führen, weil die Tab-Taste bei HTML-Formularen wie diesem standardmäßig mit einer Sonderfunktion belegt ist. Zahlen zur Zusammensetzung des verwendeten Gesamtkorpus können durch einen Schrägstrich getrennt von den recherchierten Häufigkeiten angegeben werden: recherchierte Häufigkeit/Korpusgesamtheit (z.B. 3866/535565). Die eingegebene Tabelle muss geprüft werden, bevor sie statistisch ausgewertet werden kann.

eigentlich;tatsächlich
 Gebrauch;3079/6914444;1754/691444
 Literarisch;884/182097;433/182097
 Presse;1338/291657;834/291657

Abb. 5: Ergebnis der Prüfung einer wohlgeformten Tabelle

3. Statistische Analysen

Wie oben beschrieben, basieren sämtliche Funktionen in KoGra-R auf Häufigkeitsangaben. Zu den statistischen Auswertungen ist in KoGra-R eine Dokumentation verfügbar, zu der man über diverse Hyperlinks im Tool geleitet wird. Dort wird außerdem an den entsprechenden Stellen der R-Code zur Verfügung gestellt, der zur jeweiligen Auswertung und Erstellung der Ausgabe nötig ist.

Je nach Menge und Art der Informationen wird durch das Tool ein passender Analysemodus gewählt. Der passende Modus ergibt sich aus der Anzahl der Spalten in der Tabelle sowie dem Vorhandensein von Korpusbezugsgrößen. Die meisten statistischen Analysen können mit mehrspaltigen Tabellen mit Korpusbezugsgrößen durchgeführt werden. Der typische Anwendungsfall ist hier der Vergleich von Häufigkeiten zweier Abfragen, bei denen jeweils die Größe des zugrundeliegenden Korpus bekannt ist.

Im Rahmen der Auswertung werden die Daten zunächst in verschiedenen Formaten dargestellt. In einem zweiten Schritt erfolgen tiefer gehende deskriptive und inferenzstatistische Analysen und entstehen Darstellungen, die

im Folgenden beschrieben werden. Hierzu werden zwei Abfragen mithilfe der KoGra-DB durchgeführt und in KoGra-R ausgewertet. Die beiden Abfragen beziehen sich auf zwei mögliche Realisierungen der Genitivformen des Nomens *Werk*. Die tokenbasierten Abfragen (*Werkes* und *Werks*) werden in Bezug auf die Häufigkeit der beiden Varianten in verschiedenen Jahrzehnten miteinander verglichen, wobei den Jahrzehnten unterschiedlich große Teilkorpora zugrunde liegen. Zusätzlich wird die Häufigkeit der Realisierungen in Abhängigkeit von verschiedenen Medien betrachtet, denen ähnlich große Teilkorpora zugrunde liegen. Bei den Ausgangsdaten der Analysen handelt es sich um mehrspaltige Tabellen mit entsprechenden Korpusbezugsgrößen.¹²

3.1 Darstellung der Rohdaten

Die Rohdaten werden in Form von Kontingenztabelle dargestellt (vgl. Tab. 1).

	<i>Werkes</i>	<i>Werks</i>
–1969 ¹³	83	19
1970–79	23	3
1980–89	7	5
1990–99	13629	8751
2000–09	16786	12433
2010–	16024	11353

Tab. 1: Rohdaten der Abfragen *Werks* vs. *Werkes* im DeReKo¹⁴

Die Tabelle gibt die absolute Trefferzahl für die übergebenen Abfragen an. Die Spalten repräsentieren die Art der verschiedenen Abfragen bzw. die Varianten. In den Zeilen sind die verschiedenen Ausprägungen des jeweiligen Metadatum (hier Jahrzehnte) abgetragen. Die Wortform *Werkes* kommt demnach z.B. 23-mal zwischen 1970 und 1979 vor, in den Jahren zwischen 2000 und 2009 ist sie 16.786-mal vertreten. Um die enormen Häufigkeitsschwankungen bezogen auf die Jahrzehnte beurteilen und interpretieren zu können, muss man die Teilkorpusgrößen (hier: Anzahl aller Tokens in den betreffenden Jahren) zurate ziehen.¹⁵ Die zweite Tabelle in KoGra-R (vgl. Tab. 2) zeigt diese Information für die einzelnen Abfragen an.

¹² Vgl. die Kontingenztabelle mit Beispieldaten auf der Startseite von KoGra-R.

¹³ Die ältesten Texte des in KoGra-DB abfragbaren DeReKo-Ausschnitts stammen von 1955.

¹⁴ Datengrundlage ist das DeReKo-Release 2014-II (vgl. Institut für Deutsche Sprache 2014).

¹⁵ Da in diesem Beispiel eine auf Tokens basierte Abfrage durchgeführt wurde, werden die Teilkorpusgrößen ebenfalls in Tokens (laut TreeTagger) angegeben. Je nach Abfrage- bzw. Phä-

	<i>Werkes</i> -Teilkorpus	<i>Werks</i> -Teilkorpus
-1969	8061722	8061722
1970-79	2120012	2120012
1980-89	2912554	2912554
1990-99	1715718062	1715718062
2000-09	2952571847	2952571847
2010-	2803516611	2803516611

Tab. 2: Korpusbezugsgrößen für die übergebenen Abfragen

Ein Blick auf die Tabelle zeigt, dass die Anzahl an Tokens in den Teilkorpora ab 1990 stark ansteigt. Es verwundert also nicht, dass auch die in den beiden Abfragen ermittelten Häufigkeiten der Realisierungen von *Werkes* respektive *Werks* in diesen Jahren zunehmen.

3.2 Darstellung von normierten Daten

Um Daten aus unterschiedlich großen Teilkorpora miteinander vergleichen zu können und diese deskriptiv zu untersuchen, ist es sinnvoll, die gefundenen Häufigkeiten zu normieren. Die normierten Daten werden in KoGra-R ebenfalls in Form einer Kontingenztabelle (vgl. Tab. 3) dargestellt. Dabei werden die Treffer pro eine Million Tokens angegeben. Hierzu wird die Tabelle mit der absoluten Anzahl der Treffer mit der Tabelle der Korpusbezugsgrößen entsprechend verrechnet.¹⁶

	<i>Werkes</i>	<i>Werks</i>
-1969	10.295567	2.356817
1970-79	10.848995	1.415086
1980-89	2.403389	1.716706
1990-99	7.943613	5.100488
2000-09	5.685213	4.210905
2010-	5.715679	4.049557

Tab. 3: Treffer pro eine Million Tokens

nomenart kann es sinnvoll sein, die Teilkorpusgrößen in Sätzen oder anderen Einheiten anzugeben.

¹⁶ Die normierten Werte werden deshalb in KoGra-R nur berechnet und visualisiert, wenn die Informationen über die Korpusbezugsgrößen vorliegen.

Die normierten Daten werden zusätzlich in einem Säulendiagramm abgetragen (vgl. Abb. 6). Die Säulen werden dabei nach den Ausprägungen des jeweiligen Metadatum gruppiert. Im vorliegenden Beispiel sind dies die verschiedenen Jahrzehnte, in denen die abgefragten Genitivformen von *Werk* vorkommen. Innerhalb jeder Säulengruppe befindet sich für jede Abfrage eine Säule.

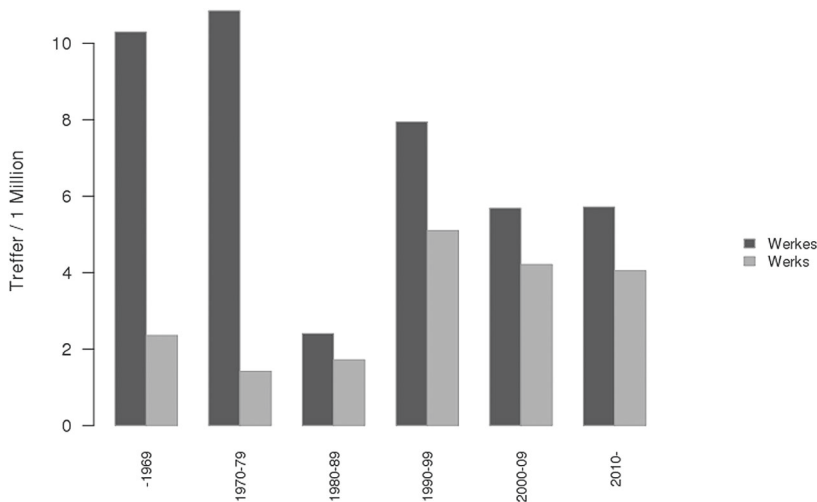


Abb. 6: Gruppierendes Säulendiagramm der normierten Werte¹⁷

Die Visualisierung der normierten Daten verdeutlicht, dass die genannten Genitivvarianten von *Werk* insgesamt rückläufig sind. Außerdem lässt sich feststellen, dass die lange Endung *-es* in Relation zur kürzeren Endung *-s* seltener wird.

3.3 Darstellung der relativen Werte

Die absoluten Werte geben an, wie oft ein Merkmal bzw. eine Variante vorkommt. Mit den relativen Häufigkeiten wird in KoGra-R angegeben, wie hoch der Anteil jedes Teilkorpus an der jeweiligen Variante ist. Sind die Teilkorpora ungefähr gleich groß, ist die Berechnung und Darstellung der relativen Werte sinnvoll und kann als alternative Analyseverfahren zur Normierung der Daten angewendet werden. Die eventuellen Affinitäten eines Teilkorpus zu einer bestimmten Variante werden so unmittelbar sichtbar gemacht. Die relativen Häufigkeiten werden berechnet, indem für jede Variante ihre absoluten Häufigkeiten in den Teilkorpora durch ihre Häufigkeit im Ge-

¹⁷ In diesem Beitrag werden alle Diagramme in Graustufen präsentiert. In KoGra-R sind die Grafiken farbig.

samtkorpus geteilt werden. Somit variiert die relative Häufigkeit immer zwischen 0 und 1. Die Summe der relativen Häufigkeiten für eine Variante beträgt 1. Multipliziert man die Zahl mit 100, erhält man den Prozentanteil des jeweiligen Teilkorpus für die entsprechende Variante.

Um dies zu illustrieren, untersuchen wir die regionale Verteilung von *Werks* und *Werkes* in unterschiedlichen Teilkorpora, die eine ungefähr gleich große Tokenanzahl aufweisen. Hierzu wurden die beiden Varianten in verschiedenen Zeitungen über COSMAS II abgefragt. Ausgewählt wurden Texte des Trierischen Volksfreunds aus den Jahren 2000–2016, der Ostthüringer Zeitung aus den Jahren 2000–2012, der Kleinen Zeitung aus den Jahren 2002–2016 und der Neuen Zürcher Zeitung von 2000–2016. Der Trierische Volksfreund ist dabei eine Zeitung aus Westmittelldeutschland, die Ostthüringer Zeitung ein Blatt aus Ostmittelldeutschland, die Kleine Zeitung eines aus Österreich, und die Neue Zürcher Zeitung wird in der Schweiz verlegt. In Tabelle 4 sind die absoluten Häufigkeiten sowie die Korpusbezugsgrößen (hier: Anzahl der Tokens) in Klammern hinter den Zeitungsnamen abgetragen.

	<i>Werks</i>	<i>Werkes</i>
Trierischer Volksfreund (347946108)	1765	1710
Ostthüringer Zeitung (335299068)	317	2317
Kleine Zeitung (338764250)	1004	1638
Zürcher Zeitung (314206790)	3793	2101

Tab. 4: Rohdaten und Korpusbezugsgrößen der Abfragen *Werks* vs. *Werkes* in den betreffenden Zeitungen

In einem ersten Schritt werden in KoGra-R die relativen Werte in Form einer Kontingenztafel dargestellt (vgl. Tab. 5).

	<i>Werks</i>	<i>Werkes</i>
Trierischer Volksfreund	25.657799	22.01906
Ostthüringer Zeitung	4.608228	29.83518
Kleine Zeitung	14.595145	21.09194
Neue Zürcher Zeitung	55.138828	27.05382

Tab. 5: Relative Werte der gefundenen Tokens pro Variante

Die Tabelle gibt pro Variante an, wie viel Prozent der Treffer einer Ausprägung des Metadatum (hier: Zeitungen aus verschiedenen Regionen) zuzuordnen sind. Somit summiert sich jede Spalte auf 100%.¹⁸

¹⁸ Relative Werte werden immer ausgegeben – mit Ausnahme von aggregierten Tabellen. In aggregierten Tabellen werden alle Zeilen zu einer summiert. Sie enthalten somit nur die jeweili-

In einem zweiten Schritt werden die prozentualen Verteilungen der Variantenrealisierungen über die Ausprägungen des Metadatum mithilfe eines gruppierten und eines gestapelten Säulendiagramms visualisiert. Im gruppierten Säulendiagramm (Abb. 7) sind die Säulengruppen über die Ausprägungen des jeweiligen Metadatum definiert.¹⁹

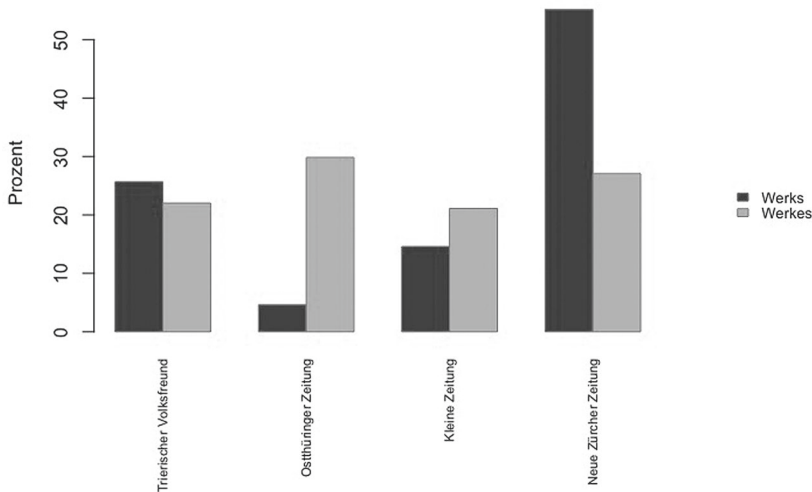


Abb. 7: Gruppiertes Säulendiagramm der relativen Werte

Diese Grafik ist u.U. schwieriger zu lesen, als es zunächst den Anschein hat. Es ist daher nochmals darauf hinzuweisen, dass eine Säule den **Anteil** der Realisierungen einer Variante bei einer Metadaten-Ausprägung am Gesamtvorkommen dieser Variante repräsentiert. Ist innerhalb einer Säulengruppe eine Säule höher, muss das also nicht bedeuten, dass hier mehr Vorkommen vorliegen. Die Darstellung der relativen Werte unterscheidet sich demnach inhaltlich von der der normierten Werte. Am vorliegenden Beispiel lässt sich anhand der relativen Werte feststellen, dass die Variante *Werkes* im Vergleich zur Variante *Werks* in der Ostthüringer Zeitung und der Kleinen Zeitung in Verhältnis überrepräsentiert ist, während sie in den anderen beiden Zeitungen unterrepräsentiert ist. In Kapitel 3.4.4 werden wir zeigen, dass die unterschiedliche Verteilung der Varianten auf die verschiedenen Zeitungskorpora höchstsignifikant ist. Vor allem in der Neuen Zürcher Zeitung kommt die Variante *Werks* relativ häufig (in 55% aller Fälle der Variante) vor. Die Variante mit der Endung *-es* scheint somit für die Presstexte aus der Ostthüringer

gen Gesamtvorkommen des Phänomens. Aggregierte Tabellen können in KoGra-R über den Reiter „aggregierte Daten“ angefordert werden.

¹⁹ Innerhalb jeder Säulengruppe befindet sich pro Abfrage eine Säule. Wird lediglich eine Abfrage zu einem Einzelphänomen analysiert, besteht jede Gruppe folgerichtig aus nur einer Säule.

Zeitung und für die österreichische Zeitung dominierend zu sein.²⁰ Diese Verteilung bestätigt die Ergebnisse aus einer Studie von Fürbacher (2015), in der die Verteilung der Genitivallomorphe *-es* und *-s* untersucht wird. In dieser Studie konnte festgestellt werden, dass Tokens mit *-es* in Texten aus dem Osten und Südosten (einschließlich Österreichs) höchstsignifikant häufiger vorkommen als Formen mit der Endung *-s* (vgl. ebd.).

Bei der Darstellung der relativen Werte in KoGra-R bietet es sich an, über die Tabelle und das gruppierte Säulendiagramm hinaus ein gestapeltes Säulendiagramm zu erstellen. Die Verhältnisse in den einzelnen Subkorpora können so auf zwei Arten visuell erfasst werden.

Das gestapelte Säulendiagramm (Abb. 8) zeigt die prozentuale Verteilung der jeweiligen Variante in gestapelten Rechtecken an, die den Ausprägungen der Metadaten (hier: regionale Zeitungen) entsprechen. Die Ausprägungen werden in KoGra-R mit unterschiedlichen Farben kodiert. Jede Säule korrespondiert mit einer Variante. Es werden in alternativer Weise exakt dieselben Werte visualisiert wie im vorigen Diagramm.

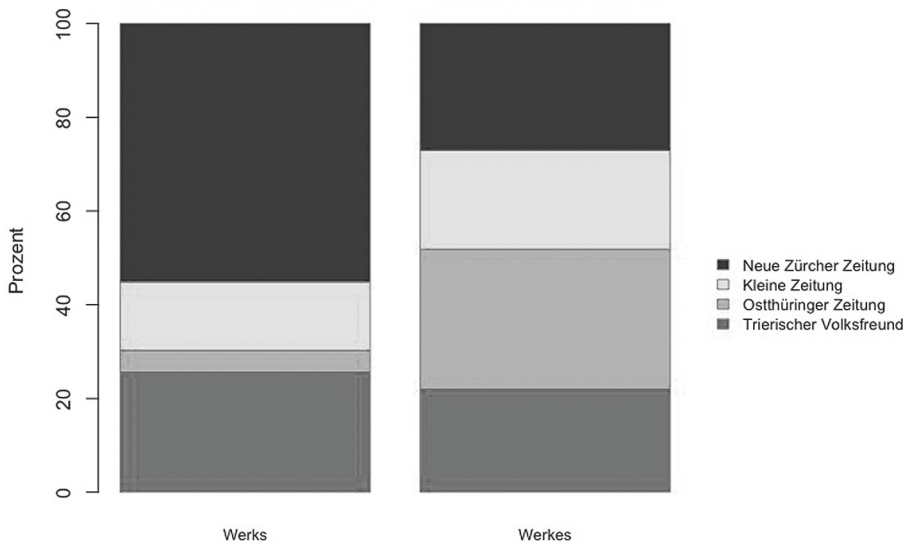


Abb. 8: Gestapeltes Säulendiagramm der relativen Werte

²⁰ Um weitere Aussagen zu treffen, müsste die Verteilung der Genitivallomorphe durch genauere quantitative und qualitative Auswertungen auf einer breiteren Datenbasis untersucht werden.

3.4 Analysen im Rahmen der Chi-Quadrat-Statistik

3.4.1 Der Chi-Quadrat-Test

Um zu prüfen, ob die Ausprägungen von zwei nominalen Variablen voneinander unabhängig sind, eignet sich der Chi-Quadrat-Test (siehe Bortz 2005, S. 168ff.).²¹ Hierzu werden die beobachteten Häufigkeiten mit erwarteten Häufigkeiten verglichen, die im Verhältnis zur jeweiligen Größe der Teilkorpora stehen.²² Als Testergebnis werden verschiedene Werte ausgegeben.

Aus dem berechneten Chi-Quadrat-Wert (Prüfgröße) und den sog. Freiheitsgraden (= *df*) kann die Wahrscheinlichkeit errechnet werden, durch Zufall eine solche Stichprobe aus einer Population zu ziehen, in der es keinen Zusammenhang zwischen den beiden Merkmalen gibt (vgl. ebd.). Ein Testergebnis wird als statistisch signifikant bezeichnet, wenn diese Wahrscheinlichkeit unterhalb einer konventionell gesetzten Schwelle liegt – i.d.R. 0,05 oder kleiner (vgl. ebd., S. 114).²³

Für die Verteilung der beiden Genitivvarianten *Werkes* und *Werks* in Abhängigkeit von Jahrzehnten ergibt der Chi-Quadrat-Test folgende Werte (so auch die Darstellung in KoGra-R):

Pearson's Chi-squared test
X-squared = 94.3741, df = 5, p-value < 2.2e-16

Der Wert der Prüfgröße (= X-squared) beträgt gerundet 94 bei 5 Freiheitsgraden, die daraus berechnete Irrtumswahrscheinlichkeit *p* ist deutlich kleiner als 0,05. Aus diesem Ergebnis ist abzulesen, dass sich die Verteilungen der Realisierungen der beiden Varianten in Abhängigkeit von Jahrzehnten höchstsignifikant unterscheiden. Allerdings ist die Anwendung des Chi-Quadrat-Tests in diesem Punkt nicht unproblematisch, weil größere absolute Häufigkeiten (unter Konstanzhaltung der Verhältnisse) direkte Auswirkungen auf die Höhe des *p*-Werts haben. Multipliziert man z.B. alle Zahlen mit 10, blei-

²¹ Es sind andere Maße und Tests denkbar, die alternativ oder zusätzlich zum Chi-Quadrat-Test berechnet werden können (siehe Wiechmann 2008). Wir haben uns dafür entschieden, den Chi-Quadrat-Test in KoGra-R zu implementieren. Der wichtigste Grund für diese Entscheidung war, dass der Chi-Quadrat-Test am allgemeinsten anwendbar ist, weil er u.a. auch auf Tabellen angewendet werden kann, die mehr als zwei Spalten/Zeilen haben. Bei Bedarf werden im Projekt Korpusgrammatik allerdings auch andere Maße berechnet, z.B. der exakte Test nach Fisher, wenn der Stichprobenumfang zu gering ist.

²² Die erwarteten Häufigkeiten sollten bei einem Chi-Quadrat-Test nicht in über 20% der Zellen unter 5 fallen (siehe Bortz 2005, S. 177).

²³ Laut Konvention spricht man ab einer Irrtumswahrscheinlichkeit kleiner 5% ($p < 0,05$) von einem signifikanten, ab 1% ($p < 0,01$) von einem hochsignifikanten und ab 0,1% ($p < 0,001$) von einem höchstsignifikanten Ergebnis.

ben die Verhältnisse zwar exakt gleich, aber der p-Wert wird kleiner. Daher sollte im Rahmen einer Chi-Quadrat-basierten Auswertung immer die Assoziationsstärke mit angegeben werden. Diese wird von der allgemeinen Größenordnung der absoluten Häufigkeiten nicht beeinflusst (z.B. 100 vs. 300 Tokens im Vergleich zu 10000 vs. 30000 Tokens).²⁴

3.4.2 Assoziationsstärke: Phi/Cramérs V

Nicht nur die statistische Signifikanz ist für die Bedeutsamkeit eines Ergebnisses ausschlaggebend. Auch die Stärke der Korrelation zweier Variablen ist relevant. Der Phi-Koeffizient ist ein Maß für die Stärke des Zusammenhangs zweier dichotomer Merkmale, die in einer Tabelle mit zwei Spalten und zwei Zeilen abgetragen werden (siehe Bortz/Lienert 2008, S. 259ff.). Cramérs V ist ein Kontingenzkoeffizient, der bei jeder Kontingenztabelle, unabhängig von der Anzahl der Zeilen und Spalten, berechnet werden kann (vgl. ebd., S. 271ff.).²⁵ Cramérs V wird berechnet mit $V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}}$, wobei n die Gesamtanzahl der Fälle angibt und k das Minimum der Anzahl der Spalten und der Anzahl der Zeilen der Kontingenztabelle. Für Tabellen mit zwei Spalten und zwei Zeilen sind die Werte von Phi und Cramérs V identisch. Phi kann mit $\Phi = \sqrt{\frac{\chi^2}{n}}$ berechnet werden, da k für Vier-Felder-Tabellen immer 2 ist.

Beide Koeffizienten weisen ein Wertespektrum zwischen 0 und 1 auf. Je höher der Wert ist, desto stärker ist der Zusammenhang. Für die Abfrage *Werkes* vs. *Werks* in Abhängigkeit von Jahrzehnten ist die Ausgabe in KoGra-R wie folgt:

```
Phi/V-Koeffizient
0.03453776
```

Zwar unterscheiden sich die Verteilungen der beiden Abfragen in Abhängigkeit von Jahrzehnten signifikant voneinander, allerdings ist die Assoziationsstärke mit einem Koeffizienten von 0,03 als sehr klein einzuschätzen.

3.4.3 Erwartete Häufigkeiten und Residuen

Um beurteilen zu können, ob die Abweichungen der beobachteten von den erwarteten Häufigkeiten in den Einzelzellen der Kreuztabelle ausschlaggebend für die Signifikanz sind oder nicht, werden sog. standardisierte Residuen nach Pearson berechnet (siehe Backhaus et al. 2016, S. 350). Die absoluten

²⁴ Der Chi-Quadrat-Test wird in KoGra-R für mehrspaltige und nicht-aggregierte Tabellen berechnet.

²⁵ Die Assoziationsstärke Phi bzw. V wird in KoGra-R für mehrspaltige, nicht-aggregierte Tabellen berechnet.

Abweichungen in jeder Zelle der Tabelle werden durch die Quadratwurzel der erwarteten Häufigkeiten dividiert. Auf diese Weise kann gezeigt werden, welche Häufigkeitsunterschiede in der Kontingenztabelle hauptsächlich für den Effekt verantwortlich sind. In unserem Beispiel lässt sich so überprüfen, in welchem Jahrzehnt die beobachteten Häufigkeiten von den erwarteten Häufigkeiten besonders auffällig abweichen.

In KoGra-R werden die erwarteten Werte und die standardisierten Residuen nach Pearson angegeben. Für unser Abfragebeispiel sind die Ausgaben in Tabelle 6 und Tabelle 7 dargestellt.

	<i>Werkes</i>	<i>Werks</i>
-1969	60.016988	41.983012
1970–79	15.298448	10.701552
1980–89	7.060822	4.939178
1990–99	13168.433187	9211.566813
2000–09	17192.513373	12026.486627
2010–	16108.677183	11268.322817

Tab. 6: Darstellung der erwarteten Häufigkeiten

Es werden zum Beispiel in den Jahren bis 1969 ca. 60 Wortformen mit *es*-Endung erwartet und ca. 42 Tokens mit *s*-Endung. In den Jahren zwischen 1990 und 1999 wird das Token *Werkes* ca. 13168-mal erwartet, das Token *Werks* allerdings 9211-mal.

Die Pearson-Residuen (Tab. 7) deuten bei einer Irrtumswahrscheinlichkeit von 5% auf die Signifikanz der Unterschiede hin, wenn sie über 1,97 oder unter -1,97 liegen (Field/Miles/Field 2012, S. 826). Ist ein standardisiertes Residuum kleiner als -1,97, weichen in dieser Zelle die beobachteten Häufigkeiten signifikant nach unten von den erwarteten Häufigkeiten ab. Ist das standardisierte Residuum größer als 1,97, weichen in dieser Zelle die beobachteten Häufigkeiten signifikant nach oben von den erwarteten Häufigkeiten ab.

	<i>Werkes</i>	<i>Werks</i>
-1969	2.96667418	-3.54707320
1970–79	1.96904017	-2.35426244
1980–89	-0.02288936	0.02736743
1990–99	4.01352145	-4.79872528
2000–09	-3.10031022	3.70685374
2010–	-0.66716993	0.79769480

Tab. 7: Darstellung der standardisierten Pearson-Residuen

An der Beispieltabelle ist abzulesen, dass alle Residuen bis auf die zwischen 1980 und 1989 und die ab 2010 signifikant sind. Bis zum Jahr 1969 ist das Token *Werkes* signifikant überrepräsentiert und das Token *Werks* signifikant unterrepräsentiert (die Residuen befinden sich bei ca. 2,97 bzw. -3,55).²⁶ Ähnlich verhält es sich in der Zeit zwischen 1970 und 1979. Zwischen 1980 und 1989 gibt es dagegen keine signifikanten Abweichungen von der erwarteten Verteilung der beiden Realisierungen. In der Zeit zwischen 1990 und 1999 wiederum sind die standardisierten Residuen am höchsten. Das heißt, dass in diesen Jahren die beobachteten Werte stark nach oben (für *Werkes*) bzw. unten (für *Werks*) von den erwarteten Werten abweichen.²⁷ In den nachfolgenden Jahren kehrt sich dieses Verhältnis um: Die Wortform mit der *es*-Endung ist signifikant unterrepräsentiert, während das Token *Werks* signifikant häufiger vorkommt als erwartet.

3.4.4 Assoziationsplot

Der Assoziationsplot visualisiert die standardisierten Pearson-Residuen in Abhängigkeit von den Metadatenausprägungen (z.B. den Jahrzehnten, vgl. Abb. 9). Die Höhe der Balken entspricht der Höhe der Abweichung: Balken oberhalb der gepunkteten Linie bedeuten höhere Werte als erwartet. Balken unterhalb der Linie bedeuten, dass die Werte niedriger sind als erwartet. Die Breite der Balken ist definiert durch die erwartete Häufigkeit der Realisierungen. Sind die Pearson-Residuen signifikant ($> 1,97$ oder $< -1,97$) werden sie in KoGra-R im Plot eingefärbt.²⁸ Wird die Schwelle von ± 4 überschritten, werden die Balken in einem stärkeren Ton eingefärbt, da hier von einer ganz besonders starken Abweichung ausgegangen werden kann (siehe Cohen 1980; Friendly 1992; Meyer/Zeileis/Hornik 2005).²⁹

²⁶ Im Allgemeinen kann davon ausgegangen werden, dass mindestens eine Zelle signifikant über- oder unterrepräsentiert ist, wenn der Chi-Quadrat-Test insgesamt eine signifikante Abweichung der beobachteten von den erwarteten Häufigkeiten anzeigt. Das gilt auch in die andere Richtung (signifikanter Chi-Quadrat-Wert bei mindestens einer signifikanten Zellenabweichung).

²⁷ Gründe dafür könnten evtl. durch genauere quantitative und qualitative Auswertung der Belege aus diesem Zeitraum ermittelt werden.

²⁸ In KoGra-R steht die Farbe Blau für signifikante Abweichungen nach oben, die Farbe Rot für signifikante Abweichungen nach unten.

²⁹ Der Assoziationsplot wird nur für mehrspaltige und nicht-aggregierte Tabellen ausgegeben.

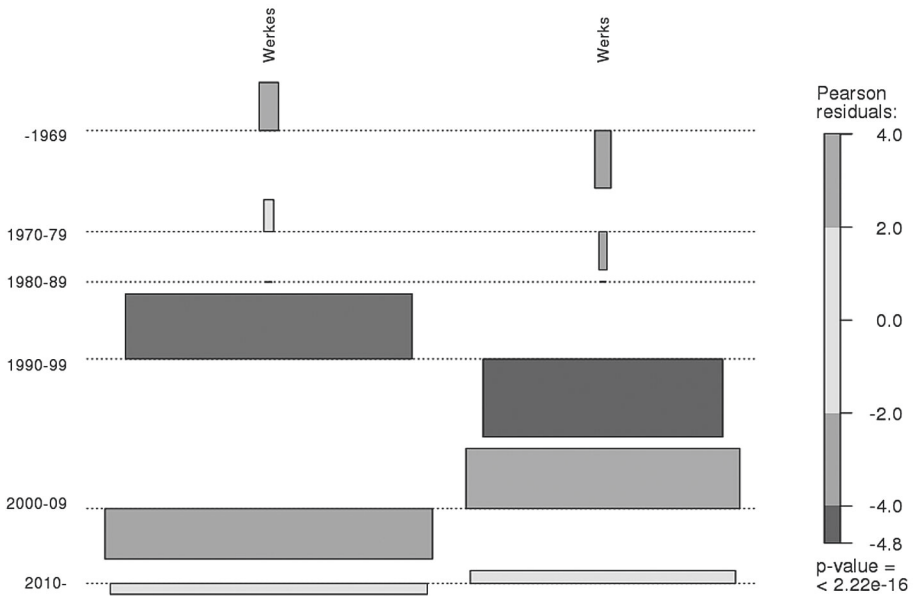


Abb. 9: Assoziationsplot für die Realisierungen *Werks* vs. *Werkes* in Abhängigkeit von Jahrzehnten

Der Plot visualisiert die Residuen, die in Tabelle 7 angegeben wurden, und verdeutlicht, dass die Realisierung des Nomens *Werk* mit der Endung *-s* in neueren Texten signifikant überrepräsentiert ist, wogegen es bei der Realisierung mit der Endung *-es* in älteren Texten der Fall ist.

Wir kommen an dieser Stelle nochmals auf die in Kapitel 3.3 (Darstellung der relativen Werte) besprochene Analyse der Variation der beiden Genitivallomorphe in Zeitungen aus verschiedenen Regionen zurück und stellen die Verteilung der Residuen dieser Analyse in einem Assoziationsplot dar (Abb. 10):

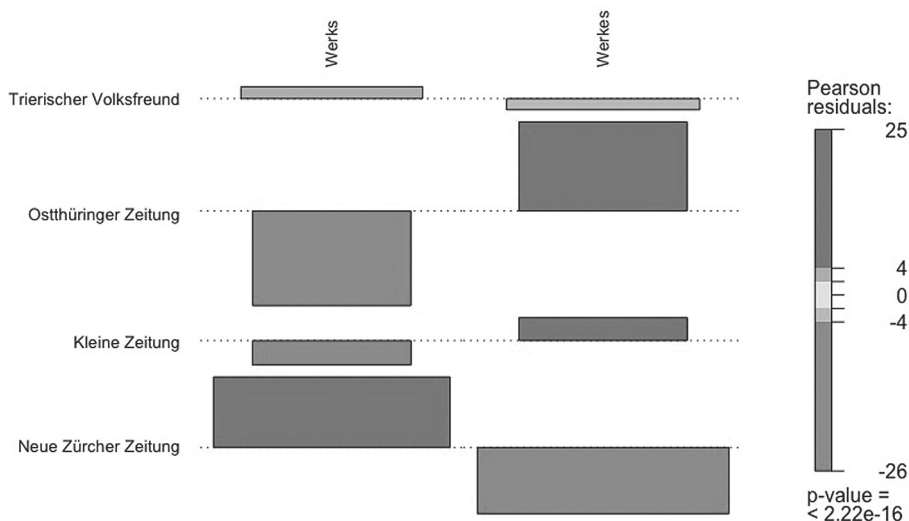


Abb. 10: Assoziationsplot für die Realisierungen *Werks* vs. *Werkes* in Abhängigkeit von verschiedenen regionalen Zeitungen

Der Assoziationsplot bestätigt die Schlussfolgerung aus der Verteilung der relativen Häufigkeiten: *Werkes* ist in der Ostthüringer Zeitung und der Kleinen Zeitung signifikant überrepräsentiert, während *Werks* in diesen Presse-texten signifikant unterrepräsentiert ist. Die Variante mit *-es* ist hingegen im Trierischen Volksfreund sowie in der Neuen Zürcher Zeitung signifikant unterrepräsentiert, während die kürzere *s*-Variante in diesen Texten signifikant überrepräsentiert ist.

3.4.5 Mosaikplot

Eine andere Möglichkeit, um Zusammenhänge zwischen Daten zu visualisieren, bietet der Mosaikplot (siehe Hartigan/Kleiner 1984; Friendly 1994; Emerson 1998; Meyer/Zeileis/Hornik 2005). Er visualisiert die Häufigkeitsverteilung in einer Kontingenztabelle über Rechtecke, deren Größe proportional zur Fallzahl in den Zellen ist (Abb. 11). Die Farbe der Rechtecke bezieht sich – ähnlich wie beim Assoziationsplot – auf die standardisierten Pearson-Residuen. Ist ein Rechteck in KoGra-R rot eingefärbt, ist die Ausprägung in der entsprechenden Zelle unterrepräsentiert. Ist ein Rechteck blau eingefärbt, ist die Ausprägung in der entsprechenden Zelle überrepräsentiert. Stärkere Abweichungen werden wiederum mit einer stärkeren Einfärbung signalisiert.

Im Mosaikplot werden im Gegensatz zum Assoziationsplot zudem die relativen Häufigkeiten dargestellt. Dabei wird jede Variable (hier: Phänomenrealisierungen der beiden abgefragten Genitivformen von *Werk* und die Informa-

tion über die Jahrzehnte) einer Achse zugeordnet. Am linken Rand ist die erste Variable abgetragen, die sich im vorliegenden Beispiel auf die beiden abgefragten Genitivvarianten bezieht. Die gesamten Daten werden zunächst in Blöcke unterteilt, die durch die Ausprägungen der Variablen definiert werden. Die Höhe einer Zelle repräsentiert die relative Häufigkeit der jeweiligen Ausprägung. Am oberen Rand wird die zweite Variable abgetragen, die sich im vorliegenden Beispiel auf das Jahrzehnt bezieht. Die Anzahl der Spalten ergibt sich aus der Anzahl der Ausprägungen der zweiten Variable. Die Breite einer Zelle repräsentiert die relative Häufigkeit der jeweiligen Ausprägung der zweiten Variable. Sind mehr als zwei Variablen im Mosaikplot abgetragen, werden die Achsen mehrfach belegt und die Flächen des Diagramms werden entsprechend weiter aufgeteilt.³⁰

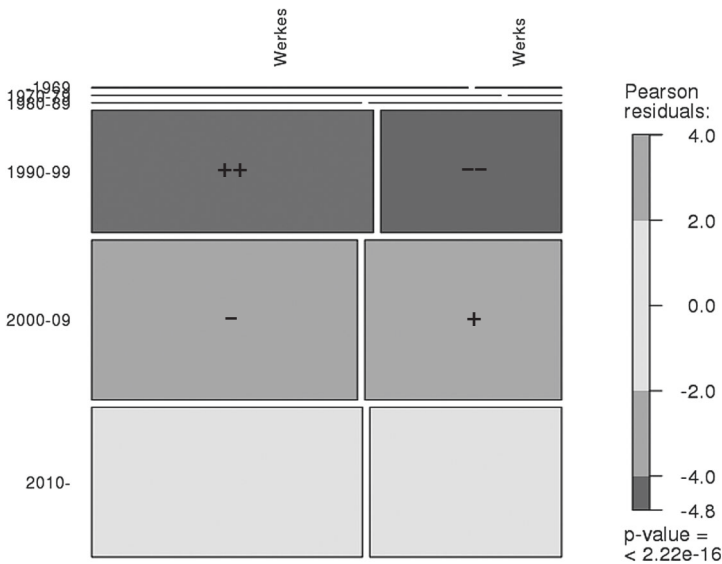


Abb. 11: Mosaikplot für die Varianten *Werks* vs. *Werkes* in Abhängigkeit von Jahrzehnten; Abweichungen nach oben bzw. unten werden für den Schwarz-Weiß-Druck über die Symbole + und – symbolisiert (bei stärkeren Abweichungen werden die Symbole gedoppelt)³¹

3.5 Konfidenzintervalle

Mit der Angabe von Konfidenzintervallen kann man von den vorliegenden Stichprobenergebnissen auf die Grundgesamtheit verallgemeinern (siehe Brunner/Munzel 2013, S. 80ff.).

³⁰ Wie der Assoziationsplot wird auch der Mosaikplot nur für mehrspaltige und nicht-aggregierte Tabellen ausgegeben.

³¹ In KoGra-R wird der Mosaikplot – wie der Assoziationsplot – farbig dargestellt.

Ein Konfidenzintervall gibt an, in welchem Bereich der Populationsparameter liegen könnte. Ein 95%-Konfidenzintervall besagt, dass bei hinreichender Wiederholung des Experiments (d.h. Neuberechnung des 95%-Konfidenzintervalls aus jeweils neuen Stichproben) 95% der so berechneten Intervalle den Populationsparameter enthalten. Liegt der obere oder untere Wert des berechneten Konfidenzintervalls für die Realisierungshäufigkeit eines Phänomens nicht im Konfidenzintervall der Realisierungshäufigkeit eines anderen Phänomens, deutet dies auf einen signifikanten Unterschied zwischen dem Anteil der beiden Realisierungen hin (siehe ebd.).

In KoGra-R werden die Konfidenzintervalle für jede Realisierung in Abhängigkeit von den verschiedenen Metadaten in einer Tabelle (vgl. Tab. 8) und mithilfe eines Diagramms (vgl. Abb. 12) dargestellt. Außerdem wird der prozentuale Anteil der Realisierungen bezogen auf die Ausprägungen der jeweiligen Metadaten angegeben, der die Grundlage für die Berechnung des Konfidenzintervalls bildet. Wie das vorliegende Beispiel zeigt, beziehen sich 100% dabei auf alle Vorkommen im betreffenden Metadatum (hier Jahrzehnt) bezogen auf eine Realisierung (*Werkes* vs. *Werks*).

=== <i>Werkes</i> ===			
	Anteil in Prozent	Konf.-Intervall unten	Konf.-Intervall oben
-1969	0.17829524	0.139972082	0.21661840
1970-79	0.04940711	0.029220367	0.06959386
1980-89	0.01503695	0.003898463	0.02617543
1990-99	29.27693762	28.863583497	29.69029174
2000-09	36.05860113	35.622412422	36.49478985
2010-	34.42172195	33.990128123	34.85331577
=== <i>Werks</i> ===			
	Anteil in Prozent	Konf.-Intervall unten	Konf.-Intervall oben
-1969	0.058346640	0.032118928	0.08457435
1970-79	0.009212627	-0.001211770	0.01963703
1980-89	0.015354379	0.001896953	0.02881181
1990-99	26.873234246	26.391755224	27.35471327
2000-09	38.180198993	37.652529346	38.70786864
2010-	34.863653114	34.346073261	35.38123297

Tab. 8: Darstellung der Konfidenzintervalle für jede Realisierung in Abhängigkeit von Jahrzehnten

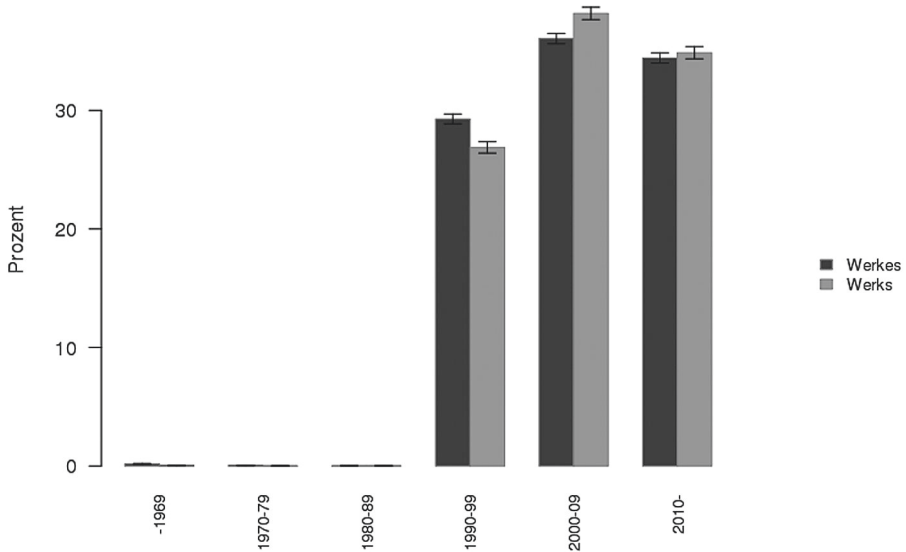


Abb. 12: Gruppierendes Säulendiagramm mit Konfidenzintervallen für jede Realisierung in Abhängigkeit von Jahrzehnten

Das Diagramm visualisiert die prozentuale Verteilung der Phänomene auf Jahrzehnte mit den entsprechenden Konfidenzintervallen, die in der Tabelle abgetragen sind.³²

3.6 Dispersion – Verteilung von Phänomenrealisierungen über das Gesamtkorpus

Die oben dargestellten statistischen Tests können verwendet werden, um Zusammenhänge zwischen verschiedenen Variablen zu erkennen. Im Analysebeispiel wurde überprüft, ob Zusammenhänge zwischen zwei Realisierungen der Genitivmarkierung des Lexems *Werk* und den Jahrzehnten, in denen die Texte entstanden sind, existieren. Dennoch kann mit den beschriebenen Analysearten noch nicht ausgeschlossen werden, dass innerhalb eines Jahrzehntes (also innerhalb eines Teilkorpus) eine sehr ungleichmäßige Verteilung der Treffer vorliegt. So kann sich beispielsweise ein Phänomen vornehmlich auf nur einen Text (z.B. einen Zeitungsartikel eines einzigen Autors) beschränken und so für eine hohe Frequenz im ganzen Teilkorpus verantwortlich sein (sog. *Clumpiness*, siehe Church/Gale 1995; Kilgarriff 2001, S. 107; Gries 2008, 2009;

³² Die Konfidenzintervalle können in der Visualisierung auch unter 0% bzw. über 100% liegen, weil sie symmetrisch berechnet werden. In solchen Fällen sind 0% bzw. 100% als die reale Grenze des Konfidenzintervalls zu betrachten.

Lijffijt/Gries 2012; Bubenhofer/Konopka/Schneider 2014, S. 134ff.). Gries (2008) diskutiert verschiedene Maße, um die Gleichmäßigkeit einer Phänomenrealisierung im Korpus zu untersuchen (z.B. Juilland et al.s D, Rosengrens S etc.). Für die Berechnung dieser Maße wird das Korpus in kleinere Einheiten, wie z.B. Texte, Abschnitte oder Teilkorpora unterteilt. Die Maße beruhen auf einer Verrechnung der erwarteten mit den beobachteten Frequenzen in den Korpusteilen.³³ Da bei den meisten Maßen davon ausgegangen wird, dass die einzelnen Korpusteile gleich groß sind, sind sie zu über- bzw. unterempfindlich gegenüber Häufigkeitsschwankungen in den Korpusteilen. Ein Maß, das speziell für Korpusdaten in Teilkorpora verschiedener Größe entwickelt wurde, ist hingegen die DP bzw. DP_{norm} (Deviation of Proportions, siehe Gries 2008, 2009).³⁴ Dabei wird für jedes Teilkorpus in Relation zur Korpusgröße die erwartete Frequenz sowie die beobachtete Frequenz eines Phänomens in Prozent berechnet. Anschließend werden die Differenzen zwischen allen erwarteten und den beobachteten Prozentwerten pro Teilkorpus berechnet, summiert und durch 2 dividiert. Der berechnete Wert wird normiert, sodass das Ergebnis zwischen 0 und 1 liegt.³⁵ Je näher es bei 0 liegt, desto gleichmäßiger ist das Phänomen über die Teilkorpora verteilt.

In Bubenhofer/Konopka/Schneider (2014, S. 134ff.) wurde gezeigt, wie die DP_{norm} verwendet werden kann, um dabei zu helfen, Realisierungen als standardnah oder standardfern einzustufen: Hierzu wurde die DP_{norm} für verschiedene Recherchebeispiele basierend auf dem DeReKo berechnet.³⁶ Da im DeReKo mehrere Texte zu Einheiten zusammengefasst und *Dokumente* genannt werden (z.B. alle Zeitungsartikel aus einem Monat einer Zeitung), wurde in der Untersuchung pro Dokument die Frequenz des Phänomens berechnet (siehe ebd.). Die abgefragten Beispiele wurden zudem als Korrelation zwischen DP_{norm} und Frequenz im Korpus in einem Streudiagramm abgebildet (vgl. Abb. 13).³⁷

³³ Siehe hierzu auch die Ausführungen in Bubenhofer/Konopka/Schneider (2014, S. 134ff.).

³⁴ Für eine Diskussion verschiedener Maße siehe Bubenhofer/Konopka/Schneider (2014, S. 134ff.).

³⁵ $DP/(1-\min(s))$, wobei $\min(s)$ die Größe des kleinsten Teilkorpus ausdrückt.

³⁶ Es wurde mit unterschiedlichen Korpora gearbeitet: Das Korpus, aus dem die hier dargestellten Abfragen sind, umfasst eine Zufallsauswahl von 10% der Wörter des gesamten DeReKo und damit 374521682 Wörter und 1547513 Texte. Zudem werden Ergebnisse aus einem Schweizer Teilkorpus von Presstexten gezeigt, das jeweils alle entsprechenden Zeitungen aus dem DeReKo beinhaltet (siehe ebd.).

³⁷ Eine detaillierte Beschreibung der Untersuchung lässt sich in Bubenhofer/Konopka/Schneider (siehe ebd.) finden.

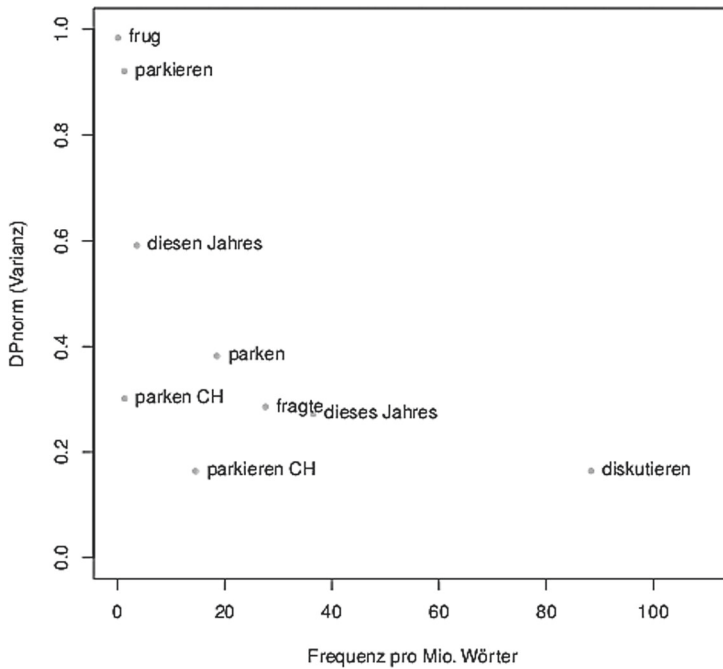


Abb. 13: Korrelationsplot zwischen DP_{norm} und Frequenz im Korpus einiger Recherchebeispiele aus dem DeReKo (Beispiel aus Bubenhofer/Konopka/Schneider 2014, S. 138)

Die Grafik verdeutlicht, dass das abgefragte Token *diskutieren* mit einer niedrigen DP_{norm} relativ gleichmäßig im vorliegenden Korpus verteilt ist. Die anderen Recherchebeispiele weisen einen höheren Wert für die DP_{norm} auf. Der Ausdruck *diesen Jahres* ist weniger gleichmäßig verteilt als *dieses Jahres*, das Token *frug* ist mit einer DP_{norm} von über 0,95 am ungleichmäßigsten im Korpus verteilt (vgl. ebd., S. 238f.). Auch der Ausdruck *parkieren* ist im gesamten Korpus mit einer DP_{norm} von über 0,9 ungleichmäßig verteilt, da das Token mit großer Wahrscheinlichkeit größtenteils in Schweizer Texten verwendet wird. Wird hingegen die DP_{norm} des Tokens *parkieren* auf der Grundlage von ausschließlich Schweizer Zeitungen berechnet, so ergibt sich ein anderer Wert: Die DP_{norm} für *parkieren* liegt hier unter 0,2 (die Recherche ist in der Grafik oben mit CH markiert). Die Variante *parkieren* ist erwartungsgemäß in den Schweizer Texten gleichmäßiger verteilt als *parken*. Werden DP_{norm} und Frequenz der abgefragten Tokens korreliert (vgl. Abb. 13), kann festgestellt werden, dass Ausdrücke, die häufiger im Korpus vorkommen, oft gleichmäßiger im Korpus verteilt sind, während Ausdrücke, die ungleichmäßiger verteilt sind (im oberen Bereich der Grafik), seltener vorkommen. Die Korrelation von der Frequenz im Korpus und der DP_{norm} kann somit dazu genutzt werden, die Standardnähe einer Realisierung zu analysieren. Auf diese Weise

kann gezeigt werden, dass Realisierungen, die im oberen Bereich der Grafik liegen und eine niedrigere Frequenz im Korpus aufweisen, eher standardfernere Varianten darstellen, während Realisierungen im unteren Bereich der Grafik häufig eine hohe Frequenz im Korpus aufweisen und eher der standardnahen Sprache zuzuordnen sind.

In KoGra-R werden die Teilkorpora durch die Metainformationen definiert, die im Rahmen der Kontingenztabelle typischerweise zeilenweise dargestellt werden. In dem Abfragebeispiel aus den vorangegangenen Abschnitten sind dies die Häufigkeiten für *Werks* und *Werkes* in den jeweiligen Jahrzehnten. Für jedes Teilkorpus (= Jahrzehnt) wird die Frequenz des Phänomens berechnet.³⁸ Folgende Ergebnisse werden für die Berechnung der DP und DP_{norm} ausgegeben:

=== *Werkes* ===

Deviation of proportions (DP): 0.06446221

Deviation of proportions, normalized (DP_{norm}): 0.06448047

=== *Werks* ===

Deviation of proportions (DP): 0.03950845

Deviation of proportions, normalized (DP_{norm}): 0.03951965

Die DP und die DP_{norm} der Abfrage *Werkes* liegen bei ca. 0,06. Die Werte der Abfrage *Werks* betragen beide ca. 0,04. Das Nomen mit der kurzen Genitivendung -s ist demnach gleichmäßiger über die Jahrzehnte verteilt als die Realisierung mit der langen Genitivmarkierung -es. Es ist aber ebenfalls anzumerken, dass beide Phänomene im Vergleich zur Gesamtspanne der DP-Werte (zwischen 0 und 1) sehr gleichmäßig im Korpus verteilt sind.³⁹

4. Implementation

KoGra-R ist ein webbasiertes Tool: Über Links, die von KoGra-DB-Rechercheergebnissen ausgehen, oder von einer Webseite aus, über die COSMAS-II-Exportdateien und frei definierte Tabellen eingegeben werden können, wird ein CGI-Programm gestartet.⁴⁰ Dieses Programm ist in der Programmierspra-

³⁸ Dispersionsmaße können in KoGra-R nur berechnet werden, wenn die jeweiligen Korpusbezugsgrößen (hier: Gesamttokenanzahl der Teilkorpora) vorliegen.

³⁹ In KoGra-R wird außerdem die DP_{norm} für jedes Phänomens bzw. für jede Abfrage mit der absoluten Trefferanzahl korreliert und in einem Streudiagramm abgetragen.

⁴⁰ CGI = Common Gateway Interface.

che Python geschrieben.⁴¹ Es liest die zu analysierenden Tabellen ein, prüft sie und wandelt sie um in das Tabellenformat der Statistiksprache R (R Core Team 2016). Dann ruft das Programm ein R-Skript auf, übergibt diesem die umgewandelten Tabellen und empfängt die Analyseergebnisse. Die Analyseergebnisse werden formatiert, in ein HTML-Template eingefügt und der Nutzerin/dem Nutzer als Webseite zurückgegeben und angezeigt.

Bei der Implementation sowohl des Python- als auch des R-Skripts wurde darauf geachtet, dass sie modular aufgebaut, leicht modifizierbar und vor allem problemlos zu erweitern sind. So können leicht zusätzliche statistische Analysen hinzugefügt und ggf. auch weitere Inputmodi ergänzt werden. Die Programme sind durchweg ausführlich dokumentiert, sodass ihre Pflege gewährleistet ist.

In diesem Abschnitt beschreiben wir in gebotener Kürze den Aufbau der KoGra-DB (Unterabschnitt 4.1), des CGI-Programms (Unterabschnitt 4.2) und des R-Programms (Unterabschnitt 4.3).

4.1 KoGra-DB

Die Korpusgrammatik-Datenbank KoGra-DB wurde als Testsystem mit dem Ziel entwickelt, die Anwendbarkeit relationaler Datenstrukturen für die Verwaltung sprachwissenschaftlich motivierter Textkorpora zu evaluieren. Sie enthält Primärdaten – also authentische Sprachbelege aus DEREKO-Subkorpora – kombiniert mit linguistischen und außersprachlichen Metadaten. Die Segmentierung erfolgt auf Wortebene, d.h., ein Datensatz besteht aus einem Textwort, angereichert um Lemma- und Wortklassenangabe. Deren Inhalte sowie die Bestimmung von Satzgrenzen basieren auf dem morphosyntaktischen Output dreier Tagging-Werkzeuge: TreeTagger, Connexor Machineese Phrase Tagger und Xerox Incremental Parser. Gemeinsam nehmen die Sprach- und Annotationsdaten ca. vier Terabyte (TB) Speicherplatz ein. Angereichert wurden die wortspezifischen Inhalte um korpuspezifische Metadaten, die in Kapitel 2.1 detaillierter beschrieben werden.

⁴¹ www.python.org (letzter Zugriff: 7.3.2017).

Connexor-Recherche auf Satzebene (Connexor-Doku)

Lemma

und gefolgt mit min. Wortabstand: mit max. Wortabstand:

Token [entfernen]

und nicht gefolgt mit min. Wortabstand: mit max. Wortabstand:

Wortklasse [entfernen]

[Suchkriterium hinzufügen]

Medium

Publikumspresse Bücher Internet Gesprochenes Sonstiges

Register

Presstextsorte Gebrauchstextsorte Literarische Textsorte

Domäne

Fiktion Kultur/Unterhaltung Mensch/Natur Politik/Wirtschaft/Gesellschaft Technik/Wissenschaft unklassifizierbar

Land

D D (Ost) D (West) A CH

Region

überregional Herkunft unbekannt Herkunft nicht zuordenbar

Mittelost Mittelsüd Mittelwest Nordost Nordwest Südost Südwest

Jahr

-1969 1970-79 1980-89 1990-99 2000-09 2010-

nur im ausgewogenen Korpus

Abb. 14: Online-Abfrage der KoGra-DB

Sämtliche Inhalte von KoGra-DB wurden zu Evaluationszwecken projektintern durch ein Online-Frontend recherchierbar gemacht (Abb. 14). Dies erlaubt die Eingabe beliebiger Kombinationen aus Wortfolgen sowie Lemma- oder Wortklassenangaben. Die außersprachlichen Metaangaben können zur Eingrenzung der Suchergebnisse herangezogen werden; vgl. auch Abschnitt 2.1.

Die nachfolgende Tabelle (Tab. 9) vermittelt einen Eindruck vom Umfang des Datenbestands. KoGra-DB basiert aktuell auf dem DEREKO-Release 2014-II; im Vergleich zum in Bubenhofer/Konopka/Schneider (2014) beschriebenen Inventar hat sich der Umfang damit auf knapp 8 Milliarden Tokens nahezu

verdoppelt. Ausgenommen von dieser Erweiterung sind bislang die auf dem Xerox-Tagger basierten Teildaten sowie das ausgewogene Korpus.⁴²

	Untersuchungskorpus	Ausgewogenes Korpus
Texte	25.426.585	20.148
Connexor-Sätze	486.874.983	1.165.198
Connexor-Tokens	7.903.963.339	20.026.757
TreeTagger-Sätze	367.143.190	815.029
TreeTagger-Tokens	7.484.900.808	18.517.447

Tab. 9: Text-, Satz- und Wortvolumen der KoGra-DB

4.2 CGI/Python

KoGra-R besteht aus Python-Skripten, R-Skripten, HTML-Dateien und -Templates und einem CSS-Skript. Die HTML-Dateien enthalten (i) die Eingabemaske für COSMAS-II-Exportdateien und Nutzer-definierte Tabellen, (ii) die Erläuterungen zur Dateneingabe und (iii) die Erläuterungen zu den statistischen Auswertungen. Die Templates dienen der Formatierung der Ausgaben, also (i) der Prüfungsergebnisse von Tabellen⁴³ und (ii) der Auswertungen wohlgeformter Tabellen. Die Darstellung aller HTML-Seiten wird durch ein CSS-Skript gesteuert.

Das eigentliche CGI-Programm ist in Python 2.7 implementiert. Zentral ist ein Modul, in dem verschiedene Funktionen zur Datenkontrolle und -transformation definiert sind. Außerdem werden globale Variablen in einer separaten Steuerdatei definiert, weshalb das Programm leicht an Änderungen der Serverstruktur oder der Input-Formate angepasst werden kann. Zwei CGI-Skripte laden die zentral definierten Funktionen und Variablen. Ein Skript dient der Prüfung von Input-Daten, das andere fungiert als Schnittstelle zu den in R programmierten statistischen Analysen und ruft diese für alle geprüften Tabellen auf.

⁴² Dieses Korpus ist aus ausgewählten Texten des Gesamtkorpus zusammengestellt und ausgewogen in Bezug auf den Parameter Medium (siehe Bubenhofer/Konopka/Schneider 2014, S. 76f.).

⁴³ Siehe oben, Abschnitt 2: Gerade von Nutzer(inne)n frei definierte Tabellen sind möglicherweise nicht wohlgeformt und können also nicht ausgewertet werden. Sie werden daher vor ihrer Weiterverarbeitung geprüft. Das Prüfungsergebnis wird der Nutzerin/dem Nutzer angezeigt.

4.3 R

Serverseitige R-Ausführung

Ein R-Steuerskript (unsere Terminologie) wird mit den übergebenen Parametern gestartet. Die ersten 6 Parameter bei der Ausführung repräsentieren Teile der übergebenen Tabellen, Parameter 7 ist technischer Natur und Parameter 8 steuert den genauen Analysemodus (in Abhängigkeit von der Anzahl vorhandener Abfragen und dem Vorhandensein von Korpusbezugsgrößen). Das Steuerskript lädt dann ein R-Paket, das alle Funktionen enthält, die für die weiteren Auswertungen notwendig sind. Je nach aktuellem Analysemodus werden dann alle oder einige der Funktionen ausgeführt.

Erweiterungen am R-Teil von KoGra-R müssen also immer an zwei Orten erfolgen: Die neue Funktion muss in das Paket aufgenommen werden und die Funktion muss im Steuerskript aufgerufen werden. Das Steuerskript enthält außerdem einige Ausgaben, die dem umgebenden CGI-Skript erlauben, mit den Outputs der R-Funktionen umzugehen. Diese Ausgaben sind in Form von XML-Tags realisiert. Das Steuerskript ist ausführlich im Programmcode dokumentiert, sodass eine Erweiterung für Personen, die mit R vertraut sind, keine große Hürde darstellen sollte.

Anwenderseitige Ausführung

Die R-Funktionen werden immer serverseitig ausgeführt und von KoGra-R per Browser zurückgegeben. Allerdings ermöglicht KoGra-R, den R-Code in einer lokalen R-Installation auszuführen. Hierzu wird im Output sowie in der Dokumentation der aktuell serverseitig ausgeführte R-Code mit ausgegeben. Wichtig ist hierbei, dass auch der Code für die Erzeugung der Tabellen bereitgestellt wird. Nur so können alle Auswertungen am eigenen PC nachvollzogen bzw. bei Bedarf modifiziert oder erweitert werden. Hierzu muss eine lokale R-Installation gestartet werden und der angezeigte Code in einen Skript-Editor kopiert werden. R ist freie Software und kann unter www.r-project.org (letzter Zugriff: 10.12.2018) für Windows, macOS und Linux heruntergeladen werden.

5. Evaluation

KoGra-R wurde zweimal evaluiert, jeweils mithilfe eines Fragebogens (siehe Anhang). Beide Evaluationen fanden nach unterschiedlichen Entwicklungsstufen des Tools statt. Die erste Evaluation erfolgte von Dezember 2014 bis Januar 2015. Sie beschränkte sich auf die Auswertung von KoGra-DB-Re-

cherchen.⁴⁴ Sechs Personen haben teilgenommen, durchweg Mitarbeiter/innen des Projekts Korpusgrammatik und Nutzer/innen der KoGra-DB, die nicht an der Konzeption und Implementierung von KoGra-R beteiligt waren. Diese Evaluation war besonders wichtig, da die Teilnehmer/innen zunächst die primäre Zielgruppe des Tools bildeten. Die zweite Evaluation erfolgte, nachdem das Tool für die Öffentlichkeit zur Verfügung gestellt wurde, von März bis Juni 2015, wieder mithilfe desselben, leicht angepassten, Fragebogens. Bei der zweiten Evaluation ging es um die statistischen Auswertungen von COSMAS-II-Exportdateien und Nutzer-definierten Tabellen. Es haben 18 Personen teilgenommen, die aber nicht durchweg alle Fragen beantwortet haben – die meisten Fragen wurden von insgesamt 13 Personen beantwortet. Ein Teil der Mitarbeiter/innen des Projekts Korpusgrammatik, die bereits an der ersten Evaluation teilgenommen hatten, hat auch an der zweiten Evaluation teilgenommen. Die übrigen Teilnehmer/innen waren Mitarbeiter/innen aus verschiedenen Abteilungen des Instituts für Deutsche Sprache, deren wissenschaftliches Interesse vom Projekt Korpusgrammatik unabhängig war.

Bei beiden Evaluationen wurden die Teilnehmer/innen gebeten, Korpusrecherchen durchzuführen und die Ergebnisse mit KoGra-R auszuwerten, um so Erfahrungen im Umgang mit dem Tool zu sammeln. Es stand den Teilnehmer(inne)n frei, eigene Fragestellungen zu bearbeiten. Darüber hinaus wurden Vorschläge gemacht, was sie untersuchen könnten. Bei der ersten Evaluation wurden sie angeregt, das Verhältnis von *wegen* + Genitiv und *wegen* + Dativ zu untersuchen. Eine Annäherung kann durch die Suchanfragen *wegen des* und *wegen dem* vorgenommen werden. Bei der zweiten Evaluation wurden zusätzlich Vergleiche von *Werks* und *Werkes, des Irak, des Iraks* und *des Irakes, parkieren* und *parken, fragte* und *frug* und *diesen Jahres* und *dieses Jahres* vorgeschlagen. Im Rahmen der ersten Evaluation wurden die Teilnehmer/innen aufgefordert, sowohl Korpusrecherchen einzeln auszuwerten als auch mehrere Recherchen miteinander zu vergleichen und gemeinsam auszuwerten. Bei der zweiten Evaluation sollten die Teilnehmer/innen die Auswertung von COSMAS-II-Exportdateien und von frei definierten Tabellen bewerten.

Die Fragebögen wurden jeweils online ausgefüllt. Beide Evaluationen waren anonym.

Die Evaluationen hatten qualitative und quantitative Aspekte. Es wurden zum einen offene Fragen gestellt, zum anderen sollten Vorkenntnisse der Teilnehmer/innen und Eigenschaften von KoGra-R auf einer Skala von 1–7 beurteilt werden. Mit den Evaluationen haben wir drei Ziele verfolgt: Erstens wollten wir mehr über die Nutzer/innen, insbesondere ihre korpuslinguisti-

⁴⁴ Die Schnittstellen zur Eingabe von COSMAS-II-Exportdateien und Nutzer-definierten Tabellen waren zu dieser Zeit noch nicht implementiert.

schen und statistischen Vorkenntnisse und ihr Interesse im Umgang mit Korpusdaten erfahren. Zweitens wollten wir die Nützlichkeit und Nutzbarkeit des Tools einschätzen. Drittens wollten wir Anforderungen zur Erweiterung und Verbesserung des Analysetools erheben.

Um mehr über die Nutzer/innen und ihre Vorkenntnisse zu erfahren, haben wir sie nach ihren Fachgebieten und ihren Erfahrungen mit Korpusrecherche und statistischer Analyse gefragt. Wir haben sie gebeten, ihre Erfahrungen mit Korpusrecherche und statistischer Analyse auf einer Skala von 1–7 (1: sehr niedrig, 7: sehr hoch) einzuschätzen. Außerdem haben wir sie gebeten zu beschreiben, welche Korpusrecherchen sie üblicherweise durchführen.

Darüber hinaus wollten wir die Akzeptanz der Nutzer/innen von KoGra-R einschätzen. Dem *Technology Acceptance Model* (TAM; Davis/Bagozzi/Warshaw 1989) zufolge sind die zentralen Faktoren der Akzeptanz die wahrgenommene Nützlichkeit und die Nutzbarkeit, d.h. Einfachheit der Nutzung (*Perceived Usefulness and Perceived Ease of Use*; Davis 1989). In Anlehnung an Davis (1989) haben wir Fragen formuliert, um die Einschätzungen bezüglich Nützlichkeit und Einfachheit der Nutzung zu erheben. Wir haben die Teilnehmer/innen auf einer Skala von 1–7 (1: sehr niedrig, 7: sehr hoch) bewerten lassen, (i) ob KoGra-R nützlich für sie ist, (ii) ob sie ihre Arbeitsziele mit dem Tool besser erreichen können, (iii) ob die Bedienung einfach, (iv) frustrierend oder (v) mühsam ist, (vi) ob das Tool motivierend wirkt, (vii) ob die Teilnehmer/innen ihre Aufgaben mit dem Tool besser eigenständig bewältigen können und (viii) ob sie das Tool öfter benutzen wollen. Außerdem haben wir detailliert für jeden Input-Modus und jede einzelne statistische Analyse und Darstellung gefragt, (i) ob der/die Nutzer/in die Daten respektive Auswertungsergebnisse *versteh*t, (ii) ob Letztere für ihn *hilfreich* sind, (iii) ob die *Erläuterungen* zu den Daten respektive Auswertungen *hilfreich* sind und (iv) ob die Erläuterungen das *Informationsbedürfnis befriedigen*. Wir haben die genannten Urteile wieder durch Bewertungen auf einer Skala von 1–7 abgeben lassen.

Schließlich haben wir zur Erhebung weiterer Anforderungen die Teilnehmer/innen explizit gefragt, welche zusätzlichen statistischen Analysen und Darstellungsformen sie sich wünschen, wie die Erläuterungen verbessert werden können, welche weiteren Input-Modi sie für sinnvoll erachten und wie man das Tool darüber hinaus noch verbessern könnte.

*Zu den Ergebnissen:*⁴⁵

Die Teilnehmer/innen beider Evaluationen gaben an, über gute Kenntnisse in Korpusrecherchen zu verfügen (auf einer Skala von 1–7, im Mittel: erste Evaluation 5,0, zweite Evaluation 4,71). Allerdings konstatierten sie, dass ihre Er-

⁴⁵ Die Ergebnisse sind – bis auf die Freitextantworten – im Anhang dargestellt.

fahrungen mit statistischen Analysen weniger als mittelgut sind (auf einer Skala von 1–7, im Mittel: erste Evaluation 3,17, zweite Evaluation 2,94). Es lässt sich daraus schließen, dass die Teilnehmer/innen Unterstützung bei der Auswahl und Interpretation statistischer Analysen und Darstellungsweisen benötigen. Diese Einschätzung wird durch explizite Antworten auf die Fragen zur Verbesserung des Tools gestützt. Nach der Auswertung der ersten Evaluation wurde aus diesem Grund ein besonderes Augenmerk auf die Erweiterungen der Dokumentation und der Erläuterungen zu den statistischen Analysen gelegt. Dabei legten wir insbesondere Wert auf die Nennung zusätzlicher weiterführender Literatur sowie eine bessere didaktische Aufbereitung der Erläuterungen zu den Analysen. Außerdem räumten wir potenzielle Missverständnisse aus, die aus der bis dato vorliegenden Dokumentation hervorgehen konnten. An einigen Stellen wurden weiterhin Hinweise zu den Voraussetzungen der einzelnen Tests eingefügt.

In beiden Evaluationen wurde KoGra-R von den Teilnehmer(inne)n auf einer Skala von 1–7 als nützlich (im Mittel: erste Evaluation 6,0, zweite Evaluation 4,29) und einfach zu bedienen (im Mittel: erste Evaluation 5,12, zweite Evaluation 4,67) eingeschätzt. Die Benutzung war weder mühsam (im Mittel: erste Evaluation 2,67, zweite Evaluation 2,19) noch frustrierend (im Mittel: erste Evaluation 2,5, zweite Evaluation 2,13). Die Teilnehmer/innen gaben an, dass sie das Tool in Zukunft öfter nutzen werden (erste Evaluation 5,84, zweite Evaluation 4,0). Das Ergebnis der ersten Evaluation kann im Allgemeinen als gut, das der zweiten als etwas verhaltener beschrieben werden. Wenn man statt des arithmetischen Mittels jeweils den Modus und den Median betrachtet, muss man die Ergebnisse beider Evaluationen noch positiver bewerten.⁴⁶ Die unterschiedlichen Ergebnisse beider Evaluationen sind dadurch zu erklären, dass KoGra-R in erster Linie für das Projekt Korpusgrammatik entwickelt wurde, dessen Mitarbeiter/innen es in der ersten Evaluation entsprechend sehr gut bewertet haben. Auf die Mitarbeiter/innen der anderen Abteilungen des IDS, die an der zweiten Evaluation teilnahmen, war das Tool nicht direkt zugeschnitten. Es ist daher nicht verwunderlich, dass sie das Tool zwar positiv, aber zurückhaltender bewerteten.

Die Teilnehmer/innen beider Evaluationen haben zahlreiche Hinweise gegeben, wie das Tool für die allgemeine Nutzung erweitert werden kann. Darunter waren kleinere Hinweise zu Details der Ergebnisansichten und Erläuterungen, die leicht umgesetzt werden konnten und unmittelbar umgesetzt wurden. Darüber hinaus betrafen die Hinweise der ersten Evaluation in erster Linie (i) die Dokumentation und Erläuterungen und (ii) die Ergänzung weiterer Input-Modi. Anforderungen zur Erweiterung der Erläuterungen wurden

⁴⁶ Vgl. Ergebnisse im Anhang, insbesondere die Fragen unter Punkt 4.

sowohl explizit als auch implizit gegeben, Letzteres durch Hinweise, die Missverständnisse in Bezug auf einzelne Analysen offenbarten. Die Erläuterungen wurden entsprechend ergänzt und umformuliert. Zur Zeit der ersten Evaluation gab es nur einen Input-Modus, nämlich die Schnittstelle zur KoGra-DB. Die Antworten der Evaluationsteilnehmer/innen wurden zum Anlass genommen, die beiden zusätzlichen Input-Modi – die Schnittstelle zu COSMAS II und das Formular zur Eingabe selbstdefinierter Tabellen – zu implementieren. In der zweiten Evaluation wurden verstärkt Hinweise zur Anpassung der statistischen Analysen und Diagramme gegeben, die in die Dokumentation aufgenommen wurden. Außerdem wurde das Farbspektrum zur Darstellung der Ergebnisse bei den Vergleichen beliebig vieler Abfragen erweitert. Darüber hinaus wurden an einigen Stellen die Labels in den Diagrammen so angeordnet, dass sie übersichtlicher erscheinen. Dennoch muss (auch im Rahmen der Dokumentation) darauf hingewiesen werden, dass ab einer Anzahl von 100 Variablenausprägungen mit Sicherheit davon ausgegangen werden kann, dass weder Übersichtlichkeit noch Responsivität des Tools gegeben sind (z.B. bei einer COSMAS-II-Quellenansicht). Nach der zweiten Evaluation wurde zusätzlich zu den Erläuterungen zur Tabelleneingabe eine Beispieltabelle auf der Startseite erstellt, die zum einen als Vorlage für eigene Tabellen verwendet werden kann, zum anderen in das Auswertungsformular kopiert werden kann, um direkt zu einer Beispielauswertung zu gelangen.

Im Ergebnis konnte durch die beiden Evaluationen bestätigt werden, dass KoGra-R ein sinnvolles und gut zu nutzendes Werkzeug für die datengestützte Linguistik ist. KoGra-R kann ausgebaut und durch zusätzliche Analysen ergänzt werden. Der Ausbau soll allerdings nicht zu weit gehen: KoGra-R muss einfach und unmittelbar zu bedienen sein und die schnelle Durchführung vordefinierter Analysen ermöglichen.

6. Zusammenfassung und Ausblick

In diesem Kapitel wurde das webbasierte Tool KoGra-R vorgestellt, über das in der Programmiersprache R implementierte statistische Auswertungen durchgeführt werden können. In Abbildung 15 wird die Struktur von KoGra-R zusammenfassend dargestellt:

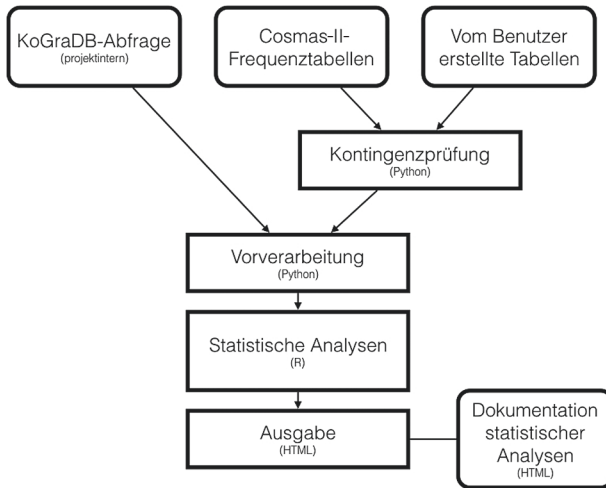


Abb. 15: Schematische Darstellung der Struktur von KoGra-R

Durch das Tool können Kontingenztabelle, die über die projektinterne Datenbank KoGra-DB recherchiert werden, ausgewertet werden. Darüber hinaus bietet KoGra-R die Möglichkeit, von COSMAS II erzeugte Frequenzlisten zu analysieren und miteinander zu vergleichen. Durch das Hochladen von (evtl. in Excel erzeugten) CSV-Tabellen und durch die direkte Eingabe von Kontingenztabelle über die Benutzeroberfläche bestehen weitere Möglichkeiten der Dateneingabe. Die eingegebenen Tabellen werden nicht direkt ausgewertet, sondern zuerst auf ihre Wohlgeformtheit geprüft. Falls die Tabelle nicht wohlgeformt ist, werden Fehlermeldungen ausgegeben und potenzielle Fehler markiert. Zu den statistischen Werkzeugen, die durch das Tool zur Verfügung gestellt werden, zählen deskriptive Analysearten, wie z.B. die Auswahl von Tabellen und Diagrammen für Rohdaten, normierte und relative Werte sowie die Anwendung von inferenzstatistischen Verfahren und Visualisierungsmöglichkeiten, die auf der Chi-Quadrat-Statistik beruhen. Darüber hinaus werden Konfidenzintervalle berechnet und visualisiert. Eine weitere Analysemöglichkeit besteht in der Berechnung der DP_{norm} mit der Aussagen über die Verteilung von Phänomenrealisierungen im Korpus getroffen werden können. Zu der Eingabe der Tabellen und den statistischen Analysearten ist in KoGra-R eine Dokumentation verfügbar, zu der man über Hyperlinks im Tool geleitet wird. Außerdem wird zu den verschiedenen Analyseschritten der verwendete R-Code zur Verfügung gestellt.

KoGra-R wurde im Rahmen der Projektarbeit zwei Evaluationsprozessen unterzogen. Das Tool wurde als nützlich und in der Bedienung einfach empfunden. Die Hinweise der Teilnehmer/innen aus beiden Evaluationen wurden je nach Möglichkeit berücksichtigt und umgesetzt. Vor allem die Verbesserun-

gen in der Dokumentation und die Erweiterung der Input-Modi um eine Schnittstelle zu COSMAS II und das Formular zur Eingabe selbstdefinierter Tabellen haben sich im Zuge der zweiten Evaluation als gewinnbringend erwiesen.

Die statistischen Analysen sowie die einbettende Infrastruktur können an mehreren Stellen erweitert werden. Mögliche Erweiterungen lassen sich grob in zwei Teilbereiche gliedern: Erweiterungen bezüglich der Präsentation der Ergebnisse und Erweiterungen bezüglich der durchgeführten Analysen. Präsentationsseitig wäre daran zu denken, noch mehr auf unterschiedliche Zielsetzungen beim Einsatz des Tools einzugehen. Führt man z.B. Analysen durch, um die resultierenden Schaubilder in Veröffentlichungen zu verwenden, wäre es hilfreich, die Schaubilder zusätzlich zur momentanen Form noch als hochauflösende Grafiken zur Verfügung zu stellen. Da sich die Erstellung hochauflösender Grafiken u.U. negativ auf die Performanz des Tools auswirken könnte, wäre es wünschenswert, diese Versionen der Grafiken nur auf Anfrage der Nutzerin/des Nutzers zu erstellen. Ebenfalls eine präsentationsbezogene Erweiterung wäre die Einführung eines Modus für Expert(inn)en, in dem die Rückgabe besonders schlank, also ohne zusätzliche Kurzkommentare, erfolgen könnte.

Auf Seiten der statistischen Analyse sind der Erweiterung auf weitere Tests oder Visualisierungen momentan nur insofern Grenzen gesetzt, als jedes zusätzliche Verfahren mit einer einfachen Kontingenztafel umgehen können muss. Ein korpuslinguistisch inzwischen breit angewendetes Verfahren wie eine logistische Regressionsanalyse, mit der man z.B. auch die Interaktion zwischen verschiedenen Faktoren untersuchen kann, ist somit nicht ohne Weiteres möglich, weil logistische Regressionen auf Einzelfallbasis operieren und somit nicht mit Kontingenztabellen umgehen können. Komplexere statistische Verfahren, die in der Regel auf Einzelfallbasis beruhen, stehen aber auch nicht im Fokus des Tools, denn das Ziel von KoGra-R ist eine erste Datenexploration, um ggfs. Hypothesen für weitere statistische Analysen zu gewinnen.⁴⁷

Auch die bereits implementierten Tests können an einigen Stellen erweitert bzw. verfeinert werden. Der Chi-Quadrat-Test kann z.B. zu verzerrten Ergebnissen führen, wenn in mehr als 20% der Zellen die erwartete Häufigkeit unter 5 fällt. Dies ist momentan lediglich in der Dokumentation vermerkt. Es

⁴⁷ Im Projekt Korpusgrammatik werden und wurden komplexere statistische Verfahren in weiteren Untersuchungsphasen durchgeführt. Zu nennen sind hier z.B. Bubenhofer/Hansen-Morath/Konopka (2014) und Konopka/Fuß (2016) sowie Brandt (unter Mitwirkung von Bildhauer; in diesem Band), in denen die Variation zum einen mithilfe eines durch maschinelles Lernen erzeugten Entscheidungsbaumes und zum anderen durch logistische Regressionsanalysen untersucht wurde.

wäre aber auch denkbar, dass dies innerhalb des statistischen Auswertungs-kripts automatisch eruiert wird. Sollte es der Fall sein, dass die kritische Zahl an schwach besetzten Zellen überschritten wird, könnte dann statt des Chi-Quadrat-Tests automatisch ein Fisher's Exact Test berechnet und ausgegeben werden.

KoGra-R ist im augenblicklichen Zustand auch durch projektexterne Mitarbeiter/innen des Instituts für Deutsche Sprache evaluiert worden. Man merkt der dazugehörigen Dokumentation an einigen Stellen noch die ursprüngliche Ausrichtung auf korpusbasierte Häufigkeitsdaten an. Im Grunde spricht aber nichts dagegen, KoGra-R für alle häufigkeitsbasierten Auswertungen anzuwenden. Dies würde den potenziellen Kreis an Benutzer(inne)n deutlich erweitern. So könnten auch experimentell arbeitende Wissenschaftler/innen ihre Daten in KoGra-R auswerten. Dabei ist z.B. an Häufigkeiten bestimmter Antwortkategorien (z.B. richtig/falsch) von Proband(inn)en innerhalb eines experimentellen Settings zu denken, wie es beispielsweise in der Psycholinguistik angewendet wird. Die Spalten einer solchen Tabelle könnten sich dann durch verschiedene Experimentalbedingungen definieren. Im Grunde kann KoGra-R schon jetzt auf diese Art von Daten angewendet werden, z.B. über die Eingabe von benutzerdefinierten Tabellen. Auch der Einsatz von KoGra-R in der Lehre, nämlich als einfach zu bedienendes Werkzeug zur häufigkeitsbasierten statistischen Auswertung, ist nicht zu unterschätzen. Hier kann die ausführliche Dokumentation schon jetzt dazu dienen, Studierenden die quantitativ-statistische Forschungsarbeit behutsam nahezubringen.

Literatur

- Backhaus, Klaus et al. (2016): *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. 14. Aufl. Berlin/Heidelberg: Springer.
- Bortz, Jürgen (2005): *Statistik für Human- und Sozialwissenschaftler*. 6., vollst. überarb. u. aktual. Aufl. Berlin u.a.: Springer.
- Bortz, Jürgen/Lienert, Gustav A. (2008): *Kurzgefasste Statistik für die klinische Forschung. Leitfaden für die verteilungsfreie Analyse kleiner Stichproben*. 3., aktual. u. bearb. Aufl. Heidelberg: Springer.
- Brunner, Edgar/Munzel, Ullrich (2013): *Nicht-parametrische Datenanalyse. Unverbundene Stichproben*. 2., überarb. Aufl. Berlin/Heidelberg: Springer.
- Bubenhofer, Noah/Hansen-Morath, Sandra/Konopka, Marek (2014): *Korpusbasierte Exploration der Variation der nominalen Genitivmarkierung*. In: *Zeitschrift für germanistische Linguistik* 42, 3, S. 379–419.
- Bubenhofer, Noah/Konopka, Marek/Schneider, Roman (2014): *Präliminarien einer Korpusgrammatik. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 4)*. Tübingen: Narr.

- Church, Kenneth W./Gale, William A. (1995): Poisson mixtures. In: *Natural Language Engineering* 1, 2, S. 163–190.
- Cohen, Ayala (1980): On the graphical display of the significant components in two-way contingency tables. In: *Communications in Statistics – Theory and Methods* 9, 10, S. 1025–1041.
- Davis, Fred D. (1989): Perceived usefulness, perceived ease of use, and user acceptance of information technology. In: *MIS Quarterly* 13, 3, S. 319–340.
- Davis, Fred D./Bagozzi, Richard P./Warshaw, Paul R. (1989): User acceptance of computer technology: A comparison of two theoretical models. In: *Management Science* 35, 8, S. 982–1003.
- Emerson, John W. (1998): Mosaic displays in S-PLUS: A general implementation and a case study. In: *Statistical Computing & Statistical Graphics Newsletter* 9, 1, S. 17–23.
- Field, Andy/Miles, Jeremy/Field, Zoë (2012): *Discovering statistics using R*. Los Angeles, CA u.a.: SAGE.
- Friendly, Michael (1992): Graphical methods for categorical data. In: SEUGI '92. Proceedings of the SAS User's Group International Conference May 19–22, 1992. (= SAS Institute 17). Heidelberg: SAS Institute, S. 190–200. www.math.yorku.ca/SCS/sugi/sugi17-paper.html (Stand: 23.11.2017).
- Friendly, Michael (1994): Mosaic displays for multi-way contingency tables. In: *Journal of the American Statistical Association* 89, 425, S. 190–200.
- Fürbacher, Monica (2015): *Spezialstudie: Regionale Verteilung*. In: *Grammis 2.0, Variation der starken Genitivmarkierung*. Mannheim: Institut für Deutsche Sprache. http://hypermedia.ids-mannheim.de/call/public/korpus.ansicht?v_id=5087 (Stand: 16.4.2018).
- Gries, Stefan Thomas (2008): Dispersions and adjusted frequencies in corpora. In: *International Journal of Corpus Linguistics* 13, 4, S. 403–437.
- Gries, Stefan Thomas (2009): Dispersions and adjusted frequencies in corpora: Further explorations. In: *Language and Computers* 71, 1, S. 197–212.
- Hansen-Morath, Sandra/Wolfer, Sascha (2017): Standardisierte statistische Auswertung von Korpusdaten im Projekt „Korpusgrammatik“ (KoGra-R). In: Konopka, Marek/Wöllstein, Angelika (Hg.): *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*. (= Jahrbuch des Instituts für Deutsche Sprache 2016). Berlin: De Gruyter, S. 345–356.
- Hartigan, John A./Kleiner, Beat (1984): A mosaic of television ratings. In: *The American Statistician* 38, 1, S. 32–35.
- Institut für Deutsche Sprache (2014): *Deutsches Referenzkorpus. Archiv der Korpora geschriebener Gegenwartssprache 2014-II* (Release vom 11.9.2014). Mannheim: Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo (Stand: 23.11.2017).
- Kilgarriff, Adam (2001): Comparing corpora. In: *International Journal of Corpus Linguistics* 6, 1, S. 97–133.
- Konopka, Marek/Fuß, Eric (2016): *Genitiv im Korpus. Untersuchungen zur starken Flexion des Nomens im Deutschen*. (= Studien zur deutschen Sprache 70). Tübingen: Narr.

- Konopka, Marek/Waßner, Ulrich Hermann, unter Mitwirkung von Sandra Hansen (2013): Standarddeutsch messen? Frequenz und Varianz negativ-konditionaler Konnektoren. In: *Korpus – gramatika – axiologie* 8, S. 12–35.
- Lijffijt, Jeffrey/Gries, Stefan Thomas (2012): Correction to „Dispersions and adjusted frequencies in corpora“. In: *International Journal of Corpus Linguistics* 17, 1, S. 147–149.
- Meyer, David/Zeileis, Achim/Hornik, Kurt (2005): The strucplot framework: Visualizing multi-way contingency tables with vcd. In: *Research Report Series – Department of Statistics and Mathematics 22*, *Wirtschaftsuniversität Wien*. http://epub.wu.ac.at/dyn/openURL?id=oai:epub.wu-wien.ac.at:epub-wu-01_8a1 (Stand: 23.11.2017).
- R Core Team (2016): The R project for statistical computing. www.R-project.org (Stand: 7.3.2017).
- Wiechmann, Daniel (2008): On the computation of collocation strength: Testing measures of association as expressions of lexical bias. In: *Corpus Linguistics and Linguistic Theory* 4, 2, S. 253–290.
- Wolfer, Sascha/Hansen-Morath, Sandra (2017): Visualisierung linguistischer Daten mit der freien Grafik- und Statistikumgebung R. <http://kograno.ids-mannheim.de/VisR-OnlinePub> (Stand: 23.11.2017).

- normierte Werte (Tabelle)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- normierte Werte (Diagramm)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- relative Werte (Tabelle)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- relative Werte (Diagramm, gestapelt)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- relative Werte (Diagramm, gruppiert)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- Chi-Quadrat-Test
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- Phi Assoziationsstärke/Cramérs V Assoziationsstärke
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- Chi-Quadrat-Test, erweiterte Häufigkeiten (Tabelle)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- Chi-Quadrat-Test, Residuen (Tabelle)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- Assoziationsplot (Diagramm)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- Mosaikplot (Diagramm)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- Konfidenzintervalle (Tabelle)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- Konfidenzintervalle (Diagramm)
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu
- Dispersionsmaße (DP_{norm})
 - stimme gar nicht zu □ □ □ □ □ □ □ stimme vollkommen zu

6. Welche zusätzlichen statistischen Analysen und Darstellungsformen würden Sie sich für Ihre Arbeit wünschen?

6.1 Für den Fall der Auswertung einer einzelnen KoGra-Abfrage:

6.2 Für den Fall der Auswertung und des Vergleichs mehrerer KoGra-Abfragen:

7. Wie könnten die Erläuterungen zu den statistischen Analysen und Darstellungsformen verbessert werden?

8. Bislang gibt es eine Schnittstelle nur zwischen KoGra und dem Modul zur statistischen Analyse. Grundsätzlich könnte das System auch für andere Daten, die nicht aus KoGra stammen, geöffnet werden. Für Daten welcher Art, meinen Sie, ist eine solche Erweiterung sinnvoll?

9. Verrichten Sie Korpusarbeit, die nicht auf statistische Analyse hinausläuft? Wenn ja, bitte beschreiben Sie die kurz.

10. Haben Sie weitere Vorschläge zur Verbesserung des Systems?

B. Ergebnisse der Evaluation (ohne Freitextantworten)

4. 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

5. Es gab zwei Benutzungsszenarien: (1) Die Daten einer einzelnen KoGra-Abfrage wurden ausgewertet, (2) die Daten mehrerer KoGra-Abfragen wurden ausgewertet und im Zuge dessen verglichen. Inwieweit stimmen Sie den folgenden Aussagen zu? (Skala 1–7, stimme gar nicht zu – stimme vollkommen zu)

5.1 Für den Fall der Auswertung einer einzelnen KoGra-Abfrage:

5.1.1 Ich verstehe die Daten respektive Auswertungsergebnisse.

- R-Code zur Tabellenerzeugung:
 - AM = 5.17, SD = 1.77, Median = 5.5, Modus = 7
- Rohdaten (Tabelle):
 - AM = 6.0, SD = 1.0, Median = 6.0, Modus = 6
- normierte Werte (Tabelle):
 - AM = 5.83, SD = 1.77, Median = 6.5, Modus = 7

- normierte Werte (Diagramm):
 - AM = 5.83, SD = 1.77, Median = 6.5, Modus = 7
- relative Werte (Tabelle):
 - AM = 4.67, SD = 1.80, Median = 5.0, Modus = 6
- relative Werte (Diagramm):
 - AM = 5.17, SD = 1.57, Median = 5.5, Modi = 5, 6
- Chi-Quadrat-Test:
 - AM = 5.17, SD = 1.21, Median = 5.0, Modus = 4
- Phi Assoziationsstärke/Cramérs V Assoziationsstärke:
 - AM = 5.0, SD = 1.63, Median = 5.5, Modus = 6
- Chi-Quadrat-Test, erwartete Häufigkeiten (Tabelle):
 - AM = 4.0, SD = 1.0, Median = 4.0, Modus = 4
- Chi-Quadrat-Test, Residuen (Tabelle):
 - AM = 3.17, SD = 1.57, Median = 3.0, Modus = 3
- Assoziationsplot (Diagramm):
 - AM = 4.0, SD = 1.67, Median = 3.0, Modus = 3 (nur 5 Antworten)
- Mosaikplot (Diagramm):
 - AM = 3.67, SD = 1.49, Median = 3.5, Modus = 2
- Konfidenzintervalle (Tabelle):
 - AM = 5.17, SD = 1.07, Median = 5.0, Modi = 4, 5
- Konfidenzintervalle (Diagramm):
 - AM = 5.33, SD = 1.11, Median = 5.5, Modi = 4, 6
- Dispersionsmaße:
 - AM = 4.83, SD = 1.77, Median = 5.5, Modus = 6

5.1.2 Die Daten respektive Auswertungsergebnisse sind für meine Arbeit hilfreich.

- R-Code zur Tabellenerzeugung:
 - AM = 4.83, SD = 2.19, Median = 5.0, Modus = 7
- Rohdaten (Tabelle):
 - AM = 6.0, SD = 1.41, Median = 6.5, Modus = 7
- normierte Werte (Tabelle):
 - AM = 5.67, SD = 1.60, Median = 6.5, Modus = 7
- normierte Werte (Diagramm):
 - AM = 5.83, SD = 1.46, Median = 6.5, Modus = 7

- relative Werte (Tabelle):
 - AM = 5.83, SD = 1.46, Median = 6.5, Modus = 7
- relative Werte (Diagramm):
 - AM = 5.83, SD = 1.46, Median = 6.5, Modus = 7
- Chi-Quadrat-Test:
 - AM = 6.0, SD = 1.53, Median = 7.0, Modus = 7
- Phi Assoziationsstärke/Cramérs V Assoziationsstärke:
 - AM = 6.0, SD = 1.53, Median = 7.0, Modus = 7
- Chi-Quadrat-Test, erwartete Häufigkeiten (Tabelle):
 - AM = 5.5, SD = 1.61, Median = 6.0, Modus = 7
- Chi-Quadrat-Test, Residuen (Tabelle):
 - AM = 5.5, SD = 1.61, Median = 6.0, Modus = 7
- Assoziationsplot (Diagramm):
 - AM = 5.67, SD = 1.49, Median = 6.0, Modus = 7
- Mosaikplot (Diagramm):
 - AM = 4.33, SD = 1.60, Median = 3.5, Modus = 3
- Konfidenzintervalle (Tabelle):
 - AM = 6.0, SD = 1.41, Median = 6.5, Modus = 7
- Konfidenzintervalle (Diagramm):
 - AM = 6.17, SD = 1.46, Median = 7.0, Modus = 7
- Dispersionsmaße:
 - AM = 5.33, SD = 1.80, Median = 6.0, Modus = 7

5.1.3 Die Erläuterungen der Daten respektive der Auswertungsergebnisse sind hilfreich.

- R-Code zur Tabellenerzeugung:
 - AM = 5.67, SD = 1.25, Median = 6.0, Modi = 4, 6, 7
- Rohdaten (Tabelle):
 - AM = 6.17, SD = 1.07, Median = 6.5, Modus = 7
- normierte Werte (Tabelle):
 - AM = 6.33, SD = 1.11, Median = 7.0, Modus = 7
- normierte Werte (Diagramm):
 - AM = 6.33, SD = 1.11, Median = 7.0, Modus = 7
- relative Werte (Tabelle):
 - AM = 5.67, SD = 1.60, Median = 6.5, Modus = 7

- relative Werte (Diagramm):
 - AM = 6.0, SD = 1.15, Median = 6.5, Modus = 7
- Chi-Quadrat-Test:
 - AM = 5.20, SD = 1.47, Median = 6.0, Modus = 6 (nur 5 Antworten)
- Phi Assoziationsstärke/Cramérs V Assoziationsstärke:
 - AM = 5.83, SD = 1.46, Median = 6.5, Modus = 7
- Chi-Quadrat-Test, erwartete Häufigkeiten (Tabelle):
 - AM = 4.83, SD = 1.95, Median = 5.0, Modus = 7
- Chi-Quadrat-Test, Residuen (Tabelle):
 - AM = 4.83, SD = 1.95, Median = 5.0, Modus = 7
- Assoziationsplot (Diagramm):
 - AM = 5.83, SD = 0.90, Median = 6.0, Modus = 7
- Mosaikplot (Diagramm):
 - AM = 5.33, SD = 1.37, Median = 6.0, Modus = 7
- Konfidenzintervalle (Tabelle):
 - AM = 5.67, SD = 1.11, Median = 5.5, Modi = 5, 7
- Konfidenzintervalle (Diagramm):
 - AM = 4.83, SD = 1.57, Median = 5.0, Modus = 5
- Dispersionsmaße:
 - AM = 5.83, SD = 1.34, Median = 6.0, Modus = 7

5.1.4 Die Erläuterungen der Daten respektive der Auswertungsergebnisse erfüllen mein Informationsbedürfnis.

- R-Code zur Tabellenerzeugung:
 - AM = 5.5, SD = 1.61, Median = 6.0, Modus = 7
- Rohdaten (Tabelle):
 - AM = 6.17, SD = 1.21, Median = 7.0, Modus = 7
- normierte Werte (Tabelle):
 - AM = 6.17, SD = 1.21, Median = 7.0, Modus = 7
- normierte Werte (Diagramm):
 - AM = 6.0, SD = 1.54, Median = 6.5, Modus = 7
- relative Werte (Tabelle):
 - AM = 6.0, SD = 1.54, Median = 6.5, Modus = 7
- relative Werte (Diagramm):
 - AM = 6.0, SD = 1.54, Median = 6.5, Modus = 7

- Chi-Quadrat-Test:
 - AM = 5.17, SD = 1.07, Median = 5.0, Modi = 4, 5
- Phi Assoziationsstärke/Cramérs V Assoziationsstärke:
 - AM = 4.83, SD = 1.21, Median = 5.0, Modus = 5
- Chi-Quadrat-Test, erwartete Häufigkeiten (Tabelle):
 - AM = 4.17, SD = 1.07, Median = 4.5, Modus = 5
- Chi-Quadrat-Test, Residuen (Tabelle):
 - AM = 4.0, SD = 1.10, Median = 4.0, Modi = 4, 5 (nur 5 Antworten)
- Assoziationsplot (Diagramm):
 - AM = 4.83, SD = 1.07, Median = 5.0, Modi = 5, 6
- Mosaikplot (Diagramm):
 - AM = 4.67, SD = 1.11, Median = 4.5, Modi = 4, 6
- Konfidenzintervalle (Tabelle):
 - AM = 5.0, SD = 1.29, Median = 5.0, Modus = 5
- Konfidenzintervalle (Diagramm):
 - AM = 4.17, SD = 1.34, Median = 4.5, Modus = 5
- Dispersionsmaße:
 - AM = 5.17, SD = 1.46, Median = 5.0, Modi = 5, 7

5.2 Für den Fall der Auswertung und des Vergleichs mehrerer KoGra-Abfragen:

5.2.1 Ich verstehe die Daten respektive Auswertungsergebnisse.

- R-Code zur Tabellenerzeugung:
 - AM = 5.33, SD = 1.89, Median = 6.0, Modus = 7
- Rohdaten (Tabelle):
 - AM = 6.17, SD = 1.07, Median = 6.5, Modus = 7
- normierte Werte (Tabelle):
 - AM = 6.0, SD = 1.41, Median = 6.5, Modus = 7
- normierte Werte (Diagramm):
 - AM = 6.0, SD = 1.41, Median = 6.5, Modus = 7
- relative Werte (Tabelle):
 - AM = 6.60, SD = 0.49, Median = 7.0, Modus = 7 (nur 5 Antworten)
- relative Werte (Diagramm, gestapelt):
 - AM = 5.83, SD = 1.46, Median = 6.5, Modus = 7

- relative Werte (Diagramm, gruppiert):
 - AM = 5.5 , SD = 1.61, Median = 6.0, Modus = 7
- Chi-Quadrat-Test:
 - AM = 5.0 , SD = 1.41, Median = 5.0, Modus = 3, 4, 5, 6, 7 (nur 5 Antworten)
- Phi Assoziationsstärke/Cramér's V Assoziationsstärke:
 - AM = 4.33, SD = 1.60, Median = 4.5, Modus = 5
- Chi-Quadrat-Test, erwartete Häufigkeiten (Tabelle):
 - AM = 4.5 , SD = 1.38, Median = 4.5, Modi = 3, 5
- Chi-Quadrat-Test, Residuen (Tabelle):
 - AM = 4.5 , SD = 1.38, Median = 4.5, Modi = 3, 5
- Assoziationsplot (Diagramm):
 - AM = 5.0, SD = 1.30, Median = 5.0, Modus = 5
- Mosaikplot (Diagramm):
 - AM = 4.5, SD = 1.61, Median = 4.0, Modus = 3
- Konfidenzintervalle (Tabelle):
 - AM = 5.17, SD = 1.21, Median = 5.0, Modus = 5
- Konfidenzintervalle (Diagramm):
 - AM = 4.83, SD = 0.90, Median = 5.0, Modus = 5
- Dispersionsmaße:
 - AM = 4.83, SD = 1.67, Median = 4.5, Modi = 3, 7

5.2.2 Die Daten respektive Auswertungsergebnisse sind für meine Arbeit hilfreich.

- R-Code zur Tabellenerzeugung:
 - AM = 5.60, SD = 1.74, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Rohdaten (Tabelle):
 - AM = 6.40, SD = 0.80, Median = 7.0, Modus = 7 (nur 5 Antworten)
- normierte Werte (Tabelle):
 - AM = 6.20, SD = 0.98, Median = 7.0, Modus = 7 (nur 5 Antworten)
- normierte Werte (Diagramm):
 - AM = 6.40, SD = 0.80, Median = 7.0, Modus = 7 (nur 5 Antworten)
- relative Werte (Tabelle):
 - AM = 6.0, SD = 1.26, Median = 7.0, Modus = 7 (nur 5 Antworten)
- relative Werte (Diagramm, gestapelt):
 - AM = 6.02, SD = 1.17, Median = 7.0, Modus = 7 (nur 5 Antworten)

- relative Werte (Diagramm, gruppiert):
 - AM = 5.60, SD = 1.74, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Chi-Quadrat-Test:
 - AM = 6.40, SD = 1.2, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Phi Assoziationsstärke/Cramérs V Assoziationsstärke:
 - AM = 6.40, SD = 1.2, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Chi-Quadrat-Test, erwartete Häufigkeiten (Tabelle):
 - AM = 5.60, SD = 1.74, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Chi-Quadrat-Test, Residuen (Tabelle):
 - AM = 5.60, SD = 1.74, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Assoziationsplot (Diagramm):
 - AM = 6.40, SD = 1.2, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Mosaikplot (Diagramm):
 - AM = 5.80, SD = 1.60, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Konfidenzintervalle (Tabelle):
 - AM = 6.20, SD = 0.98, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Konfidenzintervalle (Diagramm):
 - AM = 6.20, SD = 0.98, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Dispersionsmaße:
 - AM = 5.60, SD = 1.74, Median = 7.0, Modus = 7 (nur 5 Antworten)

5.2.3 Die Erläuterungen der Daten respektive der Auswertungsergebnisse sind hilfreich.

- R-Code zur Tabellenerzeugung:
 - AM = 5.80, SD = 1.47, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Rohdaten (Tabelle):
 - AM = 6.20, SD = 1.17, Median = 7.0, Modus = 7 (nur 5 Antworten)
- normierte Werte (Tabelle):
 - AM = 5.80, SD = 1.47, Median = 7.0, Modus = 7 (nur 5 Antworten)
- normierte Werte (Diagramm):
 - AM = 5.80, SD = 1.47, Median = 7.0, Modus = 7 (nur 5 Antworten)
- relative Werte (Tabelle):
 - AM = 5.60, SD = 1.74, Median = 7.0, Modus = 7 (nur 5 Antworten)
- relative Werte (Diagramm, gestapelt):
 - AM = 6.20, SD = 1.17, Median = 7.0, Modus = 7 (nur 5 Antworten)

- relative Werte (Diagramm, gruppiert):
 - AM = 6.0, SD = 1.55, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Chi-Quadrat-Test:
 - AM = 5.60, SD = 1.50, Median = 6.0, Modus = 7 (nur 5 Antworten)
- Phi Assoziationsstärke/Cramérs V Assoziationsstärke:
 - AM = 5.20, SD = 1.60, Median = 5.0, Modus = 7 (nur 5 Antworten)
- Chi-Quadrat-Test, erwartete Häufigkeiten (Tabelle):
 - AM = 5.0, SD = 1.67, Median = 4.0, Modi = 4, 7 (nur 5 Antworten)
- Chi-Quadrat-Test, Residuen (Tabelle):
 - AM = 5.0, SD = 1.67, Median = 4.0, Modi = 4, 7 (nur 5 Antworten)
- Assoziationsplot (Diagramm):
 - AM = 5.40, SD = 1.2, Median = 6.0, Modi = 4, 6 (nur 5 Antworten)
- Mosaikplot (Diagramm):
 - AM = 4.80, SD = 1.47, Median = 4.0, Modus = 4 (nur 5 Antworten)
- Konfidenzintervalle (Tabelle):
 - AM = 5.40, SD = 1.36, Median = 5.0, Modi = 4, 7 (nur 5 Antworten)
- Konfidenzintervalle (Diagramm):
 - AM = 4.40, SD = 1.62, Median = 4.0, Modus = 4 (nur 5 Antworten)
- Dispersionsmaße:
 - AM = 6.0, SD = 1.55, Median = 7.0, Modus = 7 (nur 5 Antworten)

5.2.4 Die Erläuterungen der Daten respektive der Auswertungsergebnisse erfüllen mein Informationsbedürfnis.

- R-Code zur Tabellenerzeugung:
 - AM = 5.60, SD = 1.74, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Rohdaten (Tabelle):
 - AM = 5.80, SD = 1.47, Median = 7.0, Modus = 7 (nur 5 Antworten)
- normierte Werte (Tabelle):
 - AM = 5.80, SD = 1.47, Median = 7.0, Modus = 7 (nur 5 Antworten)
- normierte Werte (Diagramm):
 - AM = 6.20, SD = 1.17, Median = 7.0, Modus = 7 (nur 5 Antworten)
- relative Werte (Tabelle):
 - AM = 6.20, SD = 1.17, Median = 7.0, Modus = 7 (nur 5 Antworten)
- relative Werte (Diagramm, gestapelt):
 - AM = 6.20, SD = 1.17, Median = 7.0, Modus = 7 (nur 5 Antworten)

- relative Werte (Diagramm, gruppiert):
 - AM = 6.20, SD = 1.17, Median = 7.0, Modus = 7 (nur 5 Antworten)
- Chi-Quadrat-Test:
 - AM = 5.0, SD = 1.10, Median = 5.0, Modi = 4, 5 (nur 5 Antworten)
- Phi Assoziationsstärke/Cramérs V Assoziationsstärke:
 - AM = 4.78, SD = 1.33, Median = 5.0, Modus = 5 (nur 5 Antworten)
- Chi-Quadrat-Test, erwartete Häufigkeiten (Tabelle):
 - AM = 5.0, SD = 1.10, Median = 5.0, Modi = 4, 5 (nur 5 Antworten)
- Chi-Quadrat-Test, Residuen (Tabelle):
 - AM = 5.0, SD = 1.10, Median = 5.0, Modi = 4, 5 (nur 5 Antworten)
- Assoziationsplot (Diagramm):
 - AM = 5.0, SD = 0.63, Median = 5.0, Modus = 5 (nur 5 Antworten)
- Mosaikplot (Diagramm):
 - AM = 4.80, SD = 0.75, Median = 5.0, Modi = 4, 5 (nur 5 Antworten)
- Konfidenzintervalle (Tabelle):
 - AM = 5.20, SD = 0.75, Median = 5.0, Modi = 5, 6 (nur 5 Antworten)
- Konfidenzintervalle (Diagramm):
 - AM = 4.40, SD = 1.36, Median = 5.0, Modus = 5 (nur 5 Antworten)
- Dispersionsmaße:
 - AM = 5.20, SD = 1.60, Median = 5.0, Modus = 7 (nur 5 Antworten)