Erschienen in: Konopka, Marek/Wöllstein, Angelika (Hrsg.): Grammatische Variation. Empirische Zugänge und theoretische Modellierung. - Berlin/Boston: de Gruyter, 2017. S. 345-356. (Jahrbuch des Instituts für deutsche Sprache 2016)

Sandra Hansen-Morath (Mannheim)/Sascha Wolfer (Mannheim)

Standardisierte statistische Auswertung von Korpusdaten im Projekt „Korpusgrammatik“ (KoGra-R)

Abstract: Wir zeigen anhand dreier Beispielanalysen, wie das im IDS-Projekt „Korpusgrammatik“ entwickelte Auswertungstool KoGra-R in der quantitativ-linguistischen Forschung zur Analyse von Frequenzdaten auf mehreren linguistischen Ebenen eingesetzt werden kann. Wir demonstrieren dies anhand regionaler Präferenzen bei der Selektion von Genitivallomorphen, der Variation von Relativpronomina sowie der Verwendung bestimmter anaphorischer Ausdrücke in Abhängigkeit davon, ob sich das Antezedens im gleichen Satz befindet oder nicht. Die in KoGra-R implementierten statistischen Tests sind für jede dieser Ebenen geeignet, um mindestens einen ersten statistisch abgesicherten Eindruck der Datenlage zu erlangen.

1 Einleitung

Im Projekt Korpusgrammatik werden u.a. Techniken und Werkzeuge entwickelt, um grammatische Phänomene mit Bezug auf große Korpora geschriebener Sprache zu beschreiben, Korpusdaten zu explorieren sowie eine transparente quantitativ-statistische Basis für die Validierung linguistischer Hypothesen bereitzustellen (vgl. Hansen-Morath et al. i.Vorb.). Im Rahmen von Pilotstudien (vgl. Brandt/Fuß i.Vorb.; Konopka/Fuß 2016; Bubenhofer/Konopka/Schneider 2014, S. 125 ff.; Konopka/Waßner 2013) wurden statistische Analysen erprobt, die sich für die erste Exploration der Daten als sinnvoll erwiesen haben (vgl. Hansen-Morath et al. i.Vorb.). Um die Analysen auf einfache Weise zugänglich zu machen, wurde KoGra-R entwickelt.[1] Das web-basierte Tool ist über die Internetadresse www.ids-mannheim.de/kogra-r frei zugänglich. Intern soll eine Menge an standardisierten statistischen Werkzeugen für die Forschungsarbeit im Projekt bereitgestellt werden, die bei allen Studien im Rahmen der Korpusgrammatik zur Anwendung kommen sollen (vgl. ebd.).

1 Die statistischen Analysen in KoGra-R sind in R realisiert, einer Softwareumgebung für statistische Verarbeitung und Visualisierung. Zu weiteren Details der technischen Realisierung von KoGra-R vgl. Hansen-Morath et al. (i.Vorb.).

Die verfügbaren Analysen basieren auf der Auswertung von Kontingenztabellen. Die Analysefunktionen sind prinzipiell offen und können erweitert werden. Bislang sind folgende Funktionen verfügbar:

- die Darstellung von Tabellen und Diagrammen für Rohdaten, normierte und relative Werte,
- die Berechnung des Chi²-Tests sowie die Berechnung von erwarteten Häufigkeiten und Residuen,
- die Berechnung der Assoziationsstärke Phi bzw. Cramérs V,
- die Darstellung von Assoziations- und Mosaikplots,
- die Darstellung von Tabellen und Diagrammen für Konfidenzintervalle und
- die Berechnung von Dispersionsmaßen, insbesondere der DPnorm (Gries 2008, 2009; Lijffijt/Gries 2012).

In KoGra-R können von COSMAS II[2] erzeugte Frequenzlisten für eine oder mehrere Phänomenrealisierungen sowie beliebige anderweitig erzeugte Kontingenztabellen mit Informationen über die Häufigkeit von Realisierungen statistisch ausgewertet werden (vgl. Hansen-Morath et al. i.Vorb.). Im Folgenden werden wir einen Großteil der verfügbaren Funktionen exemplifizieren.

Der Beitrag ist wie folgt aufgebaut: Zunächst werden ausgewählte Auswertungsmöglichkeiten von KoGra-R anhand der regionalen Verteilung von Genitivallomorphen dargestellt. In Abschnitt 2.2 zeigen wir am Beispiel der Variation zwischen den Relativpronomina *das* und *was*, dass die Analyseverfahren in KoGra-R sich auch dazu eignen, Fragestellungen auf syntaktischer Ebene zu untersuchen. In Abschnitt 2.3 schließlich analysieren wir mithilfe des Tools die Variation zwischen vollen Nominalphrasen und Proformen bezüglich der Referenzherstellung über eine Satzgrenze hinweg.

2 Beispielanalysen

2.1 Die regionale Verteilung von Genitivallomorphen

Fürbacher (2015) zeigt in einer Studie zu Genitivallomorphen bei starken Maskulina und Neutra, dass für die Variation der Genitivallomorphe neben sprachimmanenten Faktoren auch regional bedingte Einflüsse eine Rolle spielen können.

2 Das Corpus Search, Management and Analysis System des IDS (COSMAS II) verfügt über eine Web-Schnittstelle: https://cosmas2.ids-mannheim.de/cosmas2-web/ (Stand: 9.5.2016).

Die nominalen Genitive können dabei in der Standardsprache mehrere Genitivallomorphe zulassen (vgl. Konopka/Fuß 2016, S. 14 ff.). Zum einen existieren Nomina, die ein Genitivallomorph zulassen oder stark präferieren, zum anderen können Lemmata mit verschiedenen Genitivendungen vorkommen (vgl. Fürbacher 2015; Konopka/Fuß 2016). Fürbacher (2015) analysiert auf Basis von regional differenzierten Daten das regionale Vorkommen der Allomorphe *-es* und *-s* für ausgewählte Lemmata. Hierzu wurde die im Projekt Korpusgrammatik entstandene Datenbank GenitivDB[3] herangezogen, mit deren Hilfe die regionale Verteilung der beiden Genitivallomorphe untersucht werden kann. Die Daten liegen in der Datenbank annotiert mit verschiedenen Metadaten (u.a. zur regionalen Zuordnung) vor.[4] Es wurden insgesamt 20 Lemmata untersucht, die mindestens 100-mal im Korpus vorkommen und „die ausgewogenste Verteilung auf beide Allomorphe besitzen" (Fürbacher 2015).

Tabelle 1 zeigt eine Kontingenztabelle mit den absoluten Trefferzahlen für die Lemmata mit den Genitivendungen *-es* und *-s* in den Texten aus den Regionen Nordost, Südost (inklusive Österreich), Nordwest und Südwest (inklusive der Schweiz).

Tab. 1: Absolute Trefferzahlen für die Lemmata mit den verschiedenen Genitivendungen in Abhängigkeit von unterschiedlichen Regionen

	-es	***-s***
Nordost	425	369
Südost	1435	573
Nordwest	239	235
Südwest	435	438

Neben der normierten Darstellung der Daten (vgl. Hansen-Morath et al. i.Vorb.) werden in einem ersten Auswertungsschritt in KoGra-R die Daten relativiert, indem für jede Realisierung die absoluten Häufigkeiten der Tokens in den einzelnen Regionen durch die Gesamthäufigkeit geteilt werden. Die relativen Häufigkeiten werden als Prozentwerte ebenfalls in Form einer Kontingenztabelle (vgl. Tab. 2) dargestellt.

3 Die GenitivDB ist über folgende Seite verfügbar: http://hypermedia.ids-mannheim.de/call/public/korpus.genitivdb (Stand: 3.5.2016).

4 Zu regionalen und anderen Metadaten, die im Rahmen des Korpusgrammatik-Projektes erfasst werden, vgl. Bubenhofer/Konopka/Schneider (2014, S. 84 ff.).

Tab. 2: Relative Werte für die Lemmata mit den verschiedenen Genitivendungen in Abhängigkeit von unterschiedlichen Regionen

	-es	*-s*
Nordost	16.77	22.85
Südost	56.63	35.48
Nordwest	9.43	14.55
Südwest	17.17	27.12

Um die Daten besser interpretieren zu können, können in KoGra-R die Häufigkeiten in gruppierten und gestapelten Säulendiagrammen dargestellt werden. Darüber hinaus werden Konfidenzintervalle berechnet, mit deren Angabe man von den vorliegenden Strichprobenergebnissen auf die Grundgesamtheit schließen kann (vgl. Brunner/Munzel 2013, S. 80 ff.). Bei einem Konfidenzniveau von 95% ist davon auszugehen, dass bei einer Neuberechnung des Konfidenzintervalls aus jeweils neuen Stichproben 95% der Intervalle den Populationsparameter enthalten (vgl. Hansen-Morath et al. i.Vorb.). Liegen die berechneten Werte der Intervalle einer Realisierung (z.B. der Genitivmarkierung *-es* im vorliegenden Beispiel) nicht im Intervall der Häufigkeit einer anderen Phänomenrealisierung (z.B. der Genitivmarkierung *-s*), deutet dies auf einen signifikanten Unterschied im Vorkommen der beiden Varianten hin. In KoGra-R werden die Konfidenzintervalle für jede Realisierung tabellarisch sowie in einem Säulendiagramm (vgl. Abb. 1) dargestellt.

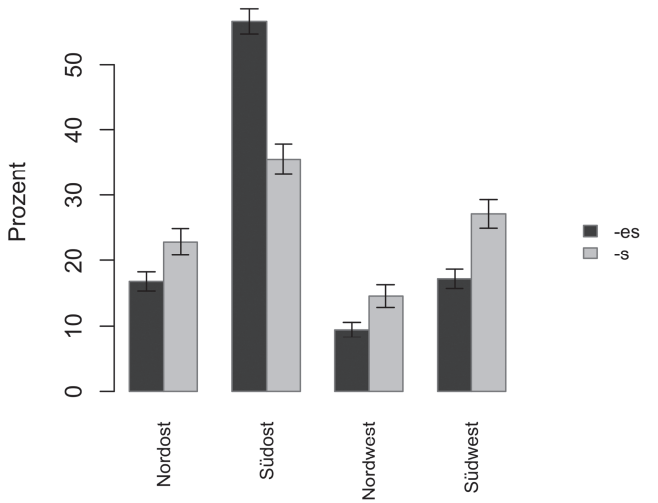

Abb. 1: Gruppiertes Säulendiagramm mit Konfidenzintervallen für jedes Genitivallomorph in Abhängigkeit von Regionen

Das Diagramm zeigt die prozentuale Verteilung der Genitivallomorphe in Abhängigkeit von der regionalen Zuteilung der Texte mit den entsprechenden Konfidenzintervallen. Beide Varianten kommen am häufigsten in Texten aus dem Südosten vor, wobei sich die Variante *-es* stärker als die Variante *-s* auf diese Texte konzentriert.

Um genauer zu prüfen, ob die Wahl zwischen den Realisierungsvarianten der Genitivmarkierung von den Regionen abhängig ist, in denen die Texte entstanden sind, eignet sich der Chi-Quadrat-Test (vgl. Bortz 2005, S. 168 ff.). Hierzu werden die beobachteten Häufigkeiten mit erwarteten Häufigkeiten verglichen, die im Verhältnis zur jeweiligen Größe der regional differenzierten Teilkorpora stehen.[5] Die Berechnung des Chi-Quadrat-Tests ist in KoGra-R implementiert und ergibt für das vorliegende Beispiel folgende Werte:[6]

X-squared = 179.27, df = 3, p-value < 2.2e-16

Die Verteilungen der beiden Realisierungsvarianten in Abhängigkeit von Regionen unterscheiden sich demnach höchst signifikant.

Da nicht nur die statistische Signifikanz für die Bedeutsamkeit eines Ergebnisses ausschlaggebend ist, sollte auch die Stärke des Zusammenhangs der Variablen betrachtet werden. In KoGra-R wird hierzu je nach Anzahl der Spalten und Zeilen in einer Tabelle der Phi-Koeffizient bzw. Cramérs V berechnet (vgl. Bortz/Lienert 2008, S. 259 ff., 271 ff.). Beide Koeffizienten weisen ein Wertespektrum zwischen 0 und 1 auf; je höher der Wert ist, desto stärker ist der Zusammenhang.[7] Für die vorliegende Analyse wird in KoGra-R folgender Wert ausgegeben:

Cramérs V / Phi: 0.21

Zwar unterscheiden sich die Verteilungen der beiden Genitivallomorphe in Abhängigkeit von Regionen signifikant voneinander, allerdings ist die Assoziationsstärke, d.h. die Stärke des Effekts, mit einem Koeffizienten von 0.21 eher als klein einzuschätzen.

---

5 Für eine genaue Beschreibung des Chi-Quadrat-Tests in KoGra-R vgl. Hansen-Morath et al. (i.Vorb.).

6 Darüber hinaus werden in KoGra-R die erwarteten Werte und deren Abweichungen als Pearson Residuen angezeigt.

7 Zu den Schwellen der Assoziationsstärke vgl. die Dokumentation zu KoGra-R: http://kograno.ids-mannheim.de/kograr-doku-statistik.html (Stand: 13.6.2016).

Eine Möglichkeit, die standardisierten Abweichungen zwischen beobachteten und erwarteten Häufigkeiten darzustellen, bietet der Assoziationsplot (vgl. Cohen 1980; Friendly 1992; Meyer/Zeileis/Hornik 2005). Abbildung 2 zeigt einen Assoziationsplot für die vorliegenden Daten.

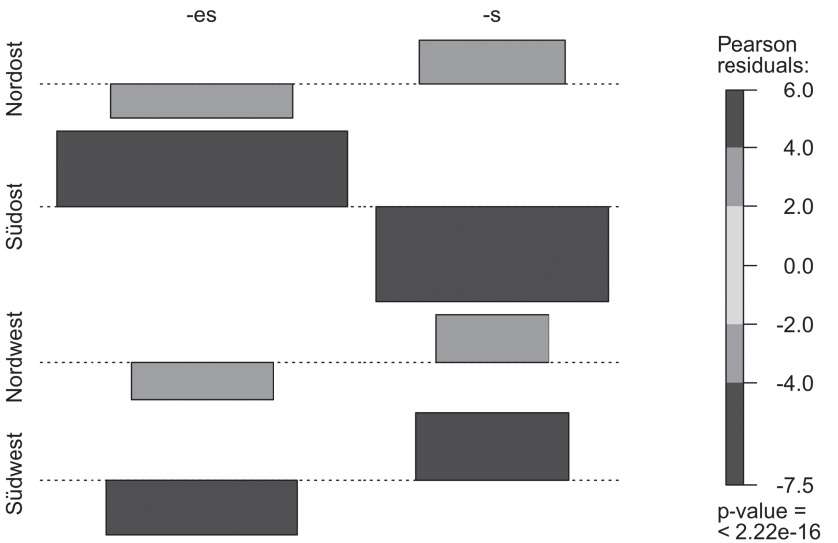

Abb. 2: Assoziationsplot für die Genitivallomorphe *-es* und *-s* in Abhängigkeit von Regionen (In der Online-Version sind Abweichungen nach oben durch eine blaue und Abweichungen nach unten durch eine rote Einfärbung dargestellt.)

Die Höhe der Balken entspricht dabei 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 spiegelt die erwartete Frequenz der Realisierungsvarianten wider. Ist der Betrag des entsprechenden Pearson-Residuums größer als 1.97, wird der Balken im Plot eingefärbt (vgl. Hansen-Morath et al. i.Vorb.).

Für das vorliegende Analysebeispiel kann festgehalten werden, dass der Wert der Residuen, d.h. der standardisierten Abweichungen der beobachteten von den erwarteten Häufigkeiten, im Südosten mit (gerundet) 6.0 für *-es* bzw. -7.5 für *-s*, aber auch im Südwesten mit -4.3 für *-es* bzw. 5.3 für *-s* auf eine besondere Bedeutung der Abweichung für die Signifikanz der Unterschiede hinweist (vgl. Fürbacher 2015). Der signifikant häufigere Gebrauch der *es*-Endung im Südosten könnte durch eine kompensierende Funktion gegenüber der gesprochenen Sprache bedingt sein, da in dieser der Schwa-Laut gemieden wird (vgl. ebd.; Tatzreiter 1988, S. 79).

2.2 Die Variation der Relativpronomina *was* vs. *das* in attributiven Relativsätzen

In einer Pilotstudie zur syntaktischen Variation im Deutschen beschreiben Brandt/Fuß (i.Vorb.) „Distribution und Eigenschaften verschiedener, z.T. miteinander konkurrierender Strategien, die im Deutschen zur Einleitung von Relativsätzen Verwendung finden". Im Mittelpunkt der umfassenden Studie steht die Alternation zwischen *das* und *was* in attributiven Relativsätzen.

Im Rahmen der Studie wurde eine Extraktion einschlägiger Belege aus dem Deutschen Referenzkorpus (DeReKo; Kupietz et al. 2010) vorgenommen. Die Beleglisten wurden nach bestimmten Kriterien sortiert und reduziert und mit verschiedenen Informationen angereichert. Das resultierende Korpus umfasst 18.307 Belegsätze (vgl. Brandt/Fuß i.Vorb.). Im Folgenden wird ein Teilergebnis der Analyse vorgestellt.

Brandt/Fuß stellten fest, dass die Präsenz quantifizierender Determinative wie *das einzig*(*e*) oder *alles* in Kombination mit verschiedenen substantivierten Adjektiven (z.B. *das einzig*(*e*)/*alles Gute*/*Schöne*/*Neue*) als Bezugselemente, einen Einfluss auf die Wahl des Relativums hat.

Für die statistische Analyse wurden die Häufigkeiten der entsprechenden Nominalphrasen zusammengefasst. Folgende Kontingenztabelle ging in die Analyse ein.

Tab. 3: Häufigkeiten der Relativpronomina in Abhängigkeit von den quantifizierenden Determinativen bei den untersuchten Adjektiven (vgl. Brandt/Fuß i.Vorb.)

	was	***das***
das einzig(*e*) *Gute/Schöne/Neue*	56	36
alles Gute/Schöne/Neue	74	96
vieles/das/... Gute/Schöne/Neue	124	681

Neben den Häufigkeiten der Realisierungen bei quantifizierenden Determinativen *das einzig*(*e*) und *alles* gehen als Vergleichsgröße die Häufigkeiten von *das* vs. *was* in allen anderen Daten mit *Gute*/*Schöne*/*Neue ein*, die im Rahmen der Analyse durch COSMAS-Recherchen erhoben wurden.

Der Chi-Quadrat Test ergibt mit einem p-Wert von $2.2*10^{-16}$ (der kleinste Wert, der hier von R ausgegeben werden kann), dass sich die Häufigkeiten der Relativpronomina in Abhängigkeit von den drei Kontexttypen signifikant unterscheiden. Die Assoziationsstärke Cramérs V beträgt 0.36. Der Zusammenhang ist somit als mittelstark einzustufen. Der Assoziationsplot (Abb. 3) zeigt, dass *was* im

Zusammenhang mit den quantifizierenden Determinativen *das einzig*(*e*) und *alles* deutlich überrepräsentiert ist, während *das* in diesen Konstruktionen deutlich unterrepräsentiert ist.

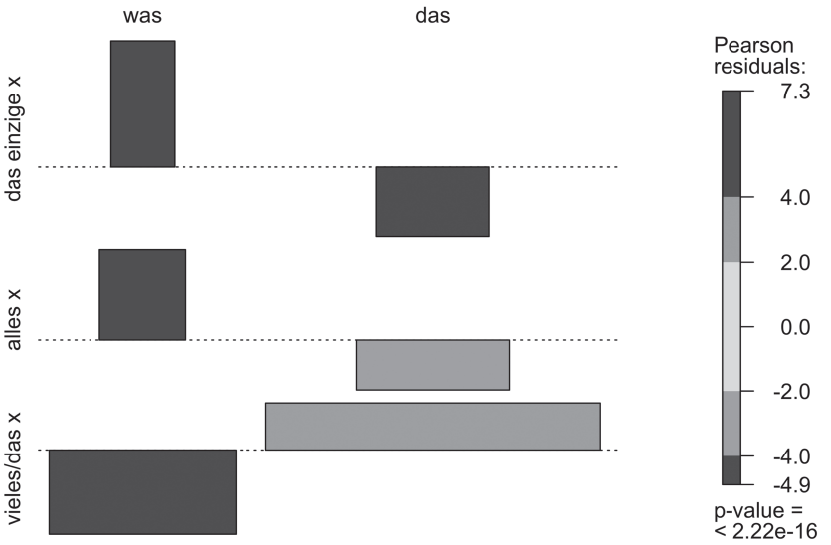

Abb. 3: Assoziationsplot für den Einfluss quantifizierender Determinative auf die *das*/*was*-Alternation (substantivierte Adjektive) (In der Online-Version sind Abweichungen nach oben durch eine blaue und Abweichungen nach unten durch eine rote Einfärbung dargestellt.)

In allen anderen Konstruktionen (*vieles*/*das*/...) kehrt sich das Verhältnis um: Das Relativpronomen *was* wird signifikant seltener verwendet, während *das* überrepräsentiert ist.

2.3 Referenzherstellung über Nominalphrasen und Proformen

Zuletzt werden wir linguistische Verteilungsdaten auf der Textebene analysieren. Datengrundlage ist ein Ausschnitt der Koreferenz-Annotation eines juristisch-fachsprachlichen Korpus (Wolfer i.Dr.). Zu Demonstrationszwecken fokussieren wir auf anaphorische Referenzen und die folgenden Informationen:

– Form der Anapher: Ist der Referenzausdruck eine volle Nominalphrase (mindestens ein Determinativ und ein Nomen) oder eine Proform (Personalpronomen oder Demonstrativpronomen)?

– Überschreiten einer Satzgrenze: Liegt das Antezedens des referenziellen Ausdrucks im gleichen Satz (satzintern) wie der referenzielle Ausdruck, oder überschreitet die Referenzbeziehung eine Satzgrenze (satzextern)?

Es liegen Informationen zu 1295 referenziellen Ausdrücken vor, die sich folgendermaßen verteilen (siehe Tab. 4).

Tab. 4: Verteilung von referenziellen Ausdrücken gemäß der beiden Variablen „Form der Anapher“ (Spalten) und „Überschreiten einer Satzgrenze“ (Zeilen)

	volle Nominalphrase	**Proform**
satzintern	178	126
satzextern	879	112

Es ist zu erwarten, dass volle Nominalphrasen eher dazu verwendet werden, um über Satzgrenzen hinweg zu referieren, als Proformen.[8] Ein Grund dafür ist, dass volle Nominalphrasen explizit sind. Daher sind sie besser dazu geeignet, Diskursreferenten zu reaktivieren, deren Aktivation in der mentalen Textrepräsentation der Leserin/des Lesers zwischenzeitlich gesunken ist. Proformen sind im Vergleich zu vollen Nominalphrasen unterspezifiziert und werden daher eher dazu eingesetzt, um auf einen Referenten Bezug zu nehmen, der noch im expliziten Fokus des augenblicklichen Verarbeitungsvorgangs steht (vgl. Garrod/Freudenthal/Boyle 1994, S. 41). Ein Blick auf Tabelle 4 legt nahe, dass sich dieses Muster auch in den erhobenen Daten zeigt. Mit Hilfe von KoGra-R können wir diesen Eindruck statistisch absichern.

Der Chi-Quadrat-Test zeigt mit einer Prüfgröße von $Chi^2 = 139.93$ und einer Irrtumswahrscheinlichkeit von $p < 2.2^{*}10^{-16}$ eine höchst signifikante Abweichung von einer Gleichverteilung. Die Assoziationsstärke zeigt mit Phi = 0.33 einen mittelstarken Zusammenhang der Variablen an. Das Muster kann außer mit den bisher verwendeten Assoziationsplots auch über einen Mosaikplot visualisiert werden (siehe Abb. 4).

8 Innerhalb der Proformen gilt dies nicht für Demonstrativpronomina, die auch häufig dazu verwendet werden, über eine Satzgrenze hinweg einen referenziellen Bezug herzustellen. Im Korpus, auf dem die Auszählung hier beruht, wird dies jedoch durch die geringe Anzahl von Demonstrativpronomina relativ zu Pronomina verschattet.

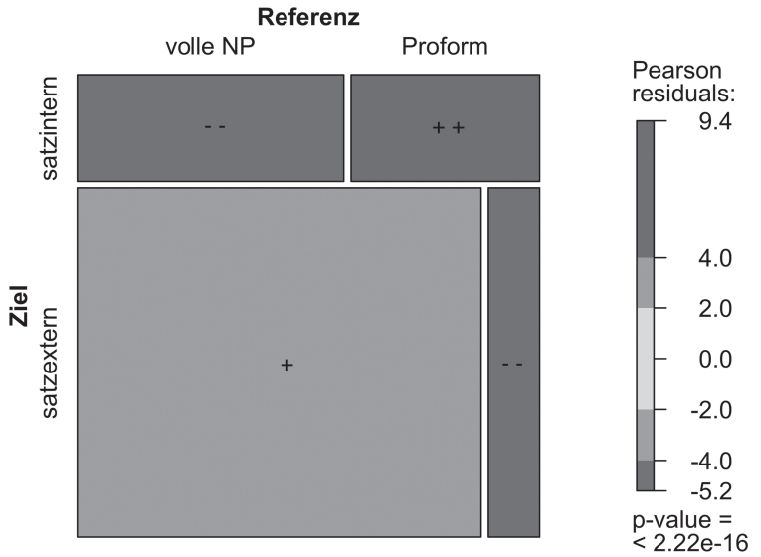

Abb. 4: Mosaikplot für den Zusammenhang zwischen der Form der Anapher (linke Beschriftung, in Zeilen) und die Referenz über eine Satzgrenze hinweg (obere Beschriftung in Spalten). (Abweichungen nach oben bzw. unten werden für den Graustufen-Druck über die Symbole + und – symbolisiert (bei stärkeren Abweichungen werden die Symbole gedoppelt). In der Online-Version sind Abweichungen nach oben durch eine blaue und Abweichungen nach unten durch eine rote Einfärbung dargestellt.)

In Mosaikplots können die Häufigkeiten über die Flächeninhalte der einzelnen Rechtecke abgelesen werden. Jedes Rechteck steht dabei für eine Zelle der Kontingenztabelle (siehe Tab. 4). Der Flächeninhalt des Rechtecks rechts oben ist bspw. proportional zur Häufigkeit von vollen Nominalphrasen, die über eine Satzgrenze hinweg referieren. Es ist leicht abzulesen, dass diese Zelle innerhalb der Kontingenztabelle die meisten Fälle enthält.

Die Einfärbung der Rechtecke symbolisiert die standardisierten Pearson-Residuen. Ist ein Rechteck (in der Online-Version von KoGra-R) blau eingefärbt, weichen die beobachteten Häufigkeiten in der entsprechenden Zelle signifikant nach oben von den erwarteten Häufigkeiten ab. Die Zelle ist überrepräsentiert. Rot eingefärbte Rechtecke bedeuten eine signifikante Abweichung nach unten. Was im Assoziationsplot über die Höhe der Abweichungsbalken symbolisiert wird, wird im Mosaikplot somit lediglich über die Farbe kodiert. Dafür werden im Mosaikplot zusätzlich die Häufigkeiten dargestellt.

Der Mosaikplot als Ganzes zeigt, dass die Kombinationen „Proform/satzintern“ (unten links) und „volle NP/satzextern“ (oben rechts) überrepräsentiert

sind. Die komplementären Kombinationen „Proform/satzextern“ sowie „volle NP/satzintern“ sind unterrepräsentiert.

Wir können somit schließen, dass sich für das zugrundeliegende Korpus die Hypothese bestätigt, dass volle Nominalphrasen eher dazu verwendet werden, über eine Satzgrenze hinweg zu referieren. Proformen hingegen werden eher dazu eingesetzt, referenzielle Bezüge innerhalb desselben Satzes herzustellen.

3 Fazit

Wir haben gezeigt, wie KoGra-R verwendet werden kann, um Häufigkeitsverteilungen, wie sie in Kontingenztabellen gefasst werden, zu beschreiben sowie inferenzstatistisch auszuwerten und zu visualisieren. Durch die Zusammenstellung der Analyseverfahren ist es möglich, Variationsphänomene unterschiedlicher grammatischer Ebenen in Abhängigkeit von sprachlichen und außersprachlichen Faktoren gewinnbringend zu analysieren. Das Tool eignet sich besonders dazu, die unterschiedlichen Verteilungen der Varianten in verschiedenen Teilkorpora zu explorieren.

In Hansen-Morath et al. (i.Vorb.) gehen wir auf weitere potenzielle Erweiterungsmöglichkeiten von KoGra-R sowie verschiedene Einsatzszenarien ein. Gerade aufgrund der ausführlichen Dokumentation der Analysen und der Funktionalität von KoGra-R ist es möglich dieses Tool auch in der akademischen Lehre einzusetzen.

Literatur

Bortz, Jürgen (2005): Statistik für Human- und Sozialwissenschaftler. 6., vollst. überarb. und aktual. Aufl. Berlin.

Bortz, Jürgen/Lienert, Gustav A. (2008): Kurzgefasste Statistik für die klinische Forschung. 3., aktual. und bearb. Aufl. Heidelberg.

Brandt, Patrick/Fuß, Eric (i.Vorb.): Relativpronomenselektion und grammatische Variation: *was* vs. *das* in attributiven Relativsätzen. In: Fuß/Konopka/Wöllstein (Hg.).

Brunner, Edgar/Munzel, Ullrich (2013): Nicht-parametrische Datenanalyse. Unverbundene Stichproben. 2., überarb. Aufl. Berlin/Heidelberg.

Bubenhofer, Noah/Konopka, Marek/Schneider, Roman (2014): Präliminarien einer Korpusgrammatik. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 4). Tübingen.

Cohen, Ayala (1980): On the graphical display of the significant components in two-way contingency tables. In: Communications in Statistics – Theory and Methods 10, S. 1025–1041.

Friendly, Michael (1992): Graphical methods for categorical data. In: Proceedings of the SAS User's Group International Conference 17, S. 1367–1373. Internet: www.math.yorku.ca/SCS/sugi/sugi17-paper.html# (Stand: 3.8.2016).

Fürbacher, Monica (2015): Variation der starken Genitivmarkierung. Spezialstudie: Regionale Verteilung. In: grammis 2.0 – Korpusgrammatik. Elektronische Ressource. Mannheim. Internet: http://hypermedia.ids-mannheim.de/call/public/korpus.ansicht?v_id=5087 (Stand: 3.8.2016).

Fuß, Eric/Konopka, Marek/Wöllstein, Angelika (Hg.) (i.Vorb.): Grammatik im Korpus [Arbeitstitel]. (= Studien zur Deutschen Sprache). Tübingen.

Garrod, Simon C./Freudenthal, Daniel/Boyle, Elizabeth (1994): The role of different types of anaphor in the on-line resolution of sentences in a discourse. In: Journal of Memory and Language 33, S. 39–68.

Gries, Stefan Th. (2008): Dispersions and adjusted frequencies in corpora. In: International Journal of Corpus Linguistics 13, S. 403–437.

Gries, Stefan Th. (2009): Dispersions and adjusted frequencies in corpora: Further explorations. In: Language and Computers 71, S. 197–212.

Hansen-Morath, Sandra et al. (i.Vorb.): KoGra-R: Standardisierte statistische Auswertung von Korpusrecherchen. In: Fuß/Konopka/Wöllstein (Hg.).

Konopka, Marek/Fuß, Eric (2016): Genitiv im Korpus. Untersuchungen zur starken Flexion des Nomens im Deutschen. (= Studien zur Deutschen Sprache 70). Tübingen.

Konopka, Marek/Waßner, Ulrich H. (2013): Standarddeutsch messen? Frequenz und Varianz negativ-konditionaler Konnektoren. In: Korpus – Gramatika – Axiologie 8, S. 12–35.

Kupietz, Marc et al. (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta: European Language Resources Association (ELRA), S. 1848–1854. Internet: www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf (Stand: 3.8.2016).

Lijffijt, Jefrey/Gries, Stefan Th. (2012): Correction to „Dispersions and adjusted frequencies in corpora". In: International Journal of Corpus Linguistics 17, S. 147–149.

Meyer, David/Zeileis, Achim/Hornik, Kurt (2005): The strucplot framework: Visualizing multi-way contingency tables with vcd. In: Research Report Series 22 – Department of Statistics and Mathematics. Internet: http://epub.wu.ac.at/dyn/openURL?id=oai:epub.wu-wien.ac.at:epub-wu-01_8a1 (Stand: 3.8.2016).

Tatzreiter, Herbert (1988): Besonderheiten der Morphologie in der deutschen Sprache in Österreich. In: Wiesinger, Peter (Hg.): Das österreichische Deutsch. (= Schriften zur deutschen Sprache in Österreich 12). Wien u.a., S. 71–99.

Wolfer, Sascha (i.Dr.): Verstehen und Verständlichkeit juristisch-fachsprachlicher Texte. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 7). Tübingen.