[Ergänzung zu:]

Deutsche Sprachwissenschaft

—

Eine Einführung

Ingo Reich & Augustin Speyer

Stuttgart: Reclam, 2020

Saarbrücken – 5. Juli 2020

© Ingo Reich

Version 1.0

Inhaltsverzeichnis

Urheberrecht und Haftungsausschluss

Das folgende Kapitel steht über den Reclam-Verlag zum kostenlosen Download zur Verfügung. Das Urheberrecht und sonstige Rechte an dem vorliegenden Text verbleiben beim Verfasser. Diese Publikation enthält Links auf Webseiten Dritter, für deren Inhalte wir keine Haftung übernehmen, da wir uns diese nicht zu eigen machen, sondern lediglich auf deren Stand zum Zeitpunkt der Erstveröffentlichung verweisen. Sie enthält weiter Hinweise zum Urheberrecht und zum Datenschutz, für deren Richtigkeit und Vollständigkeit wir ebenfalls keine Gewähr und keine Haftung übernehmen. Diese Hinweise können und wollen die rechtliche Beratung im Einzelfall nicht ersetzen.

14

Methodisches: Projektorientiertes Arbeiten

»Wie schreibe ich eigentlich eine Hausarbeit?«

In den letzten 13 Kapiteln unseres Studienbuches Deutsche Sprachwissenschaft wurde Ihnen Wissen präsentiert, das die empirische und theoretische Basis für sprachwissenschaftliche Untersuchungen darstellt. Worüber wir aber bis jetzt noch kaum ein Wort verloren haben, ist, wie man sprachwissenschaftliche Untersuchungen eigentlich durchführt. Wenn Sie über das erste oder zweite Semester hinaus sind, dann wird für Sie aber gerade diese Frage in den Vordergrund rücken, spätestens dann, wenn Ihr*e Dozent*in möchte, dass Sie Ihr Pro- oder Hauptseminar mit einer Hausarbeit abschließen. Und noch akuter wird sich diese Frage stellen, wenn Sie in der Germanistischen Linguistik Ihre Abschlussarbeit schreiben wollen. In einer Einführung in die Deutsche Sprachwissenschaft kann man dieses Thema natürlich nicht erschöpfend abhandeln. Was wir aber leisten können, ist, Ihnen einen Leitfaden an die Hand zu geben, der Ihnen einen Einstieg in methodische Verfahren gibt und von dem aus Sie sich dann Ihren Weg durch die Literatur und weitere Quellen im Internet bahnen können.*

Vorgehensweise bei empirischen Arbeiten

Theoretische oder empirische Arbeit?

In der Linguistik kann man zwei Arten von Arbeiten unterscheiden, die sich in ihrer Herangehensweise grundlegend unterscheiden. Das sind zum einen Arbeiten, die ihren Fokus auf die theoretische Diskussion eines sprachlichen Phänomens auf der Basis einschlägiger Literatur legen. So könnte man zum Beispiel im

* An dieser Stelle geht ein ganz besonderer Dank an Lisa Schäfer und Robin Lemke für ihre unermüdliche und geduldige Unterstützung bei der Ausarbeitung dieses Kapitels.

Bereich der Orthographie den akzentbasierten Ansatz von Ramers (1999) zur Beschreibung der Doppelkonsonanz im Deutschen mit dem silbenbasierten Ansatz von Eisenberg (1997) vergleichen und diskutieren, welcher dieser Ansätze die bekannten Daten am besten und am elegantesten beschreibt. Diskussionen dieser Art können sehr aufschlussreich sein und eine wirklich gute theoretische Arbeit zu schreiben, ist keineswegs einfach. In diesem Kapitel werden wir uns aber bewusst auf den anderen Typ konzentrieren: empirische Arbeiten.

Korpusbasierte Untersuchung

Das Kennzeichen empirischer Arbeiten ist, dass man zunächst Daten zu einem sprachlichen Phänomen sammelt, um diese dann unter einer bestimmten Fragestellung und in einer bestimmten Art und Weise auszuwerten. Wenn Sie sich dafür interessieren, ob *weil*-Nebensätze mit V2-Stellung ein Phänomen der gesprochenen Sprache sind, dann werden Sie in **Korpora** gesprochener und geschriebener Sprache (also in bereits existierenden größeren Belegsammlungen) nach *weil*-Nebensätzen suchen und die Häufigkeiten der verschiedenen Verbstellungen (V2, VL) relativ zum jeweiligen Kanal (mündlich, schriftlich) auswerten. Glücklicherweise gibt es bereits ziemlich viele und frei zugängliche Korpora, in denen man systematisch suchen kann. Sollten Sie jedoch daran interessiert sein, welche Partizipien in einem Dialektraum wie zum Beispiel dem Rheinfränkischen verwendet werden, dann kann es gut sein, dass Sie Ihr eigenes Korpus zusammenstellen müssen, dass Sie also Dialektsprecher auf Tonband aufnehmen und das Aufgenommene für die Auswertung verschriftlichen (transkribieren) müssen. Daher ist es vor Beginn einer Untersuchung immer ratsam, sich darüber zu informieren, ob nicht vielleicht bereits Korpora existieren, die man direkt auswerten kann.

Experimentelle Untersuchung

Möchten Sie dagegen untersuchen, ob man das direkte Objekt eines Nebensatzes erfragen kann (wie in: *Wen hat er dir erlaubt, dass du __ mitbringst?*) und ob es dabei einen Unterschied macht, ob der Nebensatz ein *dass*-Satz oder ein satzwertiger Infinitiv ist (wie in: *Wen hat er dir erlaubt, __ mitzubringen?*), dann wird man diese Fragestellung aus mehreren Gründen nicht über die Auswertung eines Korpus beantworten können. Ein Problem wird sein, dass man solche Äußerungen nicht in ausreichender Zahl in einem Korpus finden wird, um eine verlässliche Aussage über ihren Status machen zu können. Wesentlich problematischer ist, dass man in Korpora immer nur Häufigkeiten von Vorkommen (Frequenzen) zählen kann. Die ursprüngliche Frage ist aber nicht, ob diese Konstruktionen überhaupt vorkommen oder wie oft, sondern ob sie »akzeptabel« (»grammatisch«) sind und ob eine der Varianten für Muttersprachler*innen vielleicht »besser« oder »akzeptabler« ist als die andere. Wenn wir darauf eine Antwort haben wollen, wird uns nichts Anderes übrigbleiben, als Muttersprachler*innen des Deutschen Sätze dieser Art vorzulegen und sie zu fragen, wie akzeptabel sie diese Sätze finden. Eine Möglichkeit, dies in systematischer Weise zu tun, ist, einen **Fragebogen** zu entwerfen und diesen

von Muttersprachler*innen ausfüllen zu lassen. Eine kontrollierte Datenerhebung dieser Art ist eine Form der experimentellen Untersuchung.

Schritt für Schritt zum Ziel

In diesem Kapitel werden wir uns diese zwei empirischen Verfahrensweisen etwas näher anschauen. Das Schöne dabei ist, dass die Vorgehensweise in beiden Fällen (und auch in komplexeren experimentellen Verfahren) im Wesentlichen dieselbe ist. Möchte man diese Schritt für Schritt beschreiben, dann könnte das wie folgt aussehen:

1. *Themenfindung*: Welches Phänomen möchte ich untersuchen?
2. *Literaturrecherche*: Was weiß man bereits über dieses Phänomen?
3. *Fragestellung*: Welcher konkreten Fragestellung möchte ich nachgehen?
4. *Hypothesenbildung*: Wie formuliere ich meine Arbeitshypothese?
5. *Methode*: Mit welchem Verfahren kann ich meine Hypothese testen?
6. *Durchführung*: Wie führe ich meine Untersuchung genau durch?
7. *Auswertung*: Wie kann ich meine (numerischen) Resultate auswerten?
8. *Interpretation*: Wie habe ich diese Resultate zu interpretieren?
9. *Präsentation*: Wie präsentiere ich meine Ergebnisse?

Im Folgenden werden wir diese Vorgehensweise an einem Beispiel so konkret wie möglich illustrieren. Bei den Schritten 1–5 und 8–9 werden wir zwischen den beiden Methoden (Korpus, Fragebogen) nicht wesentlich differenzieren müssen. In der konkreten Durchführung und der Auswertung der Ergebnisse unterscheiden sich die beiden Untersuchungsverfahren aber natürlich. Daher wird die Darstellung in den Schritten 6 und 7 gewissermaßen kurz ›verzweigen‹ müssen.

14.1 Themenfindung

Orientierung an der Seminarkonzeption

Nicht selten werden sich mögliche Themen bereits über die Konzeption des Seminars oder auch die Wahl eines Referats anbieten. Meist sind diese Themen aber so allgemein formuliert (zum Beispiel: das Silbengelenk im Deutschen, die Besetzung des Vorfelds im Deutschen etc.), dass man sich selbst in diesem Fall noch Gedanken machen muss, wie man die Fragestellung konkretisieren kann.

Augen und Ohren offenhalten!

Interessanter und motivierender ist es natürlich, sich mit einem Gegenstand auseinanderzusetzen, den man für sich selbst entdeckt und ›erobert‹ hat. Wie kommt man zu solchen Themen? Gerade in der Sprachwissenschaft lohnt es sich, im Alltag mit offenen Augen und Ohren durch die Welt zu gehen. In Gesprächen oder beim Lesen von Texten stolpert man immer wieder über sprachliche Auffälligkeiten, die die Basis für eine empirische Untersuchung darstellen könnten. Das können, müssen aber nicht immer dialektale oder soziolektale Besonderheiten oder die Verwendung von Emojis in elektronischer Kommunikation sein. Man

kann sich auch eine Situation vorstellen, in der man ungewollt in einem Café einen Tischnachbarn beim Telefonieren belauscht, der verärgert in sein Smartphone ruft: »*Dann brauchst du gar nicht mehr kommen*«. Aber halt, warte mal, müsste das nicht eigentlich heißen »*Dann brauchst du gar nicht mehr ZU kommen*«? War das jetzt einfach ein Fehler oder kann man das so sagen: »*Dann brauchst du gar nicht mehr kommen*«? Wenn man darüber nachdenkt, hört sich der Satz gar nicht so schlecht an. Aber kann es denn sein, dass es hier tatsächlich zwei mögliche Varianten gibt? Und wenn ja, warum und unter welchen Bedingungen?

14.2 Literaturrecherche

Wie finde ich Literatur?

Wenn mir dieses Phänomen aufgefallen ist, dann ist das vermutlich auch schon jemand anderem aufgefallen. Es wird sich also lohnen zu recherchieren, was zu diesem Phänomen bereits alles in der Literatur gesagt wurde. Hier sind wir beim nächsten Problem. Wie finde ich Literatur zu einer bestimmten Fragestellung? Das ist sicher in manchen Fällen einfacher und in anderen Fällen schwieriger. Zunächst muss man sich auf jeden Fall darüber klar werden, wie das Phänomen linguistisch einzuordnen ist. In unserem Fall geht es um die Verwendung von *brauchen* als Modalverb: Im obigen Beispiel ist zum einen (*nicht*) *brauchen* im Sinne von (*nicht*) *müssen* zu verstehen, zum anderen konstruiert *brauchen* hier mit einem Infinitiv (*brauchst nicht zu kommen*) und nicht mit einem nominalen Objekt wie das Vollverb *brauchen* in *ich brauche einen Flaschenöffner*.

Einschlägige Grammatiken

Wenn man nicht weiß, wo man mit seiner Recherche anfangen soll, dann ist es immer eine gute Idee, zunächst einmal die einschlägigen **Grammatiken** des Deutschen zu konsultieren. Die vermutlich bekannteste Grammatik ist die Duden-Grammatik. Über das Register gelangt man (in Auflage 9) zu §591, in dem zu modalem *brauchen* angemerkt wird: »[B]ei *brauchen* schwankt die Rektion zwischen dem *zu*-Infinitiv und dem reinen Infinitiv«. Damit ist zunächst die Intuition bestätigt, dass beide Varianten grundsätzlich möglich zu sein scheinen. In der 8. Auflage (ebenfalls in §591) heißt es weiter: »Noch ist die Verwendung mit dem reinen Infinitiv in geschriebenen Texten seltener als die Verbindung mit dem *zu*-Infinitiv.« Diese Ergänzung legt nahe, dass die Verbindung mit dem reinen Infinitiv eher ein Phänomen der gesprochenen Sprache ist (das sich aber auch in der geschriebenen Sprache immer stärker ausbreitet). Leider wird an dieser Stelle nicht auf eine entsprechende Publikation verwiesen oder die Aussage durch Zahlen untermauert. Darauf werden wir noch zurückkommen. Empfehlenswerte Grammatiken, die man konsultieren kann und sollte, finden sich in der folgenden Liste:

- **Akademie-Grammatik**: Heidolph, K.-E. et al. (Hgg.) (1981). *Grundzüge einer deutschen Grammatik*. Berlin: Akademie-Verlag.

- **Duden-Grammatik**: Dudenredaktion (Hg.) (2016). *Duden 4. Die Grammatik*. Berlin: Dudenverlag. 9., vollständig überarbeitete und aktualisierte Auflage.
- **Eisenberg-Grammatik**: Eisenberg, P. (2013). *Grundriss der deutschen Grammatik*. Band 1: *Das Wort*. Band 2: *Der Satz*. Stuttgart, Weimar: Metzler. 4., aktualisierte und überarbeitete Auflage.
- **IDS-Grammatik**: Zifonun, G. et al. (1997): *Grammatik der deutschen Sprache*. 3 Bände. Berlin, New York: de Gruyter.

Selbständige Literatur

Deskriptive Grammatiken können aber nur eine erste Anlaufstation sein. Wenn man Glück hat und in der Grammatik weiterführende Literatur genannt wird, dann kann man bereits hier den Faden aufnehmen. Wird aber keine Literatur genannt, dann wird man systematischer suchen müssen. Hier müssen wir jetzt zwei Formen der Publikation unterscheiden: unselbständige und selbständige Literatur. Als **selbständig** bezeichnet man Publikationen, die in Buchform veröffentlicht werden. Diese erhalten eine ISBN-Nummer und werden, falls verfügbar, in den Katalogen von Bibliotheken gelistet. Gäbe es also zum Beispiel eine Monographie mit dem Titel »Untersuchungen zum modalen Gebrauch von *brauchen*«, dann würde man diese Publikation online über den OPAC (den öffentlich zugänglichen Online-Katalog) der entsprechenden Bibliothek finden, indem man zum Beispiel nach dem Wort *brauchen* im Titel sucht. Nun gibt es meines Wissens noch keine solche selbständige Publikation. Alternativ könnte man im OPAC nach Monographien oder Sammelbänden (thematisch zusammenhängende Sammlungen von Artikeln verschiedener Autor*innen) zum übergeordneten Themenbereich der Modalverben suchen und diese auf entsprechende Bemerkungen durchsehen.

Unselbständige Literatur

Artikel, die als Teil eines Sammelbandes oder in einer Zeitschrift veröffentlicht werden, werden als **unselbständige** Literatur bezeichnet. Sie bekommen keine ISBN-Nummer (Online-Artikeln wird allerdings ein DOI, ein »Digital Object Identifier« zugewiesen, um sie im Internet dauerhaft identifizieren zu können) und sind auch im Allgemeinen nicht über Online-Kataloge wie den OPAC aufzufinden. Hier werden wir uns also anderer Recherche-Tools bedienen müssen. Das Mittel der Wahl sind hier einschlägige Bibliographien (also Verzeichnisse unselbständig publizierter Literatur), die im Idealfall online zugänglich sind. Drei für die Linguistik wichtige Bibliographien sind hier mit einer URL aufgeführt. Da die Nutzung von Bibliographien zu lizenzieren ist, werden Sie sich möglicherweise für eine Recherche mittels VPN (Virtual Private Network) mit Ihrem Universitätsnetz verbinden und sich den Weg zu der jeweiligen Bibliographie über das Datenbank-Infosystem DBIS suchen müssen:

- **Bibliographie Linguistischer Literatur (BLL)**
  www.blldb-online.de

- **Linguistic Bibliography (LB)**
  bibliographies.brillonline.com/browse/linguistic-bibliography
- **Modern Language Association (MLA)**
  www.mla.org/Publications/MLA-International-Bibliography

Lin|gu|is|tik-Portal

Eine einfachere Möglichkeit, im BLL und gleichzeitig in weiteren Bibliographien und Datenbanken zu recherchieren, bietet das Lin|gu|is|tik-Portal, das derzeit zum »Fachinformationsdienst FID Linguistik« ausgebaut und von der Universitätsbibliothek Frankfurt am Main verwaltet wird. Das Lin|gu|is|tik-Portal ist über die Webseite **www.linguistik.de** frei zugänglich und bietet eine Vielzahl weiterer relevanter Informationen.

Erste Ergebnisse

Gibt man zum Beispiel die Stichworte »Modalverb« und »brauchen« in der Suchmaske des Lin|gu|is|tik-Portals ein, dann wird man sehen, dass es bereits eine nicht unbeträchtliche Anzahl an unselbständigen Publikationen zur modalen Verwendung von *brauchen* gibt. Nicht jede dieser Publikationen wird sich allerdings ausführlich mit der Infinitiv-Problematik auseinandersetzen. Um einschätzen zu können, welche Artikel am vielversprechendsten sind, braucht es etwas Erfahrung. Deswegen kann es zu diesem Zeitpunkt durchaus sinnvoll sein, sich mit seiner*m Betreuer*in oder Dozent*in zusammenzusetzen und sich beraten zu lassen. Grundsätzlich ist es immer sinnvoll, mit neuerer Literatur zu beginnen, in der (berechtigten) Annahme, dass die Autoren bereits die existierende Literatur im Blick und aufgearbeitet haben (sollten). Sichtet man die »relevante« Literatur, dann wird man häufig auf folgende Aussagen treffen:

- Die Verbindung von *brauchen* mit reinem Infinitiv ist vorwiegend (wenn auch nicht ausschließlich) ein Phänomen der gesprochenen Sprache.
- Die Verbindung von *brauchen* mit reinem Infinitiv hat sich (in Analogie zum Kernbestand der Modalverben) neben brauchen mit *zu*-Infinitiv etabliert.
- In diesem Sinne gilt *brauchen* mit reinem Infinitiv als markiert.

Bei der Sichtung der Literatur wird man aber auch feststellen, dass diese Aussagen häufig nicht mit Zahlen untermauert werden, sondern vorwiegend auf einer **qualitativen** Diskussion ausgewählter Beispiele beruhen. Das eröffnet uns die Möglichkeit, diese Aussagen mit **quantitativen** Untersuchungen zu überprüfen.

14.3 Fragestellung

Konkretisierung der Fragestellung

In der Auseinandersetzung mit der Forschungsliteratur zu dem fraglichen Phänomen wird man also mehr oder weniger direkt auf konkretere Fragestellungen

geführt werden. Man könnte jetzt zum Beispiel der Frage nachgehen, ob bzw. inwiefern die Verbindung von *brauchen* mit reinem Infinitiv tatsächlich primär ein Phänomen der gesprochenen Sprache ist. Oder man könnte untersuchen, ob die Verbindung von *brauchen* mit reinem Infinitiv tatsächlich die »markierte« Variante darstellt. Oder man könnte der Frage nachgehen, ob normative Vorstellungen, wie sie sich in dem Sprüchlein »Wer *brauchen* nicht mit *zu* gebraucht, braucht *brauchen* gar nicht zu gebrauchen« zeigen, einen Einfluss darauf haben, wie wir entsprechende Äußerungen bewerten. Oder, oder, oder, oder … Die Erarbeitung einer Fragestellung ist also gewissermaßen der Übergang von einem eher allgemeinen »Ich interessiere mich für die Selektionseigenschaften von modalem *brauchen*« zu einer Konkretisierung wie »Ist die Verbindung von *brauchen* mit dem reinen Infinitiv primär ein Phänomen der gesprochenen Sprache?«.

14.4 Hypothesenbildung

Wie sieht eine geeignete Forschungsfrage aus?

Auf welcher Granularitätsebene eine Forschungsfrage anzusiedeln ist (wie präzise sie zu formulieren ist), um als eine vernünftige Fragestellung durchzugehen, kann man so nicht pauschal sagen. Tatsächlich hätten wir auch schon unserer Ausgangsfragestellung »Kann modales *brauchen* sowohl mit dem *zu*- als auch mit dem reinen Infinitiv konstruieren?« nachgehen können, wenn sie nicht schon im Wesentlichen durch die Literatur beantwortet worden wäre. Ob eine Forschungsfrage im Grundsatz eine geeignete Fragestellung ist, hängt nicht zuletzt damit zusammen, ob man sie in eine testbare Hypothese umformulieren kann.

Formulierung einer Hypothese

Eine Hypothese ist zunächst eine Aussage, deren Wahrheit oder Falschheit noch nicht bekannt ist. So können wir die Forschungsfrage »Ist modales *brauchen* plus reiner Infinitiv primär ein Phänomen der gesprochenen Sprache?« wie folgt in eine Hypothese umformulieren: »Modales *brauchen* plus reiner Infinitiv ist primär ein Phänomen der gesprochenen Sprache«. Problematisch an dieser Formulierung ist jedoch, dass sie nicht, oder zumindest nicht ohne Weiteres, empirisch zu testen ist, also über ein empirisches Verfahren bestätigt (verifiziert) oder widerlegt (falsifiziert) werden kann. Denn wann wollen wir überhaupt davon sprechen, dass die Verbindung mit dem reinen Infinitiv primär ein Phänomen der gesprochenen Sprache ist? Wenn wir seine Verwendung in gesprochen-sprachlichen Äußerungen für angemessen(er) halten? Oder wenn die Verbindung mit dem reinen Infinitiv in der gesprochenen Sprache häufiger vorkommt als die Verbindung mit dem *zu*-Infinitiv? Oder reicht es bereits aus, dass sie in der gesprochenen Sprache häufiger vorkommt als in der geschriebenen Sprache?

Was messen wir? Und wie?

Wenn wir eine Hypothese empirisch testen wollen, dann müssen wir etwas messen. Und aus der Hypothese muss klar hervorgehen, was wir messen wollen

und wie wir es genau messen wollen. Außerdem muss aus der Hypothese hervorgehen, wovon die Messergebnisse (unserer Meinung nach) abhängen. Machen wir dazu ein Beispiel. Eine erste Annäherung an eine gute Hypothese (H1) wäre folgende Formulierung:

(H1) Die Verbindung von modalem *brauchen* mit reinem Infinitiv tritt in der gesprochenen Sprache häufiger auf als in der geschriebenen Sprache.

In dieser Formulierung wird deutlich, was wir messen wollen: Häufigkeiten. Wie oft kommt eine bestimmte Konstruktion vor? Daraus können wir auch folgern, wie wir messen werden: Wir werden die einzelnen Vorkommen einfach abzählen. Und es wird auch klar, wovon die Messung unserer Meinung nach abhängen wird: Davon, ob wir gesprochene Sprache betrachten oder geschriebene Sprache.

Abhängige und unabhängige Variablen

Da die Häufigkeit der Konstruktion (i) variieren kann und zwar (ii) in Abhängigkeit davon, ob wir gesprochene oder geschriebene Äußerungen betrachten, spricht man hier bei der Häufigkeit von einer **abhängigen Variablen**. Diese Variable kann im Prinzip jeden beliebigen Wert innerhalb der natürlichen Zahlen (0, 1, 2, 3 ...) annehmen. Auch die Art der jeweiligen Äußerung, ob gesprochen oder geschrieben, kann verschiedene Werte annehmen, eben gesprochen oder geschrieben. In diesem Sinne ist auch die Art der Äußerung eine Variable. Und da sie (im Gegensatz zur Häufigkeit) hier von keinem anderen Faktor abhängt, nennt man sie eine **unabhängige Variable**. Tatsächlich bringt letztlich jede testbare Hypothese (mindestens) eine unabhängige und (mindestens) eine abhängige Variable wie folgt in einen konditionalen Zusammenhang: »Wenn man den Wert der unabhängigen Variable verändert, dann verändert sich auch (in bestimmter Art und Weise) der Wert der abhängigen Variable.« Wenn ich also statt geschriebenen Äußerungen gesprochene Äußerungen betrachte, dann wird sich die Häufigkeit der Vorkommen von Verbindungen mit dem reinen Infinitiv (z.B.) erhöhen.

Zur Vertiefung

Gerichtete und ungerichtete Hypothesen

In der Formulierung des letzten Satzes und auch bereits in der Formulierung von (H1) wird explizit gesagt, in welcher Weise sich die Häufigkeiten verändern werden. Wenn wir uns in der gesprochenen Sprache bewegen, dann erwarten wir *mehr* Vorkommen von Verbindungen mit dem reinen Infinitiv. Eine solche Hypothese nennt man auch eine **gerichtete Hypothese**. Würde man im letzten Satz des letzten Abschnitts »erhöhen« durch »verändern« ersetzen, dann ließe die Hypothese offen, in welche Richtung sich die Häufigkeiten verändern, ob sie also mehr oder weniger werden. In diesem Fall würde man von einer **ungerichteten Hypothese** sprechen. – Warum ist das wichtig? Wenn

man die erhobenen Daten mit statistischen Verfahren auf Signifikanz (dazu später mehr) auswerten will, dann kann man zeigen, dass bei einer gerichteten Hypothese wesentlich schneller ein gegebenes Signifikanzniveau erreicht wird als bei einer ungerichteten. (Das werden wir später noch etwas näher erläutern.) Darüber hinaus wird man immer versuchen, möglichst viele theoretische Vorannahmen in die Hypothese eingehen zu lassen. Und dies alleine wird in der Regel schon zu einer gerichteten Hypothese führen.

Stichprobe und Grundgesamtheit

Die Formulierung der Hypothese in (H1) geht also auf jeden Fall in die richtige Richtung. Sie lässt aber immer noch einige wichtige Punkte offen, über die man sich sicherlich noch Gedanken machen müsste. So wird zum Beispiel nicht klar, was die Basis der Untersuchung ist. Wir werden ja schlecht alle jemals gesprochenen und geschriebenen Äußerungen auf Vorkommen von modalem *brauchen* hin untersuchen können. Wir werden uns also auf **Stichproben**, in diesem Fall auf bestimmte Korpora (Belegsammlungen) beschränken müssen. Die Hypothese ist daher zunächst auf die untersuchten Korpora zu relativieren. Die Hoffnung wäre natürlich, dass diese Korpora gewissermaßen repräsentativ für gesprochene bzw. geschriebene Sprache sind, dass wir also von diesen Korpora auf gesprochene bzw. geschriebene Sprache an sich (also auf die **Grundgesamtheiten**) verallgemeinern können. Wir sollten uns daher gut überlegen, wie die Korpora genau aussehen sollten, die wir dann später untersuchen. (So ist z.B. SMS-Kommunikation zwar schriftlich, aber eben auch mündlichkeitsnah.)

Absolute oder relative Zahlen?

Was in der Hypothese (H1) ebenfalls nicht deutlich wird, ist, ob hier absolute oder relative Zahlen gemeint sind. Machen wir ein Beispiel. Nehmen wir an, wir haben ein geeignetes Korpus gesprochener Sprache KS (mit S für »spoken«) und ein geeignetes Korpus der geschriebenen Sprache KW (mit W für »written«). Nehmen wir außerdem der Einfachheit halber an, dass beide Korpora gleich groß sind (also dieselbe Anzahl an Äußerungen oder an Wörtern enthalten). Nehmen wir schließlich an, dass wir in KS 60 Vorkommen von *brauchen* plus einfachem Infinitiv und 120 Vorkommen von *brauchen* plus *zu*-Infinitiv finden. KW enthält 30 Vorkommen von *brauchen* plus einfachem Infinitiv und 60 Vorkommen von *brauchen* plus *zu*-Infinitiv, vgl. Tabelle 14.1.

	KS	KW
brauchen plus einfacher Infinitiv	60	30
brauchen plus *zu*-Infinitiv	120	60

Tabelle 14.1: Hypothetische Ergebnisse einer Korpusrecherche

In absoluten Zahlen bestätigt das Ergebnis offenbar unsere Hypothese: Wir haben in KS doppelt so viele Vorkommen von *brauchen* plus einfachem Infinitiv gefunden wie in KW. Wenn wir diese Zahlen aber ins Verhältnis setzen zu den Vorkommen von *brauchen* plus *zu*-Infinitiv, dann stellen wir fest, dass das Verhältnis in beiden Fällen dasselbe ist, nämlich 1:2. Relativiert auf die jeweiligen Vorkommen von *brauchen* plus *zu*-Infinitiv, ist also letztlich keinerlei Veränderung festzustellen und die Hypothese ist nicht bestätigt.

Art der Konstruktion

Was mit der Hypothese (H1) natürlich gemeint ist, sind relative Häufigkeiten. Wir sind ja letztlich daran interessiert, ob die Konstruktion *brauchen* plus einfacher Infinitiv – in Relation zur Konstruktion *brauchen* plus *zu*-Infinitiv – für die gesprochene Sprache charakteristischer ist als für die geschriebene Sprache. Da die Häufigkeiten (wie in Tabelle 14.1 dargestellt) sowohl von der Art der Äußerung (gesprochen, geschrieben) als auch von der Art der Konstruktion (einfacher Infinitiv, *zu*-Infinitiv) abhängen, liegen hier genau genommen zwei unabhängige Variablen, auch **Faktoren** genannt, vor. Jeder dieser Faktoren hat dabei genau zwei **Ausprägungen** (also mögliche Werte).

Wann ist »mehr« tatsächlich mehr?

Ein letztes Problem, das die Formulierung von (H1) offenlässt, ist, was wir unter »häufiger« genau verstehen wollen. Machen wir auch hier wieder ein Beispiel. Ändern wir die Anzahl der Vorkommen von *brauchen* plus einfachem Infinitiv in KS von 60 auf 80 ab. Betrachten wir relative Verhältnisse, dann konstruiert im Korpus KW *brauchen* in rund 33% (= 30 von insgesamt 90) der relevanten Fälle mit dem einfachen Infinitiv, im Korpus KS dagegen in 40% der relevanten Fälle (= 80 von insgesamt 200). Die relative Häufigkeit liegt also in KS um 7% höher. Wir können damit für unsere Korpora klar bestätigen, dass wir in KS sowohl in absoluten wie auch in relativen Zahlen mehr Vorkommen von *brauchen* plus einfachem Infinitiv gefunden haben als im geschriebenen Korpus KW.

Signifikanz und Signifikanzniveau

Aber kann das nicht vielleicht auch Zufall gewesen sein? Denn wenn ich eine (perfekte) Münze werfe, dann stehen die Chancen auf Kopf auch 50:50. Und trotzdem kann es mir passieren, dass meine Münze bei 10 Würfen 7 Mal Kopf zeigt. Die Antwort ist: Ja, es kann Zufall sein. Damit sind wir bei der nächsten Frage: Wie groß muss die relative Veränderung denn sein, damit ich davon ausgehen kann, dass es kein Zufall ist? 10%, 20% oder mehr? Die Antwort auf diese zweite Frage hat einen pessimistischen und einen optimistischen Teil. Der eher pessimistische Teil besagt, dass man Zufall nie völlig ausschließen kann. Der andere, optimistischere Teil aber besagt, dass man (unter bestimmten Bedingungen) einschätzen kann, wie wahrscheinlich es ist, dass die Veränderung alleine auf Zufall beruht. Liegt diese Wahrscheinlichkeit unter 5% (dem Signifikanzniveau), dann geht man (wenn man nicht gerade in der Medizin arbeitet, wo es möglicherweise um Leben und Tod geht) in der Regel davon aus, dass die Daten die obige Hypothese bestäti-

gen, dass die relative Veränderung also signifikant ist. Wenn wir Pech haben, kann die Veränderung immer noch Zufall sein, aber die Wahrscheinlichkeit, dass dem so ist, ist relativ gering (eben unter 5%). Und wir legen einfach fest, dass wir damit (zumindest in der Linguistik) gut leben können.

Statistische Verfahren

Wie man einschätzt, ob in einem gegebenen Fall eine Veränderung mit großer Wahrscheinlichkeit (über 95%) nicht auf Zufall (sondern auf die in der Hypothese formulierte Vermutung) zurückzuführen ist, ist Gegenstand der Statistik. Die Statistik ist ein Teilgebiet der Mathematik und hat viel mit Wahrscheinlichkeitsrechnung zu tun. Es ist natürlich nicht möglich, in diesem knapp gefassten Kapitel in statistische Verfahren einzuführen. Trotzdem werden wir in den Vertiefungsboxen auf einschlägige Testverfahren und weiterführende Literatur hinweisen. Dazu werden wir an der einen oder anderen Stelle vereinfachen müssen (und bitten darum diejenigen, die sich schon etwas auskennen, um Nachsicht).

Zur Vertiefung

Nullhypothese und Alternativhypothese

Tatsächlich haben wir in diesem Abschnitt bereits mindestens in einer Hinsicht stark vereinfacht. Für statistische Auswertungen unterscheidet man zwei Hypothesen, die **Nullhypothese** (NH) und die **Alternativhypothese** (AH). Die AH ist die Hypothese, die uns eigentlich inhaltlich interessiert, in unserem Beispiel (H1). Die NH (H0) ist die Annahme, dass die gemessene Veränderung nur auf Zufall beruht. Die statistische Auswertung testet jetzt genau genommen, wie wahrscheinlich die NH ist, also dass die Veränderung auf Zufall beruht. Liegt diese Wahrscheinlichkeit (die als ***p*-Wert** in der Statistik angegeben wird) unter dem festgelegten Signifikanzniveau, also unter den 5% ($p < 0,05$), dann geht man davon aus, dass die NH nicht zutrifft. Da man die NH als die Negation der AH auffassen kann, folgt, dass (mit einer Wahrscheinlichkeit von $1 - p$) die Negation der NH, also die AH, zutrifft. Die Wahrscheinlichkeit p, dass wir die Alternativhypothese AH fälschlicherweise angenommen haben, bezeichnet man als ***α*-Fehler** oder auch als **Fehler 1. Art**.

Zur Vertiefung

Kopf oder Zahl?

Kommen wir in diesem Zusammenhang nochmal kurz auf obige Aussage zurück, dass es einem ja auch passieren kann, dass bei 10-maligem Werfen einer Münze die Münze 7-mal Kopf und 3-mal Zahl zeigt. Aber ist das tatsächlich noch Zufall? Oder würden wir hier schon annehmen wollen, dass die Münze gezinkt ist? Das kann man vergleichsweise einfach ausrechnen. Die Frage, die man beantworten muss, lautet: Wie wahrscheinlich ist es, dass bei 10 Würfen mindestens 7-mal Kopf kommt? (An dieser Stelle ist das »Mindestens« sehr

wichtig, denn bei noch extremeren Ergebnissen werden wir ja erst recht von einer gezinkten Münze sprechen wollen.)

Dazu müssen wir zunächst berechnen, wie viele mögliche Kombinationen es bei 10 Würfen überhaupt gibt. Da jeder Wurf nur 2 mögliche Ausgänge (Kopf oder Zahl) hat, sind das $2^{10} = 1024$. Dann müssen wir überlegen, wie viele Kombinationen es für 7-mal Kopf und 3-mal Zahl gibt. Hier können wir zum Glück auf eine Formel aus der Kombinatorik zurückgreifen, den Binomialkoeffizienten »n über k«. Für $n = 10$ Würfe und $k = 7$-mal Kopf liefert dieser uns 120 Kombinationsmöglichkeiten, für $k = 8$ sind es 45, für $k = 9$ nur noch 10 und für $k = 10$ offenbar nur eine. In Abbildung 14.2 wird für jedes k von 0 bis 10 die Anzahl der möglichen Kombinationen in einem Balkendiagramm dargestellt. Dabei sieht man schön, dass die Anzahl der möglichen Kombinationen für 7-mal Kopf dieselbe ist wie die für 3-mal Kopf (was eigentlich klar sein sollte, da 3-mal Kopf ja 7-mal Zahl bedeutet.) Die Verteilung ist also symmetrisch. Und für immer größer werdende n (100, 1.000, 10.000, 100.000, 1.000.000 etc. Würfe) nähert sich diese Verteilung einer Kurve an, die die meisten noch aus der Schule kennen dürften, der **Gauß'schen Normalverteilung**.

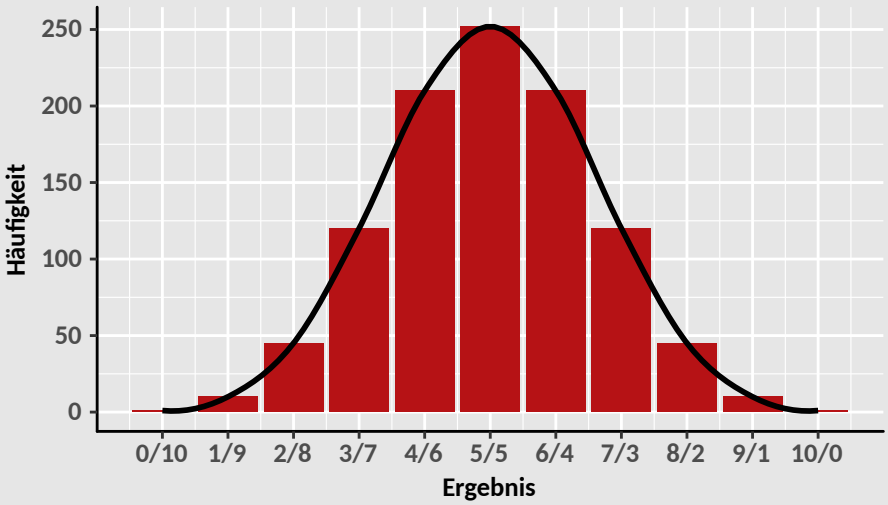


Abbildung 14.1: Anzahl der Kombinationen für k Mal Kopf ($0 \leq k \leq 10$)

Wie wahrscheinlich ist es jetzt also, dass bei 10 Würfen mindestens 7-mal Kopf kommt? Dazu müssen wir die Wahrscheinlichkeiten aufaddieren, dass genau 7-mal, genau 8-mal, genau 9-mal und genau 10-mal Kopf kommt. Die Wahrscheinlichkeit, dass genau 7-mal Kopf kommt, entspricht dem prozentualen Anteil der 120 Kombinationsmöglichkeiten an der Gesamtmenge von 1024 Möglichkeiten. Das sind gerundet $(120x100)/1024 = 11,72\%$. Für genau 8-

mal Kopf erhalten wir entsprechend 4, 39%. Für genau 9-mal Kopf 0, 98%. Und für genau 10-mal Kopf 0, 1%. Addieren wir das zusammen, dann erhalten wir insgesamt 17, 19%. Dass wir bei 10 Würfen 7-mal Kopf erhalten, ist also auch bei einer fairen Münze so wahrscheinlich, dass wir es als ein zufälliges Ereignis akzeptieren (die 17, 19% liegen weit über den angenommenen 5%).

Wenn wir bei 7-mal Kopf (nach unseren Annahmen) immer noch von Zufall ausgehen müssen, wo liegt dann die Grenze, bei der wir das nicht mehr tun? Bei 8-, bei 9- oder gar bei 10-mal Kopf? Da wir die extremsten Werte immer miteinrechnen müssen, können wir die Frage beantworten, indem wir zuerst die Wahrscheinlichkeit des extremsten Werts, also 10-mal Kopf, nehmen. Dieser liegt mit 0, 1% unter dem Signifikanzniveau von 5%. Dieses Ereignis ist also so unwahrscheinlich, dass wir hier nicht mehr von Zufall sprechen wollen. Wenn wir das zweitextremste Ereignis, also genau 9-mal Kopf, dazu zählen, kommen wir auf 1, 08%, was immer noch unter 5% liegt. Auch das kann also noch nicht Zufall sein. Erst wenn wir auch noch das drittextremste Ereignis, 8-mal Kopf, dazuzählen, kommen wir mit 5, 47% über das Signifikanzniveau. Bei 8-mal Kopf werden wir also (nach unseren Annahmen) noch ganz knapp von Zufall ausgehen müssen.

Was wir gerade gemacht haben, ist, dass wir uns dem Signifikanzniveau von 5% von unten angenähert haben, indem wir zuerst die Wahrscheinlichkeit des extremsten, dann die Wahrscheinlichkeit des zweitextremsten Ereignisses usw. berechnet und dann zusammengezählt haben. Wir haben uns also, wenn wir das auf die Abbildung 14.2 übertragen, gewissermaßen von rechts reingefressen, bis wir über die Grenze von 5% gekommen sind. Tatsächlich ist das genau das, was man auch im allgemeinen Fall macht: Verteilungen wie die Binomialverteilung oder auch die Gauß'sche Normalverteilung sind **Dichtefunktionen**, d.h. der Flächeninhalt unter der Kurve ist genau 1 (was 100% entspricht). Berechnet man die Fläche, die von ganz rechts (also von $+\infty$) 5% abschneidet, dann erhält man genau das Ereignis auf der x-Achse, für das man gerade noch annehmen würde, dass es Zufall ist. Liegt ein Wert weiter rechts, dann schneidet er weniger als 5% der Fläche unter der Kurve ab, beruht also (nach Annahme) nicht mehr auf Zufall.

Eine letzte Anmerkung: Tatsächlich gilt das nur für gerichtete Hypothesen, also wenn wir beispielsweise die Vermutung haben, dass die Münze eine starke Tendenz zu Kopf hat. Denn dann müssen wir nur die Ereignisse berücksichtigen, bei denen mehrheitlich Kopf auftritt. Wenn wir aber nur vermuten, dass die Münze gezinkt ist, dass sie also eine Präferenz für eine der beiden Seiten hat, wir aber nicht wissen, für welche, dann müssen wir auch die anderen ex-

tremen Ereignisse berücksichtigen, also dass 10-mal Zahl auftritt, dass 9-mal Zahl auftritt usw. Mit anderen Worten, wir müssen uns auch von ganz links (also von $-\infty$) reinfressen. Da die Verteilung aber symmetrisch ist, heißt das, dass wir uns auf jeder Seite maximal 2,5% reinfressen dürfen, da die Summe dann bereits 5% ergibt. Das heißt aber wiederum, dass sich die Grenze in der Regel weiter nach außen verschiebt. In unserem Beispiel macht es zwar keinen Unterschied, da die Wahrscheinlichkeit für mindestens 9-mal Kopf oder Zahl mit $2 \times 1.08\% = 2.16\%$ immer noch unter 5% liegt. Betrachtet man aber 100 Würfe, dann sieht man (mehr oder weniger schnell), dass sich die Grenze verschiebt: Bei einer gerichteten Hypothese kann man (z.B.) ab 59-mal Zahl davon ausgehen, dass die Münze nicht in Ordnung ist; bei einer ungerichteten Hypothese dagegen erst ab 61-mal Zahl (siehe z.B. Gries 2008: 47f).

14.5 Methode der Datenerhebung

Korpusstudien

Wie wir bereits angedeutet haben, scheint für die Untersuchung der Hypothese (H1) eine Korpusstudie ein geeignetes empirisches Verfahren zu sein: Wir untersuchen ein Korpus zur gesprochenen Sprache und eines zur geschriebenen Sprache, werten beide Korpora nach Vorkommen von *brauchen* plus einfachem Infinitiv und *brauchen* plus *zu*-Infinitiv aus und schauen, ob wir in unserem Korpus zur gesprochenen Sprache tatsächlich (signifikant) mehr Vorkommen von *brauchen* plus einfachem Infinitiv finden (relativ zu *brauchen* plus *zu*-Infinitiv) als in unserem Korpus zur geschriebenen Sprache.

Quantitative vs. qualitative Aussagen

Die Untersuchung von Korpora wird methodisch vor allem immer dann sinnvoll sein, wenn wir an Häufigkeiten (einem Mehr oder Weniger) interessiert sind, wenn wir also **quantitative Aussagen** machen wollen. Nicht selten sind wir aber auch an **qualitativen Aussagen** interessiert, wie die in (H2):

(H2) Muttersprachler*innen finden modales *brauchen* plus *zu*-Infinitiv (generell) natürlicher / akzeptabler als *brauchen* plus einfachen Infinitiv.

Akzeptabilitätsurteile

Um (H2) zu überprüfen, könnten wir uns (in erster Annäherung) folgendes Szenario vorstellen: Wir präsentieren Muttersprachler*innen des Deutschen Minimalpaare wie *jetzt brauchst du auch nicht mehr kommen* und *jetzt brauchst du auch nicht mehr zu kommen* und fragen sie, welchen der Sätze sie natürlicher finden. Mit anderen Worten: Wir lassen sie die Natürlichkeit von Sätzen beurteilen. Oder nochmal anders formuliert: Wir erheben in systematischer Weise **Akzeptabilitätsurteile**.

Grammatikalität und Akzeptabilität

Zur Vertiefung

Warum fragt man in Fragebögen nach der Natürlichkeit oder der Akzeptabilität von Sätzen und nicht direkt nach deren Grammatikalität? Um das zu verstehen, muss man sich erst einmal klar machen, was Linguisten unter dem Begriff »grammatisch« verstehen. Ein Satz ist grammatisch, wenn er gemäß den Regeln der Grammatik der fraglichen Sprache gebildet ist. Und mit Grammatik ist hier nicht eine Grammatik gemeint, wie sie z.B. im Duden formuliert wird, sondern unsere mentale Grammatik, unsere Sprachkompetenz, wie wir sie als Kinder erworben haben. Diese Grammatik ist uns aber nicht unmittelbar zugänglich (sonst würden wir Linguisten nicht Modelle entwickeln, mit denen wir uns dieser Grammatik annähern), folglich können wir auch keine sicheren Urteile darüber fällen, ob ein Satz grammatisch ist oder nicht. Was wir aber beurteilen können, ist, ob ein Satz für uns »gut« oder »natürlich« klingt, ob er für uns »akzeptabel« ist. Ein methodisches Problem ist dabei, dass man von der Natürlichkeit eines Satzes nicht ohne Weiteres auf seine Grammatikalität schließen kann. So kann man zeigen, dass es klar ungrammatische Sätze gibt, die dennoch von Versuchspersonen als natürlich eingestuft werden, und grammatische Sätze, die (z.B. aufgrund ihrer Komplexität) als nicht akzeptabel bewertet werden. Eine eingehende Diskussion dieses komplexen Zusammenhangs findet sich z.B. in Schütze (2016).

Fragebogenstudien

Die Erhebung solcher Urteile kann vergleichsweise einfach über **Fragebögen** erfolgen, auch wenn wir später noch sehen werden, dass das gerade skizzierte Vorgehen nicht ganz unproblematisch ist und bei der Erstellung von Fragebögen eine ganze Reihe von »Gütekriterien« zu berücksichtigen sind. Anders als bei Korpusstudien erheben wir mit Fragebögen genau die Daten, an denen wir interessiert sind. Die Erhebung erfolgt dabei in höchstem Maße kontrolliert und in diesem Sinne sind Fragebogenstudien auch als ein **experimentelles Verfahren** zu bezeichnen. Ob das ein Vorteil ist, hängt von diversen Faktoren ab, eben davon, ob wir eine qualitative Aussage machen wollen, aber auch davon, ob das interessierende Phänomen nur selten in Korpora auftritt. Umgekehrt liefern uns Korpora nicht konstruierte, natürliche Daten und führen uns immer wieder vor Augen, wieviel Variation man in natürlichen Daten beobachten kann.

Weitere Datentypen und Methoden

Natürlich sind Häufigkeiten und Akzeptabilitätsurteile nicht die einzigen Datentypen, an denen wir als Sprachwissenschaftler*innen interessiert sind. Wenn man zum Beispiel untersuchen möchte, ob längere Texte tatsächlich schwieriger zu lesen sind, wenn sie in serifenlosen Schriften gesetzt sind, dann wird man die Zeit messen wollen, in der Texte (einmal mit Serifen und einmal ohne Serifen) von Versuchspersonen gelesen werden. In diesem Fall spricht man von **Lesezeit-**

experimenten. Ist man dagegen daran interessiert, wie pronominale Anaphern interpretiert werden, dann kann es sinnvoll sein, mit einem **Eye-Tracker** Blickbewegungen aufzuzeichnen. Und sollte man etwas mehr über die neuronale Verarbeitung von sprachlichen Stimuli (in der Zeit) erfahren wollen, dann wird man entweder mit einem EEG in einer **EKP-Studie** elektrische Potenziale an der Kopfoberfläche messen oder auf ein bildgebendes Verfahren zurückgreifen. Da letztere Methoden einen nicht unbeträchtlichen technischen Aufwand erfordern, werden wir uns hier auf die Darstellung von Korpus- und Fragebogenstudien beschränken. Man sollte aber immer im Blick haben, dass andere experimentelle Verfahren möglicherweise für die eigene Fragestellung besser geeignet sind.

14.6 Korpusstudien

14.6.1 Ressourcen und Tools

Digitale Korpora

Nehmen wir also an, wir wollen unsere Hypothese (H1) anhand von Korpusdaten überprüfen. Wir haben bereits gesehen, dass wir dafür ein Korpus zur gesprochenen Sprache und eines zur geschriebenen Sprache (in ausreichender Größe) benötigen. Jetzt können wir uns natürlich solche Korpora (aufwendig) selbst zusammenstellen. Da es aber nicht unwahrscheinlich ist, dass auch andere schon mit dem gleichen Problem konfrontiert waren, ist es durchaus sinnvoll, zunächst zu schauen, ob es nicht bereits geeignete Korpora in elektronischer und damit leicht zugänglicher Form gibt. Ein weiterer Vorteil bereits verfügbarer Korpora ist, dass sich beim Aufbau von Korpora durchaus komplexe urheberrechtliche (UrhG) und datenschutzrechtliche (DSGVO) Fragen ergeben können, die man vielleicht gerne vermeiden möchte und auf die wir hier auch nicht im Einzelnen eingehen können. (Rechtliche Hinweise finden sich in den meisten Einführungen in die Korpuslinguistik. Die Deutsche Forschungsgemeinschaft DFG stellt darüber hinaus auf ihrer Webseite www.dfg.de Hinweise zum Umgang mit Forschungsdaten zur Verfügung, insbesondere zu rechtlichen Aspekten bei Sprachkorpora.)

Eine Suche im Netz nach »Korpus gesprochene, geschriebene Sprache« liefert einige erste Ergebnisse. Eine etwas systematischere Suche erlaubt dagegen die Webseite der CLARIN-D-Initiative, deren Ziel eine nachhaltige Verfügbarkeit von wissenschaftlichen Ressourcen in den Geistes-, Kultur- und Sozialwissenschaften ist. Der unten angegebene Link führt zu einer kommentierten Liste von Archiven und CLARIN-D-Zentren. Neben diesen über CLARIN-D zugänglichen Korpora gibt es eine nicht unbeträchtliche Anzahl weiterer Korpora unterschiedlichster Natur, die hier leider nicht im Einzelnen aufgeführt werden können. Hinweisen wollen wir aber zumindest auf einige Korpora, die verschiedene Aspekte unseres Fachs abdecken. Da sind zum einen die CHILDES-Korpora (*Child Language Data Ex-*

change System) zum Erstspracherwerb und die DDD-Korpora (*Deutsch – Diachron – Digital*) zu früheren Sprachstufen des Deutschen zu nennen, zum anderen aber auch das FALKO-Korpus (*fehlerannotiertes Lernerkorpus*) zu Deutsch als Fremdsprache der Humboldt-Universität zu Berlin. Das PCC (*Potsdam Commentary Corpus*) erlaubt die Untersuchung eher pragmatischer Fragestellungen und die DGD (*Datenbank für Gesprochenes Deutsch*) hat sich auf gesprochenes Deutsch spezialisiert.

- **CLARIN-D (Links zu Archiven und CLARIN-D-Zentren)**
  www.clarin-d.net/de/sprachressourcen-und-dienste/korpora/
- **CHILDES (Korpora zum Erstspracherwerb)**
  childes.talkbank.org/access/German/
- **DDD (Deutsch – Diachron – Digital) Referenzkorpus Altdeutsch**
  www.deutschdiachrondigital.de (mit Links zu anderen Projekten)
- **FALKO (Korpus zu Deutsch als Fremdsprache)**
  www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/
- **PCC (Potsdam Commentary Corpus)**
  angcl.ling.uni-potsdam.de/resources/pcc.html
- **DGD (Datenbank für Gesprochenes Deutsch)**
  dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome

Lemmatisierung und Annotation

Angenommen, wir haben jetzt ein oder mehrere potentiell geeignete Korpora gefunden, die wir gerne auswerten wollen. Wie gehen wir dann konkret vor? Wenn das Korpus lediglich als reine Text-Datei (oder in ASCII-Code) vorliegt, dann werden wir nach allen Wortformen von *brauchen* suchen müssen, also unter anderem nach *brauche*, *bräuchte*, *brauchst* und *gebraucht*. Einfacher wäre es natürlich, wenn wir einfach nach dem Lexem *brauchen* in (s)einer Grundform suchen könnten und dabei automatisch alle zu dem Lexem gehörenden Vorkommen von Wortformen finden würden. Dazu müssen aber diejenigen, die das Korpus zur Verfügung gestellt haben, dieses Korpus bereits **lemmatisiert**, also jede einzelne Wortform von *brauchen* einer Grundform (dem Lemma) zugeordnet haben. Mit anderen Worten: Das Korpus muss mit Informationen angereichert, also **annotiert** worden sein.

Suchanfragen

Haben wir Glück und das Korpus ist bereits lemmatisiert, dann wird es in der Regel auch eine Art Suchmaske oder eine **Abfragesoftware** (Query-Tool) geben, über die wir nach den verschiedenen Lemmata in systematischer Weise suchen können. So kann man etwa in **Cosmas II** mit der Anfrage &brauchen in den Korpora des IDS (Institut für Deutsche Sprache) online nach allen Wortformen von *brauchen* suchen. Das Clarin-D-Zentrum in Saarbrücken bietet (wie einige andere) mit **cqpweb** ebenfalls Online-Abfragen an, die Anfrage müsste hier aber [lemma = "brauchen"]; lauten. Wie die Anfrage konkret zu formulieren ist (ihre Syntax),

hängt also im Allgemeinen von der jeweils zur Verfügung gestellten Abfragesoftware ab. Daher ist es immer eine gute Idee, die in der Regel auch online verfügbaren Handbücher (Manuals) zu konsultieren und sich die grundlegenden Suchanfragen anzueignen. Hier gilt wie so oft: »*Learning by Doing!*«.

POS-Tagging

Den meisten wird schon klar geworden sein, dass selbst eine Suche nach dem Lemma *brauchen* noch viel zu unspezifisch formuliert ist. Da die Lemmatisierung in der Regel rein formbasiert ist, also zwischen verschiedenen Verwendungen nicht unterscheidet, werden bei einer solchen Suche auch Vorkommen der Vollverb-Variante von *brauchen* gefunden (wie in dem Beispiel *Für den Militäreinsatz braucht man große Truppenverbände*), an denen wir aber gar nicht interessiert sind. Und bei sehr großen Korpora werden wir diese Treffer nicht mehr von Hand aussortieren können. Im Idealfall sollte man also bereits bei der Suchanfrage die Vollverben ausschließen oder zumindest auf ein erträgliches Maß reduzieren können. Wie kriegen wir das hin? Überlegen wir uns dazu noch einmal, nach welcher Art von Konstruktion wir suchen, welche distinktiven Eigenschaften sie hat und wie wir diese Eigenschaften in einer Suchanfrage nutzen können. Die modale Verwendung von *brauchen* zeichnet sich dadurch aus, dass *brauchen* immer entweder mit einem einfachen oder einem mit *zu* erweiterten infiniten Verb auftritt (ein solches selegiert) und das Ganze innerhalb desselben einfachen Satzes (da der Infinitiv nicht satzwertig ist). Wenn man also zum Beispiel fordern könnte, dass im selben Satz noch ein infinites Verb stehen muss, dann könnte man zumindest Beispiele wie das obige systematisch ausschließen. Das würde aber voraussetzen, dass im Korpus alle Wörter nach ihrer Wortart kategorisiert sind, *Militäreinsatz* als Nomen, *braucht* als finites Verb, *für* als Präposition. Tatsächlich sind diese Informationen in den meisten annotierten Korpora inzwischen Standard. Die Annotation mit solchen Informationen nennt man **POS-Tagging** (wobei POS für *Part of Speech* steht). Die Label, die für die Wortarten verwendet werden (das *Tagset*), sind dabei leider meist nicht identisch mit den in der Syntax verwendeten Labeln N, A, V, P, sondern kodieren noch weitere positionelle (z.B. Präpositionen vs. Postpositionen), syntaktische (z.B. attributives vs. prädikatives Adjektiv), morphologische (z.B. finites vs. infinites Verb) oder semantische (z.B. Appellativum vs. Eigenname) Informationen. Ein weit verbreitetes Tagset ist das Stuttgart-Tübingen-Tagset (kurz STTS), das in Tabelle 14.2 in Auszügen wiedergegeben ist.

Über das Interface cqpweb könnte man damit z.B. eine Suchanfrage wie `[lemma = "brauchen"] []1,8 [pos = "VVINF|VVIZU"] within s;` stellen, die fordert, dass im selben Satz spätestens nach 8 Wörtern auf eine Wortform von *brauchen* ein Vollverb im einfachen Infinitiv oder erweitert mit *zu* folgt. In Cosmas II lassen sich für STTS-konforme Korpora entsprechende Anfragen mit dem Operator `MORPH(...)` formulieren, wie wir gleich noch an einem konkreten Beispiel sehen werden.

POS-Tag	Beschreibung
VVFIN	finites Vollverb
VVINF	Vollverb, einfacher Infinitiv
VVIZU	Vollverb, Infinitiv mit *zu*
…	…
ADJA	attributives Adjektiv
ADJD	adverbiales oder prädikatives Adjektiv
…	…
$.	satzbeendende Interpunktion
$,	Komma

Tabelle 14.2: Auszug aus dem Stuttgart-Tübingen-Tagset (STTS)

Zwischen Wort und Satz

Bei der gerade formulierten Suchanfrage haben wir von zwei strukturellen Einheiten Gebrauch gemacht: dem Wort und dem Satz. In der Syntax haben wir aber gelernt, dass man zwischen der Wortebene und der Satzebene noch weitere strukturelle Einheiten identifizieren kann: die Konstituenten eines Satzes. Dabei haben wir auch gesehen, dass die Konstituenten eines Satzes hierarchisch organisiert (ineinander verschachtelt) sind, was uns zur Annahme von Phrasenstrukturen bzw. zur X-bar-Syntax geführt hat. Wären die Sätze in unserem Korpus in diesem Sinne syntaktisch analysiert, dann könnten wir unsere Anfragen nochmals deutlich verfeinern, indem wir uns auf diese hierarchische Strukturierung und damit auf die Konstituentenstruktur eines Satzes beziehen.

Parsing

Die syntaktische Analyse und Annotation von Sätzen wird als **Parsing** bezeichnet und in großen Korpora mit Programmen (Parsern) automatisch durchgeführt. Wie eine solche Analyse eines Satzes aussehen kann, soll das Beispiel in Abbildung 14.2 illustrieren, das der Tübinger Baumbank des Deutschen / Spontansprache (TüBa-D/S) entnommen ist und mit TIGERSearch (Lezius 2002) visualisiert wurde. An diesem Beispiel wird erneut deutlich, dass auch die Bezeichnungen der Einheiten zwischen Wort und Satz nicht mit denen der theoretischen Linguistik zusammenfallen, sondern zum Teil stark davon abweichen. Wir können hier nicht in die Details gehen, wollen aber auf eine Besonderheit der TüBa-Korpora hinweisen, die gerade im Kontext der germanistischen Linguistik von Interesse ist: Die Sätze sind (neben einer flachen phrasenstrukturellen Analyse) topologisch strukturiert. So bezeichnen zum Beispiel das Label LK die linke Satzklammer, MF das Mittelfeld und VC die rechte Satzklammer.

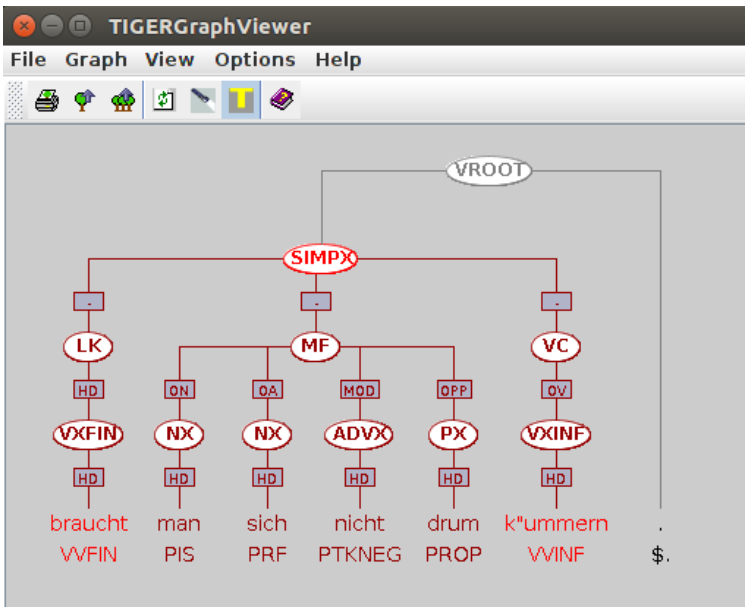


Abbildung 14.2: Syntaktische Analyse des Satzes s531 im Korpus TüBa-D/S

Such- und Visualisierungs-Tools

Das Beispiel sollte auch deutlich machen, dass geparste Korpora über eine sehr reiche Annotationsstruktur verfügen, die in geeigneter Form visualisiert werden muss. Das bereits erwähnte Programm TIGERSearch kombiniert eine solche Visualisierung mit einer (graphischen) Oberfläche für die Formulierung von Suchanfragen. Da dieses auf der Programmiersprache Java basierende Programm nicht mehr weiterentwickelt wird, läuft es leider nur noch unter älteren Java-Versionen und Betriebssystemen. Mit ANNIS (ANNotation of Information Structure) hat sich jedoch inzwischen ein browser-basiertes System etabliert, das sogar mehrdimensionale Annotationen erlaubt und ebenfalls eine graphische Visualisierung mit einem Suchanfragen-Interface kombiniert (Krause & Zeldes 2016). In Abbildung 14.3 findet sich zum Beispiel die Analyse eines *brauchen*-Satzes mit einfachem Infinitiv aus dem FraC-Korpus (Horch & Reich 2017). Links oben in der Grafik erkennt man die Suchanfrage und links unten wird der Vor- und der Nachkontext beschränkt. Der rechte Teil der Grafik zeigt die Analyse des Satzes in mehreren Ebenen: Die erste Ebene veranschaulicht (unter anderem) die POS-Tags, eine zweite Ebene zeigt die geparste Struktur, also die automatisch generierte hierarchische Analyse. Weitere Ebenen können je nach Korpus und Art der Annotation hinzukommen. ANNIS läuft auf allen gängigen Betriebssystemen und ist Open Source. Einige der oben erwähnten Korpora liegen bereits in ANNIS-Format vor (wie die DDD-Korpora), andere können mit dem Programm »Pepper« konvertiert werden. Weitere Screenshots und Informationen sind über http://corpus-tools.org/ verfügbar.

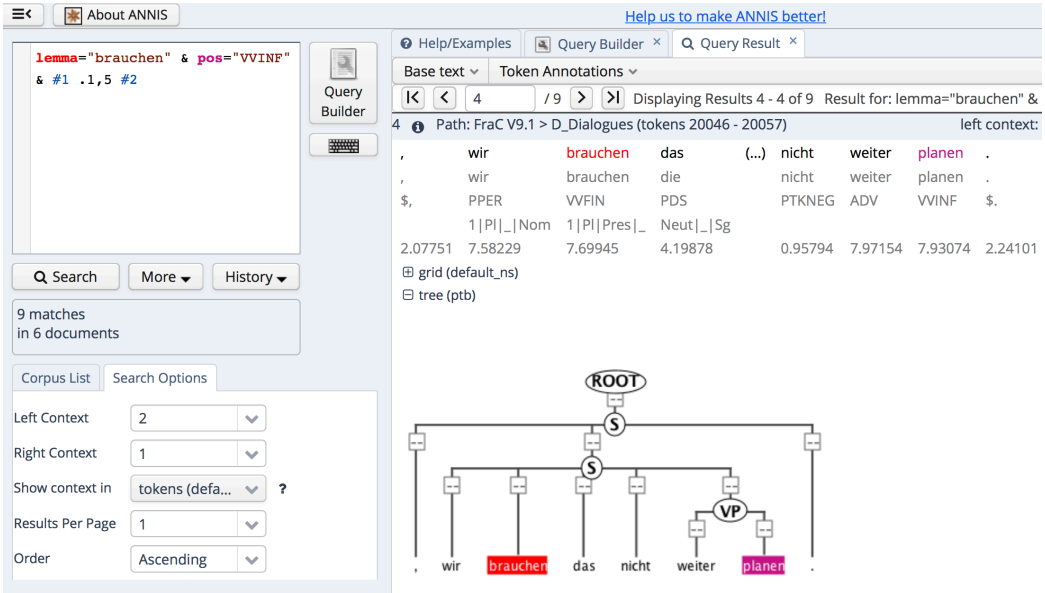


Abbildung 14.3: Analyse eines Satzes aus dem FraC-Korpus in ANNIS

14.6.2 Eine exemplarische Korpusstudie

Eine exemplarische Korpusstudie

Kommen wir damit zurück zu unserer Hypothese (H1). Um (H1) zu überprüfen, benötigen wir, wie gesagt, ein Korpus zum gesprochenen Deutsch und ein Korpus zum geschriebenen Deutsch. Da hier mehrere Korpora in Frage kommen, haben wir die Qual der Wahl. Ein Kriterium für die Wahl des Korpus kann sein, dass die Gespräche bzw. Texte möglichst unterschiedlicher Art sind. Ein weiteres Kriterium kann sein, dass die Gespräche bzw. Texte möglichst so gewichtet sind, dass sie als mehr oder weniger repräsentativ für die gesprochene bzw. geschriebene Sprache gelten können (auch wenn das natürlich immer nur Annäherungen sein können). In unserem Fall fiel die Wahl auf zwei Korpora am IDS, zum einen auf das »Forschungs- und Lehrkorpus Gesprochenes Deutsch« (FOLK) und zum anderen auf das »Deutsche Referenzkorpus« (DeReKo), das sich aus Zeitungen und Zeitschriften aus dem deutschsprachigen Raum speist. Die Größe der Korpora liegt mit gut 2,2 Millionen Wörtern beim FOLK (Stand Mai 2018) und über 40 Milliarden Wörtern beim DeReKo weit auseinander, was für uns jedoch nicht wirklich problematisch ist, da wir nicht an absoluten Zahlen, sondern an relativen Verhältnissen interessiert sind. Beim DeReKo haben wir uns dennoch aus praktischen Gründen auf ein Teilkorpus (TAGGED-T2-öffentlich mit gut 1 Milliarde laufenden Wortformen) beschränkt. Unsere Suchanfragen über Cosmas II (TAGGED-T2) und die Token-Recherche der DGD (Datenbank für gesprochenes Deutsch) liefern die in Abbildung 14.4 in einem Balkendiagramm graphisch aufbereiteten Zahlen. Die

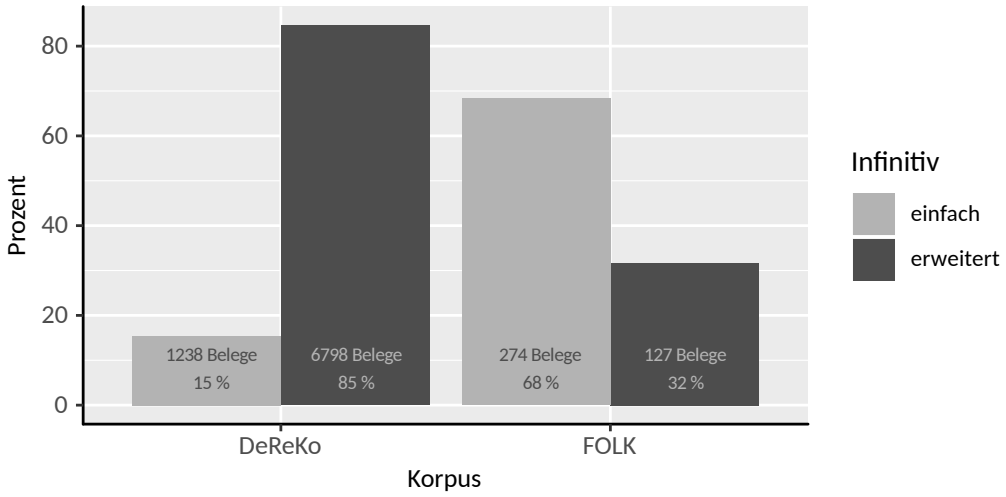


Abbildung 14.4: Frequenz von modalem *brauchen* mit einfachem Infinitiv und erweitertem Infinitiv in absoluten Zahlen und Prozentangaben

Zahlen machen deutlich, dass mit einem Verhältnis von rund 13:2 der erweiterte Infinitiv im DeReKo (also im Korpus zum geschriebenen Deutsch) deutlich öfter mit modalem *brauchen* konstruiert als der einfache Infinitiv. Im FOLK (also im Korpus zum gesprochenen Deutsch) kehren sich die Verhältnisse dagegen um: Hier kommen auf 1 mit *zu* erweiterten Infinitiv knapp 3 einfache Infinitive. Die Zahlen legen also nahe, dass unsere Korpusstudie (H1) stützt. Ob dies tatsächlich der Fall ist, muss am Ende ein geeigneter statistischer Test zeigen. Darauf werden wir hier aber nur kurz unten in einer Vertiefungsbox eingehen.

Wirklich so einfach?

So, wie wir die Ergebnisse unserer Recherche gerade dargestellt haben, könnte der Eindruck entstehen, dass eine Korpusstudie vergleichsweise einfach ist: Man sucht sich geeignete Korpora, stellt dann seine Suchanfragen und erhält schon die gewünschten Zahlenverhältnisse. Die Realität ist leider etwas komplizierter, da es in der Regel schwierig bis unmöglich ist, Suchanfragen so präzise zu formulieren, dass man die und nur die Daten erhält, an denen man auch tatsächlich interessiert ist. Illustrieren wir das an einer Suche nach modalem *brauchen* über Cosmas II im DeReKo-Korpus.

Eine exemplarische Suchanfrage

Ideal wäre es, wenn wir unsere Suchanfrage so stellen könnten, dass eine beliebige Wortform von *brauchen* gemeinsam mit einem (mit *zu* erweiterten) Infinitiv im selben Satz vorkommt, der von *brauchen* regiert (also selegiert bzw. gefordert) wird. Leider sind solche Abhängigkeitsverhältnisse im DeReKo-Korpus nicht annotiert, weshalb wir uns unserem gewünschten Ergebnis eher indirekt mit den eben verfügbaren Mitteln werden annähern müssen. Eine erste Formulierung für eine

Suchanfrage könnte wie folgt lauten: `&brauchen /s0 (MORPH(VRB inf v))` – Wir suchen nach allen Wortformen von *brauchen* (`&brauchen`), wobei im selben Satz (`/s0`) ein einfacher Infinitiv (`VRB inf v`) eines Vollverbs stehen soll. Schauen wir uns jetzt die Treffer an, dann werden diese leider auch Fälle der Art (i) *wie dringend kann ein Mensch Urlaub brauchen* oder (ii) *denn wer Sterne beobachten will, der braucht es dunkel* enthalten. Satz (i) landet in unserer Trefferliste, weil wir nicht explizit ausgeschlossen haben, dass *brauchen* und der geforderte Infinitiv auch zusammenfallen können. Das ist relativ einfach zu ändern, indem wir die erwähnte `MORPH`-Bedingung mit der Bedingung `%w0 brauchen` modifizieren. Bei Satz (ii) ist das Problem, dass wir zwar einen anderen Infinitiv gefunden haben (*beobachten*), dieser aber Teil des freien und linksversetzten Relativsatzes ist. Eine Möglichkeit, diesen Satz auszuschließen, ist, mit der Bedingung `/+s0` zu fordern, dass der Infinitiv auf *brauchen* folgen soll. Man sollte sich dann aber darüber klar sein, dass man damit gleichzeitig auch Sätze der Art *weil er dann gar nicht mehr kommen braucht* ausschließt, die man eigentlich in der Trefferliste haben möchte. Da man ein ähnliches Problem bei nachgestellten Nebensätzen hat, wäre die elegantere Möglichkeit (mit der Bedingung `%w0 ,`) zu fordern, dass zwischen *brauchen* und dem geforderten Infinitiv kein Komma stehen darf. Leider funktioniert das nicht bei mit *um* eingeleiteten satzwertigen Infinitiven, da diese (wie sich zeigt) im Korpus häufig nicht mit Komma abgetrennt werden. Außerdem zeigt sich bei einem etwas intensiveren Blick in die Resultate, dass auch Bindestriche oder koordinierende Konjunktionen wie *und* zu unerwünschten Treffern führen. Und würde man noch tiefer bohren, würden sich sicherlich noch weitere problematische Punkte finden lassen.

Was eine gute Suchanfrage ausmacht

Dieses Beispiel sollte vor allem verdeutlichen, dass das Formulieren einer guten Suchanfrage seine Zeit braucht. Denn die Kunst ist, so viele unerwünschte Treffer wie möglich auszuschließen, dabei aber gleichzeitig möglichst viele erwünschte Treffer beizubehalten. Und um sich diesem Optimum anzunähern, wird man immer wieder die Suchanfrage modifizieren und in den Ergebnissen überprüfen müssen, wie sich diese Modifikationen auswirken, welche unerwünschten Treffer immer noch enthalten sind und wie man diese sukzessive eliminieren kann, ohne über das Ziel hinauszuschießen. Nicht selten hilft hier auch ein Blick in die Annotation der Daten. Aber selbst wenn man eine vergleichsweise präzise Suchanfrage formuliert und damit schon viel Datenmüll vermieden haben sollte, kann noch eine manuelle Nachbearbeitung erforderlich sein. Je umfangreicher die Trefferliste ist, umso schwieriger wird eine solche Nachbearbeitung natürlich werden.

Zur Vertiefung

Der Chi-Quadrat-Test

In unserem Beispiel scheinen die Verhältnisse so klar zu sein, dass man nicht mehr wirklich in Frage stellen würde, dass (H1) bestätigt ist. Der Schein kann

allerdings trügen, insbesondere wenn die Anzahl der Treffer recht klein (unter 5 Datenpunkte pro Bedingung) ist. Daher ist es fast immer sinnvoll, die deskriptive Analyse mit einem statistischen Test zu untermauern. Was für einen Test kann man in diesem Fall aber anwenden? Und wie funktioniert eigentlich so ein Test? Aus Platzgründen können wir hier leider nur auf die zentrale Idee eingehen und müssen die Leser*innen für detailliertere Informationen auf die einschlägige Literatur verweisen (z.B. auf die Darstellungen in Gries 2008, 2013, Meindl 2011, Brezina 2018).

Also, was für einen Test kann man jetzt rechnen? Und wie funktioniert er? Ist die Anzahl an Treffern ausreichend, dann ist die Antwort auf die erste Frage meist: einen χ^2-Test (Chi-Quadrat-Test). Und was macht dieser Test? Die kurze Antwort ist, dass dieser Test unsere numerischen Resultate mit den unter der Nullhypothese (also bei Annahme von Zufall) erwarteten Resultaten vergleicht und berechnet, ob die Differenz zwischen beobachteten und erwarteten Werten (relativ zum Signifikanzniveau) noch auf Zufall beruhen kann. Wir brauchen also die beobachteten und die erwarteten Werte. Die beobachteten Werte haben wir bereits in Abbildung 14.4 aufgeführt. Für den χ^2-Test werden diese aber in der Regel in einer Kreuztabelle angegeben:

	einfach	*erweitert*	*Zeilensummen*
DeReKo	1238	6798	8036
FOLK	274	127	401
Spaltensummen	1512	6925	8437

Tabelle 14.3: Kreuztabelle der beobachteten Werte

Welche Werte würde man unter der Nullhypothese erwarten? Die Nullhypothese ist, dass eventuelle Differenzen auf Zufall beruhen. Mit anderen Worten: Die Erwartung ist, dass sich die Verhältnisse zwischen einfachem und erweitertem Infinitiv nicht (wesentlich) verändern, wenn wir von geschriebener zu gesprochener Sprache übergehen (oder umgekehrt). Für die zu erwartenden Werte müssen wir also die relativen Verhältnisse (13:2 und 1:3) einander angleichen. Das kann man auf der Basis einer Kreuztabelle wie der obigen tun, indem man die jeweilige Spaltensumme und Zeilensumme multipliziert und durch die Gesamtsumme von 8437 dividiert. Der erwartete Wert für einfache Infinitive im DeReKo ergibt sich also, indem man die Spaltensumme 1512 mit der Zeilensumme 8036 multipliziert und das Ganze durch 8437 dividiert. Gerundet ergibt das 1440. Analog für die anderen Zellen:

	einfach	*erweitert*	*Zeilensummen*
DeReKo	1440	6596	8036
FOLK	72	329	401
Spaltensummen	1512	6925	8437

Tabelle 14.4: Kreuztabelle der erwarteten Werte

Der χ^2-Test bildet nun für jede mögliche Kombination aus Art des Korpus und Art des Infinitivs (also für jede der vier Zellen) die Differenz aus dem beobachteten und dem erwarteten Wert, quadriert diese Differenzen (damit sie sich nicht gegenseitig aufheben) und gewichtet die Quadrate jeweils mit dem erwarteten Wert. Diese vier Werte werden aufsummiert und ergeben den χ^2-Wert. Der χ^2-Wert ist damit ein Maß für die beobachtete Abweichung von einer zufälligen Verteilung. In unserem Fall ergibt sich für die einfachen Infinitive im DeReKo eine Abweichung von (1238 - 1440)² / 1440 = 28,3. Insgesamt ergibt sich gerundet ein χ^2-Wert von 725.

Die nächste Frage lautet: Wie unwahrscheinlich ist diese Abweichung? Hier ergibt sich die Antwort, indem man schaut, wieviel Fläche der χ^2-Wert 725 rechts von der Fläche unter der χ^2-Verteilung (mit dem Freiheitsgrad 1) abschneidet. Diesen Wert kann man in einer Tabelle nachschauen oder mit einem Programm berechnen. An dieser Stelle sei die Software R empfohlen, die sich in der Linguistik weitgehend als Standard durchgesetzt hat und die zum Beispiel über das Programm RStudio gut bedienbar ist. Die einschlägige Funktion `pchisq(725, 1, lower.tail = FALSE)` gibt hier einen p-Wert von $1.1 * 10^{-159}$ zurück. (Die Dezimalzahlen sind hier nach dem englischen System notiert, das heißt, Dezimalzahlen sind durch einen Punkt und nicht durch ein Komma abgetrennt wie im deutschen System.) Dieser p-Wert wird in der Regel nicht selbst berichtet, es wird nur angegeben, unter welchem Signifikanzniveau er liegt. Liegt der p-Wert unter 0.05, dann spricht man von einem *signifikanten* Ergebnis. Liegt der p-Wert unter einem Wert von 0.01, dann wird das Ergebnis als *sehr* und unter 0.001 als *hoch* signifikant bezeichnet. In einer Publikation würde man dann in etwa wie folgt berichten: »Die Verteilung der beiden Infinitivformen weicht gemäß einem Chi-Quadrat-Anpassungstest hoch signifikant von der erwarteten Gleichverteilung ab ($\chi^2(1) = 725, p < 0.001$)«.

Anmerkung: Eine Einführung in die Programme R und RStudio kann hier leider nicht gegeben werden, lediglich Hinweise auf relevante Funktionen.

Am Ende dieses Kapitels ist jedoch Literatur angeführt, in der beide Programme systematisch dargestellt werden. Da selbst RStudio für Einsteiger anfangs etwas schwer zugänglich ist, sei hier auch das Programm Jamovi empfohlen, das ebenfalls auf R basiert, aber einfacher zu installieren ist und auch keine Programmierkenntnisse erfordert. In Jamovi können die beobachteten Daten direkt in Tabellen eingegeben und statistisch ausgewertet werden.

14.7 Fragebogenstudien

Hypothese (H2)

Wie wir bereits in Abschnitt 14.6 gesehen haben, sind Korpusrecherchen nicht für jede Fragestellung geeignet. So macht unsere Hypothese (H2) eine qualitative Aussage über die Einschätzung der Akzeptabilität der Konstruktionen modales *brauchen* plus *zu*-Infinitiv vs. einfachen Infinitiv durch Muttersprachler*innen:

(H2) Muttersprachler*innen finden modales *brauchen* plus *zu*-Infinitiv natürlicher / akzeptabler als *brauchen* plus einfachen Infinitiv.

14.7.1 Zum Design von Fragebögen

Minimalpaare und Token-Sets

Wir hatten auch schon eine erste Idee, wie man (H2) überprüfen könnte: Wir fragen einfach Muttersprachler*innen, welche Variante sie besser finden. Das heißt, wir bilden zwei Sätze wie (i) *dann brauchst du gar nicht mehr kommen* und (ii) *dann brauchst du gar nicht mehr zu kommen*, die sich nur in der Art der Konstruktion (einfacher Infinitiv vs. *zu*-Infinitiv) unterscheiden. Dieses **Minimalpaar** deckt in unserem Fall bereits alle relevanten Varianten ab und bildet damit ein so genanntes »Token-Set«. Als **Token-Set** bezeichnet man alle möglichen Ausprägungen einer unabhängigen Variable bzw. alle möglichen Kombinationen der Ausprägungen bei mehreren unabhängigen Variablen. Illustrieren wir das an unserem Beispiel. Unsere unabhängige Variable ist hier die Art des Infinitivs. Diese Variable hat zwei mögliche Werte (Ausprägungen): einfacher Infinitiv oder erweiterter Infinitiv. Damit hat unser Token-Set zwei Bedingungen, vgl. hierzu Tabelle 14.5.

Bedingung	Art des Infinitivs
1	einfach
2	erweitert

Tabelle 14.5: Schematisches Token-Set zu Hypothese (H2)

Ein 2x2-Design

Die Ergebnisse unserer Korpusstudie lassen vermuten, dass (H2) möglicherweise nicht generell gilt, sondern dass wir zwischen gesprochener und geschriebener Sprache unterscheiden müssen, dass (H2) also nur für die geschriebene Sprache gilt und für die gesprochene Sprache genau die umgekehrte Aussage. Bezeichnen wir diese erweiterte Hypothese als (H3). Mit der Unterscheidung zwischen gesprochener und geschriebener Sprache führen wir in (H3) eine weitere unabhängige Variable (nennen wir sie Kanal) mit zwei Ausprägungen ein, eben gesprochen vs. geschrieben. Damit liegen für (H3) aber insgesamt $2 \times 2 = 4$ Bedingungen vor, wie Tabelle 14.6 illustriert. Wer sich nochmals unsere Korpusstudie vor Augen führt,

Bedingung	Art des Infinitivs	Art des Kanals
1	einfach	schriftlich
2	einfach	mündlich
3	erweitert	schriftlich
4	erweitert	mündlich

Tabelle 14.6: Schematisches Token-Set zu Hypothese (H3)

wird sich erinnern, dass wir dort vier Häufigkeiten zueinander in Beziehung gesetzt haben. Jede dieser Häufigkeiten entspricht offenbar genau einer der Bedingungen in Tabelle 14.6. Wir könnten jetzt auch unsere Fragebogenstudie erweitern und (H3) in einem **2x2-Design** testen. Um die Darstellung der Fragebogenstudie möglichst einfach zu halten, werden wir aber an der einfacheren Hypothese (H2) und damit 2 Bedingungen festhalten und die Hypothese implizit auf geschriebene Sprache einschränken. Das ist auch insofern sinnvoll, als wir die zu beurteilenden Sätze schriftlich und nicht mündlich präsentieren werden.

Einfaches kompliziert?

Auf den ersten Blick scheint es, als ob wir hier etwas eigentlich sehr einfaches unnötig kompliziert dargestellt hätten. Wir werden aber gleich sehen, dass es für die Durchführung eines Experiments außerordentlich wichtig ist, zum einen die Hypothese präzise zu formulieren und sich zum anderen über ein schematisches Token-Set darüber klar zu werden, wie viele und welche unabhängigen Variablen man in seinem Experiment hat, wie viele und welche Ausprägungen diese Variablen haben und zu wie vielen Bedingungen alle möglichen Kombinationen dieser Ausprägungen führen. Nur so kann man sich wirklich sicher sein, dass man auch genau das testet, was man eigentlich testen will.

Lexikalische Varianten

Der Nutzen schematischer Token-Sets wird auch gleich an einem anderen Problem deutlich werden: Bisher haben wir nur ein Minimalpaar gebildet, eben die beiden Sätze (i) und (ii) zu Anfang dieses Abschnitts. Diese beiden Sätze werden aber sicher nicht ausreichen, um die Hypothese (H2) zu testen. Denn wir wollen

ja eine Aussage über die Konstruktionen an sich machen und nicht über konkrete einzelne Sätze. Und es könnte ja sein, dass wir das eine Beispiel so ungeschickt gewählt haben, dass es – aus welchen Gründen auch immer – für die interessierende Konstruktion gänzlich untypisch ist. Um die Wahrscheinlichkeit eines solchen Fehlgriffs möglichst gering zu halten, wird man mehrere Paare wie (i) und (ii) testen wollen, man wird sich also mehrere lexikalische Varianten überlegen, die das schematische Token-Set in Tabelle 14.5 mit Inhalt füllen. Jedes dieser Paare repräsentiert dann ein konkretes Token-Set. Und jeder einzelne Satz eines konkreten Token-Sets wird als Test-Item, oder einfach kurz Item, bezeichnet.

Störfaktoren

Bei der Konstruktion solcher lexikalischer Varianten gilt es gleichzeitig weitere potentielle Störfaktoren auszuschließen oder zumindest möglichst gering zu halten. So wird man in der Regel expressive Ausdrücke wie *idiotisch*, *krass* oder *geil* vermeiden, die bei den Versuchspersonen (VPen) eine emotionale Reaktion auslösen und so deren Beurteilung der Items negativ beeinflussen könnten. Auch eher technische oder bildungssprachliche Ausdrücke wie *artifiziell* oder *authentisch* können, wenn die VP sie nicht kennt, die Bewertung verfälschen. In manchen Fällen wird man auch die Häufigkeit von bestimmten Ausdrücken wie zum Beispiel von Verben oder von Eigennamen über korpusbasierte Vorstudien kontrollieren wollen. In unserem konkreten Fall kommt als weiteres Problem hinzu, dass modales *brauchen* ähnlich wie das Modalverb (*nicht*) *müssen* polyfunktional ist, das heißt, es lässt neben den primären zirkumstanziellen Lesarten (wie dispositionell oder deontisch) auch eine epistemische Lesart zu (vgl. hierzu z.B. Reis 2005, Maché 2019). In einem Experiment wird man daher entweder die Art der modalen Lesart konstant halten (und so andere Lesarten als Störfaktoren ausschließen) oder den Kontrast zweier (oder mehrerer) Lesarten als zusätzlichen Faktor (als weitere unabhängige Variable) in das Experiment integrieren wollen. Um nicht zu tief in eine inhaltliche Diskussion von modalen Lesarten einsteigen zu müssen, entscheiden wir uns hier für die erste Strategie und beschränken die Items im Wesentlichen auf zirkumstantielle (also nicht-epistemische) Lesarten.

Verschleiern der Fragestellung

Nehmen wir also an, wir hätten jetzt nicht nur 1, 2 oder 3 Minimalpaare in der Art von (i) und (ii) gebildet, sondern 12 oder sogar 24. Wie gehen wir jetzt konkret vor? Die erste Idee war ja, für jedes Paar zu fragen, welche Variante die Muttersprachler*innen des Deutschen als natürlicher oder akzeptabler einschätzen. Dieses Vorgehen hätte aber einen entscheidenden Nachteil: Den VPen wäre sofort klar, worum es in diesem Experiment geht. Und da man häufig schon in der Schule vermittelt bekommt, dass nur die Kombination mit dem erweiterten Infinitiv »gutes« Deutsch ist (man erinnere sich an den Spruch »wer *brauchen* ohne *zu* gebraucht, braucht *brauchen* überhaupt nicht zu gebrauchen«), wäre damit zu rechnen, dass dieses normative Wissen das Urteil der VPen dahingehend beeinflusst,

dass sie stärker zum erweiterten Infinitiv tendieren, als sie das ohne die Aktivierung dieses Wissens eigentlich tun würden. Um solche Effekte zu vermeiden, werden bei Fragebogenuntersuchungen zwei wichtige Strategien angewandt: Zum einen werden die Beispielsätze so auf mehrere Fragebögen verteilt, dass jede VP nur einen Beispielsatz aus jedem Token-Set sieht. Und zum anderen wird das Experiment noch durch weitere Sätze ergänzt, die mit der uns interessierenden Konstruktion nichts zu tun haben. Beide Strategien haben das Ziel, die VPen möglichst darüber im Unklaren zu lassen, worauf das Experiment abzielt.

Verteilen der Items auf mehrere Fragebögen

Kommen wir zur ersten Strategie, der Verteilung der zu beurteilenden Sätze auf verschiedene Fragebögen mit dem Ziel einer möglichst unabhängigen Einschätzung der Items. Wir haben bereits angedeutet, dass keine VP mehr als ein Item aus einem konkreten Token-Set sehen sollte, um Minimalpaarbildungen in Fragebögen zu vermeiden. Wenn jeder VP genau ein Item aus jedem Token-Set präsentiert wird und die VP dabei alle Bedingungen sieht, dann spricht man von einem **»Within-Subjects«-Design**. Manchmal kann es aber gute Gründe geben, die Token-Sets oder bestimmte Bedingungen über mehrere Versuchspersonen(klassen) zu verteilen (beispielsweise wenn – wie in unserem 2x2-Design – die Stimuli auf unterschiedliche Art und Weise präsentiert werden müssen). Dann spricht man von einem **»Between-Subjects«-Design**. Aus statistischen Gründen wird man aber immer ein »Within-Subjects«-Design bevorzugen.

Anzahl der Fragebögen

Wenn wir jetzt annehmen, dass wir jeder VP genau ein Item aus jedem Token-Set präsentieren, dann wird auch sofort klar, dass wir nicht mit einem einzelnen Fragebogen auskommen werden. Sonst würde ja nur je eine Bedingung pro Token-Set getestet werden. Wir brauchen also (mindestens) so viele Fragebögen, wie wir Bedingungen in einem Token-Set haben. Da wir in unserem Experiment lediglich zwei Bedingungen haben, kommen wir also mit zwei Fragebögen aus. Idealerweise sollte dabei jede VP von jeder Bedingung dieselbe Anzahl Items sehen. Um es wieder an unserem Beispiel zu verdeutlichen: Angenommen wir haben 12 Token-Sets mit je zwei Bedingungen (Bedingung 1, Bedingung 2). Dann sieht jede VP 12 Items: Je eines pro Token-Set, 6 von Bedingung 1 und 6 von Bedingung 2.

Lateinisches Quadrat

Diese Art der Verteilung der Items auf mehrere Fragebögen folgt dem Muster des **Lateinischen Quadrats**, wie man es vom Sudoku kennt. Wenn wir zwei Bedingungen in unserem Experiment haben, dann benötigen wir minimal zwei Token-Sets (TS1, TS2), um die Items nach obigem Muster auf zwei Fragebögen (FB1, FB2) zu verteilen. Mit 1 für »Bedingung 1« und 2 für »Bedingung 2« kann man die Verteilung in einer Tabelle mit 2 Zeilen und 2 Spalten wie in 14.7 darstellen.

Für 2 Bedingungen ist das noch ziemlich einfach. Etwas komplexer wird es, wenn man (wie in einem 2x2-Design) 4 Bedingungen in seinem Experiment hat. In diesem Fall besteht ein Lateinisches Quadrat aus 4 Zeilen und 4 Spalten. Mit anderen

	FB1	FB2
TS1	1	2
TS2	2	1

Tabelle 14.7: Lateinisches Quadrat bei 2 Bedingungen

Worten: Wir brauchen 4 Fragebögen und (minimal) 4 Token-Sets. Die Verteilung der Items bei einem solchen Experiment ergibt sich dann wie in Tabelle 14.8.

	FB1	FB2	FB3	FB4
TS1	1	2	3	4
TS2	2	3	4	1
TS3	3	4	1	2
TS4	4	1	2	3

Tabelle 14.8: Lateinisches Quadrat bei 4 Bedingungen

Anzahl der Token-Sets

Auch in diesem Lateinischen Quadrat kommt jedes zu beurteilende Item genau einmal vor: Zunächst verteilen wir alle Items von TS1 so auf die vier Fragebögen, dass die Bedingungen von 1 bis 4 numerisch aufsteigen. Dann verteilen wir alle Items von TS2 in derselben Art und Weise, beginnen aber nicht mit 1, sondern mit 2. Bei TS3 beginnen wir mit 3, bei TS4 mit 4. Jeder Fragebogen enthält jetzt für jede Bedingung genau ein Item, diese stammen aber alle aus unterschiedlichen Token-Sets. Macht man sich die Systematik hinter dieser Verteilung klar, dann sieht man schnell, dass aus der Art der Verteilung eine (weitere) Beschränkung für die Anzahl der Token-Sets in einem Experiment folgt: Die Anzahl der Token-Sets muss immer ein **Vielfaches der Anzahl der Bedingungen** sein. Haben wir also 2 Bedingungen in unserem Experiment, dann muss die Anzahl der Token-Sets durch 2 teilbar sein. Haben wir 4 Bedingungen in unserem Experiment, dann muss die Anzahl durch 4 teilbar sein. Und damit sich ein einzelnes Item nicht zu stark auf das Gesamtbild auswirkt, sollte sich die Anzahl der Token-Sets in einer Größenordnung von 12 bis 24 bewegen. (Zu empfehlen wären immer eher 24.) Die Verteilung der verbleibenden Token-Sets erfolgt natürlich nach demselben Schema.

Konkretes Vorgehen

Soviel zur abstrakten Beschreibung. Aber wie sieht das in der Praxis aus? Wie geht man ganz konkret vor? Zunächst einmal wird man die Items nach Token-Sets und Bedingungen geordnet in einem Tabellenkalkulationsprogramm erfassen. In

einer Spalte LISTE wird man die Items dann den Fragebögen (Listen) zuordnen, vgl. hierzu Abbildung 14.5: Die Items des ersten Token-Sets werden der Reihe nach auf die Fragebögen fb1 und fb2 verteilt, die Items des zweiten Token-Sets auf die Fragebögen fb2 und fb1, die Items des dritten Token-Sets wieder auf fb1 und fb2 und so weiter und so fort. Die Systematik ist in Abbildung 14.5 farblich angedeutet. Filtert man jetzt die Tabelle über die Spalte LISTE zum Beispiel nach fb1, dann erhält man genau die Items, die am Ende auf den Fragebogen fb1 sollen.

T-SET	BEDINGUNG	LISTE	TYP	TESTSATZ
1	mit_zu	fb1	Item	Dann brauchst du gar nicht mehr aufzuräumen!
1	ohne_zu	fb2	Item	Dann brauchst du gar nicht mehr aufräumen!
2	mit_zu	fb2	Item	Das braucht niemanden zu interessieren.
2	ohne_zu	fb1	Item	Das braucht niemanden interessieren.
3	mit_zu	fb1	Item	Mit einem eigenen Auto brauchst du nicht mehr zu laufen.
3	ohne_zu	fb2	Item	Mit einem eigenen Auto brauchst du nicht mehr laufen.
4	mit_zu	fb2	Item	Du brauchst das Ziel nicht als Erster zu erreichen.
4	ohne_zu	fb1	Item	Du brauchst das Ziel nicht als Erster erreichen.
...	...	...	...	...

Abbildung 14.5: Zuordnung von Items zu Fragebögen (2 Bedingungen)

In einem 2x2-Design sieht das nicht wesentlich anders aus. Ein Unterschied ist allerdings, dass die beiden unabhängigen Variablen mit ihren Ausprägungen in eigenen Spalten erfasst und die 4 Kombinationsmöglichkeiten ähnlich wie in Tabelle 14.6 auf 4 Zeilen verteilt sind. Beim ersten Token-Set erfolgt die Zuordnung zu den Listen in der Reihenfolge fb1, fb2, fb3, fb4. Beim zweiten Token-Set in der Reihenfolge fb2, fb3, fb4, fb1 und so weiter und so fort. Das ist in Abbildung 14.6 für die beiden unabhängigen Variablen INFINITIV und KANAL angedeutet. Fettdruck soll hier andeuten, dass das fragliche Item mündlich präsentiert wird.

T-SET	KANAL	INFINITIV	LISTE	TYP	TESTSATZ
1	schriftlich	mit_zu	fb1	Item	Dann brauchst du gar nicht mehr aufzuräur
1	schriftlich	ohne_zu	fb2	Item	Dann brauchst du gar nicht mehr aufräume
1	mündlich	mit_zu	fb3	Item	**Dann brauchst du gar nicht mehr aufzurä**
1	mündlich	ohne_zu	fb4	Item	**Dann brauchst du gar nicht mehr aufräum**
2	schriftlich	mit_zu	fb2	Item	Das braucht niemanden zu interessieren.
2	schriftlich	ohne_zu	fb3	Item	Das braucht niemanden interessieren.
2	mündlich	mit_zu	fb4	Item	**Das braucht niemanden zu interessieren.**
2	mündlich	ohne_zu	fb1	Item	**Das braucht niemanden interessieren.**
3	schriftlich	mit_zu	fb3	Item	Mit einem eigenen Auto brauchst du nicht n
3	schriftlich	ohne_zu	fb4	Item	Mit einem eigenen Auto brauchst du nicht n
...	...	...	...	...	...

Abbildung 14.6: Zuordnung von Items zu Fragebögen (4 Bedingungen)

Ablenkersätze

Wir haben jetzt also mehrere Fragebögen und die VPen sehen von jedem Token-Set genau ein Item. Damit ist Teil 1 unserer Verschleierungstaktik erfolgreich abge-

schlossen. Kommen wir zu Teil 2, dem Auffüllen des Experiments mit **Ablenkern** oder auch **Fillern**, wobei auf jedes Item mindestens 2, besser 3 Fillern kommen sollten, damit die Items unter den Fillern verschwinden. Grundsätzlich sind Filler mehr oder weniger beliebig gewählte Sätze, die im engeren Sinne nichts mit unserem Experiment zu tun haben. Mehr oder weniger deshalb, weil natürlich auch die Filler bestimmten Kriterien genügen sollten. Einerseits sollten sie den Items nicht zu ähnlich sein, um die Beurteilung der Items nicht zu verfälschen. In unserem Fall sollten die Filler beispielsweise keine Modalverben oder entsprechende Infinitkonstruktionen enthalten. Auf der anderen Seite sollten die Filler aber auch nicht zu verschieden sein, damit die Items nicht als solche erkennbar hervorstechen. So wäre zum Beispiel der Satz *hoffentlich ist das Paket gestern noch angekommen* ein in unserem Kontext geeigneter Filler, da er eine ähnliche Länge und Komplexität wie unsere Items aufweist. Nicht geeignet wäre dagegen der folgende Satz aus der SZ.de vom 19.02.2019: *Die Frage, ob Autofahrer Schadenersatz für manipulierte Dieselautos von Volkswagen bekommen, geht vor das oberste deutsche Zivilgericht.*

Akzeptabilität von Ablenkersätzen

Gleichzeitig sollten die Filler ein geeignetes **Akzeptabilitätsspektrum** abdecken, sie sollten also nicht alle völlig natürlich sein. Der Grund dafür ist offensichtlich: Wenn nur die Items (in bestimmten Bedingungen) unnatürlich wirken, dann sind sie auch wieder relativ leicht zu identifizieren. Darüber hinaus erlauben weniger akzeptable Filler auch, den Grad der Akzeptabilität der getesteten Konstruktionen besser einzuschätzen. Man wird also auch Filler wie *selbst zubereitet wurde das Essen von mir* einstreuen, die zwar grammatisch sind, aber eher spezielle sprachliche Kontexte erfordern, bis hin zu nahezu ungrammatischen Sätzen wie *das Fenster über der Heizung hat oft er geschlossen*. Sätze dieser Art sind auch recht gut dafür geeignet zu überprüfen, ob die VPen die ihnen gestellte Aufgabe auch gewissenhaft ausführen: Werden (mehrere) ungrammatische Sätze von einer VP als völlig akzeptabel bewertet, dann wird man sie bei der statistischen Auswertung ausschließen wollen. Ab wieviel solcher falsch bewerteten **Catch Filler** man eine VP ausschließt, ist vor der Durchführung des Experiments festzulegen.

Pseudo-Randomisierung

Das Auffüllen eines Experiments mit Fillern alleine reicht natürlich noch nicht aus, um die Items zwischen den Fillern verschwinden zu lassen. Man muss die Items und die Filler noch mischen und das am besten in zufälliger Reihenfolge. Man spricht hier von **Randomisierung**. Da aber auch zufällige Verteilungen zu problematischen Abfolgen führen können, wird man die zufällige Reihung überprüfen und, wenn nötig, von Hand nachbessern. Als besonders problematisch wird dabei erachtet, wenn das Experiment mit einem Item beginnt oder wenn zwei Items direkt aufeinander folgen. Da in solchen Fällen manuell nachgebessert wird, spricht man auch von **Pseudo-Randomisierung**.

Konkretes Vorgehen

Auch hier noch ein Hinweis praktischer Natur: Werden die Filler in dieselbe Tabelle aufgenommen wie die Items, dann bietet es sich an, eine Spalte ZUFALLSZAHL zu ergänzen und in jeder Zeile die Funktion für eine Zufallszahl zwischen 0 und 1 einzufügen. In LibreOffice ist dies zum Beispiel die Funktion = RAND(), in Excel gibt es dafür die Funktion = ZUFALLSZAHL(). Sortiert man schließlich die gefilterte Tabelle nach den Zufallszahlen aufsteigend oder absteigend, dann erhält man vergleichsweise einfach eine Randomisierung der Items und Filler.

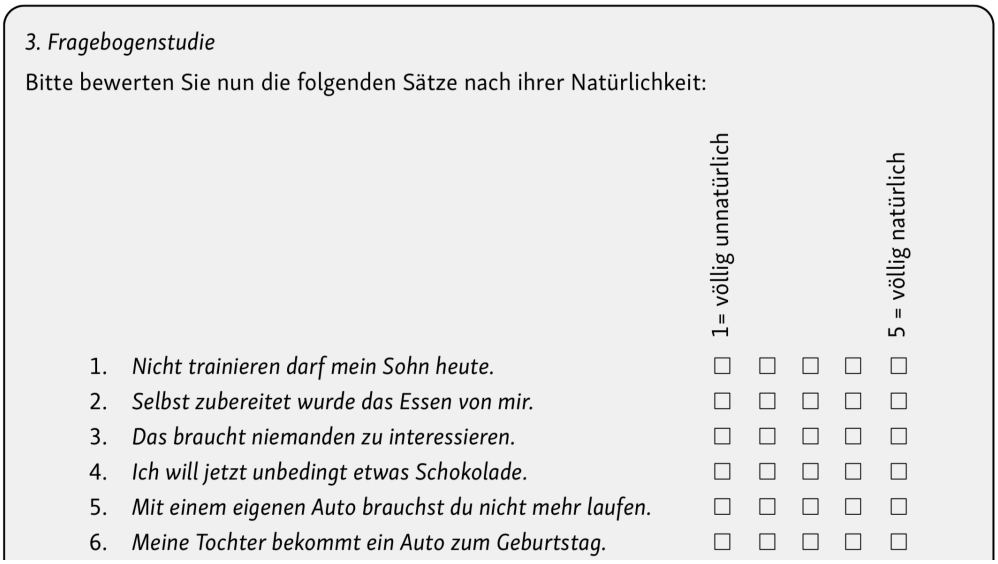
Likert-Skalen

Aufgrund der Verteilung aller Items eines Token-Sets auf verschiedene Fragebögen ist eine vergleichende Bewertung natürlich nicht mehr möglich und wir werden uns eine Alternative überlegen müssen, in der die Items nicht relativ zueinander, sondern absolut bewertet werden können. Eine Möglichkeit wäre zu fragen, ob die VPen die Items als akzeptabel oder als nicht akzeptabel beurteilen. Als Ergebnis erhielten wir damit gewissermaßen zwei Klassen von Sätzen: akzeptable und nicht-akzeptable. In manchen Fällen kann es durchaus sinnvoll sein, auf diese Art zu fragen. In unserem konkreten Fall werden wir aber sicher zwischen verschiedenen **Graden der Akzeptabilität** oder Natürlichkeit unterscheiden wollen. Wir gehen ja davon aus, dass beide Konstruktionen im Prinzip möglich sind, nur eben eine vielleicht etwas natürlicher ist als die andere. Wir sollten also grundsätzlich die Möglichkeit offen lassen, dass es noch schlechtere (also weniger natürliche) und noch bessere (natürlichere) Sätze gibt. Was wir folglich brauchen, ist eine Skala mit mindestens vier Stufen, von »völlig unnatürlich« über mehrere Zwischenstufen bis hin zu »völlig natürlich«. Skalen dieser Art werden als **Ratingskalen** oder auch (nach einem amerikanischen Soziologen) als **Likert-Skalen** bezeichnet. In der Soziologie werden Skalen häufig durchgängig verbal beschrieben (z.B. »völlig unnatürlich, eher unnatürlich, eher natürlich, völlig natürlich«). In der Linguistik werden dagegen nicht selten nur die Enden verbalisiert, da zum einen die Bezeichnung der Mitte mit »neutral« oder »weder noch« im Fall ungerader Skalen (5er- oder 7er-Skala) zu Missverständnissen führen kann und da zum anderen diese Art der Bezeichnung zu einer gleichmäßigeren Verteilung der Bewertungen über die jeweilige Skala führt. Um noch etwas Raum zwischen den Extremen zu lassen, entscheiden wir uns in unserem Experiment für eine 5er-Skala. (Eine 7er-Skala wäre in diesem Fall vielleicht sogar noch besser gewesen, da sie mehr Raum für Differenzierungen zulässt.) Ein Fragebogen, der die Beurteilung der Items beinhaltet, könnte damit in etwa wie in Abbildung 14.7 aussehen.

Zur Vertiefung

Ordinalskalen und Intervallskalen

Der Grad der Natürlichkeit von Sätzen ist in unserem Experiment die abhängige Variable, also das, was wir in Abhängigkeit von der Art der Konstruktion (einfacher vs. erweiterter Infinitiv) messen. Diese abhängige Variable hat bei



3. Fragebogenstudie

Bitte bewerten Sie nun die folgenden Sätze nach ihrer Natürlichkeit:

		1= völlig unnatürlich				5 = völlig natürlich
1.	*Nicht trainieren darf mein Sohn heute.*	☐	☐	☐	☐	☐
2.	*Selbst zubereitet wurde das Essen von mir.*	☐	☐	☐	☐	☐
3.	*Das braucht niemanden zu interessieren.*	☐	☐	☐	☐	☐
4.	*Ich will jetzt unbedingt etwas Schokolade.*	☐	☐	☐	☐	☐
5.	*Mit einem eigenen Auto brauchst du nicht mehr laufen.*	☐	☐	☐	☐	☐
6.	*Meine Tochter bekommt ein Auto zum Geburtstag.*	☐	☐	☐	☐	☐

Abbildung 14.7: Ausschnitt aus einem Fragebogen mit einer 5-Punkt-Likert-Skala

einer 5er-Skala offenbar 5 numerische Ausprägungen (von 1 bis 5). Da diese Ausprägungen linear geordnet sind (5 ist natürlicher als 4, 4 ist natürlicher als 3 …) spricht man hier von Ordinalskalen und **ordinalskalierten Variablen**. Neben ordinalskalierten Variablen kennt man auch **intervallskalierte Variablen**, deren Ausprägungen ebenfalls linear geordnet sind, bei denen aber die Abstände zwischen benachbarten Skalenwerten immer gleich groß sind (Equidistanz). Klassische Beispiele sind für **Intervallskalen** die Zentimeterskala und für **Ordinalskalen** das Notensystem mit den Noten 1 (sehr gut) bis 6 (ungenügend). Dass bei der Zentimeterskala die Abstände zwischen zwei benachbarten Zentimeterangaben gleich sind, sollte klar sein. Bei den Schulnoten ist das aber nicht der Fall, da für 4 (ausreichend) im Allgemeinen mindestens 50% der möglichen Punkte erbracht sein müssen. Damit verteilen sich aber bereits 50% der Punkte auf den Bereich zwischen 4 und 6 und weitere 50% auf das größere Intervall zwischen 1 und 4, was eine Gleichverteilung ausschließt. Das Notensystem wäre nur dann intervallskaliert, wenn für jeden Notensprung genau 20% der möglichen Punkte erforderlich wären. Warum ist das wichtig? Intervallskalierte Variablen verhalten sich aufgrund der identischen Abstände zwischen benachbarten Skalenwerten besonders regelmäßig und erlauben daher in der Statistik Verfahren, die bei ordinalskalierten Variablen nicht (ohne Weiteres) verfügbar sind. Darunter fällt genau genommen sogar die Berechnung von Mittelwerten. In der Praxis ist es allerdings nicht unüblich Ratingskalen

als Intervallskalen zu behandeln, auch wenn man sich nicht völlig sicher sein kann, dass sie tatsächlich intervallskaliert sind. Wir werden diesem Usus hier ebenfalls folgen (auch wenn man die Statistik immer absichern sollte).

Übungsphase

In der Regel ist es sinnvoll, der eigentlichen Studie eine kürzere Übungsphase voranzustellen, in der die Skala erläutert wird und in der sich die VPen an die Skala gewöhnen können, indem sie einige Sätze bewerten, die nichts mit dem Experiment zu tun haben und auch nicht in die Auswertung des Experiments eingehen. Eine solche Übungsphase könnte in etwa wie in Abbildung 14.8 aussehen.

2. Illustration des Fragebogens

Auf der nächsten Seite werden wir Ihnen nach dem untenstehenden Muster Sätze präsentieren. Bitte geben Sie auf einer Skala von 1 (*völlig unnatürlich*) bis 5 (*völlig natürlich*) an, wie natürlich die Sätze für Sie klingen. Kreuzen Sie bitte immer genau eine der Antwortmöglichkeiten an und folgen Sie dabei ganz Ihrer Intuition. Es gibt hier kein Richtig oder Falsch. Die folgenden drei Übungssätze sollen Ihnen einen ersten Eindruck von der Aufgabe vermitteln.

		1 = völlig unnatürlich				5 = völlig natürlich
1.	*Gerhard besucht morgen seine Tante.*	☐	☐	☐	☐	☐
2.	*Gestern in Paris kaufte ich ein Abendkleid mir.*	☐	☐	☐	☐	☐
3.	*Grünes hat er drei Gemüse gegessen.*	☐	☐	☐	☐	☐

Abbildung 14.8: Mögliche Gestaltung der Übungsphase

Persönliche Daten

Neben den Akzeptabilitätsurteilen (den Experimentaldaten) wird man schließlich vielleicht auch noch persönliche Daten der VPen erheben wollen, die relevant für die Hypothese sein könnten. In unserem Experiment sollten wir zum Beispiel absichern können, dass nur Muttersprachler*innen in die Bewertung eingehen, da L2-Lerner*innen möglicherweise stärker normativ geprägt sind und daher den erweiterten Infinitiv bevorzugen. Und unter den Muttersprachler*innen könnte es dialektale oder altersbedingte Unterschiede geben, die man ebenfalls gerne berücksichtigen würde. Derartige persönliche Daten (Alter, Geschlecht etc.) wird man unmittelbar vor oder nach der Beurteilung der Sätze erheben.

Datenschutz

Man sollte sich vor einem Experiment jedoch gut überlegen, ob persönliche Daten wirklich erhoben werden müssen. Und dies aus einem ganz praktischen Grund: Wenn persönliche Daten erhoben werden, dann stellen sich automatisch datenschutzrechtliche Fragen. Was hier im Einzelnen zu beachten ist, wird unter anderem in der Datenschutz-Grundverordnung (DSGVO) der EU geregelt. Sind die erhobenen persönlichen Daten personenbezogene Daten im Sinne der DSGVO (das heißt »[…] Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person […] beziehen. […]«, Artikel 4, Absatz 1), dann sind bei der Durchführung von Experimenten unter anderem folgende Punkte zu beachten: (i) Die Versuchsperson muss in schriftlicher Form über den Zweck und die Art der Erhebung, der Verarbeitung, der Speicherung und die mögliche Weitergabe der persönlichen Daten an Dritte informiert und eine für das Experiment verantwortliche Person benannt werden. (ii) Die Versuchsperson muss mit ihrer Unterschrift in diese Verfahrensweise ausdrücklich einwilligen. (iii) Die Versuchsperson muss in schriftlicher Form über ihre Rechte in Bezug auf die Verarbeitung ihrer persönlichen Daten (Betroffenenrechte) informiert werden und eine Kontaktperson für Rückfragen ist zu benennen. (iv) Die erhobenen Daten dürfen nur in einer bestimmten Form (»pseudonymisiert«) weiterverarbeitet werden und müssen (v) in angemessener Weise vor dem Zugriff Dritter geschützt werden. Schließlich muss die für das Experiment verantwortliche Person die Einhaltung einer DSGVO-konformen Verfahrensweise nachweisen können. Das legt nahe, möglichst auf die Erhebung persönlicher Daten zu verzichten oder diese zumindest zu minimieren.

Tatsächlich entspricht dies auch den Grundsätzen der DSGVO: Daten sollten nur dann erhoben werden, wenn es unbedingt erforderlich ist. Und sie sollten nur so lange gespeichert werden, wie es erforderlich ist. Auch aus diesem Grund sollte man sich vor der Durchführung eines Experiments immer genau überlegen, ob man bestimmte persönliche Daten wirklich zur Auswertung benötigt. Bei vielen Fragestellungen kann es völlig ausreichend sein, alleine Experimentaldaten (also z.B. die Bewertung der Beispielsätze) zu erheben. Für den Fall, dass die Erhebung persönlicher Daten nicht ganz vermieden werden kann, soll die folgende Vertiefungsbox eine erste Orientierung geben, was hier im Einzelnen zu beachten ist. Da diese Thematik im Rahmen einer Einführung aber weder vollständig noch rechtssicher abgehandelt werden kann, sollten Sie Ihre Verfahrensweise immer rechtzeitig mit Ihrer Betreuerin bzw. Ihrem Dozenten absprechen, die wiederum Rücksprache mit der Datenschutzbeauftragten halten können.

Zur Vertiefung

Experimente und Datenschutz (DSGVO)

Bei der Erhebung persönlicher Daten sind eine Reihe datenschutzrechtlicher Fragen zu beachten, die in der Datenschutz-Grundverordnung (DSGVO) ge-

regelt sind. Ein zentraler Punkt ist, dass die Versuchspersonen zunächst in schriftlicher Form über den Zweck der Studie (z.B. wissenschaftliches Experiment im Rahmen eines Seminars), die Art und Dauer der Speicherung ihrer Daten (z.B. anonymisiert und dauerhaft) sowie darüber informiert werden müssen, ob ihre Daten an Dritte weitergegeben werden (und wenn dies der Fall sein sollte, unter welchen Bedingungen). Außerdem ist eine verantwortliche Person anzugeben und das Experiment eindeutig zu kennzeichnen. Die persönlichen Daten der Versuchsperson dürfen nur dann verwendet werden, wenn die Versuchsperson mit ihrer Unterschrift in diese Verfahrensweise einwilligt. Gleichzeitig muss die Versuchsperson erklären, dass sie über ihre Rechte in Bezug auf die Datenverarbeitung (Betroffenenrechte) aufgeklärt wurde. Dazu gehören unter anderem das Recht auf Auskunft, auf Berichtigung und auf Löschung der persönlichen Daten. Diese Aufklärung erfolgt über ein Informationsblatt – das auch eine Kontaktperson für Rückfragen benennt –, das der Versuchsperson beim Experiment ausgehändigt wird (und das sie dann mitnehmen kann und auch mitnehmen soll) oder das – bei einem Online-Experiment – als PDF archiviert werden kann. Wie so eine Einverständniserklärung konkret zu formulieren ist und was alles auf das Informationsblatt muss, hängt zum Teil von Ihrer Verfahrensweise ab. Folglich gibt es nicht die eine perfekte Formulierung. Fragen Sie daher Ihre Betreuerin / Ihren Seminarleiter nach Vorlagen, die Sie adaptieren können, und lassen Sie sich Ihre Anpassungen von Ihrem Betreuer / Ihrer Seminarleiterin absegnen. Da die Einverständniserklärung zum späteren Nachweis archiviert werden muss, ist es ratsam, sie vom Fragebogen getrennt, auf einem eigenen Blatt einzuholen.

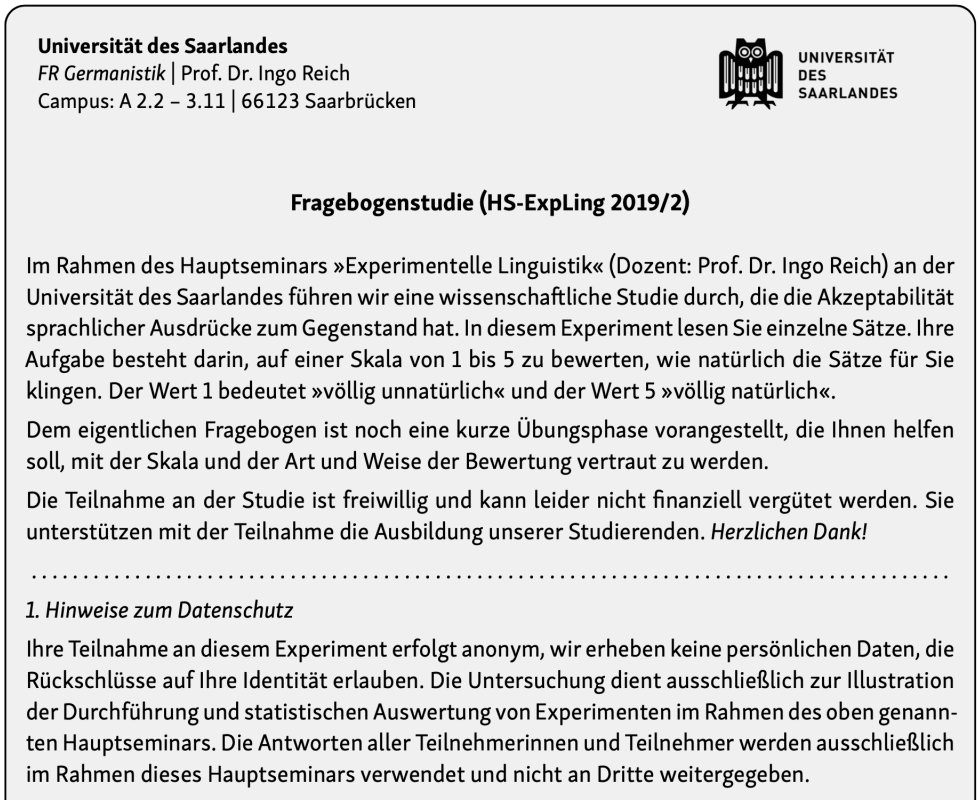
Damit kommen wir zu einem weiteren wichtigen Punkt: Die persönlichen Daten einer Versuchsperson sind gemäß der DSGVO immer so zu archivieren (ob elektronisch auf dem PC oder im Schrank), dass sie »ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können« (Artikel 4, DSGVO). Diese Art der Speicherung nennt man »Pseudonymisierung«. Ist eine Versuchsperson zum Beispiel über den Namen, eine Anschrift (auch Email), eine IP-Adresse (bei Online-Experimenten) oder eine Kontoverbindung (bei vergüteten Experimenten) eindeutig identifizierbar, dann sind diese Daten getrennt von den weiteren personenbezogenen Daten und den Experimentaldaten zu speichern (z.B. auf dem PC in unterschiedlichen, am besten mit verschiedenen Passwörtern geschützten Ordnern). Eine Zuordnung kann, falls erforderlich, indirekt über einen getrennten Schlüssel (die »zusätzliche Information«) erfolgen, indem man zum Beispiel den Versuchspersonen eindeutig eine Nummer zuweist. Dieser

Schlüssel muss getrennt von den personenbezogenen und experimentellen Daten aufbewahrt und vor dem Zugang Dritter geschützt werden.

Auch hier gilt: Fragen Sie immer Ihre Betreuerin bzw. Ihren Dozenten, wie gemäß DSGVO, landesrechtlichen und universitären Richtlinien zu verfahren ist. Im Zweifelsfall wenden Sie sich an die Datenschutzbeauftragte bzw. den Datenschutzbeauftragten Ihrer Universität. Diese bzw. dieser kann auch im Einzelfall beurteilen, ob die von Ihnen erhobenen persönlichen Daten tatsächlich personenbezogene Daten im Sinne der DSGVO sind oder nicht.

Deckblatt

Auf dem Deckblatt des Fragebogens sollte immer über den Zweck der Studie (z.B. wissenschaftliche Studie im Rahmen eines Seminars) und darüber informiert werden, ob die Teilnahme finanziell vergütet wird. Außerdem ist immer eine verantwortliche Person zu benennen und das Experiment eindeutig zu kennzeichnen. Und auch wenn keine persönlichen Daten erhoben werden oder diese nicht unter die DSGVO fallen, sind dennoch entsprechende Hinweise zum Datenschutz zu formulieren. Das könnte wie in Abbildung 14.9 skizziert aussehen.



Universität des Saarlandes
FR Germanistik | Prof. Dr. Ingo Reich
Campus: A 2.2 – 3.11 | 66123 Saarbrücken

UNIVERSITÄT DES SAARLANDES

Fragebogenstudie (HS-ExpLing 2019/2)

Im Rahmen des Hauptseminars »Experimentelle Linguistik« (Dozent: Prof. Dr. Ingo Reich) an der Universität des Saarlandes führen wir eine wissenschaftliche Studie durch, die die Akzeptabilität sprachlicher Ausdrücke zum Gegenstand hat. In diesem Experiment lesen Sie einzelne Sätze. Ihre Aufgabe besteht darin, auf einer Skala von 1 bis 5 zu bewerten, wie natürlich die Sätze für Sie klingen. Der Wert 1 bedeutet »völlig unnatürlich« und der Wert 5 »völlig natürlich«.

Dem eigentlichen Fragebogen ist noch eine kurze Übungsphase vorangestellt, die Ihnen helfen soll, mit der Skala und der Art und Weise der Bewertung vertraut zu werden.

Die Teilnahme an der Studie ist freiwillig und kann leider nicht finanziell vergütet werden. Sie unterstützen mit der Teilnahme die Ausbildung unserer Studierenden. *Herzlichen Dank!*

1. Hinweise zum Datenschutz

Ihre Teilnahme an diesem Experiment erfolgt anonym, wir erheben keine persönlichen Daten, die Rückschlüsse auf Ihre Identität erlauben. Die Untersuchung dient ausschließlich zur Illustration der Durchführung und statistischen Auswertung von Experimenten im Rahmen des oben genannten Hauptseminars. Die Antworten aller Teilnehmerinnen und Teilnehmer werden ausschließlich im Rahmen dieses Hauptseminars verwendet und nicht an Dritte weitergegeben.

Abbildung 14.9: Mögliche Gestaltung des Deckblatts

14.7.2 Eine exemplarische Fragebogenstudie

Rahmendaten

Im Rahmen eines projektorientierten Seminars an der Universität des Saarlandes haben wir die Hypothese (H2) mit einer Fragebogenstudie in der oben skizzierten Art getestet: 12 Token-Sets (mit je 2 Bedingungen) und 36 Filler wurden auf 2 Fragebögen verteilt und in pseudo-randomisierter Abfolge online (über die Software LimeSurvey) von 64 VPen (32 pro Fragebogen) auf einer 5-Punkt-Likert-Skala (mit den Skalenenden 5 = *völlig natürlich* und 1 = *völlig unnatürlich*) bewertet. Drei VPen mussten ausgeschlossen werden, da sie keine Muttersprachler*innen des Deutschen sind. Weitere drei VPen wurden ausgeschlossen, da sie zwei klar ungrammatische Sätze als völlig natürlich bewertet haben. Die verbleibenden 58 VPen gingen in die Auswertung ein.

Online oder klassisch?

Ob man eine Fragebogenstudie besser online oder klassisch vor Ort durchführen sollte, hängt von vielen Faktoren ab. Untersucht man ein dialektales Phänomen, dann wird man zur klassischen Variante tendieren, bei der man VPen einen ausgedruckten Fragebogen in die Hand gibt. Möchte man dagegen dialektale Einflüsse ausschließen und benötigt daher VPen aus dem ganzen Bundesgebiet, dann bietet sich eine Online-Umfrage schon eher an. Über »Crowdsourcing«-Plattformen wie »Clickworker«, »Amazon Mechanical Turk« oder »Prolific« können sehr leicht und sehr schnell VPen nach bestimmten Auswahlkriterien rekrutiert und auch vergleichsweise unkompliziert für die Teilnahme am Experiment entlohnt werden. Ein weiterer, nicht zu unterschätzender Vorteil einer Online-Erhebung ist, dass die Ergebnisse gleich in elektronischer Form vorliegen.

Datenerfassung

Werden die Daten dagegen klassisch auf Papier erhoben, dann wird man die ausgefüllten Fragebögen für die Auswertung in eine Tabelle übertragen müssen. Viele werden dazu Tabellenkalkulationsprogramme wie *Excel*, *Numbers* oder auch *LibreOffice* verwenden. Wie man die Daten am besten in einem solchen Programm erfasst, hängt nicht zuletzt davon ab, wie die erfassten Daten weiterverarbeitet werden sollen. Wertet man die Daten in R oder Jamovi aus, dann sollte die Tabelle so gestaltet werden, dass jede Bewertung mit allen dazugehörigen Informationen (Welches Token-Set? Welche Bedingung? Welche Versuchsperson? ...) eine eigene Zeile bekommt. Dazu legt man Spalteneinträge für alle relevanten Informationen (also Token-Set, Bedingung, Versuchsperson ...) sowie die Bewertungen (Rating) an und füllt die Tabelle Zeile für Zeile aus, vgl. zur Illustration Abbildung 14.10. Hier wurden zunächst alle Bewertungen für das Token-Set 1 erfasst: Im ersten Fragebogen fb1 haben die ersten 32 Probanden die Bedingung 1 (»ohne_zu«) aus dem Token-Set 1 bewertet. Die Bewertung wird in der Spalte RATING eingetragen. Im zweiten Fragebogen fb2 wurde die 2. Bedingung (»mit_zu«) aus dem Token-Set 1 bewertet, und zwar von den Probanden mit der ID 33-64. Die Anzahl der Zeilen in Abbildung 14.10 summiert sich also alleine für die Bewertungen der beiden Items

A	B	C	D	E	F
T-SET	BEDINGUNG	RATING	LISTE	PROBAND-ID	MS
1	ohne_zu	5	fb1	1	1
1	ohne_zu	4	fb1	2	1
1	ohne_zu	4	fb1	3	1
…	…	…	…	…	…
1	mit_zu	4	fb2	33	1
1	mit_zu	5	fb2	34	0
1	mit_zu	5	fb2	35	1
…	…	…	…	…	…
2	mit_zu	5	fb1	1	1
2	mit_zu	4	fb1	2	1
2	mit_zu	5	fb1	3	1
…	…	…	…	…	…
2	ohne_zu	4	fb2	33	1
2	ohne_zu	3	fb2	34	0
2	ohne_zu	4	fb2	35	1
…	…	…	…	…	…

Abbildung 14.10: Zeilenweise Datenerfassung (nach Token-Set)

»ohne_zu« und »mit_zu« aus dem Token-Set 1 auf 64. (Das ist natürlich genau die Anzahl der Versuchspersonen.) Dies wird jetzt wiederholt für alle weiteren Token-Sets (in diesem Fall 11), was $12 \times 64 = 768$ Zeilen ergibt. Und wenn wir auch noch unsere Filler erfassen, kommen nochmal $24 \times 64 = 1536$ Zeilen dazu.

Geht man bei der Erfassung einigermaßen systematisch vor, dann können sehr viele Informationen nach einmaliger Eingabe kopiert werden und der Aufwand hält sich in Grenzen. So wurde zum Beispiel in Abbildung 14.10 auch erfasst, ob die Probanden Muttersprachler*innen sind (MS = 1) oder nicht (MS = 0). Das muss man natürlich nur einmal für die ersten 64 Zeilen (Probanden) erfassen und kann das dann kopieren. Und wenn unser Experiment $n = 2$ Bedingungen hat, dann decken die ersten $n = 2$ Token-Sets bereits alle relevanten Verteilungen der Items auf die Fragebögen ab und dieser Block von $n = 2$ Token-Sets kann (ohne die Ratings, damit sich keine Fehler einschleichen) schematisch kopiert werden.

Statt nach Token-Sets und Bedingungen können die Daten natürlich auch pro Proband erfasst werden, was in der Praxis deutlich einfacher ist, da man jeden Fragebogen nur einmal in die Hand nehmen muss. Dazu wird man die Testsätze (Items und Filler) zunächst nach ihrer Position im Fragebogen aufsteigend durchnummerieren und ihre Position in einer Spalte POS eintragen. Außerdem wird man die Art des Testsatzes (Item vs. Filler) in einer weiteren Spalte TYP festhalten, damit man später die Filler herausfiltern und über die Items rechnen kann. Wer diese Spalten bereits bei der Erstellung des Fragebogens angelegt hat, kann

hier einfach auf diese Tabelle zurückgreifen und muss sie lediglich so oft untereinander kopieren, wie VPen den Fragebogen ausgefüllt haben. Die Datenerfassung würde dann in etwa wie in Abbildung 14.11 aussehen:

A	B	C	D	E	F	G	H
POS	TYP	T-SET	BEDINGUNG	RATING	LISTE	PROBAND-ID	MS
1	Filler	0	0	4	fb1	1	1
2	Filler	0	0	5	fb1	1	1
3	Item	11	ohne_zu	3	fb1	1	1
4	Filler	0	0	5	fb1	1	1
5	Filler	0	0	4	fb1	1	1
6	Filler	0	0	3	fb1	1	1
7	Item	3	mit_zu	5	fb1	1	1
8	Filler	0	0	4	fb1	1	1
9	Item	8	ohne_zu	4	fb1	1	1
10	Filler	0	0	4	fb1	1	1
11	Filler	0	0	2	fb1	1	1
…	…	…	…	…	…	…	…

Abbildung 14.11: Datenerfassung nach Proband und Position

Hat ein Proband einen Testsatz nicht bewertet, dann wird die fehlende Bewertung in der fraglichen Zelle in der Tabelle mit NA gekennzeichnet. Auf diese Weise kann man systematisch fehlende Bewertungen (fehlende Angaben) identifizieren und gegebenenfalls bei Auswertungen in R, Jamovi, Excel & Co. ausschließen.

Zur Vertiefung

Tabellen in R und Jamovi

Ein (anfänglicher) Nachteil von R besteht darin, dass weder die Erstellung von Tabellen in R noch deren Import nach R selbsterklärend ist. Hat man erst einmal verstanden, wie es geht, kann man bei der Auswertung und der Erstellung von Grafiken alle Vorteile und die ganze Flexibilität von R nutzen. Möchte man diese Probleme aber zunächst vermeiden und sich auf andere Dinge konzentrieren, dann bietet sich hier einmal mehr die auf R basierende Software Jamovi an. In Jamovi können wie in Excel Tabellen angelegt werden, Jamovi stellt aber auch auf R basierende Statistiken und Grafiken zur Verfügung.

Mittelwerte

Damit sind wir bei der vielleicht wichtigsten Frage angelangt: Wie wertet man die erhobenen Daten am besten und am saubersten aus? Der erste Gedanke, der den meisten hier vermutlich durch den Kopf schießt, ist: **Mittelwerte** bilden! Summieren wir also alle Bewertungen (Ratings) für die Bedingung »ohne_zu« auf und teilen sie dann durch die Gesamtanzahl der Ratings für diese Bedingung. Auf diese Weise erhalten wir das arithmetische Mittel für die Bedingung »ohne_zu«. Für die Bedingung »mit_zu« gehen wir analog vor. In unserer Fragebogenstudie erhalten

wir auf diese Weise (gerundet auf zwei Stellen nach dem Komma) einen Mittelwert von 3,97 für den einfachen Infinitiv (»ohne_zu«) und 4,14 für den erweiterten Infinitiv (»mit_zu«). Der erweiterte Infinitiv wird in unserer Untersuchung also im Schnitt leicht besser bewertet als der einfache Infinitiv.

Streuung

Mittelwerte alleine sind unter Umständen jedoch nicht sehr aussagekräftig. Machen wir auch hier ein Beispiel: Angenommen, von 8 VPen beurteilen 4 einen Satz S1 mit dem Wert 2 und die anderen 4 mit dem Wert 3. Dann liegt der Mittelwert (MW) bei 2,5. Nehmen wir weiter an, dass von diesen 8 VPen 2 einen anderen Satz S2 mit dem Wert 1, 4 mit dem Wert 2 und 2 mit dem Wert 5 beurteilen. Auch in diesem zweiten Fall liegt der Mittelwert bei 2,5, vgl. Tabelle 14.9. Aber würden

Satz	VP1	VP2	VP3	VP4	VP5	VP6	VP7	VP8	MW
S1	2	2	2	2	3	3	3	3	2,5
S2	1	1	2	2	2	2	5	5	2,5

Tabelle 14.9: gleiche Mittelwerte bei unterschiedlichen Streuungen

wir daraus schließen wollen, dass die VPen die beiden Sätze »gleich gut« finden? Vermutlich eher nicht. Denn im ersten Fall sind sich die VPen in ihrer Bewertung relativ einig, sie liegen alle nahe beieinander. Im zweiten Fall dagegen liegen sie zum Teil sehr weit auseinander: Zwei finden den Satz völlig natürlich (5), zwei völlig unnatürlich (1). Die VPen sind sich in diesem Fall offenbar eher uneinig in ihren Bewertungen. Mit anderen Worten: Ihre Urteile variieren sehr stark, ihre Bewertungen von Satz S2 »streuen« sehr stark um den Mittelwert.

Standardabweichung

Die Streuung um den Mittelwert wird in der deskriptiven Statistik über den Begriff der Standardabweichung und den der Varianz erfasst. Die Idee ist relativ einfach: Zuerst berechnet man den Mittelwert $\bar{x}$. Dann berechnet man für jede Bewertung x_i den Abstand zum Mittelwert (also $\bar{x} - x_i$). Da die Summe aller dieser Abstände 0 ergäbe (manche sind negativ, andere positiv), werden die Differenzen zunächst quadriert, das heißt, man bildet für jede Bewertung $(\bar{x} - x_i)^2$. Die quadrierten Differenzen werden dann aufsummiert und bei $1 \leq i \leq n$ Datenpunkten durch n geteilt. Mit diesem letzten Schritt wird der Mittelwert der quadrierten Abweichungen gebildet, die **Varianz** s^2 in unserem Experiment. (Möchte man die Varianz σ^2 der Grundgesamtheit schätzen, muss man durch $n-1$ teilen. -1 ist hier eine Art Korrekturfaktor. Auf diese Weise wird im Allgemeinen die Varianz in Statistikprogrammen berechnet.) Um schließlich zu berechnen, wie stark die Urteile im Mittel vom Mittelwert abweichen, muss man die Quadratur wieder umkehren, also die Quadratwurzel ziehen. Das Resultat bezeichnet man als **Standardabweichung** s. In unserem konstruierten Beispiel würde die Standardabweichung

$$s^2 = \sum_{1 \leq i \leq n} \frac{(\bar{x} - x_i)^2}{n}$$

$$s = \sqrt{s^2}$$

bei Satz S1 gerundet 0.54 betragen, bei Satz S2 dagegen immerhin 1.60. Das gibt uns die wichtige Information, dass der Mittelwert bei Satz S2 die Daten offenbar weniger gut repräsentiert als bei Satz S1. Aus diesem Grund ist es von zentraler Bedeutung, neben den berechneten Mittelwerten $\bar{x}$ immer auch die Standardabweichung s anzugeben. Bei unserer Studie ergeben sich für die beiden fraglichen Bedingungen konkret die in Tabelle 14.10 angegebenen Werte.

Bedingung	Mittelwert $\bar{x}$	Standardabweichung s
brauchen plus *zu*-Infinitiv	4.14	1.05
brauchen plus einfacher Infinitiv	3.97	1.15

Tabelle 14.10: Mittelwerte und Standardabweichungen

Berücksichtigt man neben dem Mittelwert auch die Standardabweichung, dann sieht es so aus, als ob die Konstruktion mit einfachem Infinitiv tatsächlich in beiden Parametern etwas »schlechter« abschneiden würde als die mit dem erweiterten Infinitiv. Allerdings liegen die Werte doch recht nahe beieinander, so dass man sicher nicht ohne Weiteres behaupten möchte, dass der einfache Infinitiv klar schlechter beurteilt wird. Denn etwas Varianz ist immer in den Daten und diese Varianz könnte durchaus zufällig sein. Ob das hier tatsächlich der Fall ist, können nur statistische Tests zeigen, auf die wir hier aber nur in den Vertiefungskästen eingehen können. In einer deskriptiven Statistik würde man lediglich die in der Tabelle angegebenen Werte berichten und diskutieren.

Zur Vertiefung

Verteilung der Mittelwertdifferenzen

Wie kann man überprüfen, ob sich die Differenz zweier Mittelwerte noch im Rahmen des Zufälligen bewegt oder bereits unsere Hypothese bestätigt? Aus Platzgründen können wir hier nicht auf Details eingehen, aber zumindest der zentrale Gedanke ist relativ einfach vermittelt: Nehmen wir an, wir würden zwei Sätze S1 und S2 nicht nur in einem, sondern in 1.000 oder 10.000 oder (theoretisch) sogar in beliebig vielen Experimenten beurteilen lassen. In jedem Experiment würde man für die Sätze S1 und S2 unterschiedliche Mittelwerte bekommen und auch die Differenzen dieser Mittelwerte würden von Experiment zu Experiment variieren. Die Differenzen der Mittelwerte bilden folglich eine Verteilung. Wenn unsere Stichproben groß genug sind (über 30 Datenpunkte), dann kann man mit dem Zentralen Grenzwertsatz der Statistik annehmen, dass sich diese Verteilung einer Normalverteilung annähert. Eine Normalverteilung ist dabei durch ihren Erwartungswert und ihre Stan-

dardabweichung eindeutig festgelegt. Was sind also der Erwartungswert und die Standardabweichung der Verteilung der Mittelwertdifferenzen? Da wir die Nullhypothese testen, die ja besagt, dass es keinen (signifikanten) Unterschied zwischen den beiden Mittelwerten gibt, erwarten wir eine Differenz von 0. Bleibt die Standardabweichung. Hier wird es etwas komplexer. Die Standardabweichung der Mittelwertdifferenzen kann (ebenfalls nach dem Zentralen Grenzwertsatz) über den **Standardfehler SE** auf der Basis der in unserem (einzigen) Experiment beobachteten Standardabweichungen und den Größen unserer Stichproben geschätzt werden. Wie das genau zu berechnen ist, können wir hier weitgehend ignorieren. Der zentrale Punkt ist, dass wir auf der Basis der von uns beobachteten Daten diese Verteilung der Mittelwertdifferenzen schätzen können. Und da diese Verteilung eine Normalverteilung ist, können wir die alles entscheidende Frage stellen: Wie groß ist unter der Nullhypothese die Wahrscheinlichkeit, dass eine Mittelwertdifferenz von $|\bar{x}_{S1} - \bar{x}_{S2}|$ oder größer auftritt? Das ist (bei einer gerichteten Hypothese) genau die Fläche, die diese Mittelwertdifferenz auf der x-Achse nach rechts unter der geschätzten Verteilung der Mittelwertdifferenzen abschneidet. Und wie berechnet man diese Fläche? Da Normalverteilungen die sehr schöne Eigenschaft haben, dass die Standardabweichungen immer exakt denselben Prozentsatz an Fläche umschließen, ist es sinnvoll, die Mittelwertdifferenz zunächst als ein Vielfaches der Standardabweichung (hier des Standardfehlers) darzustellen $|\bar{x}_{S1} - \bar{x}_{S2}| = z * SE$. Mit der Funktion `qnorm(0.05, lower.tail = FALSE)` kann man in *R* bestimmen, dass ein Wert von 1,645 ziemlich genau 5% an Fläche nach rechts abschneidet. Wenn also der oben berechnete z-Wert über dem Wert 1,645 liegt, dann werden wir das Ergebnis als signifikant betrachten.

Zur Vertiefung

Der *t*-Test

In der letzten Vertiefungsbox haben wir aus Darstellungsgründen so getan, als ob man bei der Bestimmung von Signifikanzen immer mit Normalverteilungen rechnen könnte. Das ist so aber leider nicht richtig: Je weniger Datenpunkte man in einem Experiment hat, desto problematischer wird die Schätzung mit Normalverteilungen. Daher wird in der Regel nicht mit der Familie der Normalverteilungen, sondern mit der Familie der t-Verteilungen gerechnet, die zum einen die Anzahl der Datenpunkte berücksichtigen, sich zum anderen aber immer mehr der Normalverteilung annähern, je größer die Anzahl der Datenpunkte wird. Illustrieren wir das an unserem kleinen Beispiel. Satz S1 wurde von 8 VPen beurteilt, ebenso Satz S2. Damit hätten wir eigentlich 16 Datenpunkte. Die t-Verteilung hängt nun aber nicht einfach von der Gesamt-

heit der Datenpunkte ab, sondern von den Freiheitsgraden. Die Berechnung der Freiheitsgrade kommt wie folgt zustande: Wurde S1 von n VPen beurteilt, dann ist $(n - 1)$ der Freiheitsgrad für S1. Wurde S2 von m VPen beurteilt, dann ist $(m - 1)$ der Freiheitsgrad für S2. Beim Vergleich der Mittelwerte von S1 und S2 ergeben sich die Freiheitsgrade der t-Verteilung als Summe der Freiheitsgrade der beiden Stichproben der Größen n und m, also $(n + m - 2)$. Da in unserem Fall $n = m = 8$ gilt, erhalten wir einen Freiheitsgrad von 14 für unsere t-Funktion. Wie bereits bei der Normalverteilung können wir auch hier die Mittelwertdifferenz $|\bar{x}_{S1} - \bar{x}_{S2}|$ als ein Vielfaches des Standardfehlers darstellen: $|\bar{x}_{S1} - \bar{x}_{S2}| = t * SE$. Mit der Funktion `qt(0.05, 14, lower.tail = FALSE)` kann man in R jetzt bestimmen, dass ein t-Wert von 1,762 bei der t-Funktion mit 14 Freiheitsgraden ziemlich genau 5% an Fläche nach rechts abschneidet. Liegt der t-Wert der Mittelwertdifferenz über diesem Wert von 1,762, dann können wir die Nullhypothese nach Annahme verwerfen.

Zur Vertiefung

Die Funktion t.test() in *R*

Kommen wir damit zu der Frage zurück, ob die in unserem Experiment beobachtete und in Tabelle 14.10 dargestellte Mittelwertdifferenz höchstwahrscheinlich auf Zufall beruht oder ob wir unsere gerichtete Alternativhypothese (H2) annehmen können, dass der erweiterte Infinitiv als natürlicher beurteilt wird. In einem Programm wie R können wir diese Frage ganz einfach beantworten, indem wir eine Funktion benutzen, die uns die in den letzten beiden Boxen angedeuteten Berechnungen abnimmt. Wenn die Daten des Experiments in einer Tabelle wie in Abbildung 14.10 erfasst und nach R importiert wurden, dann führt die Funktion `t.test(RATING ~ BEDINGUNG, var.equal = TRUE, alternative = "less")` einen t-Test durch, das heißt, R bestimmt die Freiheitsgrade der t-Funktion, den t-Wert der Mittelwertdifferenz und den entsprechenden p-Wert, also die Fläche, die der t-Wert nach rechts (oder links) unter der t-Funktion abschneidet. Da wir 696 Bewertungen erhoben haben, kommt R auf 694 Freiheitsgrade, einen (negativen) t-Wert von −2,03 und einen p-Wert von 0,021. Damit liegt der p-Wert unter dem Signifikanzniveau von 5% und wir könnten unsere Alternativhypothese als bestätigt betrachten. Es gibt hier nur ein Problem. Der t-Test ist an bestimmte Voraussetzungen gekoppelt: Die abhängige Variable muss intervallskaliert sein und die Daten (idealerweise) normalverteilt. Dass eine 5er-Skala nicht notwendigerweise intervallskaliert ist, haben wir bereits gesagt. Und dass die Daten nicht normalverteilt sind, wird sich in Abbildung 14.17 andeuten. Wir werden daher das Ergebnis mit einem Test für ordinale Daten absichern müssen.

Zur Vertiefung

ANOVA und weitere Testverfahren

Aus Gründen der Darstellung sind wir hier nur auf den Fall eingegangen, dass man zwei Mittelwerte miteinander vergleichen möchte. Tatsächlich tritt aber nicht selten der Fall ein, dass ein Faktor mehr als zwei Bedingungen aufweist oder dass in einem Experiment mehr als zwei Faktoren zu berücksichtigen sind. Da man aber mehrere Mittelwerte mit einem *t*-Test nicht ohne Weiteres paarweise vergleichen darf (ohne das Signifikanzniveau anzugleichen), wird man in solchen Fällen auf ein anderes Verfahren zurückgreifen: die **Varianzanalyse** oder kurz **ANOVA**. Hier müssen wir leider auf die einschlägige Literatur verweisen (Gries 2008, Meindl 2011). Ebenso nur angedeutet sei an dieser Stelle, dass sich in der Linguistik inzwischen die Arbeit mit Gemischt Linearen Modellen (Linear Mixed Effects Models) mehr oder weniger als Standard etabliert hat. Einen ersten Überblick bietet hier z.B. Winter (2013).

Median

Bei der Diskussion der Variablentypen (ordinal- vs. intervallskaliert) haben wir bereits angedeutet, dass bei einer Likert-Skala die Berechnung des Mittelwerts nicht ganz unproblematisch ist, da nicht gesichert ist, dass benachbarte Skalenwerte immer gleich weit voneinander entfernt sind (Equidistanz). Ein weiteres Problem ergibt sich, wenn die Verteilung der Werte »schief« ist oder wenn sich Ausreißer unter die Daten mischen. Machen wir dazu ein extremes Beispiel: Angenommen zwei Sätze S1 und S2 werden auf einer Skala von 1 bis 100 von 10 VPen bewertet. 9 bewerten Satz S1 mit dem Wert 1, aber nur einer mit dem Wert 91. Satz S2 wird dagegen von allen 10 VPen mit dem Wert 10 bedacht, vgl. Tabelle 14.11.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	**MW**	**Median**
S1	1	1	1	1	1	1	1	1	1	91	**10**	**1**
S2	10	10	10	10	10	10	10	10	10	10	**10**	**10**

Tabelle 14.11: Mittelwert und Median bei problematischen Verteilungen

Man sieht hier sofort, dass in beiden Fällen der Mittelwert 10 ist. Würde man aber die VP P10 ausschließen, dann wäre S1 eine klare 1 (und S2 bliebe eine 10). Damit ist aber klar, dass der von P10 verursachte Ausreißer den Mittelwert extrem beeinflusst. Aber natürlich können wir die VP P10 nicht nur deswegen ausschließen, weil ihre Bewertung nicht ins Bild passt. Wie können wir also den Einfluss dieser einzelnen Bewertung auf das Gesamtbild möglichst gering halten? Die Antwort ist: Wir suchen uns eine bessere Mitte und diese Mitte ist der **Median**. Um den Median zu bestimmen, werden zuerst die numerischen Daten ihrer Größe nach

geordnet. Bei einer ungeraden Anzahl ($2n+1$) von Datenpunkten zählt man ($n+1$) Datenpunkte in aufsteigender Reihenfolge ab. Der Wert des ($n+1$)-ten Datenpunktes ist dann der Median. Bei einer geraden Anzahl ($2n$) geht man im Prinzip gleich vor. Da es aber keine natürliche Mitte gibt, nimmt man als Median den Mittelwert der Werte des n-ten und des ($n+1$)-ten Datenpunktes. In unserem Beispiel wären dies die Bewertungen von P5 und P6, die im Mittel bei S1 einen Wert von 1 und bei S2 einen Wert von 10 ergeben.

Quartile

Wie beim Mittelwert benötigt man auch beim Median zusätzlich immer noch ein geeignetes Streuungsmaß. Wir haben gerade gesehen, dass unter dem Median 50% der Datenpunkte liegen und über dem Median ebenfalls 50% der Datenpunkte. Diese beiden Hälften kann man jetzt über das erste **Quartil** (25%) und das dritte Quartil (75%) nochmals in zwei gleich große Hälften von je 25% teilen. Damit können wir einen symmetrischen Bereich von 50% der Daten definieren, die unmittelbar unterhalb und oberhalb des Medians liegen. Diesen Bereich nennt man den **Interquartilsabstand** (IQR). Mit dem IQR kann man also angeben, in welchem Bereich sich die mittleren 50% der Daten bewegen. Und diesen Bereich kann man in einem **Boxplot** visualisieren, wie in Abbildung 14.12 exemplarisch dargestellt. In einem solchen Boxplot sind die Mediane durch die dicken Querlinien markiert

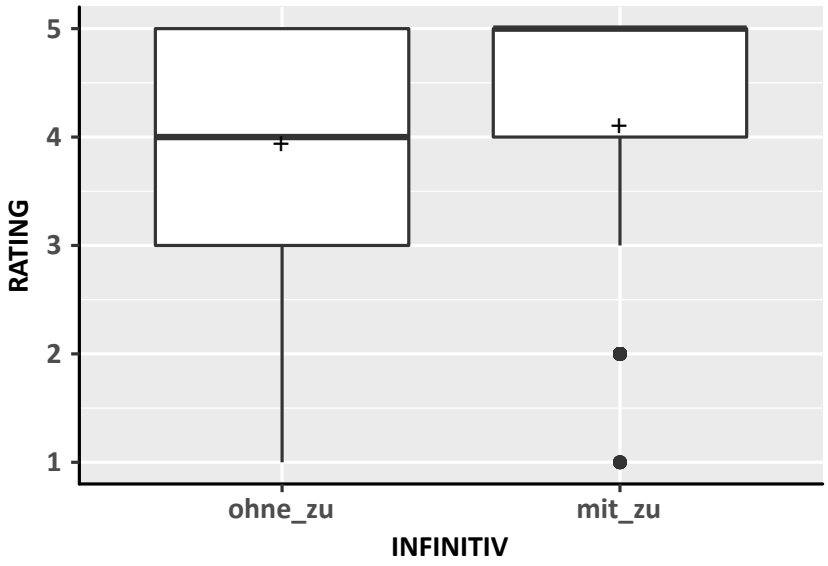


Abbildung 14.12: Boxplot zum *brauchen*-Experiment

und man sieht in diesem konkreten Fall, dass der Median beim einfachen Infinitiv bei 4 und beim erweiterten Infinitiv bei 5 liegt. Außerdem zeigen die hell unterlegten Boxen, dass sich beim einfachen Infinitiv 50% der mittleren Daten zwischen

3 und 5 bewegen. Da der Wert 5 der höchste Wert auf der Skala ist, sind dies tatsächlich sogar 75% der Daten. Beim erweiterten Infinitiv dagegen konzentrieren sich die mittleren 50% (tatsächlich sind dies aus demselben Grund ebenfalls 75%) auf den Bereich zwischen den Werten 4 und 5. Die vertikalen Striche nennt man **Whiskers**. Sie deuten an, in welchem Bereich noch ein weiterer Teil der Daten liegt (wieviel genau ist uneinheitlich definiert). Die einzelnen Punkte (hier beim erweiterten Infinitiv) charakterisieren schließlich einzelne Ausreißer.

Interpretation des Boxplots

An dem Boxplot in Abbildung 14.12 kann man jetzt mehrere Dinge beobachten. Erstens zeigt sich, dass beide Konstruktionen ziemlich gut bewertet werden. Der erweiterte Infinitiv sogar so gut, dass der Median mit dem Wert 5 an die Decke stößt. Zweitens kann man sowohl am IQR als auch an den Whiskers ablesen, dass die Bewertungen beim einfachen Infinitiv doch deutlich stärker streuen. Das bestätigt im Wesentlichen den Eindruck, den wir bereits bei der Mittelwertberechnung gewonnen haben. Aber trotzdem ist auch hier Vorsicht geboten. Denn ein Blick in die Daten zeigt uns, dass beim erweiterten Infinitiv genau 50,2% der Bewertungen den Wert 5 aufweisen. Wenn also nur eine Handvoll von Urteilen anders ausgefallen wäre, dann würden wir in dieser Bedingung einen Median von 4 beobachten und die Boxplots der beiden Bedingungen wären (bis auf die Whiskers und die Ausreißer) im Wesentlichen identisch. Daher ist es durchaus sinnvoll, in den Boxplots auch noch die Mittelwerte zu markieren. In einer deskriptiven Statistik würde man jetzt diese Mitte(l)werte und Streuungen berichten und graphisch veranschaulichen. Ob die beobachteten Differenzen aber den Schluss zulassen, dass die Unterschiede nicht auf Zufall beruhen, sondern als Evidenz für unsere Alternativhypothese (H2) gewertet werden können, muss letztlich auch hier ein geeigneter statistischer Test zeigen (vgl. dazu die folgende Vertiefungsbox).

Zur Vertiefung

Der *U*-Test

Der Boxplot in Abbildung 14.12 lässt bereits vermuten, dass unsere Daten nicht normalverteilt sind (was aber noch mit dem Shapiro-Wilk-Test zu überprüfen wäre). In diesem Fall ist der *t*-Test aber kein geeigneter Test (und selbst im positiven Fall ist er es nur bedingt, wenn die abhängige Variable nicht intervallskaliert ist) und man benötigt ein statistisches Verfahren, das auf der Grundlage von ordinalen Werten arbeitet. Ein solches Verfahren ist der Wilcoxon-Mann-Whitney-Test, auch **U-Test** genannt. Der U-Test ist ein Rangsummen-Test, das heißt, für die Berechnung der relevanten statistischen Werte, der U-Werte, werden alle Bewertungen in numerisch aufsteigender Reihenfolge geordnet und für jeden einzelnen Wert wird seine Position, sein **Rang**, festgehalten. Im Wesentlichen kann man dabei sagen, dass der Rang umso höher ist, je höher der Wert ist. Und je mehr höhere Werte eine Bedingung aufweist, umso größer

wird am Ende die Summe aller Ränge dieser Bedingung werden. Aus diesen Rangsummen werden dann (hier: zwei) *U*-Werte berechnet, die die Größen der Stichproben berücksichtigen. Über den kleinsten dieser *U*-Werte (den *W*-Wert) lässt sich dann wiederum eine Wahrscheinlichkeit, ein *p*-Wert berechnen. (Details findet man in der Literatur, z.B. in Gries 2008, 224ff.) Auch solche Berechnungen können wir zum Glück Programmen wie *R* überlassen. Für unsere gerichtete Hypothese würden wir in *R* `wilcox.test(RATING ~ BEDINGUNG, alternative = "less")` eingeben und würden einen *W*-Wert von 55.777 sowie einen *p*-Wert von 0,027 erhalten. Der *U*-Test bestätigt damit im Wesentlichen das Bild, das bereits der *t*-Test gezeichnet hat.

14.8 Interpretation der Ergebnisse

Zahlen interpretieren

In den beiden letzten Abschnitten haben wir eine Korpus- und eine Fragebogenstudie zur modalen Verwendung von *brauchen* durchgeführt, die verfügbaren bzw. erhobenen Daten statistisch ausgewertet und die Resultate der Auswertungen berichtet. Damit sind wir aber noch lange nicht am Ende unseres kleinen Projekts angelangt. Denn natürlich wollen und müssen wir diese Ergebnisse noch inhaltlich interpretieren: Können wir aus den (numerischen) Resultaten der beiden Studien Schlüsse ziehen, wie das Phänomen der modalen Verwendung von *brauchen* grammatisch und pragmatisch zu beschreiben ist? Wenn ja, wie könnte eine grammatische Modellierung aussehen, die im Einklang mit den Ergebnissen der beiden Studien steht? Widersprechen unsere Ergebnisse möglicherweise Aussagen in der einschlägigen Literatur? Und wenn ja, worauf sind diese Widersprüche letztlich zurückzuführen? Vielleicht auf das Design der Experimente?

Einordnung der brauchen-Resultate

Hier ist nicht der richtige Ort, um die Daten der oben vorgestellten Korpus- und Fragebogenstudien im Einzelnen zu diskutieren, aber wir sollten deren Resultate doch zumindest grob einordnen. Ausgehend von Ausführungen in der einschlägigen Literatur haben wir im Wesentlichen zwei Hypothesen überprüft: (H1) Die Konstruktion mit dem einfachen Infinitiv ist ein Phänomen der gesprochenen Sprache (Korpus). (H2) In der geschriebenen Sprache ist die Verbindung mit dem einfachen Infinitiv in dem Sinne die markierte Variante, als Muttersprachler*innen des Deutschen den erweiterten Infinitiv (in schriftlichen Kontexten) als natürlicher bewerten (Fragebogen).

Syntaktische Modellierung

Beginnen wir mit dem Offensichtlichen, aber nicht Selbstverständlichen. Bei der Untersuchung der Hypothese (H2) haben wir zunächst einmal festgestellt, dass (selbst in diesem schriftlichen Kontext) beide Varianten als sehr natürlich betrachtet werden und sie daher tatsächlich beide als grammatische Varianten aufzufassen sind. Für die syntaktische Beschreibung bedeutet dies, dass im Lexi-

koneintrag von *brauchen* (in modaler Verwendung) beide Realisierungsformen des Komplements (also erweiterter und einfacher Infinitiv) als grundsätzlich möglich zugelassen werden müssen.

Präferenz für eine Variante

Dass beide Varianten grundsätzlich grammatisch sind, wird in der Literatur und haben wir in unserem Experiment genau genommen vorausgesetzt, konnten es jetzt aber gewissermaßen nebenbei nochmal bestätigen. Eigentlich abgezielt hat unsere Hypothese (H2) darauf, dass die Muttersprachler*innen des Deutschen eine Präferenz für eine der beiden Varianten haben, und zwar für den erweiterten Infinitiv. Diese Hypothese wurde zum einen über Grammatiken nahegelegt, zum anderen über Aussagen in der Literatur, dass der *zu*-Infinitiv historisch zuerst zu beobachten ist und sich der einfache Infinitiv (in Analogie zu den »echten« Modalverben) erst später entwickelt hat. Unsere Fragebogenstudie hat diese Hypothese tendenziell bestätigt: Der einfache Infinitiv wird tatsächlich etwas schlechter bewertet und weist eine größere Streuung auf. Und diese Differenz hat sich in statistischen Tests auch als signifikant erwiesen. Man sollte aber nicht verschweigen, dass beide Tests alleine deswegen zu einem signifikanten Ergebnis geführt haben, weil die Hypothese gerichtet formuliert wurde.

Natürlichkeit und Normativität

Selbst in eher schriftlichen Kontexten liegen die beiden Varianten also sehr nahe beieinander. Das ist insofern überraschend, als wir in der Korpusstudie in schriftlichen Korpora quantitativ eine deutliche Präferenz für den erweiterten Infinitiv festgestellt haben. Hier scheinen die beiden Studien auseinanderzulaufen. Wie kann das sein? Offenbar können wir hier von den quantitativen Daten nicht auf die qualitativen Einschätzungen schließen. Ein Grund hierfür könnte sein, dass die Texte in den untersuchten Korpora stark normativ geprägt sind. (Das DeReKo besteht wie gesagt vorwiegend aus Zeitungstexten.) Man kann sich also durchaus vorstellen, dass trotz einer sehr ähnlichen Einschätzung der Grammatikalität der beiden Varianten sich Schreiber*innen aus normativen Gründen dennoch häufiger für eine der Varianten entscheiden. Diese Überlegung könnte Grundlage für eine Folgestudie sein.

Mündlichkeitsphänomen

Was die Korpusstudie aber deutlich zeigt, ist, dass sich Sprecher*innen in der gesprochenen Sprache und damit in informellen Kontexten deutlich häufiger für den einfachen Infinitiv entscheiden, dass sich die Verhältnisse sogar weitgehend umkehren. Damit kann man die Verbindung mit dem einfachen Infinitiv sicherlich als eine Art Mündlichkeitsphänomen bezeichnen, aber eben nicht in dem Sinne, dass diese Variante ausschließlich im Mündlichen Verwendung finden würde. Auch hier könnte man sich interessante Folgeuntersuchungen vorstellen. So ist zum Beispiel noch nicht ganz klar, ob diese Umkehrung der Präferenzen tatsächlich eine Frage des Kanals (mündlich oder schriftlich) ist oder eine Frage des Registers (formell vs. informell). Das könnte man vielleicht testen, indem man sich in-

formelle schriftliche Kommunikation anschaut, wie man sie zum Beispiel in Kurznachrichten oder bei Facebook, Twitter & Co. vermutet.

14.9 Präsentation der Ergebnisse

Form der Präsentation

Diese knappe Diskussion soll lediglich andeuten, wie eine inhaltliche Einordnung der empirischen Daten aussehen könnte. Ist man sich darüber klar geworden, wie man diese Resultate zu interpretieren hat, dann fehlt eigentlich nur noch Eines: das Präsentieren der Ergebnisse. Im wissenschaftlichen Kontext erfolgt die Präsentation typischerweise zunächst über einen Vortrag auf einer Konferenz, in einem Seminar vielleicht in einem Referat. Am Ende steht (sollte stehen) aber meist die schriftliche Ausarbeitung in einem wissenschaftlichen Aufsatz oder (im studentischen Kontext) in einer Hausarbeit. Auch hier können wir nicht in allen Details erläutern, wie man Hausarbeiten schreibt, auch hier müssen wir auf die Literatur verweisen (vgl. z.B. Rothstein 2011). Dennoch wollen wir hier einige wenige Punkte ansprechen, die häufig falsch gemacht werden.

Der formale Rahmen

Wie jede Hausarbeit benötigt natürlich auch eine empirische Arbeit mindestens ein Deckblatt, ein Inhaltsverzeichnis, eine Bibliographie und eine Selbständigkeitserklärung. Wie das im Einzelnen auszusehen hat, hängt nicht zuletzt von den Präferenzen der Betreuerin bzw. des Dozenten ab. Jeder Betreuer bzw. jede Dozentin hat meist eigene Vorstellungen davon, wie groß der Korrekturrand sein muss, wie viel Zeilenabstand man einstellen sollte und welcher Zeichensatz in welcher Schriftgröße am geeignetsten ist. Daher ist es immer ratsam, diese Formalia mit der Betreuerin bzw. dem Dozenten rechtzeitig abzusprechen.

Eine kanonische Struktur

Inhaltlich folgen empirische Arbeiten meist einem kanonischen Aufbau, der im Wesentlichen der in Abschnitt 14.1 vorgestellten Struktur entspricht:

1. **Einführung:** In der Einführung sollten Gegenstand und Ziel der Arbeit kurz umrissen und in der ein oder anderen Form motiviert werden.
2. **Theoretischer Hintergrund:** Im darauffolgenden Kapitel ist die für die Fragestellung relevante Literatur *mit Blick auf die Fragestellung* darzustellen. Was wurde zu dem Thema, das ich untersuchen möchte, bereits in der Literatur gesagt? Inwieweit ist das relevant für meine Fragestellung? Kann ich daraus eine Hypothese ableiten?
3. **Empirische Untersuchung:** Dann folgt der empirische Teil der Arbeit. Hier sind alle relevanten Aspekte der Studie darzustellen, von der Hypothesenbildung über die Methode bis hin zur statistischen Auswertung:

   Hypothese und Methode: Zunächst muss klar werden, welche Hypothese in der folgenden Untersuchung überhaupt getestet werden soll. Dann sollte kurz

thematisiert werden, warum die gewählte Methode (Fragebogen- oder Korpusstudie) gerade für diese Fragestellung geeignet erscheint.

Methode und Durchführung: Als Nächstes sind dann die zentralen Parameter der empirischen Untersuchung darzustellen: Welche Faktoren (unabhängige Variablen) gehen mit welchen Ausprägungen in das Design ein? Und was wird genau gemessen (abhängige Variable)? Bei einer *Fragebogenstudie* sollte ein komplettes Token-Set (abstrakt und konkret) präsentiert werden, das das Design der Untersuchung illustriert. Darüber hinaus sind zentrale Informationen zum Aufbau (Übungsphase, Art und Anzahl von Items und Fillern, ihr quantitatives Verhältnis, Randomisierung) und zur Anzahl der Fragebögen zu geben sowie zur Stichprobe (Stichprobengröße) und zu den Versuchspersonen (Alter, Geschlecht, Herkunft etc.). Auch die Art und Weise der Durchführung (händisch oder online) wird hier dargestellt. Die Fragebögen (unausgefüllt), die Item- und die Filler-Liste sollten der Arbeit in einem Anhang beigelegt werden. Bei einer *Korpusstudie* sollten zentrale Aspekte der Korpora (Größe in Tokens, Ebenen der Annotation, Art der Texte, weitere spezielle Eigenschaften) dargestellt und die Auswahl mit Blick auf die Hypothese diskutiert werden. Auch die Suchanfrage sollte thematisiert werden: Mit welchem Query-Tool wurde gesucht und wie sah die Suchanfrage konkret aus? Welche Probleme haben sich bei der Formulierung der Suchanfrage gezeigt? Gibt es Daten, die nicht erfasst oder vielleicht sogar systematisch ausgeschlossen wurden (werden mussten)?

Auswertung und Resultate: Darauf aufbauend ist darzustellen, wie die erhobenen Daten erfasst und statistisch ausgewertet wurden. Mussten Treffer oder Versuchspersonen aussortiert werden? Und wenn ja, warum und auf welcher Grundlage? Wie wurde bei Fragebögen mit fehlenden Angaben umgegangen? Wie wurden die Daten am Ende kategorisiert und tabellarisch erfasst? Die erhobenen Daten (also die Mitte(l)werte, Streuungen und/oder Häufigkeiten) werden statistisch beschrieben und die Resultate immer [!] in möglichst transparenter Weise graphisch aufbereitet. Falls weiterführende Tests gerechnet wurden, sind auch diese Ergebnisse hier darzustellen: Welche Art von Test wurde gerechnet? Wie wurde gerechnet? Mit welcher Software (R, *SPSS* etc.)? Mit welchen Paketen? Mit welchen Funktionen und welchen Parametern? Welche statistischen Werte (z.B. χ^2-, t- oder W-Wert) und welche Signifikanzen haben sich gegebenenfalls ergeben?

4. **Interpretation:** Zuletzt sind die Daten noch zu interpretieren. Widersprechen sie der getesteten Hypothese oder können sie als positive Evidenz gewertet werden? Welche weiteren Überlegungen ergeben sich? Gibt es besondere Auffälligkeiten?

5. **Zusammenfassung:** Am Schluss sollten die wesentlichen Ergebnisse der Arbeit natürlich knapp und prägnant zusammengefasst werden.

Objekt- und Metasprache

Soviel zum kanonischen Aufbau empirischer Arbeiten. Kommen wir schließlich noch zu einigen sprachwissenschaftlichen Konventionen, die auch bei eher theoretisch ausgerichteten Hausarbeiten (und generell bei sprachwissenschaftlichen Präsentationen) zu beachten sind. Wir beginnen mit der zentralen Unterscheidung zwischen Objektsprache und Metasprache. Als Objektsprache bezeichnet man die Sprache, die Gegenstand der Untersuchung ist, und als Metasprache bezeichnet man die Sprache, in der die Untersuchung dargestellt wird. Da in der Germanistik das Deutsche untersucht wird, fallen hier Objektsprache und Metsprache zusammen und umso wichtiger ist es, die Objektsprache im Text deutlich zu kennzeichnen. Hier gibt es im Wesentlichen zwei Methoden: Kursivierung oder Absetzen als nummeriertes Beispiel. Werden nur einzelne Ausdrücke im laufenden Text diskutiert (wie z.B. modales *brauchen*), dann werden die objektsprachlichen Ausdrücke kursiv gesetzt. (Daher wählt man am besten eine Schriftart, die über eine gut erkennbare Kursive verfügt.) Möchte man sich im Text wiederholt auf einzelne objektsprachliche Beispiele beziehen, dann ist es meist sinnvoll, diese Beispiele wie in (1) mit Nummerierung vom Text abzusetzen:

(1) *Frieder braucht jetzt nicht mehr kommen.*

Werden mehrere Beispiele in einem gemeinsamen Zusammenhang diskutiert, dann werden sie häufig auch gemeinsam abgesetzt, vgl. (2-a) und (2-b).

(2) a. *Frieder braucht jetzt nicht mehr kommen.*
b. *Frieder braucht jetzt nicht mehr zu kommen.*

Glossierung

Werden fremdsprachliche Ausdrücke wie das englische Modalverb *must* (»muss«) diskutiert, dann wird dessen Übersetzung in runden Klammern mit Anführungszeichen wiedergegeben. Abgesetzte Beispielen erhalten in diesem Fall wie in (3) eine **glossierte Übersetzung**. Glossierte Übersetzungen zeichnen sich dadurch aus, dass unter dem eigentlichen Beispiel zunächst eine Wort-für-Wort-Übersetzung angegeben wird, in der die Übersetzung linksbündig zu dem zu übersetzenden Wort ausgerichtet ist. (Das kann man in Word oder LibreOffice über eng gesetzte Tabulatoren oder eine Tabelle ohne Zellenränder erreichen.) Die Wort-für-Wort-Übersetzung wird schließlich in einer getrennten Zeile durch eine zielsprachlich korrekte Übersetzung in einfachen Anführungszeichen ergänzt:

(3) *Frieder must not leave the country*
Frieder darf nicht verlassen das Land
›Frieder darf das Land nicht verlassen.‹

Sprecherurteile

Neben eher unauffälligen Daten werden in der Linguistik auch Beispiele diskutiert, die von Muttersprachler*innen entweder als ungrammatisch oder doch zumindest als mehr oder weniger auffällig beurteilt werden. Um solche Sprecherurteile zu markieren, wird den Beispielen ein Asterisk * (für ungrammatische Sätze) oder ein Fragezeichen ? (oder zwei für stark markierte Sätze) vorangestellt:

(4) a. *Frieder glaubt nicht, er hat ein Einhorn gesehen.
b. ??Frieder glaubt fest, dass ein Einhorn er gesehen hat.
c. #Frieder glaubt, dass der Kobold$_1$ ihn$_1$ verhext hat.

Das Beispiel in (4-c) ist eigentlich unauffällig – vorausgesetzt, das Pronomen *ihn* bezieht sich inhaltlich auf *Frieder*. Über Koindizierung der NP *der Kobold* und des Pronomens *ihn* wird jedoch ausgedrückt, dass (4-c) in der Lesart bewertet werden soll, in der *ihn* koreferent mit der NP *der Kobold* ist (sich beide Ausdrücke auf dasselbe Objekt beziehen). Das Zeichen # markiert hier, dass (4-c) in genau dieser Lesart ungrammatisch ist (da *ihn*$_1$ hier durch *sich*$_1$ ersetzt werden müsste).

Literaturverweise

Egal, ob Inhalte aus der Literatur nur sinngemäß oder im Wortlaut wiedergegeben werden, es ist immer [!] die relevante Literatur anzugeben. Das kann auf verschiedene Weise erfolgen: Bei sinngemäßer Wiedergabe über direkte Bezugnahme: »Wie bereits Montague (1974: 188) zeigt, gibt es keinen wesentlichen Unterschied zwischen formalen und natürlichen Sprachen« oder als nachgestellter Hinweis: »Es wurde in der Literatur bereits überzeugend argumentiert, dass es keinen wesentlichen Unterschied zwischen formalen und natürlichen Sprachen gibt (vgl. Montague 1974: 188)«. Anders als in der Literaturwissenschaft wird in der Linguistik bei Literaturverweisen nie gleichzeitig die vollständige Literaturangabe in einer Fußnote angegeben. Vollständige Angaben finden sich alleine in der Bibliographie am Ende der Arbeit. Fußnoten sind in der Linguistik inhaltlichen Anmerkungen und Verweisen vorbehalten, die nicht zur eigentlichen Argumentation gehören bzw. die eigentliche Argumentation nur ergänzen.

Zitate

Kleinere Zitate werden in den fortlaufenden Text integriert und mit Anführungszeichen markiert: Montague (1974: 188) behauptet, dass »[no] important theoretical difference exists between formal and natural languages«. Längere Zitate werden in einem eingerückten Block vom Fließtext abgesetzt:

> I reject the contention that an important theoretical difference exists between formal and natural languages. On the other hand, I do not regard as successful the formal treatments of natural languages attempted by certain contemporary linguists. I regard the construction of a theory of truth – or rather, of the more general notion of truth under an arbitrary interpretation – as the basic goal of serious syntax and semantics. (Montague 1974: 188)

Sonderzeichen

In sprachwissenschaftlichen Arbeiten steht man nicht selten vor dem Problem, ganz spezielle Zeichen benutzen zu müssen. In der Semantik trifft man beispielsweise häufig auf Quantoren wie ∀ und ∃ oder Junktoren wie ∧ und ∨. Auf diese Zeichen kann man in Word entweder über *Einfügen > Erweitertes Symbol* und den Zeichensatz *Symbol* zugreifen oder über den Formel-Editor. Der Formel-Editor arbeitet allerdings mit dem Zeichensatz *Cambria Math*. Möchte man mathematische Symbole mit dem Zeichensatz *Palatino* kombinieren, sollte man sich *Asana Math* installieren. Zu *Times New Roman* passen neben *Symbol* auch die *Stixfonts*. Phonetische Zeichensätze aller Art (*Gentium Plus*, *Charis SIL* für Palatino, *Doulos SIL* für Times New Roman) finden sich auf der Webseite der gemeinnützigen Organisation *SIL International*. Dort finden sich auch Hinweise zur Installation und Eingabehilfen. Die Arbeit mit alt-, mittel- oder frühneuhochdeutschen Texten erfordert ebenfalls spezielle Zeichensätze. Eine erste Anlaufstelle ist hier die Webseite der *Medieval Unicode Font Inititative*, die auch auf alternative Fonts (*Junicode*, *Cardo*, *Titus Cyberbit Font*) aufmerksam macht.

Literaturverzeichnis

Wie die Einträge im Literaturverzeichnis zu formatieren sind, ist nicht wirklich einheitlich geregelt. Eine Richtlinie bietet der »Unified Style Sheet«, der über den Link https://linguistlist.org/pubs/tocs/JournalUnifiedStyleSheet2007.pdf zugänglich ist. Man kann sich aber natürlich auch an speziellen Zeitschriften (wie der *Zeitschrift für Sprachwissenschaft*) oder Einführungen orientieren. Wichtig ist aber vor allem, dass die Formatierungen in der Bibliographie konsistent sind.

Empfohlene Literatur

Albert, Ruth & Nicole Marx. 2016. *Empirisches Arbeiten in Linguistik und Sprachlehrforschung. Anleitung zu quantitativen Studien von der Planungsphase bis zum Forschungsbericht*. Tübingen: Narr. 3., überarbeitete und aktualisierte Auflage.

Gries, Stefan Th. 2008. *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht.

Gries, Stefan Th. 2013. *Statistics for Linguists with R: A Practical Introduction*. Berlin, Boston: Walter de Gruyter. 2., überarbeitete und erweiterte Auflage.

Kübler, Sandra & Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. London, New Delhi, New York, Sydney: Bloomsbury.

Lemnitzer, Lothar & Heike Zinsmeister. 2006. *Korpuslinguistik*. Tübingen: Narr.

Meindl, Claudia. 2011. *Methodik für Linguisten: Eine Einführung in Statistik und Versuchsplanung*. Tübingen: Narr.

Navarro, Danielle J. 2016. *Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners*. Unpublished Manuscript, University of Adelaide. Version 0.6. [Verfügbar über: http://learningstatisticswithr.com]

Navarro Danielle J. and David R. Foxcroft. 2019. *Learning Statistics with Jamovi: a Tutorial for Psychology Students and Other Beginners*. Version 0.70. DOI: 10.24384/hgc3-7p15 [Verfügbar über: https://learnstatswithjamovi.com]

Rothstein, Björn. 2011. *Wissenschaftliches Arbeiten für Linguisten*. Tübingen: Narr.

Sachregister

Literatur

Albert, Ruth & Nicole Marx. 2016. *Empirisches Arbeiten in Linguistik und Sprachlehrforschung. Anleitung zu quantitativen Studien von der Planungsphase bis zum Forschungsbericht*. Tübingen: Narr. 3. Auflage.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Dudenredaktion & Angelika Wöllstein (eds.). 2016. *Duden. Die Grammatik*. Berlin: Dudenverlag. 9. Auflage.

Eisenberg, Peter. 1997. Die besondere Kennzeichnung der kurzen Vokale. Vergleich und Bewertung der Neuregelung. In Gerhard Augst, Karl Blüml, Dieter Nerius & Horst Sitta (eds.), *Zur Neuregelung der deutschen Orthographie. Begründung und Kritik*, Germanistische Linguistik 179, 323–336. Berlin: De Gruyter.

Eisenberg, Peter. 2013. *Grundriss der deutschen Grammatik*, Band 1: Das Wort. Stuttgart: Metzler. 4. Auflage.

Gries, Stefan Th. 2008. *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht.

Gries, Stefan Th. 2013. *Statistics for Linguists with R: A Practical introduction*. Berlin, Boston: Walter de Gruyter.

Heidolph, K. E., Walter Flämig & Wolfgang Motsch (eds.). 1981. *Grundzüge der deutschen Grammatik*. Berlin: Akademie-Verlag.

Horch, Eva & Ingo Reich. 2017. The Fragment Corpus (FraC). In *Proceedings of the 9th International Corpus Linguistics Conference*, Birmingham (UK).

Krause, Thomas & Amir Zeldes. 2016. Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31. 118–139.

Kübler, Sandra & Heike Zinsmeister. 2015. *Corpus Linguistics and Syntactically Annotated Corpora*. London: Bloomsbury.

Lemnitzer, Lothar & Heike Zinsmeister. 2015. *Korpuslinguistik*. Tübingen: Narr.

Lezius, Wolfgang. 2002. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*: Universität Stuttgart Dissertation.

Maché, Jakob. 2019. Braucht das epistemisch zu sein? Wie Status, Negation und lexikalische Semantik die Interpretation von Notwendigkeitsverben in Korpusdaten bestimmen. Unveröffentlichtes Manuskript.

Meindl, Claudia. 2011. *Methodik für Linguisten. Eine Einführung in Statisitk und Versuchsplanung*. Tübingen: Narr.

Montague, Richard. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. New Haven and London: Yale University Press.

Navarro, Danielle. 2016. Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners. Unpublished Manuscript, University of Adelaide.

Ramers, Karl Heinz. 1999. Vokalquantität als orthographisches Problem. Zur Funktion der Doppelkonsonanzschreibung im Deutschen. *Linguistische Berichte* 177. 350–360.

Reis, Marga. 2005. Wer *brauchen* ohne *zu* gebraucht … Zu systemgerechten ›Verstößen‹ im Gegenwartsdeutschen. *Cahiers d'Études Germaniques* 48. 101–114.

Rothstein, Björn. 2011. *Wissenschaftliches Arbeiten für Linguisten*. Tübingen: Narr.

Schütze, Carson. 2016. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Berlin: Language Science Press.

Winter, Bodo. 2013. Linear Models and Linear Mixed Effects Models in R with Linguistic Applications. Unpublished Manuscript, University of California. [Verfügbar über https://arxiv.org/pdf/1308.5499.pdf]

Zifonun, Gisela, Ludger Hoffmann & Bruno Strecker. 1997. *Grammatik der deutschen Sprache*. Berlin / New York: M. de Gruyter. 3 Bände.