Korpusbasierter Vergleich von Varietäten

Tutorium der Sektion CL

Felix Golcher, Anke Lüdeling

Institut für Deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin

DGfS-Jahrestagung, Universität Leipzig, 3. März 2015

Vergleich von Varietäten

- Wie verändern sich Kompositionsmuster zwischen dem Mittelhochdeutschen und dem Frühneuhochdeutschen?
- Wie unterscheiden sich Satzlängen in gesprochener Spontansprache von Satzlängen in Chatdaten?
- Nutzen LernerInnen des Deutschen als Fremdsprache V-N-Konstruktionen anders als MuttersprachlerInnen des Deutschen?
- Sprechen Frauen anders als Männer?
- Wie unterscheiden sich Register?

→ Man vergleicht zwei Varietäten bezüglich *eines* oder mehrerer vorher definierter Merkmale quantitativ. Eine mögliche Datengrundlage: Korpora. Da kann man einfach zählen, denn „corpora contain nothing but frequencies/probabilities — of occurrence or of co-occurrence“ Gries (2015, S. 93)

Voraussetzungen

So einfach ist es nicht: Solche Fragen setzen verschiedene Klärungen voraus

- Welche *Forschungsfrage* soll beantwortet werden? Geht es um Exploration oder geht es um ein hypothesengeleitetes Experiment? Oder ein Modell?
  → Exploration und Experiment / beschreibende und schließende Statistik
- Was ist eine Varietät? Wie kann ein Korpus eine Varietät abbilden?
  → Korpusdesign
- Was ist ein Merkmal? Wie kann man das definieren und annotieren?
  → Kategorisierung, Annotation und Urteilerübereinstimmung
- Welche statistischen Verfahren können angewendet werden? Wie ist das Merkmal verteilt?
  → Verteilungen und Verfahren

Grundannahmen

- Mit Korpora kann man sprachliches Verhalten untersuchen.
- Man untersucht nicht nur den Korpustext selbst, sondern auch Kategorisierungen von sprachlichen Mustern (Wortarten, Phrasentypen, Satztypen, rhetorische Strukturen, ...)
- Sprachliches Verhalten ist abhängig von vielen Faktoren (viele kennen wir wahrscheinlich noch gar nicht)
  - Sprecherfaktoren/sozio-ökonomische Faktoren wie Alter, Ausbildung, Geschlecht etc. des Sprechers, siehe z.B. Labov (2008)
  - funktionale Faktoren wie Zweck des Diskurses, Einbettung, siehe z.B. Biber und Conrad (2009)
  - Modalität (gesprochen, geschrieben, Umgebungsfaktoren)
  - ...

Grundannahmen

- Sprachliches Verhalten (und also die Zählungen von Kategorien in Korpusdaten) in verschiedenen Varietäten kann sich auf *jeder sprachlichen Ebene* unterscheiden. Die Unterschiede sind meist nicht kategorial, sondern quantitativ. Das bedeutet
  - → die Varietäten müssen sauber unterschieden werden
  - → die Kategorien müssen in den zu vergleichenden Korpora *auf dieselbe Art* zugewiesen werden

Hintergrund: Replizierbarkeit vs. Reproduzierbarkeit

Die *Begriffe* werden unterschiedlich verwendet, die *Konzepte* müssen unbedingt klar definiert werden. Unsere Verwendung:

Replizierbarkeit: Man erreicht mit denselben Daten und denselben Verfahren dieselben Ergebnisse. Replizierbarkeit zeigt, ob jemand sauber gearbeitet und dokumentiert hat. Die Replikation eines Experiments ist auch oft sinnvoll für das Verständnis der Daten und Ergebnisse.

Reproduzierbarkeit: Man erreicht mit vergleichbaren Daten und vergleichbaren Verfahren vergleichbare Ergebnisse. Reproduzierbarkeit von Ergebnissen weist darauf hin, dass das Phänomen, das man untersucht, existiert.

Technisches zur Replizierbarkeit: Die Daten

Genaue Dokumentation der Daten

Ohne Zugriff auf die exakt selben Daten ist Replizierbarkeit von vornherein unmöglich.

- **Versionierung** des Korpus, Bezugnahme auf die verwendete Version.
- **Dokumentation** von Korpusdesign und Annotation.
- Freie **Veröffentlichung** des Korpus.

Veröffentlichung von Daten und Skripten

zusammen mit den Veröffentlichungen, oder als Paper Packages.
Beispiel: *Mind Research Repository*
`http://openscience.uni-leipzig.de/index.php/mr2`

Technisches zur Replizierbarkeit: Skripte

Definition (Skript)

Eine Textdatei mit Arbeitsanweisungen für ein Computerprogramm.
Folgt einer programmspezifischen Syntax.

automatisiertes Arbeiten = replizierbares Arbeiten

So gut wie alle Arbeitsschritte lassen sich in Skripten beschreiben und automatisiert ausführen.

- Extreme Beschleunigung bei kleinen, repetitiven Aufgaben.
- Fehler im Resultat lassen sich auf Fehler im Skript zurückführen.
- Ein (nicht zu wirres) Skript dokumentiert/beschreibt sich selbst.
- Weiter dokumentierbar und kommentierbar.

Technisches zur Replizierbarkeit: Die Sprache

Lingua Franca!

Nur, wer Kommentare und Dokumentation lesen kann, ist in der Lage, die Ergebnisse zu replizieren.

- Am Anfang ist vielleicht nicht klar, was veröffentlicht wird.
- „schnell, schnell!“ $\Rightarrow$ Schreiben wir erst mal auf Deutsch, Polnisch,...
- Spätere Kollaborateure? Leser einer Veröffentlichung?
  $\Rightarrow$ Pech gehabt.
- Es kann enorm Zeit sparen, von vornherein in der *lingua franca* der community zu schreiben.

Technisches zur Replizierbarkeit: Open Source

Freie Software

Nur freie Software ermöglicht jedem jederzeit die Replizierung.

Frei wie „für umsonst“. Nicht jeder hat ausreichend Geld oder den Zugang zu Campuslizenzen.

Frei wie „offen und transparent“. Ein Wissenschaftler sollte die Möglichkeit haben, nachzuvollziehen, was ein Programm tut.

Frei wie „nicht fest“: Freie Software ist oft durch jedermann erweiterbar.

- für die Statistiksoftware `R` derzeit 6387 Erweiterungen (1. März 2015 17:05:46)
- in der Regel viel schnellere Implementierung neuer Methoden.

Technisches zur Replizierbarkeit: Plattformunabhängigkeit

Die Linux-Windows-Grenze

Verwenden Sie nur Programme, die überall laufen.

- Nicht so wenige Wissenschaftler verwenden Linux
- Kollaboration wird stark erleichtert.
- Replizierbarkeit für jeden.
- Tendenziell freie Software (Libreoffice, `R`, etc.)

Technisches zur Replizierbarkeit: textbasiertes Arbeiten

plain text

Nur Daten, die als reine Textdateien vorliegen, werden immer sicher gelesen werden können.

- Natürlich trotzdem Strukturierungsmöglichkeiten (Stichwort `xml`)
- Auch populäre Binärformate (Word, Excel) können nicht garantieren, dass spätere Programmversionen alte Dateien lesen können.
- Erleichterte Kollaboration durch **Versionierungstools**.

beispielhafter Workflow

- Das versionierte Korpus liegt frei in gut kommentiertem, standardisiertem Format vor (ähnlich für Nicht-Korpusdaten)
- Nötige (?) manuelle Schritte sind dokumentiert.
- Jegliche Vor- und Nachverarbeitung mittels dokumentierter Skripte.
- Statistische Auswertung mit `R` (frei, skriptfähig)
- Textfassung in LaTeX (frei, textbasiert)
- Verbindung mittels `knitr` (Xie 2014)

$\Rightarrow$ statistische Analyse und Erzeugung der Veröffentlichung fallen zusammen.

Stichprobenziehung

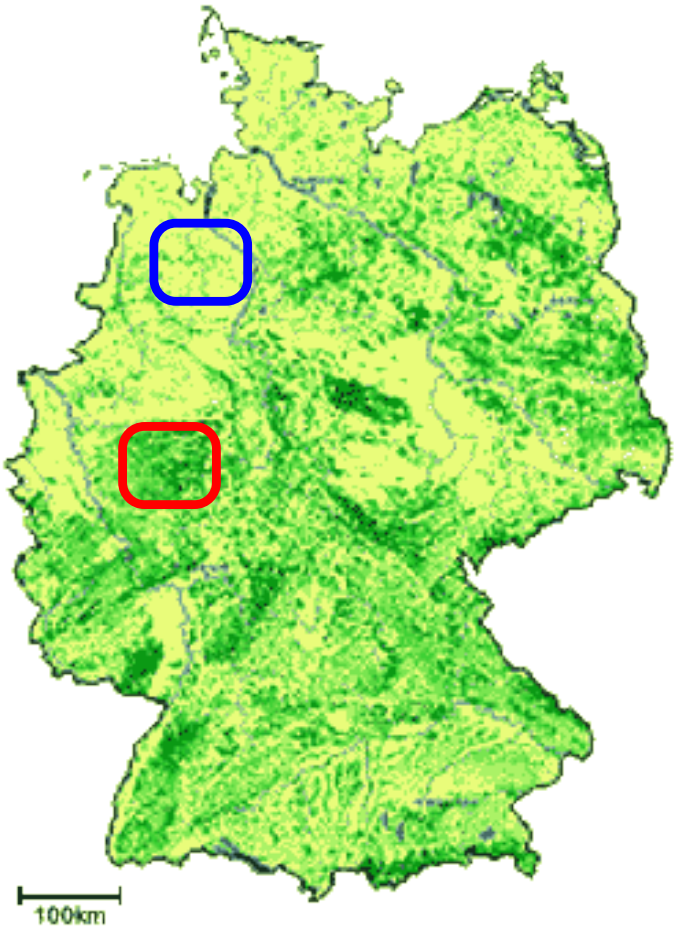
Man möchte eine gegebene Varietät untersuchen, die zu groß ist für eine exhaustive Analyse. Daher sind Korpora meist Stichproben (Samples) aus einer Grundgesamtheit (Population). Man kann aus einer Grundgesamtheit unterschiedliche Stichproben ziehen (zufällig, stratifiziert, repräsentativ). Die Art der Stichprobenziehung ist relevant für die Möglichkeiten, aus den Ergebnissen der Stichprobe auf die Grundgesamtheit zu schließen (Extrapolation).
Terminologische Anmerkungen:

- zufällige Stichproben in der Korpuslinguistik werden oft *opportunistisch* genannt
- zufällige Stichproben sind keine *Zufallsstichproben* oder *random samples* (dies sind technische Begriffe, die im Rahmen der Statistik klar definiert sind)

- Forschungsfrage: Wieviel % von Deutschland sind Wald?
- Wir können nicht ganz Deutschland bereisen.
- Was tun?

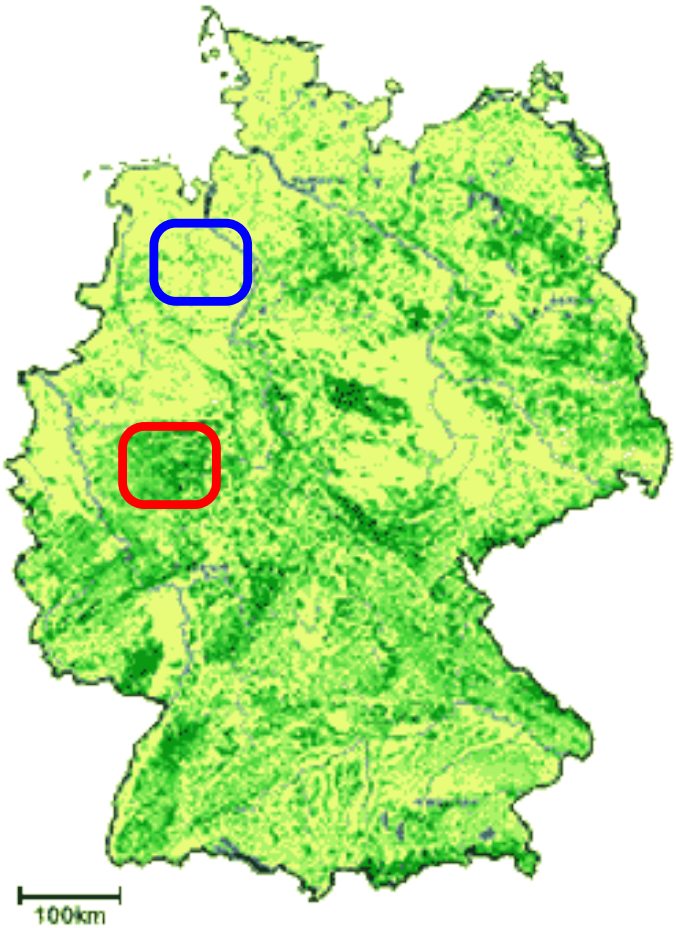
Quelle: Bundesforschungsanstalt für Forst- und Holzwirtschaft, Institut für Ökonomie, Hamburg, 1997

- Forschungsfrage: Wieviel % von Deutschland sind Wald?
- Wir können nicht ganz Deutschland bereisen.
- Was tun?



Quelle: Bundesforschungsanstalt für Forst- und Holzwirtschaft, Institut für Ökonomie, Hamburg, 1997

- Forschungsfrage: Wieviel % von Deutschland sind Wald?
- Wir können nicht ganz Deutschland bereisen.
- Was tun?



Quelle: Bundesforschungsanstalt für Forst- und Holzwirtschaft, Institut für Ökonomie, Hamburg, 1997

- Forschungsfrage: Wieviel % von Deutschland sind Wald?
- Wir können nicht ganz Deutschland bereisen.
- Was tun?

Extrapolation aus diesen zufälligen Stichproben würde zu falschen Ergebnissen führen.

Stichprobenziehung

- wenn man Parameter kennt, die die Variation in der Grundgesamtheit beeinflussen, kann man diese in die Stichprobenziehung einbeziehen
  → stratifizierte Stichprobe
  - für die Waldfrage vielleicht Parameter wie Stadt/Land, Berg/kein Berg, Privatbesitz/öffentlicher Besitz, Bodenbeschaffenheit etc.
  - für die Korpusauswahl (abhängig von der Forschungsfrage) vielleicht Parameter wie Modus, sozio-ökonomische Faktoren, Dialekt, Zeit etc.
- man kann dann ein Korpus bauen (eine Stichprobe ziehen), das Ausschnitte aus allen (Kombinationen von) Parametern enthält
- damit kann man zumindest klären, *ob* ein Parameter einen Einfluss hat

Stichprobenziehung

- wenn man die Parameter kennt *und deren Verteilung in der Grundgesamtheit*, kann man eine repräsentative Stichprobe ziehen (und es gibt natürlich Verfahren, die es unter diesen Voraussetzungen erlauben, aus stratifizierten Stichproben zu extrapolieren)
- beim Waldbeispiel wäre eine repräsentative Stichprobe vielleicht möglich
- bei Korpora ist eine repräsentative Stichprobe fast nie möglich - man kennt die Zusammensetzung der Grundgesamtheit ja eigentlich fast nie
- Nebenbemerkung zur Korpusgröße: Größe ersetzt nicht sorgfältige Stichprobenziehung; es ist nicht möglich, allgemein eine minimale sinnvolle Größe für ein Korpus anzugeben, da die Größe sich nach der Häufigkeit des zu betrachtenden Phänomens richtet

Stichprobenziehung

Kann man eine repräsentative Stichprobe ziehen aus

- dem Deutschen?
- dem Sächsischen?
- dem Althochdeutschen?
- der althochdeutschen Überlieferung?
- der internetbasierten Kommunikation?
- Chatdeutsch?
- allen auf Papier geschriebenen Dokumenten, die sich jetzt in diesem Raum befinden?
- Frauensprache?

Parameter

Ideal: Man vergleicht zwei Varietäten, die sich nur in einem Designparameter unterscheiden, denn nur dann können Varietätenunterschiede *einfach* auf den betrachteten Unterschied zurückgeführt werden. Wenn sich zwei Varietäten in mehreren Designparametern unterscheiden, muss man komplexe statistische Modelle anwenden.

Wirklichkeit: Korpusdaten sind keine Labordaten. Es ist oft schwierig, zwei Korpora zu finden, die sich nur in einem Designparameter unterscheiden - man muss das wissen und entsprechend in der Analyse berücksichtigen.

Parameter: Beispiel Lernerkorpora

Angenommen, man möchte korpusbasiert herausfinden, ob Lerner des DaF in Essays andere syntaktische Strukturen verwenden als Muttersprachler. Dann kann man ein syntaktisch annotiertes Lernerkorpus und ein vergleichbares Muttersprachlerkorpus analysieren.[1]

Ideal: Das Lernerkorpus unterscheidet sich nur in dem Punkt L1 vom Muttersprachkorpus.

Wirklichkeit: Das ist nicht möglich - zusammen mit einer anderen Muttersprache gibt es immer auch ein anderes Schulsystem (unterschiedliche Lehrmethoden, verschiedene Textbegriffe, Schreibkompetenzen, nicht nur im Fremdspracherwerb, sondern auch im Muttersprachunterricht) und auch sonst andere Erfahrungen

[1] Es gibt sehr viele solche Studien, siehe zum Beispiel die Bibliographie unter `http://www.uclouvain.be/en-cecl-lcbiblio.html`.

Parameter: Beispiel diachrone Korpora

Angenommen, man möchte eine korpusbasierte Studie zum Sprachwandel machen[2]

Ideal: Ein diachrones Korpus, in dem alle Parameter (Dialekt, Textsorte, Schreiber, Rezipient etc.) gleich sind und nur die Zeit sich unterscheidet.

Wirklichkeit: Überlieferungsproblem (oft zufällig), Informationsproblem (man kennt manchmal nicht alle Parameter)

[2] Auch dazu gibt es viel Literatur (siehe für einen Überblick Hilpert und Gries erscheint) - bei historischen Korpora gibt außer den hier angesprochenen Themen noch viel mehr zu beachten.

Zusammenfassung Korpusdesign

Stichprobe Ein Korpus ist (fast immer) eine Stichprobe aus einer nicht gut bekannten Grundgesamtheit. Man muss wissen, wie die Stichprobe gezogen wurde, um zu wissen, wie man extrapolieren kann.

Parameter Es kann unterschiedliche Parameter geben, die die sprachliche Variation beeinflussen (abhängig von der Forschungsfrage) - solche Parameter sollten bei der Stichprobenziehung einbezogen werden. Im Korpus werden sie als *Metadaten* kodiert; bei gutem Korpusdesign und guter Korpusarchitektur ist es möglich, ad hoc Subkorpora zu erstellen.

Vergleich Bei dem Vergleich von Korpora ist es wichtig, dass man die relevanten Parameter kennt und, soweit möglich, stabil hält. Wenn das nicht geht, muss man in den Schlüssen vorsichtig sein (hier wäre Reproduzierbarkeit gut).

Kategorisierung: Kompositionsmuster in frühneuhochdeutschen Daten

Kleines Experiment: In der Excel-Datei *KompositaExperiment.xlsx* sollen Komposita annotiert werden. Dazu müssten Sie eine Spalte anlegen. (Bitte die Tokenspalte unter keinen Umständen verändern.) In der neuen Spalte sollten immer da, wo Sie ein Kompositum finden, dessen neuhochdeutsche Entsprechung eingetragen werden. Wenn ein Kompositum über mehrere Zeilen geht, muss in jeder Spalte dasselbe Kompositum eingetragen werden (also keine Spannenannotation, das ist sonst nicht so einfach automatisch zu verarbeiten). Beispiele sind in den obersten Zeilen angegeben.

- Die Datei findet sich unter `http://korpling.german.hu-berlin.de/~felix/dgfs/`
- Der Upload der fertigen Datei erfolgt über die selbe Seite.

Ridges

Die Daten für das Experiment stammen aus dem Ridges-Korpus. Ridges (für **R**egister **i**n **G**erman **S**cience, unter CC_BY verfügbar unter `http://korpling.german.hu-berlin.de/ridges/index_en.html`) ist ein Korpus mit (Ausschnitten) aus Kräuterbüchern von 1487 bis 1870 (das Korpus wächst noch). Die Beispiele in dem Experiment stammen aus

- Johannes von Cuba (1487) Gart der Gesundheit, Ulm
- Otto Brunfels (1532) Contrafayt Kreüterbuch, Straßburg
- Adam von Bodenstein (1557) Wie sich meniglich, Basel

Heute geht es nicht um die Entwicklung der Komposition, sondern um die Urteilerübereinstimmung.

Annotation

- Jede Form von Kategorisierung (Zuordnung von Korpusdaten zu Kategorien) involviert Interpretation. Oft kann man dieselben Daten unterschiedlich interpretieren (vgl. zum Beispiel die Amalgam-Studie zu Wortarten[3], vgl. auch Lüdeling 2011)
- Zunächst bestimmt man eine Menge von zuweisbaren Kategorien (→ Tagset). Dann gibt man an, wie die Kategorien den Korpusdaten zugewiesen werden (→ Richtlinien, Guidelines). Daneben gibt es 'freiere' Annotationsebenen wie die Lemmaebene oder die Komposita in unserem Beispiel - auch hier muss man Richtlinien angeben (auch hier: Kategorisierung - man hat in der Annotationsebene *weniger* verschiedene Kategorien als im Original).

[3] `http://www.comp.leeds.ac.uk/ccalas/amalgam/amalghome.htm`

Urteilerübereinstimmung (Inter-Rater Agreement)

Urteilerübereinstimmung (auch: Inter-Annotator Agreement, Inter-Rater Reliability, Inter-Rater Agreement) beschreibt die Zuverlässigkeit, mit der bestimmte Tags bestimmten Korpusexponenten zugewiesen werden können. Es gibt verschiedene Maße (z. B. Cohen's Kappa, siehe Carletta 1996; Artstein und Poesio 2008). Dabei kann man zwei Fragen stellen:

- Wurden dieselben Exponenten kategorisiert?
  (Problem: False Negatives!)
- Wurden dieselben Kategorien zugewiesen?

Urteilerübereinstimmung (Inter-Rater Agreement)

- Quantifizierung der Übereinstimmung von (erstmal 2) Annotatoren:

$\kappa = 1$: Beide Annotatoren stimmen überall überein.
$\kappa = 0$: Es gibt **nur zufällige** Übereinstimmung.

- Recht einfache Definition (Cohen 1960):

$$\kappa = \frac{\textit{relative Übereinstimmung} - \textit{Zufallsbaseline}}{1 - \textit{Zufallsbaseline}}$$

- Verallgemeinert auf beliebig viele Annotatoren von Fleiss (1971)
- Implementiert von Gamer u. a. (2012) im `R`-Paket `irr`.

Auswertung des Experiments - das Vorgehen

- Die Daten werden über das Uploadscript gesammelt.
- Konversion in ein textbasiertes Format (batch)
- Zusammenführung mittels `R` (R Core Team 2014)
  - skriptfähig
  - plattformunabhängig
  - frei ...
- Aufbereitung & graphische Darstellung mittels dreier `R`-Pakete:
  - `plyr`: Elegantes Umsortieren von Daten (Wickham 2011)
  - `ggplot2`: Schöne, klare Graphiken (Wickham 2009)
  - `manipulate`: Interaktives Plotten (RStudio 2011)

Fazit: Jederzeit replizierbarer Ablauf ohne manuelle Schritte.

Zusammenfassung Annotation

- Die Zuweisung von Kategorien sollte immer explizit sein ($\rightarrow$ Annotation)
- Annotation ist oft ein iterativer Prozess (Evaluation, Überarbeitung der Richtlinien, erneute Annotation); Annotation ist Forschung.
- Für den Vergleich von Annotationen muss man sich fragen: Sind die Kategorien auf dieselbe Weise vergeben worden?

Exploration

- gegeben gut designte und verstandene Daten und dokumentierte und evaluierte Annotationen
- muss als erstes das Korpus qualitativ und quantitativ exploriert werden

„failure to understand relevant characteristics of data can lead to results and interpretations that are distorted or even worthless“ (Moisl 2009, S. 876)

„If you publish results when you have not [gotten to know your corpus, AL], it is like a drug company publishing and saying ‚use this drug‘ although they have not noticed that the group of subjects who they tested the drug on were largely under 25, with a big cluster who had travelled round South America, and none of them were pregnant. We need to guard against such bad science, and, if we intend to continue to be empiricist, and to work with data samples – corpora – in linguistics, we need to get to know our corpora.“ (Kilgarriff 2012, S. 14)

Immer der erste Schritt: Konsistenzcheck!

Kilgarriff 2012

[...] corpora are mostly too big to read (and not designed to be read)

- Dasselbe gilt meist für die Daten, in die man sie komprimiert.
  - Gerade Korpusdaten sind anfällig für Formatfehler.
  - Diese „fressen“ leicht große Teile des Korpus.
  - Statistikprogramme zeigen das oft nicht von selbst an.

    $\Rightarrow$ Man muss sie fragen.
- Konsistenzcheck für jede Variable:
  - Bewegt sie sich im erwarteten Rahmen?
  - Zeigt ein Plot die erwartete Verteilung?

Immer der erste Schritt: Konsistenzcheck!

Kilgarriff 2012

[...] corpora are mostly too big to read (and not designed to be read)

77
77
88
77
77
77
77
77

- Dasselbe gilt meist für die Daten, in die man sie komprimiert.
  - Gerade Korpusdaten sind anfällig für Formatfehler.
  - Diese „fressen“ leicht große Teile des Korpus.
  - Statistikprogramme zeigen das oft nicht von selbst an.

    $\Rightarrow$ Man muss sie fragen.
- Konsistenzcheck für jede Variable:
  - Bewegt sie sich im erwarteten Rahmen?
  - Zeigt ein Plot die erwartete Verteilung?

Immer der erste Schritt: Konsistenzcheck!

Kilgarriff 2012

[...] corpora are mostly too big to read (and not designed to be read)

- Dasselbe gilt meist für die Daten, in die man sie komprimiert.
  - Gerade Korpusdaten sind anfällig für Formatfehler.
  - Diese „fressen“ leicht große Teile des Korpus.
  - Statistikprogramme zeigen das oft nicht von selbst an.
    - ⇒ Man muss sie fragen.
- Konsistenzcheck für jede Variable:
  - Bewegt sie sich im erwarteten Rahmen?
  - Zeigt ein Plot die erwartete Verteilung?

```
77
77
88
77
77
77
77
Janusz
Damian
Romek
Kazik
Irek
Zbyszek
```

Normalisierung

- Korpuszählungen aus verschiedenen Korpora müssen auf eine gemeinsame Größe normalisiert werden. Welche Größe geeignet ist, hängt von der Forschungsfrage ab.
- Beispiel Entwicklung von Relativsätzen (Lüdeling, Hirschmann u. a. 2011), Datengrundlage Deutsche Diachrone Baumbank: kleines syntaktisch annotiertes Korpus
  `http://korpling.german.hu-berlin.de/ddb-doku/index.htm`

Beispiel: Relativsätze

pos	OHG	MHG	ENHG	NHG
PDAT	0.046131	0.011679	0.007105	0.008954
PPER	0.083545	0.052759	0.075916	0.075825
ART	0	0.07934	0.065445	0.061835
VVINF	0.01120	0.015707	0.018325	0.022104
PRELS	0.009444	0.011679	0.013837	0.016788
VAFIN	0.03705	0.035038	0.047868	0.045887
VAINF	0.001453	0.001208	0.004113	0.003078
PDS	0.023974	0.026983	0.013089	0.004757

Overuse/underuse-Statistik - zur Exploration, Normalisierungsgrundlage: Tokens

Beispiel: Relativsätze

Subkorpus	PRELS pro 100 Clauses
AHD	4,62**
MHD	10,25
FNHD	12,85
NHD	13,35

Subkorpus	NPs und PPs
AHD	0,141
MHD	0,208
FNHD	0,184
NHD	0,196

Tokens Kontinuierliche Entwicklung

Clauses signifikanter Unterschied zwischen AHD und danach

NPs und PPs weniger 'Anbindungsmöglichkeiten' in AHD

Normalisierung

- Je nach Normalisierungsgrundlage (im Beispiel: Tokens, Clauses, NPs/PPs) erfährt man unterschiedliche Dinge.
- Man muss also über die Normalisierungsgrundlage(n) gut nachdenken.
- Welche Fragen werden beantwortet durch folgende Normalisierungen (in einem beliebigen Korpus)?
  - Verben pro Token
  - finite Verben pro Token
  - PPs pro finites Verb

Ein Standardproblem

- Der Vergleich von Mittelwerten in Abhängigkeit von kategorialen Variablen taucht überall als statistisches Standardproblem auf.
- Das ist in der Korpuslinguistik nicht anders.
- Hier ein lehrreiches Beispiel
  - anspruchsvolles Beispielkorpus
  - leicht verständliche abhängige Variable
  - klare Ausgangsfrage
  - Wir sehen am Beispiel die besprochene Methodik.

Beispiel: Satzlängen in NoSta-D

kleines Korpus mit Texten aus unterschiedlichen Registern, die auf dieselbe Art annotiert worden sind (Dipper u. a. 2013; Dietterle u. a. eingereicht). Hier nutzen wir drei Subkorpora

NoSta-D BeMaTaC: Ein Ausschnitt aus dem BeMaTaC (**Be**rlin **Ma**p **Ta**sk **C**orpus, Sauer 2013), das **gesprochene** Map-Task-Dialoge enthält. BeMaTaC besteht aus einem Lernerkorpus und einem Muttersprachlerkorpus; für NoSta-D werden nur Muttersprachlerdaten verwendet.

NoSta-D Unicum: Ein Ausschnitt aus dem Dortmunder **Chat**-Korpus (Beißwenger 2013), das Plauderchatdaten enthält

NoSta-D TüBa-D/Z: Ein Ausschnitt aus dem TüBa-D/Z-Korpus (Hinrichs 2014), das **Zeitungstexte** enthält.

Satzlängenverteilung

Forschungsfrage

Unterscheidet sich die Satzlänge zwischen `unicum` und `bematac`?

Naive Herangehensweise:

- Satzlängen sind kontinuierliche Variablen.
- Wir haben zwei Varietäten (=Subkorpora).
- Vergleichen Mittelwerte.

$\Rightarrow$ Wir machen einen t-Test!

$$df = 2651$$

$$t = 0.50$$

$$p = .62 > .05$$

Forschungsantwort

Die Satzlängen in beiden Korpora sind gleich.

Exploration!

Wie sehen die Daten eigentlich aus?

Exploration!

Wie sehen die Daten eigentlich aus?

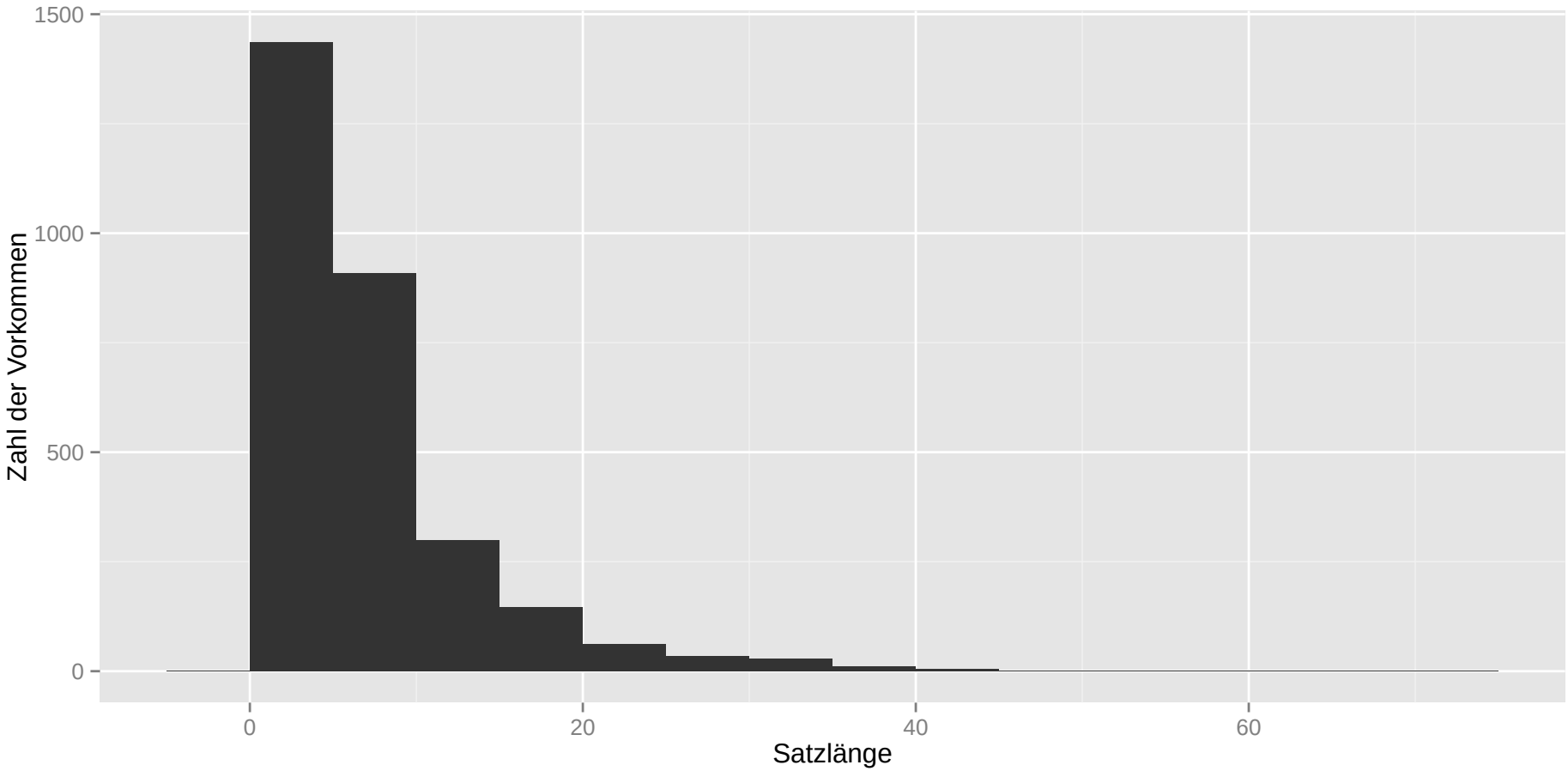


Abbildung: Satzlängenverteilung als Histogramm.

Exploration!

Wie sehen die Daten eigentlich aus? Sicher nicht normalverteilt.

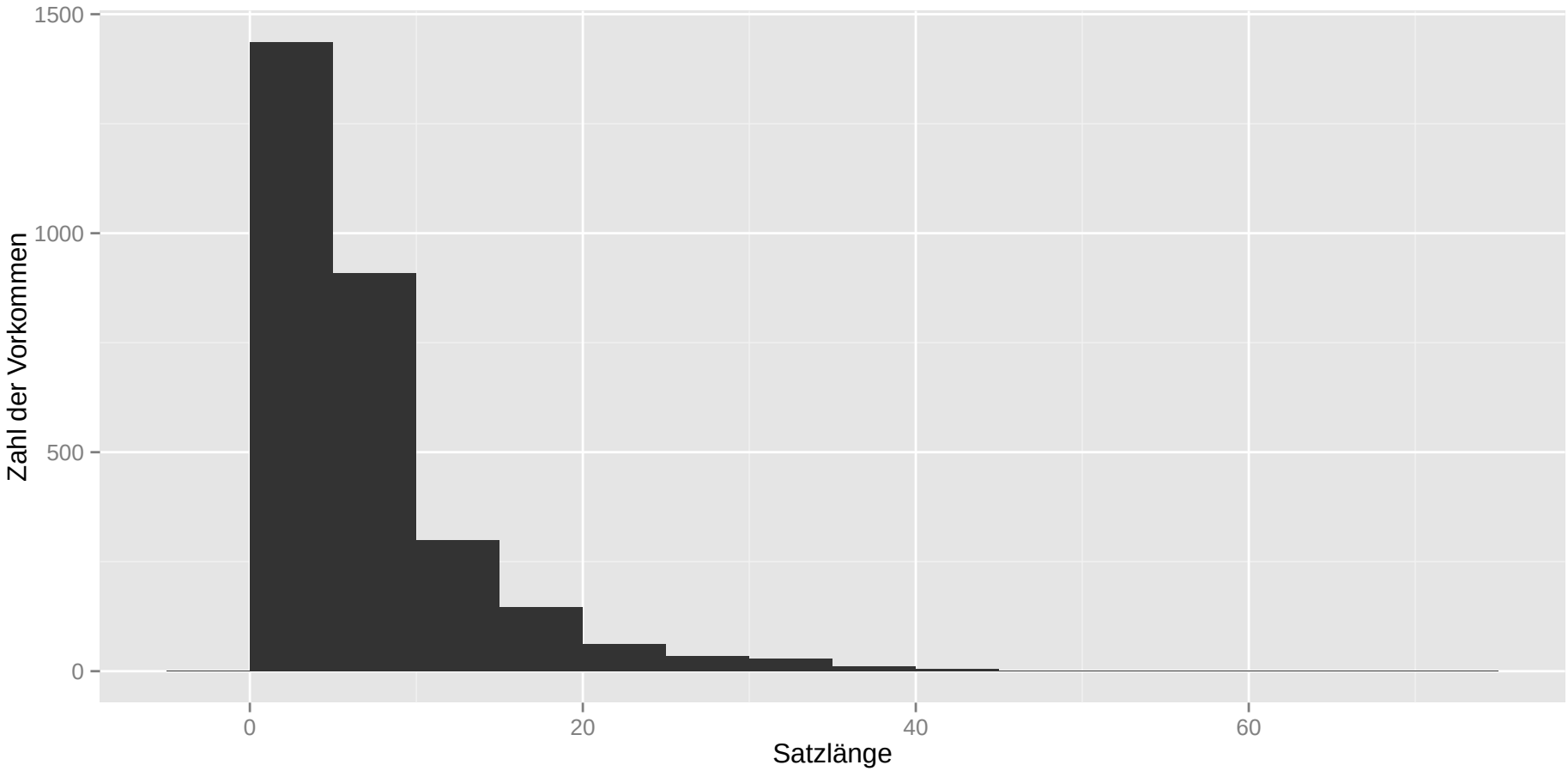


Abbildung: Satzlängenverteilung als Histogramm.

Wir zoomen hinein

Wir machen die Balkenbreite kleiner.

Wir zoomen hinein

Wir machen die Balkenbreite kleiner.

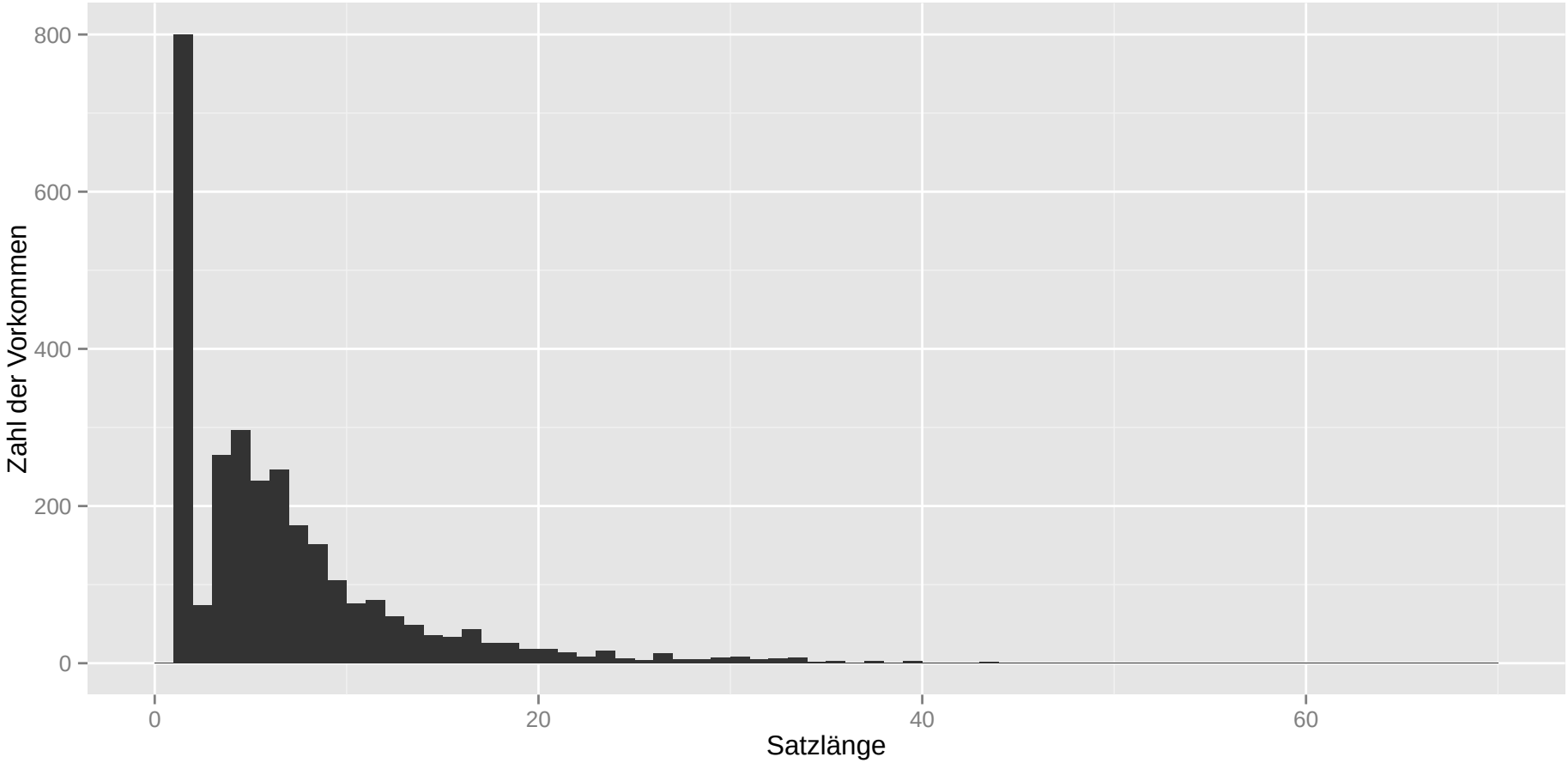


Abbildung: Satzlängenverteilung als Histogramm. Balkenbreite=1.

Wir zoomen hinein

Wir machen die Balkenbreite kleiner. Ein peak bei Satzlänge=1 erscheint.

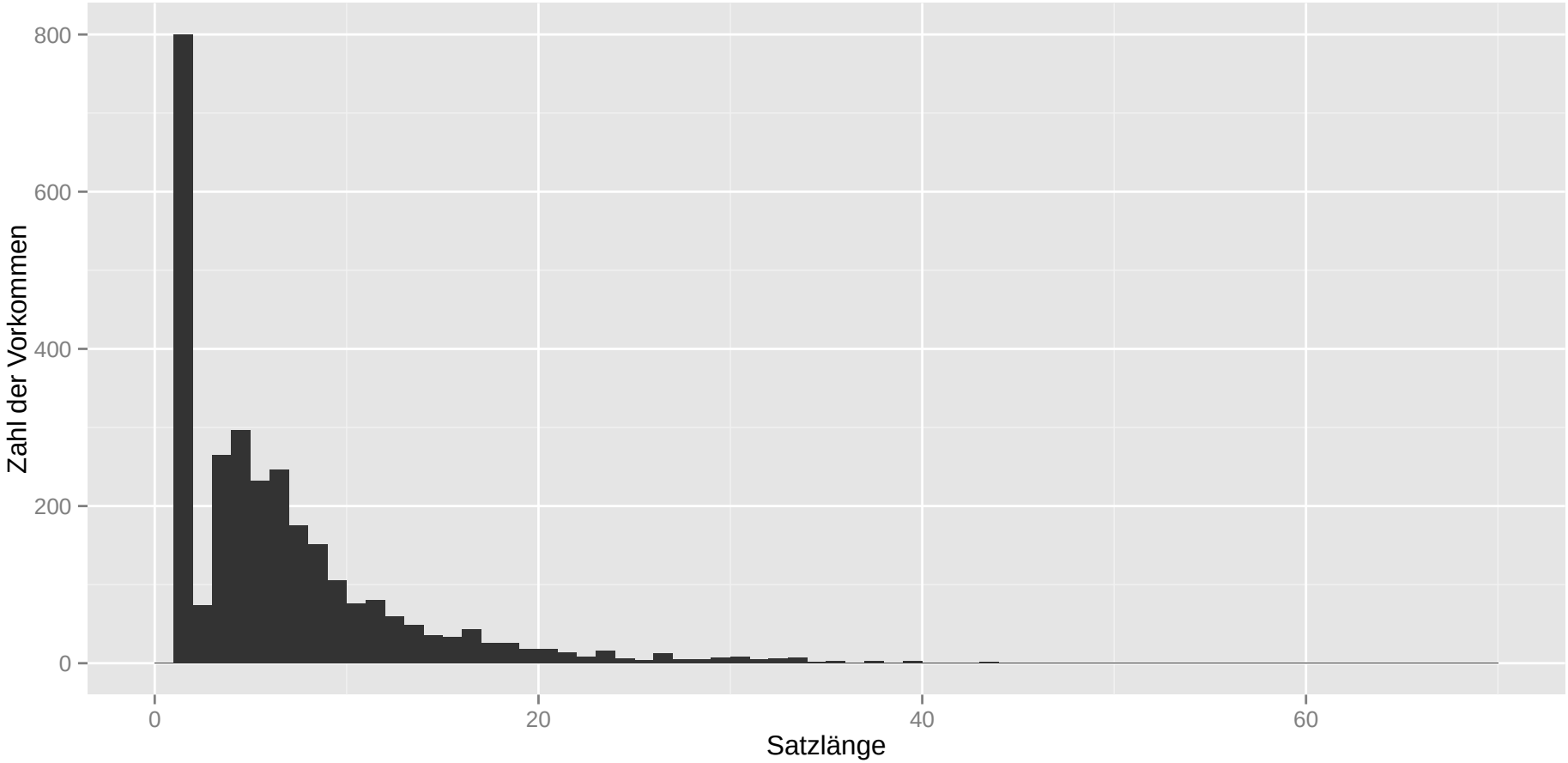


Abbildung: Satzlängenverteilung als Histogramm. Balkenbreite=1.

Der interessante Bereich nach Korpora

Nur kleine Satzlängen.

Der interessante Bereich nach Korpora

Nur kleine Satzlängen.

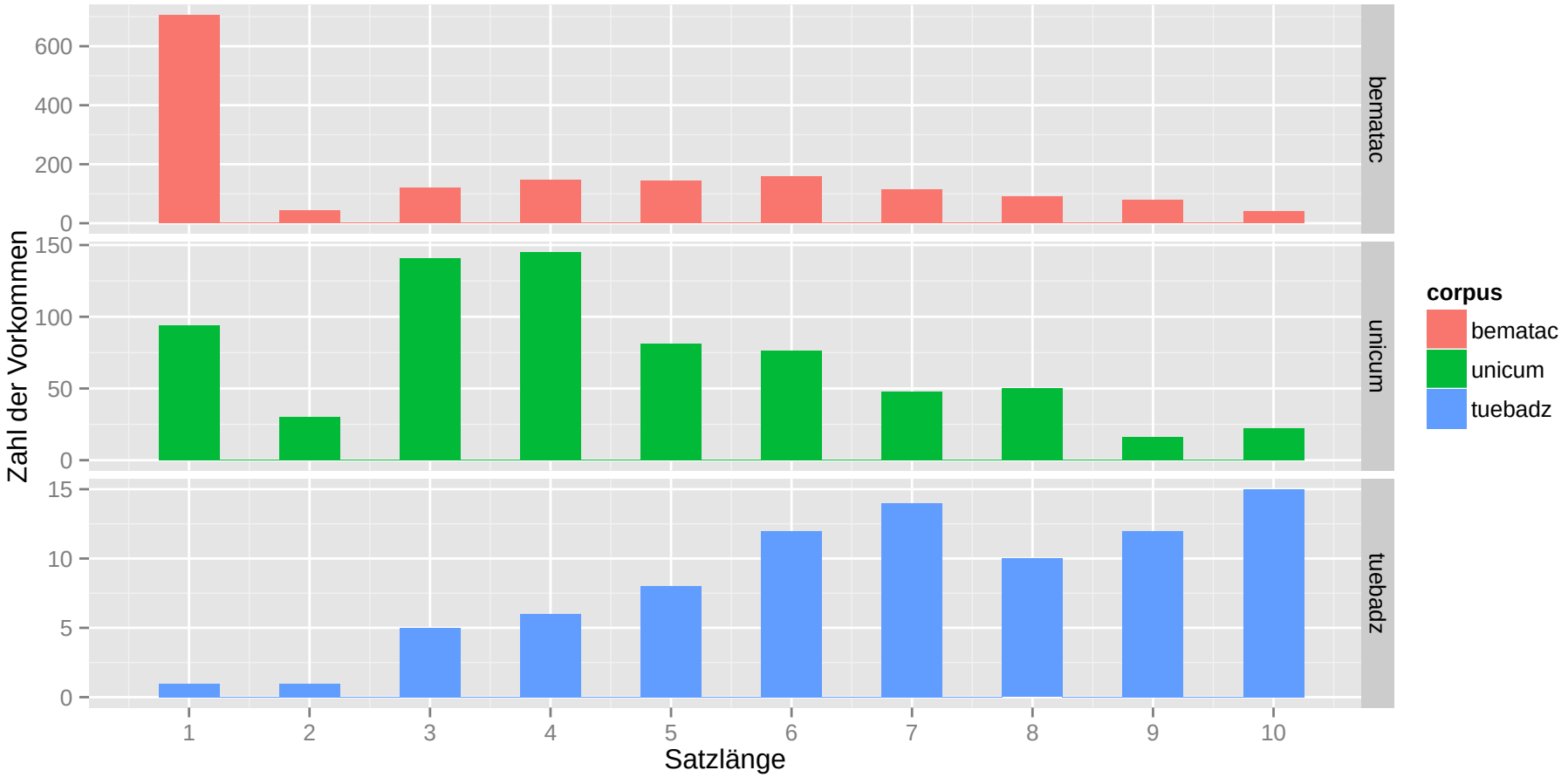


Abbildung: Korpora kodiert nach Farbe. tuebadz zum Vergleich.

Der interessante Bereich nach Korpora

Nur kleine Satzlängen. Bimodale Verteilung für bematac und unicum

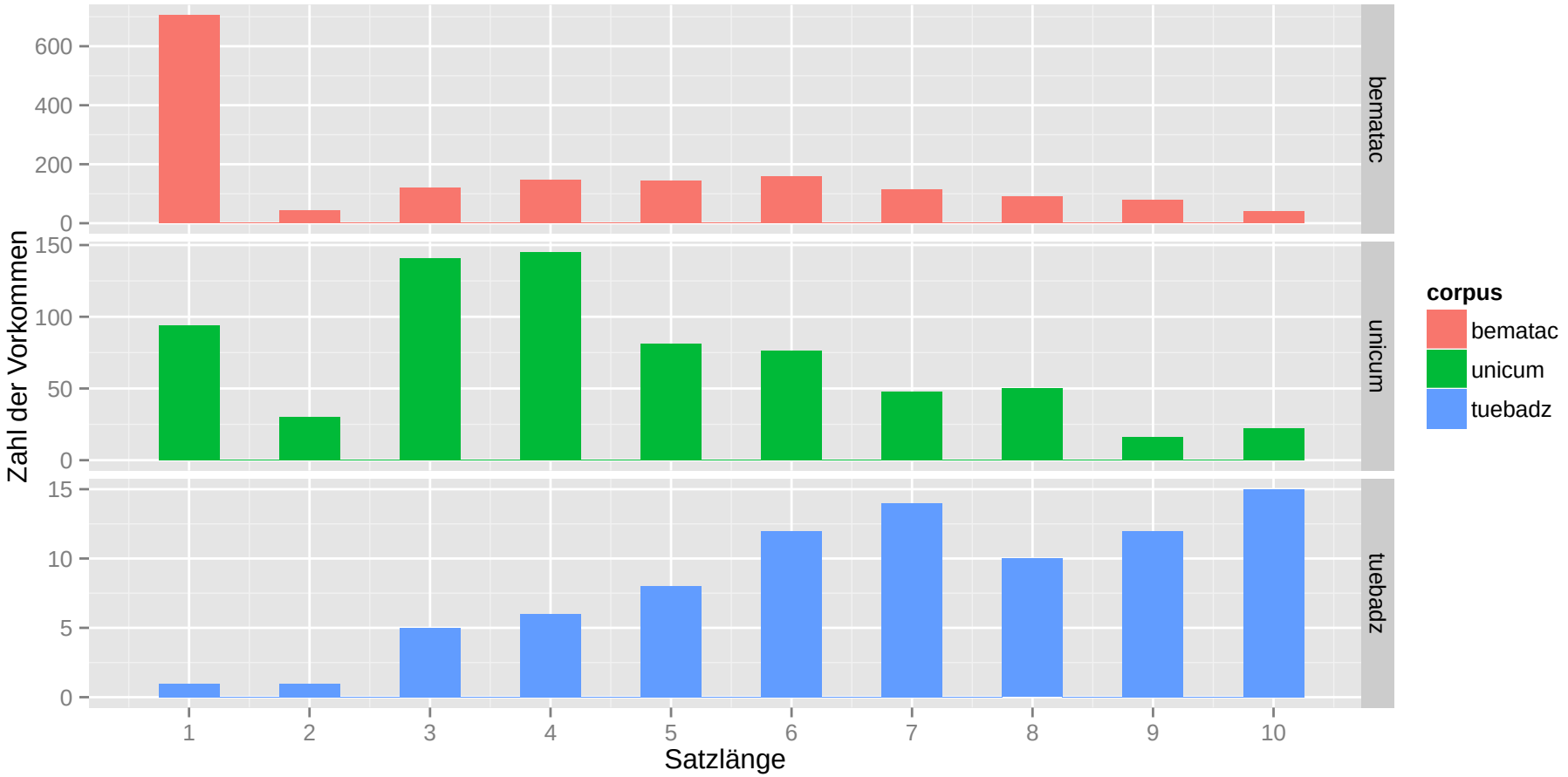


Abbildung: Korpora kodiert nach Farbe. tuebadz zum Vergleich.

Resultat der Exploration

Die gezeigten Graphiken legen Aufspaltung der Forschungsfrage nahe:

Forschungsfrage *A*

Unterscheiden sich in `bematac` und `unicum` die Anteile der Sätze der Länge 1?

Forschungsfrage *B*

Unterscheiden sich in `bematac` und `unicum` die Längen der **übrigen** Sätze?

Beantwortung Forschungsfrage *A*

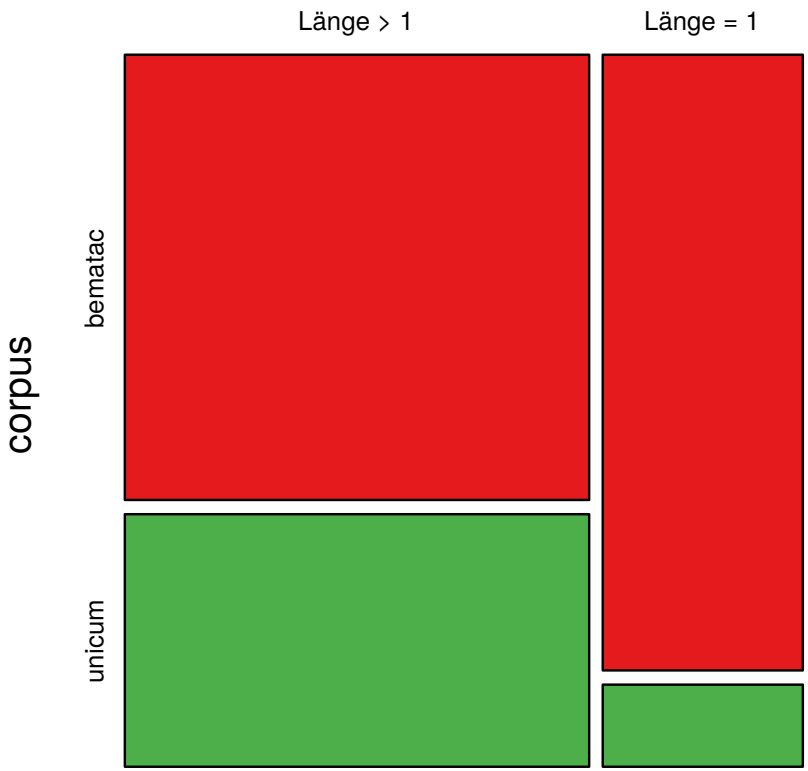


Abbildung: bematac hat wesentlich mehr Länge-1-Sätze als unicum.

Ein χ^2-Test sagt:

$$\chi^2 = 161.2$$
$$df = 1$$
$$p \approx 0$$

bematac	unicum
Ja	Nein
Okay	Bochum
Also	brml
Genau	Bye

- bematac: gesprochene Dialoge.
- unicum: geschriebene Dialoge.

Beantwortung Forschungsfrage *B*

Wir betrachten nur Sätze der Länge > 1.

- Nun kennen wir die Verteilung.
- Sie hat nur noch einen Peak (unimodal).
- *t*-Test ist anwendbar.

Ergebnis:

$$\overline{L}_{bematac} = 7.8 \ token$$
$$\overline{L}_{unicum} = 5.8 \ token$$
$$t = 9.49$$
$$df = 1852$$
$$p \approx 0 < 0.05$$

Auf einmal ein signifikanter Unterschied.

Zusammenfassung Satzlängenbeispiel

Wenn wir eine saubere Datenexploration durchführen, kommen wir

1. ... zu anderen Fragen.
2. ... auf ähnliche Fragen zu völlig anderen Ergebnissen.

Hintergrund zu Satzlängen in NoSta-D

Wie kann man 'Satzlängen' messen in Varietäten, die 'nichtkanonische' Strukturen (bspw. Äußerungen ohne finites Verb) enthalten? Was auch immer man entscheidet - es ist wichtig, dass für diese Vergleichsmessungen Sätze immer *auf dieselbe Art* definiert wurden.
Hier wurde in der Annotation eine 'kanonische' Struktur annotiert, nach der dann Sätze definiert und annotiert wurden. Die Zählungen sind aber für die entsprechenden Originaltexte.[4]

Original	also			quasi	einmal	über	das	ganze	Blatt	rüber
'Kanonische' Struktur	also	gehst	du	quasi	einmal	über	das	ganze	Blatt	rüber

[4] Für NoSta-D ist das Vorgehen ausführlich beschrieben unter (Dipper u. a. 2013; Dietterle u. a. eingereicht)

Varianz

Bisher haben wir zwei Korpora verglichen - die implizite Grundannahme war, dass jedes Korpus eine Stichprobe aus einer einzigen Grundgesamtheit ist. Das ist oft nicht gegeben: auch innerhalb eines Korpus kann es enorme Varianz geben. Man muss also auch prüfen, wie homogen die Daten innerhalb des Korpus sind (schön dazu auch Gries 2006).

Warum ist Varianz so wichtig?

Definition (Varianz)

Varianz ist der Mittelwert der quadrierten Abweichungen vom Mittelwert (so ungefähr zumindest).

$\Rightarrow$ Ein Maß für die Streuung.

Standardabweichung: $\sigma = \sqrt{Var}$

- Je größer die Streuung, desto größer die Abweichung vom Mittelwert.
- Umso eher sind große Abweichungen durch den Zufall zu erklären.
- Umso weniger braucht es einen *Effekt* zur Erklärung des Unterschieds.
- Grundlage jeden Hypothesentests:

$$\text{Abweichung vom Mittelwert} \Leftrightarrow \text{Varianz}$$

Erläuterung am Beispiel

- Beispieltext: *Alice in Wonderland*
- Das häufigste Wort: *the* (1639 *the*/27269 *token* = 0.06)
- Sollte doch gleichmäßig über den Text verteilt sein?!

Das Vorgehen

- Zerhacken des Textes in 272 Stücke á 100 Token.
- Jeweils Zählen der Vorkommen von *the*
- Zählen: Wie oft kamen in 100 Token 1, 2, 3, ..., 6, 10 *the* vor?
- Erstellen einer Graphik.

Das Ergebnis

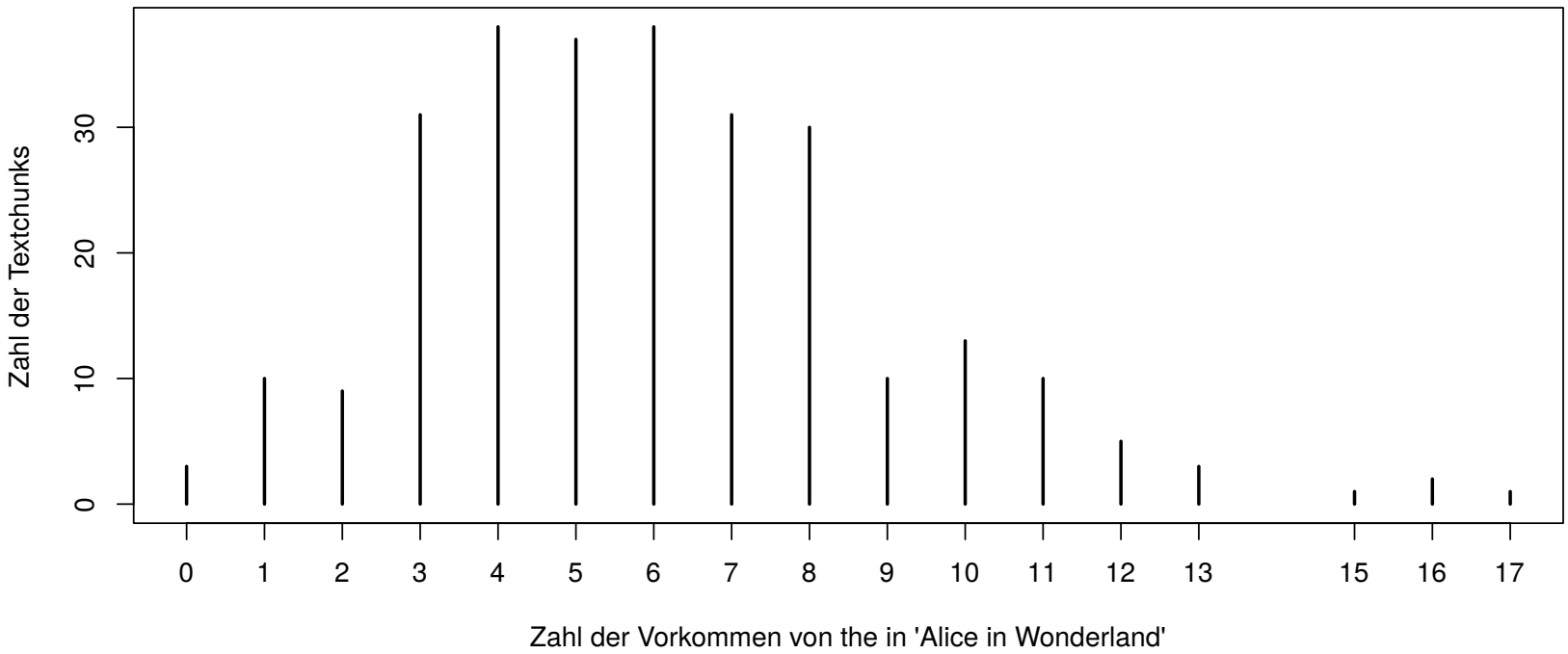


Abbildung: Vorkommensverteilung von *the*

Das Ergebnis

blau : Tatsächliche Standardabweichung in *Alice in Wonderland*

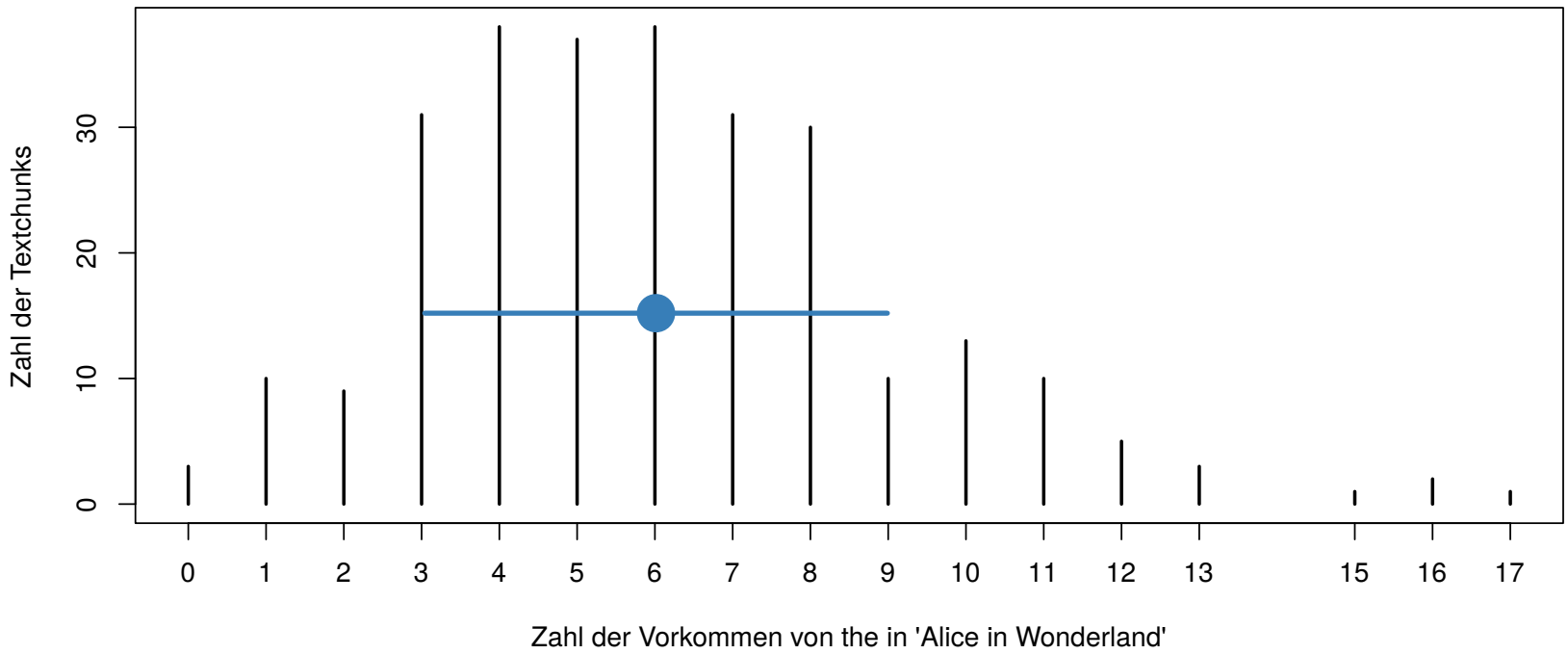


Abbildung: Vorkommensverteilung von *the*

Das Ergebnis

blau : Tatsächliche Standardabweichung in *Alice in Wonderland*

rot : Konfidenzintervall für die Standardabweichung bei Gleichverteilung

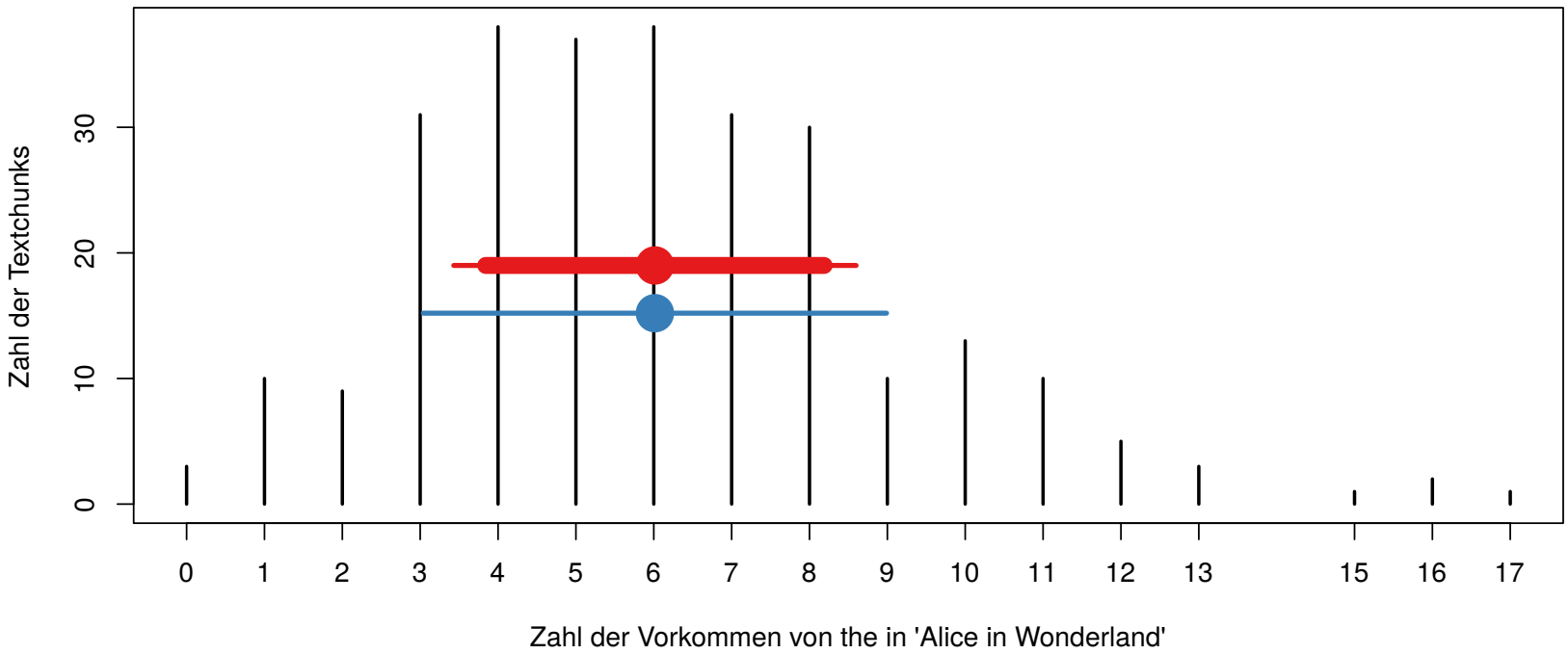


Abbildung: Vorkommensverteilung von *the*

Das Ergebnis

blau : Tatsächliche Standardabweichung in *Alice in Wonderland*

rot : Konfidenzintervall für die Standardabweichung bei Gleichverteilung

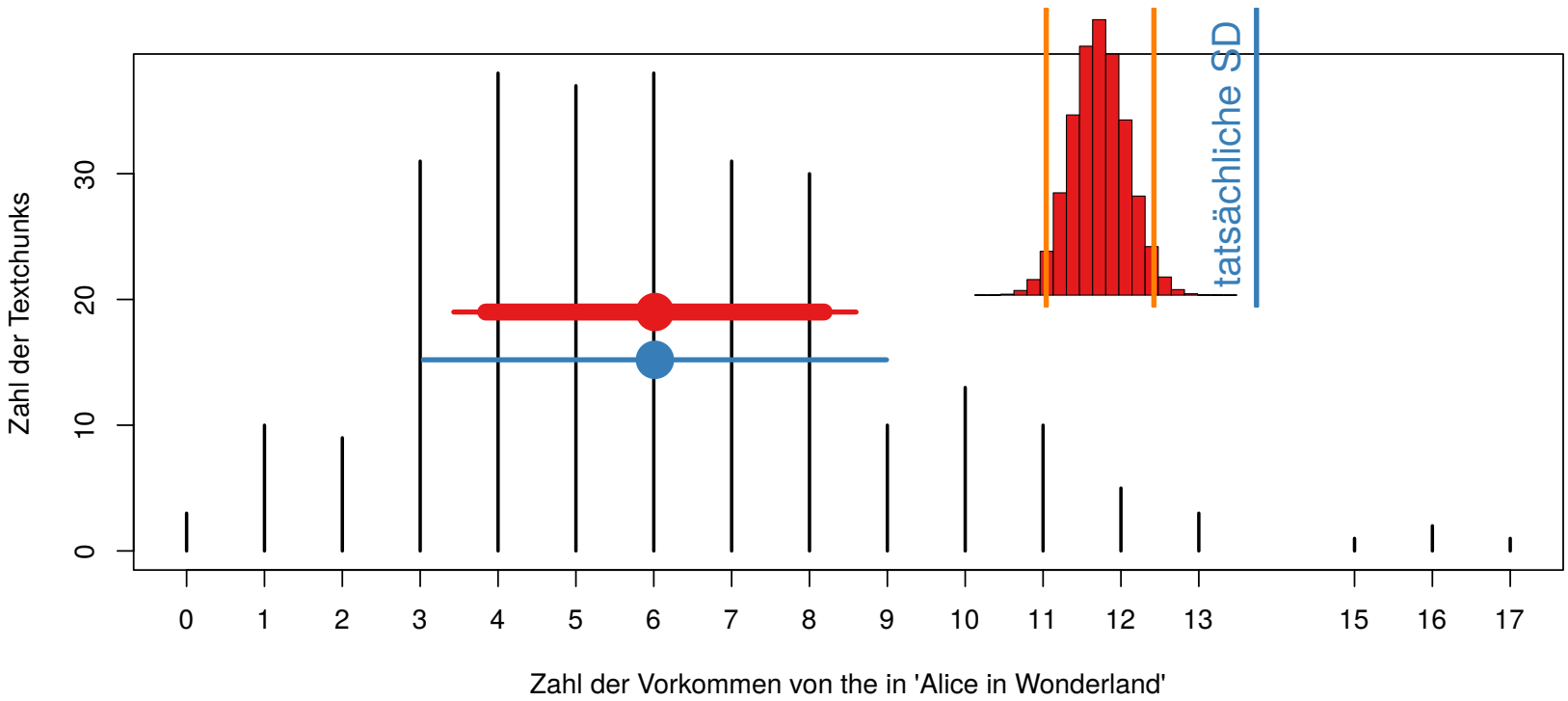


Abbildung: Vorkommensverteilung von *the*

Das Ergebnis

blau : Tatsächliche Standardabweichung in *Alice in Wonderland*

rot : Konfidenzintervall für die Standardabweichung bei Gleichverteilung

would change to tinkling sheep- bells and the queen s shrill cries to the voice of the shepherd boy and the sneeze of the baby the shriek of the gryphon and all the other queer noises would change she knew to the confused clamour of the busy farm-yard while the lowing of the cattle in the distance would take the place of the mock turtle s heavy sobs lastly she pictured to herself how this same little sister of hers would in the after-time be herself a grown woman and how she would keep through all her riper years the

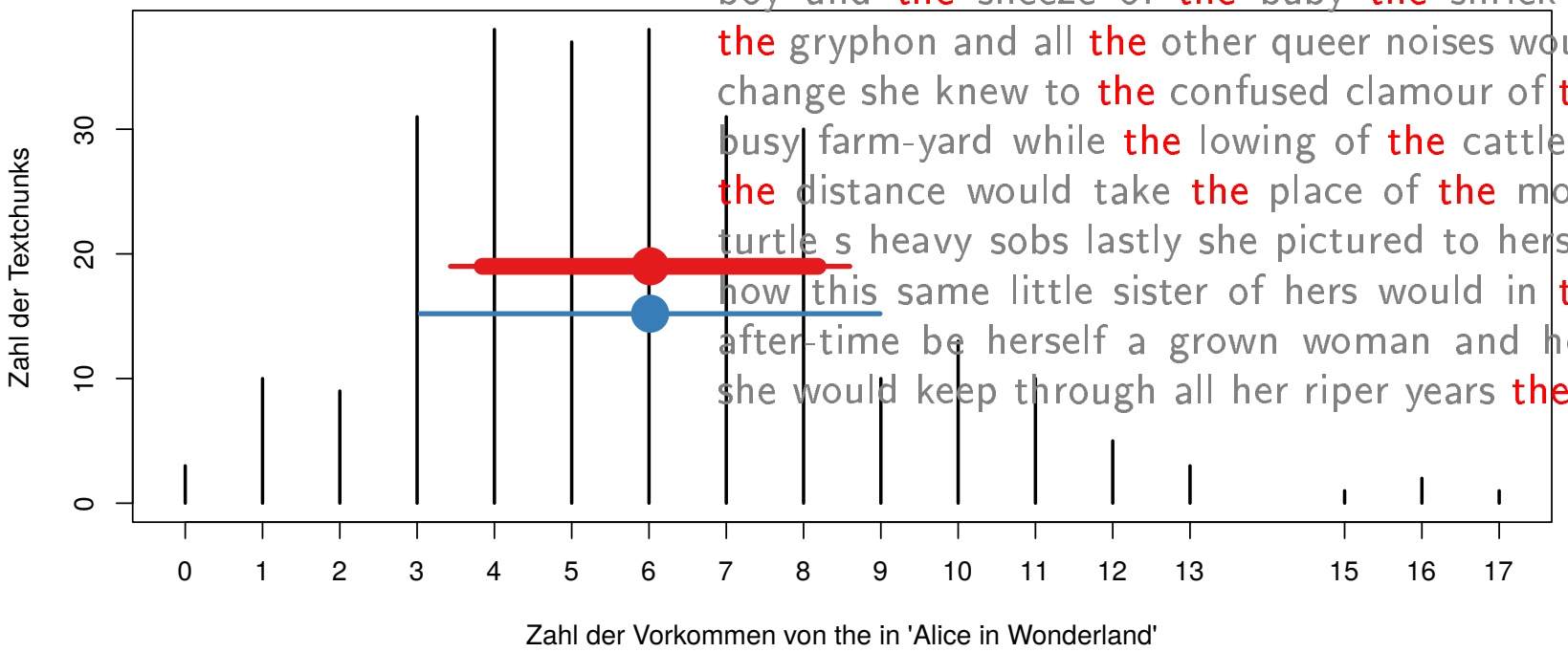


Abbildung: Vorkommensverteilung von *the*

Zusammenfassung Varianz

- Untersucht: Das häufigste Wort in einem einzigen Text.

! Zeigt mehr Varianz, als durch Gleichverteilung erklärbar.

- Wie ist das dann ...
  - ... selteneren Wörtern/Phänomenen
  - ... in verschiedenen Texten
  - ... von verschiedenen Schreibern/Sprechern
  - ... mit verschiedenen Muttersprachen,
  - ... die zu unterschiedlichen Themen schreiben?

$\Rightarrow$ Dieser Varianz muss auch im ausgeglichensten Korpus Rechnung getragen werden!

Ein anderes Beispiel

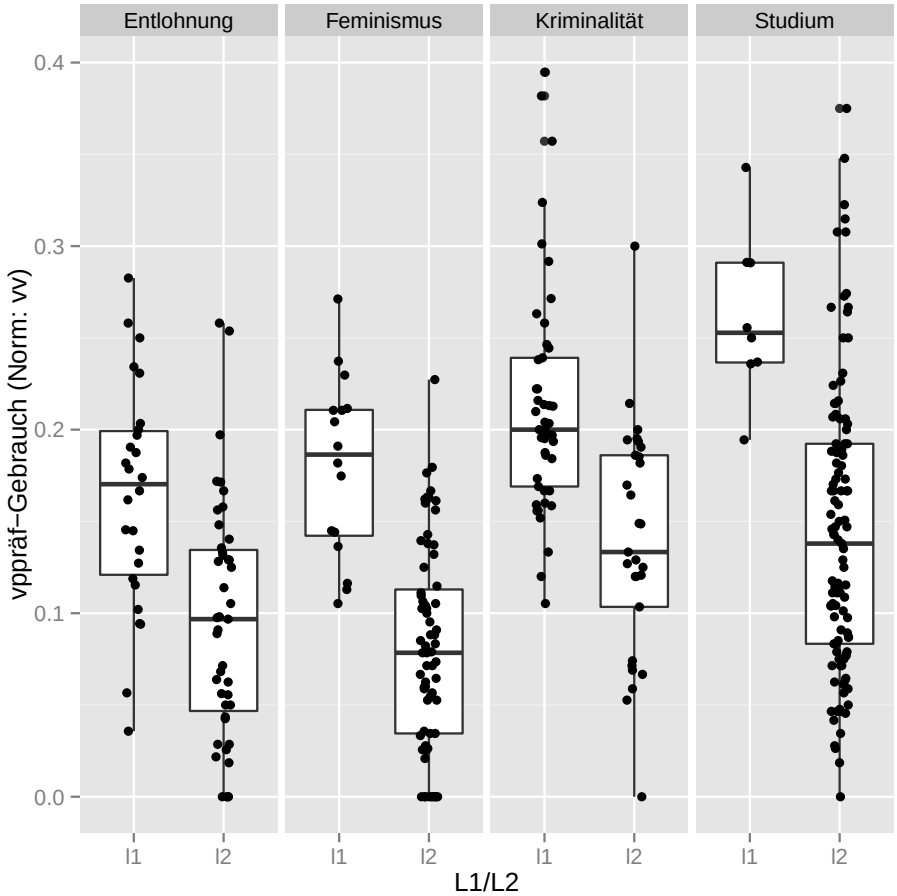
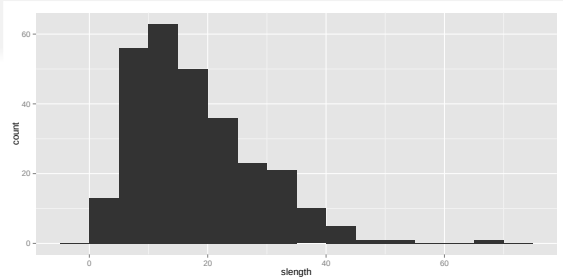


Abbildung: Wortartenverteilung in Falko (Reznicek u. a. 2010)

Nicht ganz normal

- Bisher explizit erwähnt: Normalverteilung.
- Bisher gesehen: keine Normalverteilung.

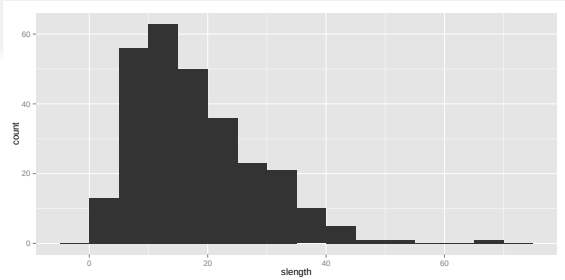
Nicht ganz normal



Satzlängen TüBa-D/Z

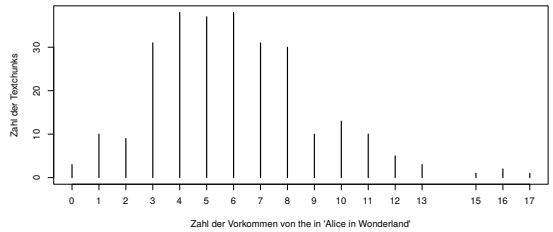
- Bisher explizit erwähnt: Normalverteilung.
- Bisher gesehen: keine Normalverteilung.

Nicht ganz normal



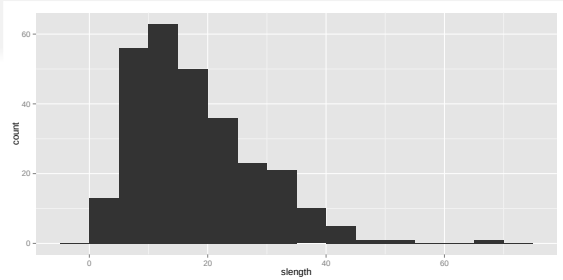
Satzlängen TüBa-D/Z

- Bisher explizit erwähnt: Normalverteilung.
- Bisher gesehen: keine Normalverteilung.



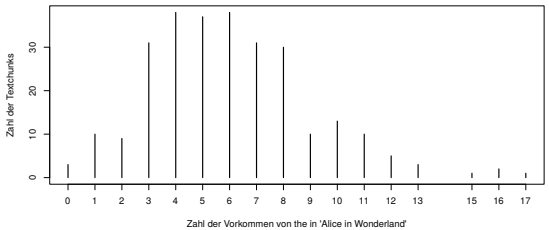
the-Verteilung Alice

Nicht ganz normal

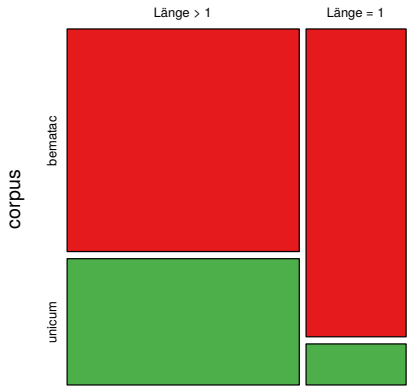


Satzlängen TüBa-D/Z

- Bisher explizit erwähnt: Normalverteilung.
- Bisher gesehen: keine Normalverteilung.



the-Verteilung Alice



kurze Sätze

Nach unten abgeschnitten

- Viele Zufallsgrößen beschränkt auf > 0
  - Satzlängen
  - Reaktionszeiten
  - Vorkommenzählungen
- Häufig (nicht immer und nie ganz) logarithmisch normalverteilt
  - Logarithmus zählt die Zahl der Nullen:
    - $\log 1 = 0$
    - $\log 100 = 2$
    - $\log 50 \approx 1.69897$
- Die Frage ist dann

  nicht mehr Wie viele Wörter länger ist ein Satz in Unicum?

  sondern Wie viel % länger ist ein Satz in Unicum?
- Diese Änderung macht oft sachlich großen Sinn.
- Man erhält eine (ungefähr) normalverteilte Größe
- Unterschiedliche Größenordnungen werden vergleichbar.

X Vorkommen in Y Fällen

- Wir hatten zwei solche Fragestellungen:
  - Wie viele von 100 Wörtern sind *the* in *Alice*?
  - Welcher Anteil der Sätze hat die Länge 1?
- Solche Größen nennt man binomial verteilt.
- Hier wird häufig der χ^2-Test angewandt.
  - Wir haben das auch gemacht. (Anteil Satzlängen 1)
- Nur unter wichtigen Prämissen erlaubt.
- Die wichtigste: Keine Cluster in den Daten.
  - Häufigste Cluster:
    - Korpuslinguistik: Texte.
    - Laborlinguistik: Versuchspersonen.
- Nichtbeachtung führt in diesem Fall zu erheblichen Fehlinterpretationen.

Noch einmal: Normierung

Was ist die maximal mögliche Zahl an Vorkommen?

- Wir würfeln 10 mal. 3 mal fällt die 6.
  - → 3 Vorkommen von 10 Möglichkeiten.
- Aber: Wie viele von 100 Sätzen können die Länge 1 haben?
  - → Sicher nicht alle.
- Fast immer ist die Normierungsgröße höchstens eine Näherung.
- Ist vielleicht die Poissonverteilung angemessener?
  - Wie viele Autos halten stehen während einer Rotphase an der Ampel?
  - Wie viele Atome zerfallen in einer Sekunde?
  - Wie viele Sätze eines Dialogs haben die Länge 1?

Signifikanz: Wiederholung des Konzeptes

1. Wir stellen 2 Hypothesen auf:
   - H_0: Es existiert kein wirklicher Effekt. Alle beobachteten Unterschiede sind Zufall.
     Die Satzlängen > 1 in bematac und unicum sind gleich.
   - H_1: Ein wirklicher Unterschied liegt vor.
     Sie sind es nicht.
2. Wir berechnen, wie wahrscheinlich der beobachtete Unterschied von H_0 aus ist.
3. Ist diese Wahrscheinlichkeit zu klein ($< 5\%$), weisen wir H_0 zurück. Wir sagen: „Ein signifikanter Unterschied liegt vor."

Kritik: Nicht das, was man möchte

Definition (p-Wert)

Der p-Wert ist die **bedingte** Wahrscheinlichkeit, dass die Abweichung von der Nullhypothese H_0 mindestens so groß ist wie die tatsächlich festgestellte Abweichung, unter der Voraussetzung, dass die Nullhypothese gilt.

- Nicht nur die Beschreibung ist kompliziert. Der Begriff selbst ist es!
- Uns interessiert oft eher: Wie wahrscheinlich sind H_0 oder H_1?
- Nicht ineinander überführbar, auch nicht ansatzweise.

  !!! Auch, wenn Sie das oft lesen, explizit oder implizit, es ist **falsch**!

$$P(H_0|\text{Wir kennen die Daten}) \neq P(\text{die Daten}|\text{Wir setzen } H_0 \text{ an})$$

Kritik: Ein Maß für die Größe der Daten.

- Irgendeinen Unterschied wird es immer geben:
  - → nicht einmal zwei Würfel sind wirklich identisch.
  - → H_0 ist immer falsch.
- Ist die Datenmenge groß genug, wird sich der Unterschied zeigen.
  - → man muss nur oft genug würfeln.
  - → dann ist p irgendwann unter 5%.
- Auch ein großer Unterschied in der Population wird sich nicht mit einer zu kleinen Datenmenge zeigen lassen.

p zeigt nicht an, ob der interessierende Effekt vorliegt, sondern nur, ob die Daten groß genug waren, p unter 5% zu drücken.

Kritik: 5% ist viel.

Ein Abbruchkriterium $\alpha = 0.05$, also

$$p < 0.05 \Rightarrow \textit{Signifikanz} \Rightarrow \textit{Zurückweisung } H_0$$

bedeutet, dass, falls H_0 gilt, jedes 20. Ergebnis signifikant ist.
Das ist sehr weit davon, dass man sagen könnte:

Zufall kann meine Daten in gar keinem Fall erklären.

Lösungsansatz 1: Sprachregelung

Statt

> *There is no difference between the relative clause usage of native speakers and learners.*

Besser

> ***Our data*** *showed no* ***statistically significant*** *difference between the relative clause usage of native speakers and learners.*

Das ist wenigstens eine korrekte Beschreibung.

Lösungsansatz 2: $p = 0$

- Falls p null ist, ist die Wahrscheinlichkeit, mit der H_0 die Daten produzieren würde, null.
- Damit kann H_0 die Daten nicht erklären.
- Damit ist H_0 **falsifiziert**.
- Das ist schonmal was...

In der Physik ganz gerne genommen: $\alpha = \frac{1}{1\,000\,000}$ oder kleiner.

Lösungsansatz 3: Effektstärke und Konfidenzintervall

Definition (Effektstärke)

Setzt den beobachteten Unterschied zur Standardabweichung der Daten in Bezug. Gibt es in 1001 Varianten.

Ein uninteressant kleiner Unterschied wird erkennbar, unabhängig von p.

Definition (Konfidenzintervall)

Bereich um den Mittelwert, so berechnet, dass er den wahren Wert mit einer gewissen Wahrscheinlichkeit (zb. 95%) enthält.

Vorteile

- Auf der selben Skala wie die Daten selbst.
  - → Graphisch darstellbar.
- Signifikanz direkt ablesbar.
- Es gibt schöne Filme zur Erläuterung:
  `https://www.youtube.com/watch?v=50L1RqHrZQ8`

Weiterführend: Multidimensionale Methoden

- Bisher: Immer nur Einfluss oder Verteilung einer Variable:
  - Zahl der *the* Vorkommen.
  - Zahl der Sätze der Länge 1
  - Länge der übrigen Sätze
- Oft eher: Was ist die Wahrscheinlichkeit, ...
  - ... dass der nächste Satz die Länge 1 hat
  - ... falls der letzte nicht die Länge 1 hatte,
  - ... der Sprecher der Instructee ist und
  - ... das Gespräch schon 3*min* dauert?

Vieles ist möglich

Es gibt Verfahren

- die beliebig viele (prinzipiell) Faktoren berücksichtigen können.
- Verschiedene Arten von Faktoren angemessen berücksichtigen
- Und mit verschiedenen Verteilungen umgehen können
  - Normalverteilung
  - Binomialverteilung
  - Poisssonverteilung
  - ... einige andere

S. zB. Baayen (2008), Zuur u. a. (2009) und Pinheiro und Bates (2000)

Zusammenfassung

Der korpusbasierte Vergleich von Varietäten lohnt sich! Man sieht viele Muster und Zusammenhänge, die man anders nicht finden kann. Dabei helfen statistische Verfahren und gute Visualisierungen.

Zusammenfassung

Um ein Phänomen in zwei (oder mehr) Korpora quantitativ zu vergleichen, muss man sich Gedanken machen über viele Aspekte. Hier konnten diese Aspekte nur angerissen werden (und es gibt noch mehr zu beachten).

Fragestellung Viele der Analyseschritte hängen von der Forschungsfrage ab. Daher ist es essentiell, die Forschungsfrage so präzise wie möglich zu stellen (gerne als Hypothese). Aber: Manchmal kann man die Forschungsfrage mit Korpusdaten nicht beantworten, dann muss man sie anpassen. Eine präzise Forschungsfrage soll aber nicht verhindern, dass man die Daten exploriert!

Stichprobenziehung Fast jedes Korpus ist eine Stichprobe aus einer Grundgesamtheit. Dabei muss man das Korpusdesign (d.h. die Parameter, nach denen die Stichprobe gezogen wurde) kennen und in die Analyse einbeziehen. Außerdem muss man wissen, was man aus den vorliegenden Daten schließen darf.

Zusammenfassung

Kategorisierung Die Daten müssen kategorisiert werden (d.h. jede quantitative Analyse beruht auf einer vorherigen qualitativen Analyse). Dabei ist zu beachten, dass Daten ganz unterschiedlich kategorisiert werden können und man Tagset, Richtlinien und Evaluation angeben sollte. Außerdem sollte die Kategorisierung immer *in den Daten* stattfinden ($\rightarrow$ Annotation) und mit veröffentlicht werden. Wenn Korpora verglichen werden sollen, müssen die Kategorien auf vergleichbare Weise vergeben worden sein.

Exploration Vor der eigentlichen Analyse muss das Korpus exploriert werden - das bedeutet zum einen Konsistenzchecks und zum anderen verschiedene quantitative Explorationen.

Zusammenfassung

Normalisierung Zählungen von Kategorien aus verschiedenen Korpora müssen normalisiert werden. Unterschiedliche Normalisierungsgrundlagen beantworten unterschiedliche Fragen.

Vergleiche Die fiesesten Fallen aus statistischer Sicht: unpassende Verteilung ($\rightarrow$ Exploration!) modelliert, Vernachlässigung von Clustern in den Daten.

Varianz Faustregel: Nichts in Korpora ist gleich(mäßig) verteilt. Die Varianz ist immer größer. Das muss mindestens für Hypothesentests beachtet werden.

Zusammenfassung - technisch

Replikation Die Replikation von Forschungsergebnissen ist ein wichtiger erster Schritt für die Wissenschaftlichkeit. Im zweiten Schritt sollte man dann versuchen, vorherige Ergebnisse an anderen Daten zu reproduzieren.

Transparenz und Offenheit Alle Ressourcen (Korpora mit Annotationen, Skripte, Programme) sollten nachhaltig veröffentlicht werden. Jeder Erstellungsschritt muss dokumentiert werden.

Wiederverwendbarkeit Alle Ressourcen sollten so aufgebaut sein, dass sie möglichst breit wiederverwendet werden können.

Allgemeine Einführungen in Statistik für Linguisten

Baayen 2008 Überblick zu statistischen Verfahren in verschiedensten Bereichen der Linguistik. Manchmal etwas unsystematisch.

Baayen 2008

R. Harald Baayen (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge University Press

Gries 2009a; Gries 2009b Systematischer. Zugänglicher. Gries (2009a) korpuslinguistischer. Geht statistisch nicht sehr weit.

Gries 2009a; Gries 2009b

Stefan Th. Gries (2009a). *Quantitative Corpus Linguistics with R. A Practical Introduction.* New York: Routledge

Stefan Th. Gries (2009b). *Statistics for linguistists with R. A Practical Introduction.* Mouton de Gruyter

Weitere Literatur

- Einführungen in die Korpuslinguistik/Artikel zu grundlegenden Themen in der Korpuslinguistik Kübler und Zinsmeister (2014), Lemnitzer und Zinsmeister (2010), Lüdeling und Kytö (2008) und Lüdeling und Kytö (2009); es gibt inzwischen auch viele Handbücher zu korpuslinguistischen Methoden in bestimmten Bereichen
- Zur ersten Datenaufbereitung ist Biber und Jones (2009) hilfreich
- zu Stichprobenziehung und Verteilung siehe zum Beispiel Baroni und Evert (2007), Biber (1993), Evert (2006) und Kilgarriff (2005)
- zur Modellierung siehe die Einführung von Gries (2012)
- zu multidimensionalen Vergleichen und Vergleichsstudien siehe zum Beispiel Biber (2009), Grieve u. a. (2011) und Szmrecsanyi (2011)

Literatur I

Artstein, Ron und Massimo Poesio (2008). „Inter-Coder Agreement for Computational Linguistics“. In: *Computational Linguistics* 34.4, S. 555–596. ISSN: 0891-2017.

Baayen, R. Harald (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baroni, Marco und Stefan Evert (2007). „Words and Echoes: Assessing and Mitigating the Non-randomness Problem in Word Frequency Distribution Modeling“. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, S. 904–911.

Beißwenger, Michael (2013). *Das Dortmunder Chat-Korpus: ein annotiertes Korpus zur Sprachverwendung und sprachlichen Variation in der deutschsprachigen Chat-Kommunikation*. URL: LINSE:%20Linguistik%20Server%20Essen:%20%5Curl%7Bhttp://www.linse.uni-due.de/tl_files/PDFs/Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf%7D.

Literatur II

- Biber, Douglas (1993). „Representativeness in Corpus Design“. In: *Literary and Linguistic Computing* 8.4, S. 243–257.
- – (2009). „Multi-dimensional Approaches“. In: *Corpus Linguistics. An International Handbook*. Hrsg. von Anke Lüdeling und Merja Kytö. Berlin: Mouton de Gruyter, S. 822–855.
- Biber, Douglas und Susan Conrad (2009). *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press, S. 1–188.
- Biber, Douglas und James K. Jones (2009). „Quantitative Methods in Corpus Linguistics“. In: *Corpus Linguistics. An International Handbook*. Hrsg. von Anke Lüdeling und Merja Kytö. Bd. 2. Berlin: Mouton de Gruyter, S. 1286–1304.
- Carletta, Jean (1996). „Squibs and Discussions. Assessing Agreement on Classification Tasks: The Kappa Statistic“. In: *Computational linguistics* 22.2, S. 249–254.

Literatur III

Cohen, Jacob (1960). „A coefficient of agreement for nominal scales“. In: *Educational and Psychological Measurement* 20.1, S. 37–46.

Dietterle, Burkhard, Anke Lüdeling und Marc Reznicek (eingereicht). „Zur Syntax von Plauderchats“. In: *Empirische Erforschung internetbasierter Kommunikation*. Hrsg. von Michael Beißwenger.

Dipper, Stefanie, Anke Lüdeling und Marc Reznicek (2013). „NoSta-D: A Corpus of German Non-Standard Varieties“. In: *Non-Standard Data Sources in Corpus-Based Research*. Hrsg. von Marcos Zampieri. Shaker Verlag. URL: https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/nosta-d.

Evert, Stefan (2006). „How Random is a Corpus? The Library Metaphor“. In: *Zeitschrift für Anglistik und Amerikanistik* 54.2, S. 177–190.

Fleiss, Joseph L. (1971). „Measuring nominal scale agreement among many raters“. In: *Psychological Bulletin* 76.5, S. 378–382.

Literatur IV

Gamer, Matthias, Jim Lemon und Ian Fellows Puspendra Singh (2012). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84. URL: http://CRAN.R-project.org/package=irr.

Gries, Stefan Th. (2006). „Exploring variability within and between corpora: some methodological considerations“. In: *Corpora* 1.2, S. 109–151.

– (2009a). *Quantitative Corpus Linguistics with R. A Practical Introduction*. New York: Routledge.

– (2009b). *Statistics for linguistists with R. A Practical Introduction*. Mouton de Gruyter.

– (2012). „Statistische Modellierung“. In: *Zeitschrift für Germanistische Linguistik* 40.1, S. 38–67.

– (2015). „Some Current Quantitative Problems in Corpus Linguistics and a Sketch of Some Solutions“. In: *Language and Linguistics* 16.1, S. 93–117.

Literatur V

- Grieve, Jack, Douglas Biber, Eric Friginal und Tatiana Nekrasova (2011). „Variation Among Blogs: A Multi-dimensional Analysis“. In: *Genres on the Web*. Springer, S. 303–322.
- Hilpert, Martin und Stefan Th. Gries (erscheint). „Quantitative Approaches to Diachronic Corpus Linguistics“. In: *The Cambridge Handbook of English Historical Linguistics*. Hrsg. von Merja Kytö und Päivi Pahta. Cambridge: Cambridge University Press.
- Hinrichs, Marie, Hrsg. (2014). *TüBa-D/Z Release 9.1 (12/2014)*. URL: http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html (besucht am 01. 03. 2015).
- Kilgarriff, Adam (2005). „Language is never, ever, ever, random“. In: *Corpus Linguistics and Linguistic Theory* 1.2, S. 263–275.
- – (2012). „Getting to know your corpus“. In: *Text, Speech and Dialogue*. Springer.

Literatur VI

- Kübler, Sandra und Heike Zinsmeister (2014). *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury.
- Labov, William (2008). „Quantitative Reasoning in Linguistics“. In: *Linguistics* 563.
- Lemnitzer, Lothar und Heike Zinsmeister (2010). *Korpuslinguistik: Eine Einführung*. 2., durchges. und aktualis. Aufl. Narr-Studienbücher. Tübingen: Narr. ISBN: 9783823365556.
- Lüdeling, Anke (2011). „Corpora in Linguistics: Sampling and Annotation“. In: *Going Digital. Evolutionary and Revolutionary Aspects of Digitization.* Hrsg. von Karl Grandin. Nobel Symposium 147. New York: Science History Publications/USA, S. 220–243.
- Lüdeling, Anke, Hagen Hirschmann und Amir Zeldes (2011). „Variationism and Underuse Statistics in the Analysis of the Development of Relative Clauses in German“. In: *Corpus Analysis and Diachronic Linguistics.* Bd. 1. Kawaguchi, Yuji; Minegishi, Makoto; Viereck, Wolfgang. John Benjamins, Amsterdam., S. 37–57.

Literatur VII

Lüdeling, Anke und Merja Kytö, Hrsg. (2008). *Corpus Linguistics: An International Handbook*. Bd. 1. Berlin: Mouton De Gruyter.

– Hrsg. (2009). *Corpus Linguistics: An International Handbook*. Bd. 2. Berlin: Mouton De Gruyter.

Moisl, Hermann (2009). „Exploratory Multivariate Analysis“. In: *Corpus Linguistics. An International Handbook*. Hrsg. von Anke Lüdeling und Merja Kytö. Bd. 2. Mouton De Gruyter, S. 874–899.

Pinheiro, José C. und Douglas M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Berlin, Heidelberg: Springer.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Literatur VIII

- Reznicek, Marc, Maik Walter, Karin Schmid, Anke Lüdeling, Hagen Hirschmann, Cedric Krummes und Thorsten Andreas (2010). *Das Falko-Handbuch: Korpusaufbau und Annotationen*. Version 1.0.1. Humboldt-Universität zu Berlin – Institut für deutsche Sprache und Linguistik – Korpuslinguistik.
- RStudio (2011). *manipulate: Interactive Plots for RStudio*. R package version 0.98.976.
- Sauer, Simon, Hrsg. (2013). *BeMaTac. Ein tief annotiertes multimodales Map-Task-Korpus gesprochener Lerner- und Muttersprache*. URL: `http://u.hu-berlin.de/bematac` (besucht am 01. 03. 2015).
- Szmrecsanyi, Benedikt (2011). „Corpus-based dialectometry: a methodological sketch“. In: *Corpora* 6.1, S. 45–76.
- Wickham, Hadley (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.
- – (2011). „The Split-Apply-Combine Strategy for Data Analysis“. In: *Journal of Statistical Software* 40.1, S. 1–29.

Literatur IX

Xie, Yihui (2014). „knitr: A Comprehensive Tool for Reproducible Research in R“. In: *Implementing Reproducible Computational Research*. Hrsg. von Victoria Stodden, Friedrich Leisch und Roger D. Peng. Chapman und Hall/CRC.

Zuur, Alain F., Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev und Graham M. Smith (2009). *Mixed Effects Models and Extensions in Ecology with R*. Berlin, Heidelberg: Springer.