Annotation

Medienlinguistische Methodik

Arne Rubehn
arne.rubehn@uni-passau.de

25.11.2025

Zusammenfassung

In dieser Sitzung werden zunächst einige statistische Gesetzmäßigkeiten von linguistischen Korpora beleuchtet, insbesondere das Zipfsche Gesetz. Danach wird die Annotation von wissenschaftlichen Daten behandelt, wobei typische Verfahren und Zwecke erörtert werden.

1 Zipfsches Gesetz

Das **Zipfsche Gesetz**, benannt nach George K. Zipf, besagt, dass die Frequenz eines Wortes in einem Korpus invers proportional zu seinem Rang steht, wenn die Wörter absteigend anhand ihrer Frequenz in einem Korpus angeordnet werden (Zipf, 1935). Weiß man also, auf welchem Rang ein Wort steht, kann man dessen Frequenz in Abhängigkeit zu der des häufigsten Wortes abschätzen – und umgekehrt auch. Eine sehr vereinfachende Faustformel wäre, dass das zweithäufigste Wort in einem Korpus 1/2 mal so häufig vorkommt wie das häufigste, das dritthäufigste 1/3 mal so häufig, usw.. Dadurch ergibt sich eine exponentielle Verteilung, die sich in einem Balkendiagramm typischerweise als „L-Form“ zeigt: Links sind wenige, hochfrequente Wörter zu sehen; dem folgt ein langer „Schwanz“ (*long tail*) and vielen, sehr seltenen Wörtern (Abb. 1, links). Ein mathematischer „Trick“ kann eine solche Verteilung zu einer linearen Korrelation transformieren, indem beide Werte (also sowohl Frequenz, als auch Rang) im Logarithmus genommen werden. Intuitiv bedeutet das, dass nicht mehr die absolute Zahl, sondern die Größenordnung gemessen wird (im Zehnerlogarithmus ist der Abstand zwischen 1, 10, 100, 1000 jeweils gleich groß) – die Exponentialfunktion wird praktisch „umgekehrt“, wodurch sich eine lineare Korrelation ergibt (Abb. 1, rechts). Da sich lineare Funktionen deutlich einfacher berechnen und modellieren lassen, ist eine solche logarithmische Transformation (*double-log* bzw. *log-log*) typisch für die Analyse von Daten, die dem Zipfschen Gesetz folgen.

Das Zipfsche Gesetz gilt auch für viele natürlich auftretende Verteilungen außerhalb der Linguistik. Die Größe von Städten, die Häufigkeit von Segmenten in Tierkommunikation oder die Umsatzverteilung in Unternehmensbeziehungen sind alles Beispiele, bei denen das Zipfsche Gesetz bereits beobachtet wurde. Ein verwandter Effekt ist das Pareto-Prinzip (Pareto, 1897), auch bekannt als 80-20-Regel, demzufolge 80% der Menge an Objekten 20% der Masse ausmachen (z.B. verteilen sich 20% der Bevölkerung auf 80% der Städte und Gemeinden).

Eine verwandte, aber nicht identische Gesetzmäßigkeit ist das **Zipfsche Gesetz der Abkürzung** (*Zipf's Law of Abbreviation*). Demnach sind hochfrequente Wörter tendenziell kürzer als seltene Wörter. Dieser Effekt ist über sämtliche Sprachen hinweg (Bentz & Ferrer-i Cancho, 2016, vgl. Abb. 2) stabil und trifft auch auf z.B. Ortsnamen zu (Haspelmath, 2024).

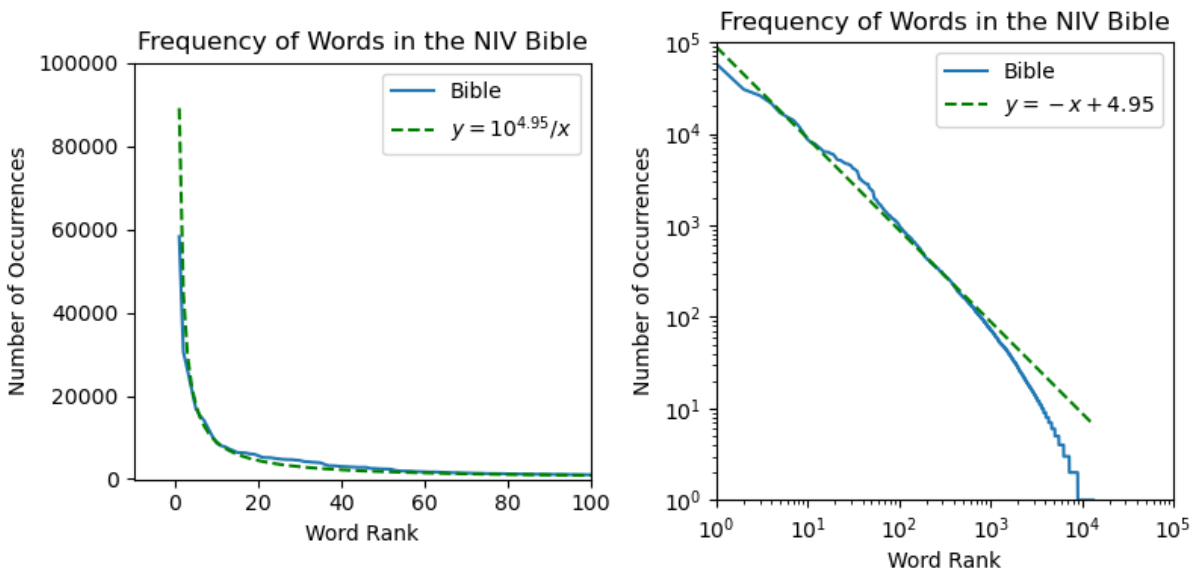


Abbildung 1: Beispiel für das Zipfsche Gesetz, vor (links) und nach (rechts) logarithmischer Transformation beider Achsen.

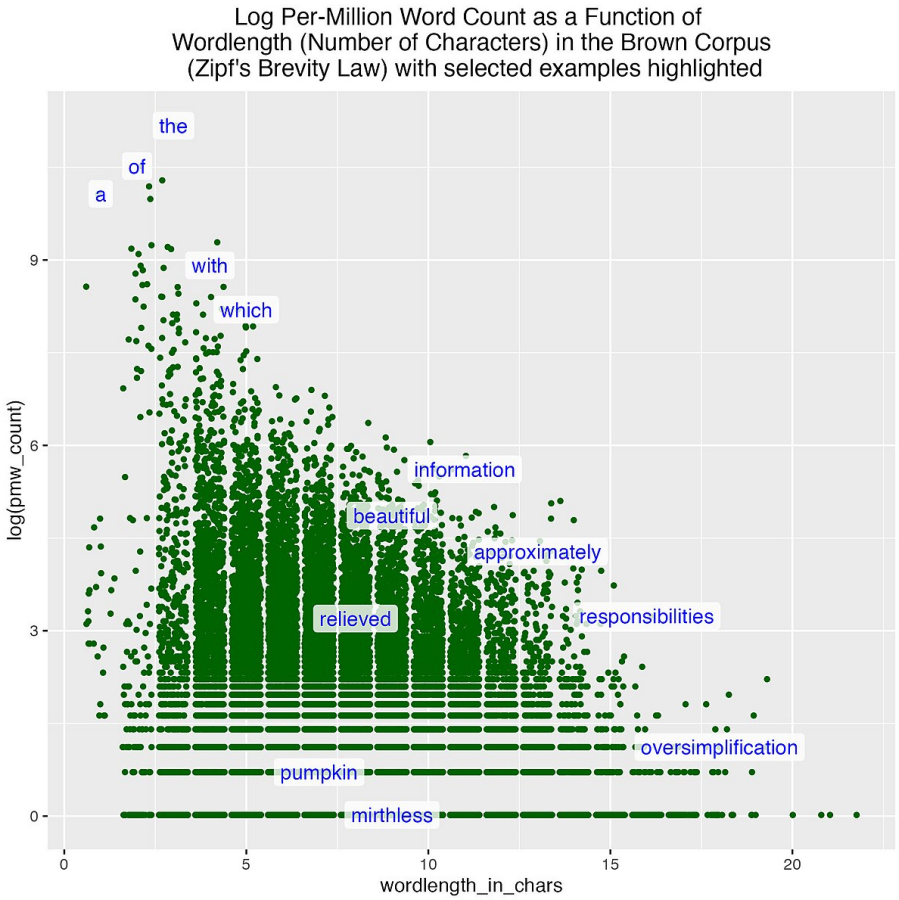


Abbildung 2: Illustration des Zipfschen Gesetz der Abkürzung.

2 Annotation im Allgemeinen

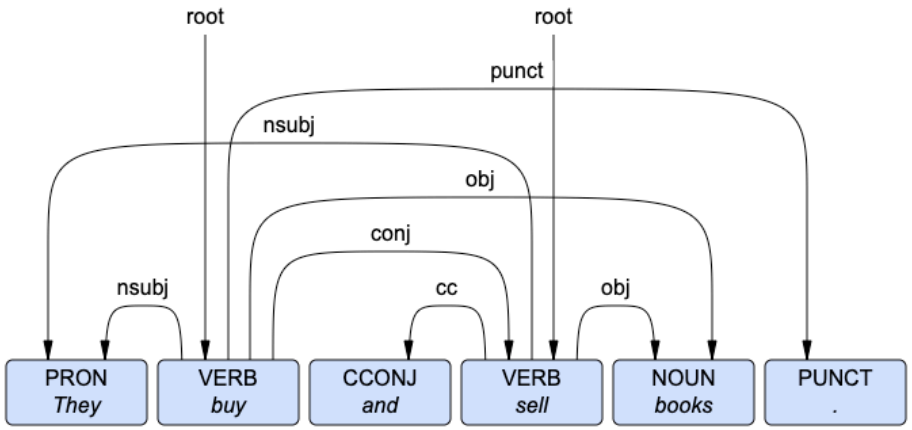
Unter Annotation kann man ganz allgemein das Anreichern oder Strukturieren von Information verstehen, wobei die Information in Form von Daten vorliegt (Horstmann & Seltmann, 2019, 2/8). Im weiten Sinne zählen handgeschriebene Notizen am Rand von Texten genauso zur Annotation wie typische Kommentarstile für die Edition klassischer Werke in lateinischen oder chinesischen Texten, oder der Hinweis auf der Rückseite eines Fotos, das angibt, wer dort zu sehen ist, wann, und wo.

In einem engeren Sinne würden wir vor dem Hintergrund von computergestützten Ansätzen in den Geisteswissenschaften unter Annotation vor allem die Anreicherung und Strukturierung von Information verstehen, die selbst maschinenlesbar ist, also in digitaler Form vorliegt, und der eine einheitliche Semantik zugrundeliegt. Es geht also um das gezielte Erstellen von **strukturierten Daten**, durch die implizite Informationen explizit gemacht werden. In diesem Sinne wird meistens die eigentliche **Annotation**, also das Anreichern der Daten, von den **Metadaten**, also Daten über die Daten, unterschieden.

Man kann sich die Idee, die hinter der Annotation in der Wissenschaft steckt, im Prinzip als augmented reality vorstellen. Wenn man die Annotationsbrille aufsetzt, dann werden all die Objekte um einen herum plötzlich automatisch beschriftet und klassifiziert, und bei Interesse kann man bei einem Objekt innehalten, in der virtuellen Realität die Beschreibung anklicken und sich ähnliche Objekte anzeigen lassen. Die Idee, die hinter der Annotation steckt, ist es, die Welt messbar zu machen, ganz wie in dem bekannten Satz „Measure what is measurable, and make measurable what is not so", der oft fälschlicherweise Galileo Galilei (1564-1641) zugeschrieben wird (Kleinert, 2009).

Der Wunsch, Dinge messbar machen zu wollen, begegnet uns nicht nur in der Wissenschaft, sondern in vielen Bereichen unseres Lebens. Wir messen unser Alter, wir messen unsere Körpergröße, man misst unsere Sehkraft, und all diese Messungen haben direkte Konsequenzen, da sie entscheiden, ob wir in die Schule gehen dürfen oder Bundespräsident werden dürfen, oder ob wir ohne Brille Auto fahren dürfen. In der Wissenschaft spielt das Messbarmachen aber eine entscheidende Rolle, da wir ohne dieses viele Untersuchungen gar nicht würden durchführen können.

Die Annotation im Allgemeinen, die man als den Versuch ansehen kann, bestimmte Aspekte der Welt mit Semantik anzureichern (was sowohl symbolisch als auch numerisch erfolgen kann, „messbar machen" muss sich hierbei nicht auf „zählbar machen" beschränken) kann daher als eine der grundlegendsten Methoden der Wissenschaft angesehen werden, auch wenn die Annotation in der Wissenschaftstheorie kaum als so bedeutsam angesehen wird. Wie bereits erwähnt, ist ein zentraler Aspekt das gezielte Erstellen von strukturierten Daten, dem normalerweise eine gezielte Erstellung eines Datenmodells vorausgeht, welches entscheidet, wie Daten verwendet werden, um die Untersuchungsgegenstände zu repräsentieren.




```
# sent_id = 1
# text = They buy and sell books.
1   They     they    PRON    PRP    Case=Nom|Number=Plur              2   nsubj   2:nsubj|4:nsubj   _
2   buy      buy     VERB    VBP    Number=Plur|Person=3|Tense=Pres   0   root    0:root            _
3   and      and     CCONJ   CC     _                                 4   cc      4:cc              _
4   sell     sell    VERB    VBP    Number=Plur|Person=3|Tense=Pres   2   conj    0:root|2:conj     _
5   books    book    NOUN    NNS    Number=Plur                       2   obj     2:obj|4:obj       SpaceAfter=No
6   .        .       PUNCT   .      _                                 2   punct   2:punct           _
```

Abbildung 3: Ein syntaktisch annotierter Satz im Projekt Universal Dependencies (de Marneffe et al., 2021), mit Illustration der syntaktischen Abhängigkeiten.

3 Annotation in computergestützten Ansätzen

Schränken wir den Annotationsbegriff auf computergestützte Ansätze ein, geht es in erster Linie darum, einen Untersuchungsgegenstand semantisch anzureichern, indem wir eine maschinenlesbare Zuordnung von Information vornehmen. In den meisten Fällen bezieht sich die Annotation dabei direkt auf die Annotation von Texten (die wir als Zeichensequenzen verstehen können, was dann auch erlauben würde, die Annotation von Gen- und Proteinsequenzen als Textannotation anzusehen). Darüber hinaus spielt aber auch die Annotation von Bildern (Bildannoation, image annotation), Videos und Audiodateien eine große Rolle, die inzwischen über die rein wissenschaftliche Anwendung weit hinausreicht.

Da Annotation unsere Untersuchungsgegenstände mit Information anreichert, ist es wichtig, zu wissen, wie wir die Untersuchungsgegenstände strukturieren. Bei Texten und anderen Formen von Untersuchungsgegenständen, die in sequentieller Form repräsentiert werden können, können wir eine Segmentierung vornehmen, auf die wir unsere Annotation beziehen. So können wir zum Beispiel sagen, dass wir die Annotation von einem Text auf der Ebene der Wörter ansetzen (vgl. Abb. 3). Während dies bei Sprachen wie dem Deutschen oder dem Englischen relativ einfach ist, da hier Wörter bis auf wenige Ausnahmen ja auch durch die Schrift repräsentiert werden, ist das bei chinesischen Texten weniger leicht zu bewerkstelligen, da im Chinesischen Wörter nicht explizit im Text dargestellt werden. Ein Schriftzeichen ist eine Silbe und denotiert meist ein Morpheme (also eine bedeutungstragende Einheit der Sprache), viele Wörter sind aber Komposita, weshalb man in einem solchen Falle zunächst eine Segmentierung vornehmen müsste.

Bei Bildern und Videos wird die Annotation noch komplizierter, ganz zu schweigen von

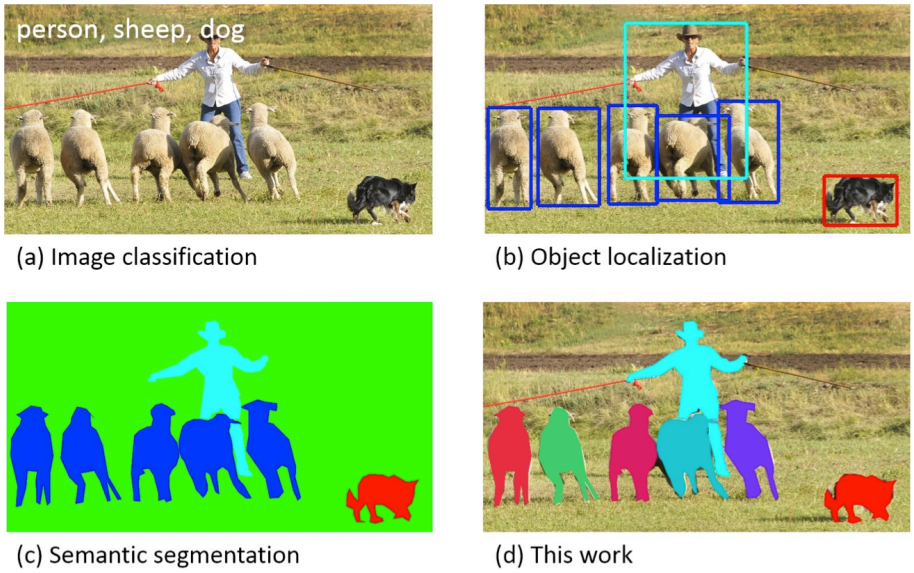


Abbildung 4: Beispiel einer annotierten Bilddatei aus dem Datensatz COCO (Lin et al., 2015).

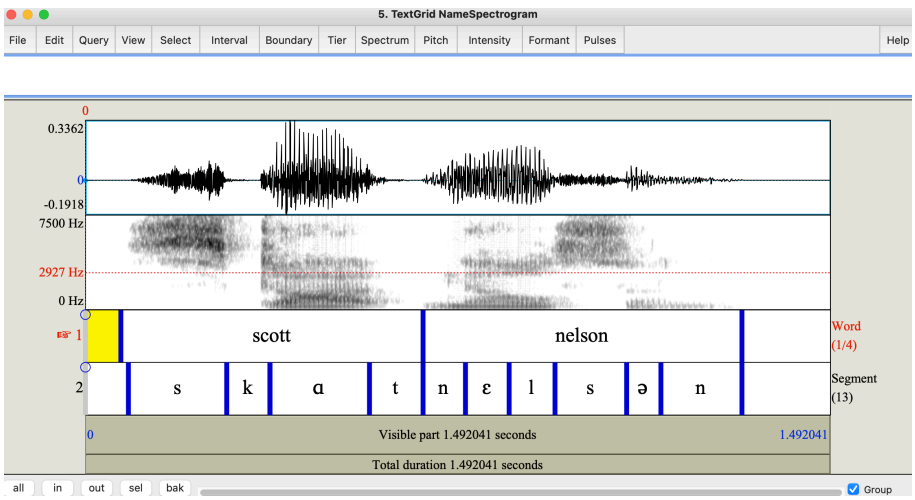


Abbildung 5: Beispiel einer annotierten Audiodatei mit der Software Praat (Boersma & Weenink, 2025).

dreidimensionalen Artefakten, die zunächst aufwendig digitalisiert werden müssen. Wenn wir jedoch von Bildern ausgehen, dann ist die Grundlage meist ein Polygon (meist ein Rechteck, in bestimmten Fällen aber ein komplexer Ausschnitt, der die komplette Kontur eines Objekts in einem Mehreck nachzeichnet), welches ein Objekt auf dem Bild verortet, welches in einem weiteren Schritt annotiert werden soll (vgl. Abb. 4). Bei der Videoannotation besteht die einfachste Annotation aus einer Untertitelspur, welche die gesprochen Sprache im Video als Text wiedergibt und zeitlich mit dem Video aliniert ist. In der Linguistik gibt es des weiteren sehr komplexe Annotationen von Videos oder Texten, die als Aufnahme vorliegen, die versuchen, einzelne Sprachlaute in phonetischer Transkription mit der Lautdatei zu alinieren, wofür typischerweise Programme wie *Praat* (Boersma & Weenink, 2025; Abb. 5) oder *ELAN* (ELAN, 2025) verwendet werden.

Wie wir sehen, können Daten anhand von Annotationen um unterschiedlichste Zusatzinformationen erweitert werden, wobei je nach Modalität der Primärdaten und Ziel der Annotation verschiedene Repräsentationen notwendig sind. Sehr häufig werden bei Annotationen Teile der Primärdaten kategorisiert, also einem vordefinierten Inventar an Tags zugeordnet (daher auch Tagging genannt). Ebenfalls zu sehen ist, dass das selbe Objekt gleichzeitig auf verschiedenen Ebenen annotiert werden kann, um verschiedene Informationen parallel zu repräsentieren.

Die Annotation von Daten kann grob gesagt zu zwei Zwecken erfolgen. Einerseits können Zusatzinformationen für Leser:innen bereitgestellt werden, die das Verständnis der Daten erleichtern. Andererseits werden für das Training von Computermodellen häufig größere Mengen an hochstrukturierten und annotierten Daten benötigt. In beiden Fällen gilt der Grundsatz, dass eine korrekt ausgeführte Annotation niemals die Primärdaten ändern darf (Hirschmann, 2019, 29).

4 Evaluation von Annotationen

Um gute Annotationen zu erzeugen, werden wissenschaftliche Daten typischerweise von mehreren Annotator:innen bearbeitet. Hierbei sollten die Annotationen der unterschiedlichen Personen möglichst gut übereinstimmen. Je nach Art der Annotation und Anzahl der Annotator:innen gibt es verschiedene statistische Maße, um das **Inter-Annotator-Agreement** zu bestimmen. Ein beliebtes Maß ist hierbei *Cohen's Kappa*, das die Übereinstimmung zwischen zwei Annotator:innen bei Klassifikationsaufgaben mit verschiedenen Kategorien misst. Annotationsprojekte verlaufen typischerweise zyklisch in mehreren Durchläufen: Jedem Durchlauf liegen Guidelines zugrunde, an denen sich die Annotator:innen orientieren. Nach erfolgter Annotation wird die Übereinstimmung zwischen den Annotator:innen ausgewertet. Hierbei sollen vor Allem Muster entdeckt werden, die für Unklarheit (und dadurch geringer Übereinstimmung) in der Annotation sorgen. Basierend

darauf werden die Annotationsguidelines angepasst, um Zweifelsfälle klarer zu gestalten; teilweise können sogar mögliche Kategorien angepasst werden, sollte sich beispielsweise herausstellen, dass sich zwei Kategorien nicht klar voneinander trennen lassen. Dieser Zyklus wird dann so lange wiederholt, bis das Annotationsschema ausreichend klar ist und verschiedene Annotator:innen die Daten möglichst übereinstimmend annotieren.

Literatur

Bentz, C. & Ferrer-i Cancho, R. (2016). *Zipf's law of abbreviation as a language universal*. Tübingen: Universitätsbibliothek Tübingen.

Boersma, P. & Weenink, D. (2025). *Praat [software, v.6.4.47]*. Amsterdam: University of Amsterdam. Zugriff auf `https://www.fon.hum.uva.nl/praat/`

de Marneffe, M.-C., Manning, C., Nivre, J. & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, *47* (2), 255-308.

ELAN. (2025). *ELAN [Software, v.7.0]*. Nijmegen: Max Planck Institute for Psycholinguistics, the Language Archive. Zugriff auf `https://archive.mpi.nl/tla/elan`

Haspelmath, M. (2024). *Trading off informativeness and length in lexicon and grammar*. Geneva: Zenodo. doi: https://zenodo.org/records/10958622

Hirschmann, H. (2019). *Korpuslinguistik: Eine Einführung*. J.B. Metzler.

Horstmann, J. & Seltmann, M. E.-H. (2019). Annotation. In A. D. H. T. des Verbandes Digital Humanities im deutschsprachigen Raum e. V. (Hrsg.), *Begriffe der Digital Humanities. Ein diskursives Glossar* (S. 1-8). Wolfenbüttel: Zeitschrift für digitale Geisteswissenschaften.

Kleinert, A. (2009). Der messende Luchs. Zwei verbreitete Fehler in der Galilei-Literatur. *NTM Zeitschrift für Geschichte der Wissenschaften, Technik und Medizin*, *17* (2), 199-206.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollár, P. (2015). *Microsoft coco: Common objects in context*. Zugriff auf `https://arxiv.org/abs/1405.0312`

Pareto, V. (1897). *Cours d'Économie politique* (Bd. 2). Lausanne & Paris: Rouge & Pichon.

Zipf, G. K. (1935). *The psychobiology of language*. New York: Houghton-Mifflin.