Cyril Belica

Von absoluten Häufigkeiten zum Differenzenkoeffizienten

Für die Veranschaulichung der Gebrauchshäufigkeit von Wörtern in zeitlich gegliederten Korpora können Tabellen mit absoluten Zahlen verwendet werden. Die Zellen geben Auskunft über die absolute Anzahl der Vorkomnisse eines Wortes im Teilkorpus für das jeweilige Jahrzehnt:

40er	50er	60er	70er	80er	90er	
100	258	368	115	530	667	*weiß*
36	97	192	79	313	220	*schwarz*

Da aber die zugrundeliegende Textmenge für jedes Jahrzehnt unterschiedlich ist, ist die Interpretation der Zahlen schwierig. Noch problematischer wäre ein Versuch, die absoluten Gebrauchshäufigkeiten eines Wortes in zwei verschiedenen Korpora oder mit unterschiedlichen Sampling-Intervallen (mit unterschiedlicher Zeitgliederung) direkt miteinander zu vergleichen.

Eine teilweise Abhilfe stellt die Verwendung von relativen Häufigkeiten dar:

40er	50er	60er	70er	80er	90er	
0,000270%	0,000268%	0,000324%	0,000219%	0,000402%	0,000449%	*weiß*
0,000097%	0,000101%	0,000169%	0,000150%	0,000237%	0,000148%	*schwarz*

Nun kann man zwar die Entwicklungstendenz eines Wortes einigermaßen gut aus der Tabelle erschließen, doch der Vergleich zwischen den zwei Wörtern ist immer noch schwierig. Zum Beispiel geht aus der Tabelle immer noch nicht klar hervor, ob sie Ähnlichkeiten im zeitlichen Verlauf ihrer Gebrauchshäufigkeit aufweisen.

Ein weiterer Schritt könnte sein, die relativen Häufigkeiten aus der Tabelle in graphischer Form zu veranschaulichen. Diese Vorgehensweise hätte aber zwei große Nachteile. Erstens, da es sowohl sehr häufige wie auch sehr seltene Wörter gibt, müssten wir in solchen Graphiken relative Häufigkeiten darstellen, die sich voneinander um viele Größenordnungen unterscheiden. Zweitens, relative Häufigkeiten bieten keine Bezugsbasis für Bewertungen wie z.B. "Wort X ist im Teilkorpus (Jahrzehnt) Y überproportional belegt". Und eben solche Bewertungen sind uns in diesem Zusammenhang wichtig. Damit wir aber etwas als "überproportional" bezeichnen können, müssen wir erst "das Normale" kennen.

Bei der Suche nach einem besseren Maß für die Veranschaulichung von Gebrauchshäufigkeiten wollen wir auf das wohl bekannteste Beispiel der elementaren Statistik zurückgreifen: auf das Würfeln.

Beispiel

> Wir haben 600 mal gewürfelt und die erzielten Punktezahlen aufgeschrieben. Wir wollen, ohne in unsere Aufzeichnungen zu schauen, folgende Frage möglichst genau beantworten: "Wie oft ist zwischen dem einundsechszigsten und dem einhundertzwanzigsten Versuch die Sechs (hier "Treffer" genannt) gefallen?" Da wir annehmen, dass bei 600 Versuchen die Sechs etwa 100 mal gefallen ist, erwarten wir auch, dass sie bei den 60 Versuchen (61. bis 120.) etwa zehnmal beobachtet wurde. Wir bitten dann jemand, die Anzahl der tatsächlich erzielten Treffern (Punktezahl 6) zwischen dem 61. und 120. Versuch aus den Aufzeichnungen zu ermitteln. Wird uns eine Zahl viel kleiner oder viel größer als zehn genannt, so werden wir diese **Feststellung** intuitiv als **überraschend** empfinden, gemessen an unserer **Erwartung**. Wir setzen also unbewusst die Ergebnisse einer Beobachtung in Verhältnis zu unseren Erwartungen.
>
> Was ist aber in diesem Experiment eigentlich unsere Erwartung? Um das herauszufinden, bitten wir, dass uns noch die tatsächliche Zahl der Treffer im gesamten Experiment genannt wird (wir hatten mit etwa 100 Treffern gerechnet). Nach der Auskunft, dass in unseren Aufzeichnungen insgesamt nur 90 Treffer belegt sind, wollen wir sicherlich unsere Antwort auf die gestellte Frage präzisieren. Da in 600 Versuchen insgesamt nur 90 Treffer beobachtet wurden, erwarten wir in einem Zehntel der Versuche (61. bis 120.) etwa ein Zehntel der Treffer, also etwa neun. Entsprechend der nun geänderten Erwartung ändert sich auch unsere Bewertung der tatsächlich beobachteten Anzahl Treffer zwischen dem 61. und 120. Versuch. "Nur sieben Treffer" ist keine so große Überraschung mehr wie früher: wir wissen ja, dass es insgesamt nur 90 Treffer gab. Ähnlich könnten wir sagen, dass es wahrscheinlich ist, dass es unter den letzten 300 Versuchen etwa 45 Treffer gab, proportional zur Größe der Teilmenge.

Wie können wir nun diese Überlegungen auf das Zählen von Wörtern in diachronisch gegliederten Texten übertragen? Wir betrachten jedes Textwort eines Korpus als eine Punktezahl eines Würfels (als hätten wir einen

"Würfel" mit sovielen Seiten, wieviele unterschiedliche Wörter es in unserem Korpus gibt). Das Korpus entspricht dann unseren Aufzeichnungen über den Verlauf des Würfelns und der Rechner übernimmt die Rolle unseres Assistenten, indem er unsere Fragen über die Ergebnisse beliebiger Beobachtungen (d.h. tatsächliche Anzahl der Vorkommnisse) anhand der Aufzeichnungen (d.h. des Korpus) beantwortet.

Zuerst lassen wir die Gesamtzahl der Vorkommnisse eines konkreten Wortes im Korpus feststellen. Wir wissen, dass diese Zahl massgeblich ist für unsere Erwartung darüber, wieviele Beobachtungen dieses Wortes in einem beliebigen Teilkorpus gemacht würden (analog zu unserer Erwartung, innerhalb von 60 Versuchen etwa 9 Treffer anzutreffen, wenn es in 600 Versuchen insgesamt 90 Treffer gab). Wir **erwarten** nämlich in jedem Teilkorpus einen der Größe des Teilkorpus proportionalen Anteil aller Vorkommnisse. Da wir aber in der Lage sind, auch die **tatsächliche** Anzahl der Vorkommnisse des Wortes in allen Teilkorpora durch einfaches Zählen zu ermitteln, können wir - ähnlich wie beim Würfeln - diese "Beobachtung" ins Verhältnis zu unserer Erwartung setzen. Die resultierende Größe gibt an, inwiefern die Erwartung durch die Beobachtung bestätigt wurde.

Sei N die Anzahl Textwörter im Korpus K und N_i die Anzahl Textwörter in einem Teilkorpus K_i von K. Sei F die Anzahl Vorkommnisse des Wortes W in K. Die erwartete Häufigkeit des Wortes W in K_i ist dann

$$f_e = F \frac{N_i}{N}$$

Zum Beispiel war unsere Erwartung im o.g. Experiment $f_e=90*(60/600)=9$ Treffer. Sei f_o die Anzahl tatsächlicher Vorkommnisse des Wortes W im Teilkorpus K_i. Dann kann die Größe

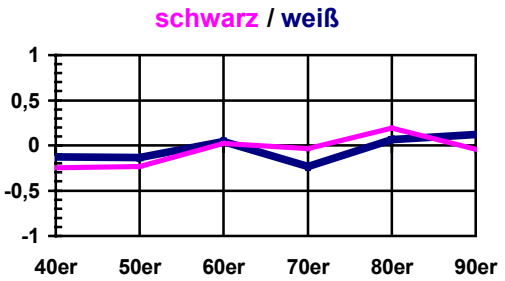
$$d = \frac{f_o}{f_e}$$

als Maß für die Übereinstimmung der Beobachtung mit unserer Erwartung verwendet werden. Zum Beispiel erhielten wir für 7 Treffer den Wert von $d=7/9=0.777$ und für 10 Treffer den Wert von $d=10/9=1.111$. Für die mit der Erwartung übereinstimmende Häufigkeit von 9 Treffern ergibt sich $d=9/9=1$. Das ist eine angenehme Feststellung, denn wir wünschen, dass d intuitiv leicht nachvollziehbare Werte annimmt und wir empfinden den Wert "eins" als gut geeignet zur Beschreibung einer vollkommenen Übereinstimmung. Man sieht auch, dass für Häufigkeiten kleiner/größer als erwartet der Wert von d sinkt/steigt. Wir wollen noch zwei Extremfälle untersuchen. Für nichtbelegte Wörter erhalten wir den Wert $d=0/f_e=0$, was ebenfalls als "intuitiv gut" bezeichnet werden kann. Wörtern, die in einem Teilkorpus viel öfter vorkommen als erwartet, wird hingegen ein Wert von d zugeordnet, der keine obere Grenze hat und theoretisch ins Unendliche steigen kann. Das erschwert die Handhabung und Interpretation von d, insbesondere wenn Werte aus unterschiedlich großen Teilkorpora miteinander verglichen werten sollen. In der Mathematik wird in solchen Fällen ein Trick namens Normalisierung angewandt. Er ändert nichts an der "Funktionsweise" einer Größe, sondern verschiebt lediglich deren Werte auf eine besser geeignete Skala. So wollen auch wir statt $d=f_o/f_e$ die normalisierte Größe

$$D = \frac{f_o - f_e}{f_o + f_e}$$

als Differenzenkoeffizienten für die Veranschaulichung der Gebrauchshäufigkeit in verschiedenen Teilkorpora eines Korpus verwenden. Wir prüfen noch die Extremfälle auf ihre Werte: für nichtbelegte Wörter ($f_o=0$) erhalten wir $D=(0-f_e)/(0+f_e)=-1$, im Falle einer Übereinstimmung ($f_o=f_e$) ist $D=(f_e-f_e)/(f_e+f_e)=0$ und für Häufigkeiten, die viel größer sind als erwartet ($f_o>>f_e$), nähert sich der Wert des Differenzenkoeffizienten gegen eins. Negative Werte bedeuten, dass das Wort in dem untersuchten Teilkorpus weniger frequent ist als sich aufgrund seiner Gesamthäufigkeit erwarten liesse und positive Werte von D deuten auf dessen überproportionale Vertretung im untersuchten Teilkorpus hin. Die Werte von D aus verschiedenen Untersuchungen können miteinander verglichen werden, da der Differenzenkoeffizient die unterschiedlichen Ausgangsparameter (Größe des Korpus und des Teilkorpus, absolute Worthäufigkeiten) auf einen "gemeinsamen Nenner" bringt.

Wie lässt sich mithilfe des Differenzenkoeffizienten die Gebrauchshäufigkeit der Wörter *schwarz* und *weiß* aus unserem obigen Beispiel graphisch darstellen? Nach der Berechnung von D für alle Jahrzehnte erhalten wir:

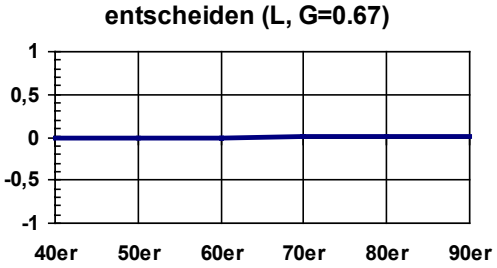


Wie aus dem Graph hervorgeht, weisen beide Wörter generell einen ähnlichen zeitlichen Verlauf ihrer Gebrauchshäufigkeiten auf.

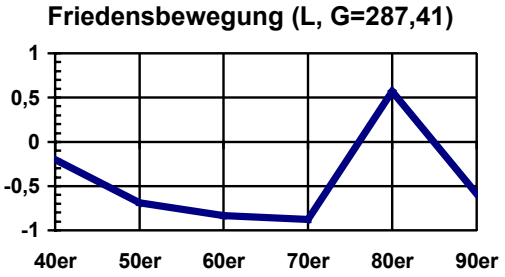
Aus der Tabelle

40er	50er	60er	70er	80er	90er	
93	249	284	138	344	393	*entscheiden*

erhalten wir



Die Gebrauchshäufigkeit des Wortes *entscheiden* ist in den zugrundeliegenden Korpora sehr stabil, die Belege sind proportional auf alle Jahrzehnte verteilt (der *D*-Wert ist überall fast gleich null). Die Graphik



verrät, dass das Wort *Friedensbewegung* im gesamten Zeitraum belegt ist (kein *D*-Wert ist gleich minus eins), dass es in den 80er Jahren viel häufiger gebraucht wurde als davor und dass seine Gebrauchshäufigkeit in den 90er Jahren wieder sehr deutlich gesunken ist.

Anmerkung.

Wir verwenden den Differenzenkoeffizienten zur **Veranschaulichung** der Gebrauchshäufigkeiten von Wörtern in zeitlich gegliederten Korpora, **nicht** zur statistischen **Validierung** von Hypothesen über die Eigenschaften der Häufigkeitsverteilung. Dazu werden andere statistische Maße verwendet.