DOI: *https://doi.org/10.14618/ids-pub-11586*

Holger Keibel
Programmbereich Korpuslinguistik, IDS Mannheim

Mathematische Häufigkeitsmaße in der Korpuslinguistik Eigenschaften und Verwendung

Einleitung

In der *Korpuslinguistik* und der *Quantitativen Linguistik* werden ganz verschiedenartige formale Maße verwendet, mit denen die Gebrauchshäufigkeit eines Wortes, eines Ausdrucks oder auch abstrakter oder komplexer sprachlicher Elemente in einem gegebenen Korpus gemessen und ggf. mit anderen Gebrauchshäufigkeiten verglichen [1] werden kann. Im Folgenden soll für eine Auswahl dieser Maße (absolute Häufigkeit, relative Häufigkeit, Wahrscheinlichkeitsverteilung, Differenzenkoeffizient, Häufigkeitsklasse) zusammengefasst werden, wie sie definiert sind, welche Eigenschaften sie haben und unter welchen Bedingungen sie (sinnvoll) anwendbar und interpretierbar sind – dabei kann eine Rolle spielen, ob das Häufigkeitsmaß auf ein Korpus als Ganzes angewendet wird oder auf einzelne Teilkorpora. Zusätzlich zu den bei den einzelnen Häufigkeitsmaßen genannten Einschränkungen gilt generell der folgende vereinfachte Zusammenhang: Je seltener ein Wort im gegebenen Korpus insgesamt vorkommt und je kleiner dieses Korpus ist, desto stärker hängt die beobachtete Gebrauchshäufigkeit des Wortes von zufälligen Faktoren ab, d.h., desto geringer ist die statistische Zuverlässigkeit der Beobachtung. [2]

Ausgangssituation

Bei der Beschreibung der Maße wird die folgende Ausgangssituation vorausgesetzt: Es wurde ein festes Textkorpus ausgewählt und (z.B. mithilfe einer geeigneten Korpusrecherche-Software) eine Suchanfrage an dieses Korpus gestellt. Zudem wird an einigen Stellen davon ausgegangen, dass das gegebene Korpus in einzelne Teilkorpora zerlegt wurde, so dass jeder Korpustext in genau einem Teilkorpus liegt. Eine solche Zerlegung lässt sich definieren z.B. nach einem zeitlichen Kriterium wie *Entstehungsjahr* [3] oder nach einem Kriterium wie *Textquelle* [4] usw.).
Zu einigen Häufigkeitsmaßen werden konkrete Beispielanalysen genannt – diesen Beispielen liegt ein *virtuelles Korpus* zugrunde, das alle bundesdeutschen Zeitungskorpora im DEUTSCHEN REFERENZKORPUS (DEREKO) von 1990 bis Mitte 2008 enthält. Im Folgenden wird dieses virtuelle Korpus kurz als *MDCA-Korpus* bezeichnet, es umfasst insgesamt ca. 2,04 Milliarden laufende Textwörter.

Art der Suchanfrage

Die Verwendung einiger Häufigkeitsmaße hängt von der Art der Suchanfrage ab – wo dies der Fall ist, wird dies explizit dazu gesagt, in allen anderen Fällen wird der Einfachheit halber davon ausgegangen, dass die Suchanfrage nach einem einzelnen Wort sucht (ggf. nach allen flektierten Formen und anderen orthografischen Varianten dieses Wortes). Hierfür müssen die folgenden zwei Suchanfrage-Typen unterschieden werden:

- Eine *einfache Suchanfrage* ist dadurch gekennzeichnet, dass jeder einzelne Treffer aus exakt einer Wortform besteht. Eine solche Suchanfrage liegt z.B. dann vor, wenn die Instanzen einer konkreten Wortform (d.h. einer spezifischen Zeichenkette) gesucht werden, aber auch wenn mithilfe eines Lemmatisierers die Instanzen aller flektierten Wortformen und orthografischen Varianten eines Lexems gesucht werden. Ebenso liegt eine einfache Suchanfrage dann vor, wenn nach den Instanzen verschiedener Wortformen oder Lexeme gesucht wird. [5]
- Eine *komplexe Suchanfrage* dagegen ist dadurch gekennzeichnet, dass einzelne Treffer zu dieser Suchanfrage evtl. aus mehreren Tokens bestehen können. Eine komplexe Suchanfrage liegt beispielsweise immer dann vor, wenn die Instanzen einer konkreten Wortfolge gesucht wird (z.B. die Wortfolge *immer öfter*), aber auch dann, wenn die Instanzen einer konkreten Wortkombination gesucht werden, deren einzelne Wörter nicht unmittelbar aufeinander folgen müssen (z.B. das Wortpaar *insofern … als*). Ein besonderes Beispiel sind die sog. *Präverbgefüge* (trennbare Präfixverben) wie *auftreten* oder *nachdenken*: Eine Suchanfrage, die nach allen flektierten Formen eines solchen Verbs sucht – also nach seinen zusammengeschriebenen ebenso wie nach seinen getrennt geschriebenen Formen –, ist zwangsläufig eine komplexe Suchanfrage. [6]

Häufigkeitsmaße

Absolute Häufigkeit

Absolute Häufigkeiten sind das elemantare Häufigkeitsmaß. Die absolute Gebrauchshäufigkeit eines Wortes in einem Korpus gibt an, wie oft das Wort insgesamt in allen Texten dieses Korpus vorkommt.

Die absolute Häufigkeit eines Wortes in einem Korpus hängt erheblich vom Umfang dieses Korpus ab: Halbiert man das Korpus, so erwartet man in dem halb so großen Korpus ungefähr halb so viele Treffer zu diesem Wort. Aus diesem Grund sind absolute Gebrauchshäufigkeiten zu verschiedenen Korpora nur dann vergleichbar, wenn diese Korpora ungefähr gleich groß sind. Da dies i.A. nicht der Fall ist, sind absolute Gebrauchshäufigkeiten zur Analyse und Interpretation von Häufigkeitsunterschieden in den meisten Fällen ungeeignet.

Relative Häufigkeit

Die relative Gebrauchshäufigkeit eines Wortes in einem Korpus gibt an, welchen Anteil das Wort an diesem gesamten Korpus ausmacht. Dieser Anteil ist eine Dezimalzahl zwischen 0,0 und 1,0, wird manchmal aber auch in Prozent oder Promille angegeben, vorzugsweise jedoch als *Instanzen pro Million Wörter* (kurz: *pMW* oder *IpM*, im Englischen entsprechend: *pmw* bzw. *ipm*). Einige Wörter kommen sehr häufig vor (so macht der bestimmte Artikel mit den Formen *der, die, das, des, dem, den* insgesamt 9,2% (= 92.000 Instanzen pMW) des MDCA-Korpus aus, d.h., beim Lesen dieses Korpus ist durchschnittlich fast jede zehnte Wortform ein bestimmter Artikel), andere Wörter sind weniger häufig (z.B. liegt die relative Häufigkeit des Wortes *hingegen* in demselben Korpus bei ca. 74,0 Vorkommen pMW) und wieder andere Wörter kommen nur vergleichsweise selten vor (z.B. *Dadaismus*: ca. 0,23 pMW).

Die absolute Häufigkeit eines Wortes in einem Korpus hängt erheblich von der jeweiligen Korpusgröße ab. Die naheliegendste und verbreitetste Weise, absolute Häufigkeiten bezüglich der Korpusgrüße zu normieren, führt zu relativen Häufigkeiten. Genauer: Die relative Gebrauchshäufigkeit eines Wortes in einem Korpus gibt an, welchen Anteil an allen Worttokens im Korpus die Tokens des gegeben Wortes ausmachen.
Weil dieser Anteil für die meisten Wörter sehr klein ist, werden relative Häufigkeiten oft in Prozent oder Promille oder auch als *Instanzen pro Million Wörter* (kurz: *pMW* oder *IpM*, im Englischen entsprechend: *pmw* bzw. *ipm*) angegeben. Die Einheit *Instanzen pro Million Wörter* ist insbesondere dann vorteilhaft, wenn relative Häufigkeiten für sehr unterschiedlich häufige Wörter angegeben werden sollen, und etabliert sich in der Korpuslinguistik zunehmend als Standard. Wichtig: Prozent-, Promille- und *pMW*-Angaben dienen nur der Skalierung von relativen Häufigkeiten auf einen vernünftigen Zahlbereich – nicht aber der Projektion auf ein Korpus mit der *Einheitsgröße* 100 bzw. 1.000 bzw. 1.000.000 Worttokens (vgl. die vertiefenden Informationen unten).

Formale Definition: Bezeichnet $F(W)$ die absolute Anzahl der Vorkommen des Wortes W und N die Größe des zugrunde liegenden Korpus (in laufenden Textwörtern, also Tokens), dann ist

(1) $$f(W) = \frac{F(W)}{N}$$

die relative Gebrauchshäufigkeit von W in diesem Korpus. Für eine Prozent-, Promille- oder *pMW*-Angabe wird dieser Quotient natürlich noch mit 100, 1.000 bzw. 1.000.000 multipliziert.

Abhängigkeit von der Suchanfrage: Diese Definition gilt nicht nur für Suchanfragen zu einem konkreten Wort, sondern analog für beliebige einfache Suchanfragen. Für komplexe Suchanfragen gilt sie so jedoch nicht. Grundsätzlich werden relative aus absoluten Häufigkeiten abgeleitet, indem die Anzahl der tatsächlich beobachteten Treffer geteilt wird durch die Anzahl der Treffer, die bei der gegebenenen Suchanfrage in dem gegebenen Korpus theoretisch maximal möglich wären. Bei einfachen Suchanfragen ist diese theoretisch maximale Trefferzahl identisch mit der Anzahl Worttokens im Korpus, bei komplexen

Suchanfragen gilt dies jedoch i.A. nicht: Beispielsweise könnte bei einem Korpus mit insgesamt *N* laufenden Textwörtern die Wortfolge *immer öfter* theoretisch maximal *N*/2 mal vorkommen[7], während die Wortfolge *nicht enden wollende* theoretisch maximal *N*/3 mal vorkommen könnte. Die o.g. Formel (1) müsste also für komplexe Suchanfragen so angepasst werden, dass im Nenner nicht die Korpusgröße, sondern die für die konkrete Suchanfrage spezifische theoretisch maximal mögliche Trefferanzahl steht. Eine Formel für diese maximal mögliche Trefferanzahl lässt sich leider nicht allgemeingültig formulieren, sie muss für jeden Typus von komplexer Suchanfrage neu hergeleitet werden. [8]
Dennoch wenden viele Studien die unveränderte Formel (1) auch auf komplexe Suchanfragen an. In solchen Fällen ist es ratsam, die resultierenden relativen Häufigkeiten als *pMW*-Werte anzugeben, da diese sich weiterhin auf sinnvolle Weise intuitiv interpretieren lassen: Ergibt Formel (1) für eine komplexe Suchanfrage beispielsweise den Wert 2,64 pMW, dann kommt die gesuchte Wortkombination oder Struktur in dem Korpus durchschnittlich 2,64 Mal je eine Million Wörter großem Korpusausschnitt vor. Die Interpretierbarkeit von *pMW*-Werten bleibt also auch für komplexe Suchanfragen bestehen, ihre unmittelbare Vergleichbarkeit jedoch nicht: Werden *pMW*-Werte zu zwei unterschiedlich komplexen Suchanfragen verglichen, dann muss weiterhin berücksichtigt werden, wie viele Treffer für jede der beiden Suchanfragen maximal möglich wäre. Ohne solche Überlegungen ist es wenig informativ zu erfahren, dass etwa der ermittelte *mpw*-Wert der Wortfolge *nicht enden wollende* kleiner ist als der der Wortfolge *immer öfter*.

Vertiefende Informationen:

- Intuitiv würde man vermuten, dass relative Häufigkeiten unabhängig von der Korpusgröße sind – schließlich werden ja genau zu diesem Zweck in der obigen Formel (1) die absoluten Häufigkeiten durch die jeweilige Korpusgröße geteilt. Grundsätzlich ist die Intuition hier auch zutreffend, allerdings mit der folgenden Einschränkung: Je seltener ein Wort (in der relevanten Sprachdomäne/Sprachausschnitt) bzw. je kleiner das verwendete Korpus, desto stärker hängt die in diesem Korpus beobachtete relative Gebrauchshäufigkeit von zufälligen Faktoren ab, d.h., desto weniger zuverlässig ist diese beobachtete Häufigkeit (s.o.). In der Praxis heißt das: Wenn ein Wort im verwendeten Korpus absolut nur einige wenige Male vorkommt, dann ist dieses Korpus zu klein, um zuverlässig die relative Gebrauchshäufgkeit dieses Wortes ermitteln zu können.
- Wenn man Korpushäufigkeiten als Zufallsvariablen in einem Zufallsexperiment auffasst, lässt sich diese abnehmende Zuverlässigkeit von relativen Häufigkeiten statistisch wie folgt veranschaulichen. Hierfür nehmen wir vereinfachend an, ein gegebenes Wort *W* hat in der relevanten Sprachdomäne eine feste, intrinsische Vorkommenswahrscheinlichkeit *p*. Das Zufallsexperiment besteht nun darin, eine (adäquate) Stichprobe aus dieser Sprachdomäne zu ziehen – in diesem Fall ist die Stichprobe natürlich ein Korpus. Die relative Häufigkeit $f(W)$ von *W* in dieser Stichprobe ist dann eine Zufallsvariable *X*. Der Erwartungswert für diese Zufallsvariable ist genau seine Vorkommenswahrscheinlichkeit *p*.
  Wie bei allen Zufallsexperimenten variieren aber die tatsächlich beobachteten Werte der Variablen *X* um ihren Erwartungswert *p* – das Ausmaß dieser Variation wird näherungsweise durch die empirische Streuung (Standardabweichung) σ beschrieben. In der Praxis kennt man den Erwartungswert bei einem Zufallslexperiment nicht im Voraus – in vielen Fällen ist diese Unkenntnis ja gerade die Motivation für das Zufallsexperiment, und man möchte von den tatsächlich beobachteten Werten auf diesen Erwartungswert schließen. Je größer aber die

Streuung, desto weniger zuverlässig geben die tatsächlich beobachteten Werte den Erwartungswert – in unserem Fall also die Vorkommenswahrscheinlichkeit von *W* – wieder. Je seltener aber das Wort absolut in einem konkreten Korpus auftaucht, desto stärker wird seine in diesem Korpus beobachtete relative Häufigkeit in vergleichbaren Korpora derselben Größe variieren, desto größer also ist die Streuung dieser beobachteten relativen Häufigkeiten. Denn ob ein Wort absolut 8 statt 7 Mal vorkommt, hat größere Auswirkungen auf seine beobachtete relative Häufigkeit, als wenn es 108 statt 107 Mal vorkommt.

- Aus diesen Gründen ist es ratsam, neben relativen Häufigkeiten immer auch absolute Häufigkeiten zu betrachten. Zwar ist man in vielen Fällen primär an der relativen Häufigkeit eines Wortes interessiert, aber man sollte dieser relativen Häufigkeit nur dann Vertrauen schenken, wenn das Wort im zugrunde liegenden Korpus absolut hinreichend häufig vorkommt. Andernfalls ist das Wort zu selten bzw. das Korpus für eine hinreichend zuverlässige Häufigkeitsmessung zu klein. Entsprechendes gilt auch für den Vergleich von relativen Häufigkeiten auf der Basis verschieden großer Korpora: In dem größeren Korpus sind die beobachteten relativen Häufigkeiten grunsätzlich zuverlässiger – ein Vergleich zweichen beiden ist nur für Wörter gerechtfertigt, die in auch im kleineren Korpus absolut hinreichend häufig vorkommen. Formal lässt sich diese Zuverlässigkeit solcher Vergleiche durch Berechnung von Konfidenzintervallen bewerten.
- Ein solcher Vergleich von relativen Häufigkeiten liegt implizit auch dann vor, wenn man absolute Häufigkeiten von einem kleineren Korpus auf ein deutlich größeres hochrechnen/hochskalieren möchte (oder umgekehrt) – auch hier reicht die die hochgerechnete Häufigkeit nicht, sondern sollte um die Angabe des Konfidenzintervalls (um diesen Wert herum) ergänzt werden. Formal kann man übrigens die Angabe von relativen Häufigkeiten in Prozent bzw. Promille bzw. *pMW* ebenfalls als eine Skalierung oder Projektion von absoluten Häufigkeiten interpretieren: nämlich auf ein Korpus mit der *Einheitsgröße* 100 bzw. 1.000 bzw. 1.000.000 Worttokens. Auch diese implizite Skalierung ist ohne Angabe von Konfidenzintervallen unvollständig und führt leicht zu Fehlschlüssen.

Wahrscheinlichkeitsverteilung

Hier wird vorausgesetzt, dass das gegebene Korpus in einzelne Teilkorpora zerlegt ist (wie oben skizziert). Die Wahrscheinlichkeitsverteilung eines Wortes ist nicht eine einzelne Zahl, sondern eine Zahlenfolge: Die einzelnen Zahlen geben an, zu welchen Anteilen sich die Vorkommen dieses Wortes im gesamten Korpus auf die einzelnen Teilkorpora verteilen. Diese Werte können auch als (bedingte) Wahrscheinlichkeiten interpretiert werden: Der Wert für ein Teilkorpus gibt an, wie wahrscheinlich es ist, dass eine zufällig aus dem Gesamtkorpus entnommene Instanz des Wortes in diesem Teilkorpus liegt. Die einzelnen Werte sind also Zahlen zwischen 0,0 und 1,0, und ihre Summe ist stets 1,0 (bzw. 100%).

Eine Wahrscheinlichkeitsverteilung stellt – nach relativen Häufigkeiten – eine zweite Möglichkeit dar, die in den einzelnen Teilkorpora beobachteten absoluten Häufigkeiten zu normieren. Während bei relativen Häufigkeiten nach Größe der Teilkorpora normiert wird, wird hier jedoch nach der Gesamthäufigkeit des Wortes im gesamten Korpus normiert. [10] Man erhält die Wahrscheinlichkeitsverteilung eines Wortes also aus der Verteilung seiner absoluten Häufigkeiten, indem man letztere durch die Gesamthäufigkeit des

Wortes im Korpus teilt – d.h. indem man jede Einzelhäufigkeit für jedes Teilkorpus durch die Gesamthäufigkeit teilt. Dadurch wird die Verteilung der absoluten Häufigkeiten auf die *Einheitsmasse* 1,0 skaliert – das bedeutet, dass die skalierten Einzelwerte sich zu 1,0 summieren. Weil diese Einzelwerte zudem allesamt Zahlen zwischen 0,0 und 1,0 sind, bilden sie zusammen formal eine Wahrscheinlichkeitsverteilung (über die Menge der Teilkorpora). Wahrscheinlichkeitsverteilungen haben gegenüber Verteilungen absoluter oder relativer Häufigkeiten den Vorteil, dass die Verteilungen zu unterschiedlich häufigen Wörtern leichter auf derselben *y*-Skala dargestellt werden können. [11] So werden Unterschiede zwischen den Wörtern sichtbar, die nicht ihre Gesamthäufigkeit betreffen, sondern nur den Verlauf ihrer Häufigkeit über die Teilkorpora. Dies ist beispielsweise dann sinnvoll, wenn die zeitliche Verteilung von verschiedenen Wörtern über die Zeit betrachtet und verglichen werden soll.

Formale Definition: Bezeichne $F_j(W)$ die absolute Anzahl der Vorkommen des Wortes W im Teilkorpus T_j, dann ist

(2) $$P_j(W) = \frac{F_j(W)}{\sum_i F_i(W)}$$

der (gewichtete) Wahrscheinlichkeitswert von W im Teilkorpus T_j. Dabei ist der Quotient $\Sigma F_i(W) = F(W)$ die absolute Anzahl aller Vorkommen von W im gesamten Korpus K.

Abhängigkeit von der Suchanfrage: Diese Definition gilt grundsätzlich für alle Arten von Suchanfragen – d.h. für einfache Suchanfragen ebenso wie für komplexe Suchanfragen. Es werden absolute Häufigkeiten durch absolute Häufigkeiten derselben Suchanfrage geteilt, so dass sich die Komplexität der Suchanfrage gleichsam *wegkürzt*.

Vertiefende Informationen:

- Bei einer (gewichteten) Wahrscheinlichkeitsverteilung sind die einzelnen Wahrscheinlichkeitswerte in Bezug auf das gegebene Korpus exakt. Werden diese Werte aber in Bezug auf die sprachliche Grundgesamtheit interpretiert, die durch das Korpus repräsentiert sein soll, dann sind sie lediglich Schätzwerte. Das zugrunde liegende Schätzverfahren wird als Maximum-Likelihood-Methode bezeichnet.
- Bei einer wie oben definierten Wahrscheinlichkeitsverteilung werden die einzelnen Teilkorpora – oder besser: die Sprachausschnitte, die durch die einzelnen Teilkorpora repräsentiert sind – proportional zur Größe dieser Teilkorpora gewichtet. Die einzelnen Wahrscheinlichkeitswerte hängen also, ebenso wie die zugrunde liegenden absoluten Häufigkeiten, von der Größe der Teilkorpora ab. Ist dies nicht erwünscht, so kann man alternativ eine *ungewichtete Wahrscheinlichkeitsverteilung* verwenden, die sich analog aus relativen Häufigkeiten ableitet. Genauer: Ersetzt man in Formel (2) die absoluten Häufigkeiten $F_j(W)$ durch die entsprechenden relativen Häufigkeiten $f_j(W) = F_j(W) / N_j$, dann liefert die Formel den ungewichteten Wahrscheinlichkeitswert von W im Teilkorpus T_j. [12]
  Die einzelnen Werte dieser (ungewichteten) Verteilung lassen sich als geschätzte (bedingte) Wahrscheinlichkeiten auffassen, bezogen auf ein fiktives zweites Korpus, dessen Teilkorpora

alle etwa gleich groß sind, dabei aber ähnlich zusammengestellt sind wie die des gegebenen Korpus: Der ermittelte (ungewichtete) Wert für ein Teilkorpus gibt dann an, wie wahrscheinlich es ist, dass eine zufällig aus dem fiktiven Gesamtkorpus entnommene Instanz des Wortes in dem entsprechenden fiktiven Teilkorpus liegt. Diese ungewichtete Variante ist besonders dann zu bevorzugen, wenn eine beobachtete Wahrscheinlichkeitsverteilung vom Korpus auf eine Grundgesamtheit extrapoliert werden soll, bei der man annimmt, dass die den Teilkorpora entsprechenden Teil-Grundgesamtheiten etwa gleich große Anteil an der gesamten Grundgesamtheit haben. Sind jedoch bereits im tatsächlich vorliegenden Korpus alle Einzelkorpora ungefähr gleich groß, dann sind beide Arten von Wahrscheinlichkeitsverteilungen (gewichtete und ungewichtete) identisch.

- Alle oben gemachten Aussagen über gewichtete Wahrscheinlichkeitsverteilungen gelten analog auch ungewichtete Wahrscheinlichkeitsverteilungen. Insbesondere können sie ebenfalls auf alle Arten von Suchanfragen angewendet werden: In diesem Fall werden relative Häufigkeiten durch relative Häufigkeiten geteilt – auch hier kürzt sich die Komplexität gegenseitig weg, sofern alle einzelnen relativen Häufigkeiten auf dieselbe Weise ermittelt wurden. Die in im Abschnitt Abhängigkeit von der Suchanfrage unter dem Punkt *Relative Häufigkeit* dargestellten Einschränkungen können also ignoriert werden, wenn die relativen Häufigkeiten nur dazu dienen, ungewichtete Wahrscheinlichkeitsverteilungen aus ihnen abzuleiten.

Differenzenkoeffizient

Der Differenzenkoeffizient liegt immer zwischen den Werten -1,0 und +1,0. Wie auch Wahrscheinlichkeitsverteilungen ist er jedoch nur dann sinnvoll zu berechnen, wenn das gegebene Korpus in einzelne Teilkorpora zerlegt ist (wie oben skizziert). In Teilkorpora, in denen der Differenzenkoeffizient eines Wortes positiv ist, ist die Gebrauchshäufigkeit des Wortes überdurchschnittlich hoch. In Teilkorpora mit einem negativen Differenzenkoeffizienten dagegen liegt die Gebrauchshäufigkeit des Wortes unter dem Durchschnitt. Kommt das Wort in einem Teilkorpus gar nicht vor, so hat der Differenzenkoeffizient hier den Wert -1,0.

Relative Häufigkeiten und Wahrscheinlichkeitsverteilungen – insbesondere in grafischer Darstellung – erlauben einen schnellen, intuitiven Zugang zur Verteilung der Gebrauchshäufigkeit eines Wortes über die einzelnen Teilkorpora. Allerdings haben beide Häufigkeitsmaße einige Nachteile, wenn es darum geht, diese Verteilung für ein Wort isoliert zu interpretieren oder als Ganzes mit der entsprechenden Verteilung anderer Wörter zu vergleichen. Zu diesen Zwecken besser geeignet ist der Differenzenkoeffizient (vgl. Belica 1999 [13] und Belica 1996/1998 [14]), denn er leitet die (ggf. optische) Wahrnehmung und damit die spontane Interpretation einer Häufigkeitsverteilung und erleichtert zudem den Vergleich der Häufigkeitsverteilungen von verschiedenen Wörtern.

Vertiefende Informationen:

Diese Vorzüge treffen v.a. auf eine grafische Darstellung des Differenzenkoeffizienten zu, und sie ergeben sich aus den folgenden Eigenschaften des Differenzenkoeffizienten [15] :

- Mit dem Differenzenkoeffizienten können die Häufigkeitsverteilungen verschiedener Wörter sinnvoll auf derselben standardisierten *y*-Skala (Werte von -1,0 bis +1,0) angezeigt werden – dies gilt für relative Häufigkeiten nicht. Wahrscheinlichkeitsverteilungen können zwar grundsätzlich auf derselben Skala von 0,0 bis 1,0 dargestellt werden, jedoch würden dann viele Häufigkeitsverteilungen sehr flach (d.h., nahe 0,0) dargestellt werden – dies gilt besonders für gleichmäßiger verteilte Wörter und umso mehr, je mehr Teilkorpora vorhanden sind. In der Praxis würde man Wahrscheinlichkeitsverteilungen daher auf einer kleineren Skala von 0,0 bis zu einem gewissen Wert y_{max} darstellen, und dieser Wert würde abhängig von der Anzahl Teilkorpora und der größten Einzelwahrscheinlichkeit unter allen darzustellenden Wörtern gewählt werden. Beim Differenzenkoeffizienten hingegen lässt sich jede Häufigkeitsverteilung sinnvoll auf derselben *y*-Skala anzeigen und studieren.
- Eine Grafik, die die Häufigkeitsverteilung eines Wortes bzgl. des Differenzenkoeffizienten anzeigt, leitet sich aus der entsprechenden Grafik mit relativen Häufigkeiten ab, indem die *y*-Achse (durch eine nichtlineare Transformation) reskaliert wird, so dass der niederfrequente Bereich gestreckt, der höherfrequente hingegen gestaucht wird. Dieser Effekt ist besonders dann erwünscht, wenn die Teilkorpora zeitlich definiert sind: Beispielsweise würde ein Anstieg der relativen Häufigkeit von 0,1 Instanzen pMW (im Jahr 2000) auf 0,2 pMW (2001) genau so groß aussehen wie der anschließende Anstieg auf 0,3 pMW (2002) – im ersten Fall verdoppelt sich die Gebrauchshäufigkeit jedoch, während sie im zweiten nur um den Faktor 1,5 steigt. Daher ist der erste Anstieg für eine Untersuchung der zeitlichen Entwicklung der Gebrauchshäufigkeit (z.B. in Bezug auf Neologismen) vermutlich der relevantere, und der Differenzenkoeffizient wird dem gerecht, indem er den ersten Anstieg stärker betont als den zweiten. Die grafische Darstellung von relativen Häufigkeiten hingegen ist in dieser Hinsicht optisch irreführend.
- Die beschriebene Reskalierung der *y*-Achse wird erreicht durch einen Vergleich der relativen Häufigkeit im jeweiligen Teilkorpus mit einem (geschätzten) Erwartungswert, unter der Annahme, dass die insgesamt beobachteten Vorkommen über alle Teilkorpora gleichverteilt sind. Durch diesen Vergleich werden Teilkorpora, in denen das Wort unter- bzw überrepräsentiert ist, visuell hervorgehoben. Sind die Teilkorpora zeitlich definiert, so treten insbesondere die wesentlichen Aspekte der zeitlichen Entwicklung der Gebrauchshäufigkeit visuell hervor – eine Eigenschaft, die beispielsweise in Untersuchungen zu Neologismen oder aussterbenden Wörtern sehr erwünscht ist. Von Vorteil ist diese Eigenschaft jedoch nicht nur beim Studium einzelner Häufigkeitsverteilungen, sondern auch bei ihrem Vergleich untereinander: wenn beispielsweise eine Typologie der zeitlichen Häufigkeitsentwicklung von Neologismen erarbeitet werden soll, deren Klassen abstrahiert von der Grundhäufigkeit dieser Wörter charakterisiert sind.

Formale Definition: Bezeichne $F_j(W)$ die absolute Anzahl der Vorkommen des Wortes W im Teilkorpus T_j und $F(W) = \Sigma F_j(W)$ die absolute Anzahl alle Vorkommen von W im gesamten Korpus K. Sei ferner N_j die Größe des Teilkorpus T_j (in laufenden Textwörtern) und $N = \Sigma N_j$ die Größe des gesamten Korpus. Dann ist $g_{obs} = f_j(W) = F_j(W) / N_j$ die beobachtete relative Gebrauchshäufigkeit von W im Teilkorpus T_j, die zugehörige erwartete relative Gebrauchshäufigkeit g_{exp} wird geschätzt (unter Annahme einer Gleichverteilung aller Vorkommen über die Teilkorpora) durch die beobachtete globale relative Gebrauchshäufigkeit im gesamten Korpus K, also $g_{exp} = f(W) = F(W) / N$. Mit diesen Notationen ist

(3) $$D_j(W) = \frac{g_{obs} - g_{exp}}{g_{obs} + g_{exp}} = 1 - \frac{2}{g_{obs} / g_{exp} + 1}$$

der Differenzenkoeffizienten von *W* im Teilkorpus T_j. [16] Äquivalent hierzu lässt sich der Differenzenkoeffizient mit Formel (3) auch direkt aus den beobachteten absoluten Gebrauchshäufigkeiten ableiten, indem als beobachtete und erwartete Häufigkeit $g_{obs} = F_j(W)$ bzw. $g_{exp} = F(W) \cdot (N_j / N)$ eingesetzt werden.

Abhängigkeit von der Suchanfrage: Diese Definition gilt grundsätzlich für alle Arten von Suchanfragen – d.h. für einfache Suchanfragen ebenso wie für komplexe Suchanfragen. Weil der Differenzenkoeffizient von relativen Häufigkeiten abgeleitet wird und diese für komplexe Suchanfragen anders berechnet werden sollten als für einfache Suchanfragen (vgl. Abhängigkeit von der Suchanfrage unter dem Punkt *Relative Häufigkeit*), stellt sich die Frage, ob sich dieses Problem auch auf die Berechnung des Differenzenkoeffizienten vererbt. Grundsätzlich sollten in Formel (3) für g_{obs} und g_{exp} die exakt berechneten relativen Häufigkeiten eingesetzt werden. Hierbei wäre Formel (1) (unter dem Punkt *Relative Häufigkeit*) auf einfache Suchanfragen anwendbar, während sie für komplexe Suchanfragen angepasst werden müsste. Diese Anpassung geschieht i.A. in Form eines Korrekturfaktors, der für die jeweilige Suchanfrage spezifisch ist, jedoch nicht vom Korpus abhängt. Bei der Transformation einer relativen Häufigkeit zum Differenzenkoeffizienten kürzt sich dieser Korrekturfaktor wieder weg, so dass der Differenzenkoeffizient in den meisten Fällen auch für komplexe Suchanfragen mit Formeln (1) und (3) berechnet werden kann.

Interpretation: Der Differenzenkoeffizient $D = D_j(W)$ nimmt Werte zwischen -1,0 und +1,0 an. Dabei bedeutet $D = 0$, dass die beobachtete Häufigkeit g_{obs} mit der erwarteten Häufigkeit g_{exp} übereinstimmt. $D > 0$ bedeutet, dass die beobachtete Häufigkeit größer ist als die erwartete – sie ist sogar sehr viel größer, falls *D* nahe am maximalen Wert +1,0 ist. [17] Hingegen bedeutet $D < 0$, dass die beobachtete Häufigkeit kleiner ist als die erwartete – sie ist sogar sehr viel kleiner, falls *D* nahe am minimalen Wert -1,0 ist. Der Extremfall $D = -1{,}0$ tritt genau dann ein, wenn das Wort *W* im entsprechenden Teilkorpus gar nicht vorkommt. Zusammengefasst lässt sich sagen: Ein Wort *W* ist überrepräsentiert in den Teilkorpora, in denen sein Differenzenkoeffizient positiv ist, und unterrepräsentiert in Teilkorpora mit negativem Differenzenkoeffizienten.

Vertiefende Informationen:

- Wie bereits erwähnt, ist der Differenzenkoeffizient nur dann sinnvoll, wenn das gegebene Korpus in einzelne Teilkorpora zerlegt ist. Er kann dann für jedes einzelne Teilkorpus berechnet werden – für das Korpus als Ganzes wäre eine solche Berechnung hingegen nicht sinnvoll – es ist unklar, wofür die benötigte *erwartete Häufigkeit* im gesamten Korpus stehen soll und auf welcher Grundlage man sie sinnvoll schätzen könnte. Legt man für diese Schätzung wie bei den Teilkorpora die Gesamthäufigkeit zugrunde, so erhielte man als Differenzenkoeffizient für das gesamte Korpus stets den Wert 0.

- Weil der Differenzenkoeffizient eine (monotone) Transformation der relativen Häufigkeit ist, gelten die dort gemachten Aussagen zur statistischen Zuverlässigkeit auch für ihn (vgl. die vertiefenden Informationen unter dem Punkt *Relative Häufigkeit*).

Häufigkeitsklasse

Mit Häufigkeitsklassen werden alle Wörter des gesamten Vokabular nach ihrer Häufigkeit in Klassen aufgeteilt, wobei Wörter derselben Klasse ungefähr gleich häufig sind. In der Praxis werden hierfür bis zu 30 Häufigkeitsklassen unterschieden, und sie tragen die Nummern 0, 1, 2, 3, usw. Zu beachten ist hierbei: Je niedriger die Nummer der Häufigkeitsklasse, desto häufiger sind die darin befindlichen Wörter. Die häufigsten Wörter befinden sich in Klasse 0 (hier findet man in Korpora der deutschen Schriftsprache i.A. nur den bestimmten Artikel), die zweithäufigste Gruppe von Wörtern in Klasse 1 (diese Klasse ist meistens leer), die dritthäufigste Gruppe von Wörtern in Klasse 2 (in diese Klasse gehören für das o.g. MDCA-Korpus z.B. die Konjunktion *und*, das Verb *sein* mit allen seinen Formen, sowie die Präposition *in*). Von den oben bereits verwendeten Beispielen gehört für das MDCA-Korpus das mittelmäßig häufige Wort *hingegen* zur Häufigkeitsklasse 10 und das eher seltene Wort *Dadaismus* zur Häufigkeitsklasse 19.

Die Häufigkeitsklasse K eines Wortes W ist ein ganzzahliger Wert, der angibt, dass das häufigste Wort R im gegebenen Korpus etwa 2^K mal so häufig vorkommt wie W. Dabei sollte dieses Referenzwort R auf derselben linguistischen Ebene bestimmt sein wie das gegebene Wort W: Ist W z.B. eine konkrete Wortform/Zeichenkette, dann sollte für R die im gegebenen Korpus häufigste Wortform/Zeichenkette gewählt werden; ist W hingegen ein Lexem (mit flektierten Wortformen und ggf. orthografischen Varianten), dann sollte für R das im gegebenen Korpus häufigste Lexem gewählt werden. In Korpora zur deutschen Schriftsprache ist die häufigste Wortform typischerweise das Wort *der* und das häufigste Lexem der bestimmte Artikel (mit den flektierten Formen *der, die, das, des, dem, den*). Wurde zudem die Suchanfrage zum Wort W so formuliert, dass auch großgeschriebene Varianten am Satzanfang gefunden werden, dann sollte idealerweise auch das Referenzwort R und seine Gesamthäufigkeit mit allen klein- und großgeschriebenen Varianten bestimmt werden.

Formale Definition: Bezeichnet $F(W)$ die absolute Anzahl der Vorkommen des Wortes W im gesamten Korpus und $F(R)$ die absolute Anzahl der Vorkommen des Referenzwortes in demselben Korpus, dann ist

(4) $$K(W) = \left\lfloor 0{,}5 + \log_2\left(\frac{F(R)}{F(W)}\right) \right\rfloor$$

die Häufigkeitsklasse von W im gegebenen Korpus, wobei die Gaußklammer $\lfloor x \rfloor$ eine Dezimalzahl x auf den nächstkleineren ganzzahligen Wert abrundet.

Abhängigkeit von der Suchanfrage: Diese hier vorgestellte Definition von Häufigkeitsklassen ist nur für einfache Suchanfragen sinnvoll. Für komplexe Suchanfragen müsste zunächst

festgelegt werden, wie eine sinnvolle Referenzsuchanfrage theoretisch zu charakterisieren wäre und wie unter allen möglichen Kandidaten praktisch der häufigste gefunden werden kann.

Vertiefende Informationen:

- Wie bereits erwähnt, hängt die statistische Zuverlässigkeit von relative Häufigkeiten und dem Differenzenkoeffizient entscheidend von der absoluten Häufigkeit eines Wortes ab: Für seltenere Wörter (bzw. kleinere Korpora) ist sie geringer (vgl. die vertiefenden Informationen unter dem Punkt *Relative Häufigkeit*). Im Gegensatz zu diesen beiden Maßen ist die Häufigkeitsklasse auch bei selteneren Wörtern und kleineren Korpora deutlich stabiler -- wenngleich auch hier die Zuverlässigkeit bei häufigeren Wörtern größer ist. Insbesondere lassen sich Häufigkeitsklassen auch bei sehr verschieden großen Korpora zuverlässiger vergleichen. Wegen der Abhängigkeit der Häufigkeitsklasse vom häufigsten Wort *R* sollten die Korpora allerdings nur Daten derselben Sprache und desselben Sprachregisters enthalten – die Häufigkeitsklasse eines Wortes in einem Korpus zur deutschen Schriftsprache lässt sich also nicht mit der Häufigkeitsklasse eines Wortes in einem englischsprachigen Korpus vergleichen, und i.A. auch nicht mit einem Wort in einem Korpus gesprochener deutscher Sprache.
- Mit diesen Eigenschaften ist die Häufigkeitsklasse besonders dann vorteilhaft – und zumindest als ergänzende Information zu empfehlen –, wenn Häufigkeitsangaben aus Korpusrecherchen publiziert werden sollen. Die so publizierten Häufigkeitsangaben können dann von anderen Wissenschaftlern interpretiert und mit Häufigkeitsklassen aus anderen Studien verglichen werden, selbst wenn die zugrunde liegenden Korpora erheblich unterschiedliche Größen haben. [18] Hierfür ist aber dringend notwendig, dass mit einer publizierten Häufigkeitsklasse auch das für ihre Berechnung verwendete verwendete Referenzwort mit angegeben wird.

Bitte zitieren Sie bei Bedarf wie folgt:

Keibel, Holger (2008, 2009): *Mathematische Häufigkeitsmaße in der Korpuslinguistik: Eigenschaften und Verwendung.* Mannheim: Institut für Deutsche Sprache.

Fragen und Kommentare bitte an: korpuslinguistik{at}ids-mannheim.de

---