– Ergänzungen zu –

Korpuslinguistik

Rainer Perkuhn / Holger Keibel / Marc Kupietz (2012):
Korpuslinguistik. Paderborn: Fink.
(Reihe LIBAC – Linguistik für Bachelor 3433).

Stand: 18. Juni 2012

Inhalt

E6 Korpusfrequenzen analysieren

E6.3 Häufigkeitsverteilung des Vokabulars: Lexikalische Vielfalt

In dem Kasten auf S. 86 im Buch werden verschiedene Maße erwähnt, die messen, wie reichhaltig der in einem Korpus verwendete Wortschatz (unabhängig von der Korpusgröße) ist. Diese Reichhaltigkeit wird meist als *Wortschatzvarianz* oder *lexikalische Vielfalt* im Korpus bezeichnet. Dieses Konzept spielt besonders in angewandten Disziplinen wie der (korpusbasierten) Spracherwerbsforschung, der Sprachpathologie sowie der Stilometrie eine wichtige Rolle, um z. B. den wachsenden Wortschatz eines Kindes oder den verminderten Wortschatz von Personen mit Sprachstörungen zu messen oder um bei ungesicherter Autorenschaft eines älteren Textes Evidenzen zu sammeln. Im vorliegenden Begleitabschnitt sollen die genannten Maße kurz vorgestellt werden.

lexikalische Vielfalt

Das am meisten verbreitete Maß für lexikalische Vielfalt ist das so genannte *Type-Token-Verhältnis* (Englisch: *type-token ratio*, kurz: *TTR*), das manchmal auch mathematisch unsauber als *Type-Token-Relation* bezeichnet wird.

Type-Token-Verhältnis
TTR

Formal ist das Type-Token-Verhältnis definiert als der Quotient aus der Anzahl aller Worttypes und der Anzahl aller Worttokens im Korpus. Je nach Erkenntnisinteresse können die Worttypes hierbei auf der Ebene von Wortformen oder von Lexemen unterschieden werden (vgl. S. 27 im Buch). Will man TTR-Werte aus verschiedenen Studien miteinander vergleichen, so sollten beide für dieselbe linguistische Ebene (Wortform vs. Lexem) berechnet sein. In unseren Beispielen betrachten wir Worttypes stets auf der Ebene von Wortformen. Typischerweise wird das Type-Token-Verhältnis in Prozent angegeben. Da die Anzahl Types natürlich nicht größer sein kann als die Anzahl Tokens, ist ein TTR-Wert immer eine Zahl zwischen 0 und 100%.

TTR-Berechnung

Beispiel
Die Beispielkorpora BRZ08 und NUN08 (vgl. S. 79 im Buch) umfassen 25.938.449 bzw. 11.509.961 Textwörter (d. h. Worttokens) und 537.416 bzw. 374.495 Worttypes. Das Type-Token-Verhältnis dieser Korpora berechnet sich also wie folgt (in %):

BRZ08: $\frac{537.416 \cdot 100}{25.938.449} \approx 2{,}072$ NUN08: $\frac{374.495 \cdot 100}{11.509.961} \approx 3{,}254$

Das Type-Token-Verhältnis gibt an, mit welcher Rate die Tokens im Korpus zu unterschiedlichen Types gehören. Anders ausgedrückt: Das Type-Token-Verhältnis ist der Anteil der verschiedenen Tokens an allen Tokens. Ein größerer TTR-Wert deutet auf einen differenzierteren und reichhaltigeren Wortschatz hin, ein niedrigerer TTR-Wert hingegen auf ein größeres Maß an Wiederholung und auf eine formelhaftere Sprachverwendung.

Interpretation

TTR-Nachteil

Leider aber muss gleich eine Einschränkung nachgeschoben werden: Wissenschaftler haben früh bemerkt, dass das Type-Token-Verhältnis von der Korpusgröße abhängt: Wenn sonst alles gleich bleibt, nimmt der TTR-Wert mit steigender Korpusgröße ab (s. z. B. Richards 1987). Als Maß, um die lexikalische Vielfalt zweier Korpora zu vergleichen, eignen sich TTR-Werte damit nur bei gleich großen Korpora.

Beispiel
Diese Abhängigkeit von der Korpusgröße demonstrieren wir hier anhand des Beispielkorpus BRZ08 (aus dem Beispiel von S. E6-2). Bildet man aus diesem Korpus durch Zufallsauswahl von Texten ein kleineres Teilkorpus, dann kann man annehmen, dass BRZ08 und das Teilkorpus ähnlich gesampelte Stichproben derselben Grundgesamtheit darstellen, die sich lediglich in ihrer Größe unterscheiden und sonst in allen Faktoren ungefähr gleich sind. Unterschiedliche TTR-Werte dieser Korpora sollten also nahezu ausschließlich eine Folge der unterschiedlichen Korpusgrößen sein. Die folgende Tabelle zeigt die TTR-Werte (in %) für BRZ08 und Teilkorpora, die etwa 1/10, 1/100 und 1/1.000 der ursprünglichen Größe von BRZ08 haben.

Teilkorpus	BRZ08	1/10 · BRZ08	1/100 · BRZ08	1/1.000 · BRZ08
TTR	2,07	6,14	15,31	32,17

Mit abnehmender Korpusgröße steigt der TTR-Wert in diesem Beispiel tatsächlich deutlich an.

Vertiefung: Abhängigkeit von der Korpusgröße
Dass das TTR-Maß von der Korpusgröße abhängt, ist zunächst überraschend, denn bei TTR-Werten wird ja durch die Anzahl Tokens geteilt und damit hinsichtlich der Korpusgröße normiert. Die Ursache hierfür ist – wieder einmal – die Form der Zipf-Kurve (vgl. Abschn. 6.3 im Buch).
Zur Veranschaulichung dieser Ursache stellen Sie sich vor, Sie gehen ein bestehendes Korpus Worttoken für Worttoken durch und berechnen dabei nach jedem Token den TTR-Wert für das Teilkorpus, das alle bisher gesehenen Tokens enthält. Nach einigen hundert Tokens werden Sie bereits den meisten der hochfrequenten (d. h. häufigen) Worttypes mindestens einmal begegnet sein. Je größer Ihr Teilkorpus wird, desto mehr werden Sie auch von den mittelfrequenten Wörtern mindestens einmal gesehen haben, und es wird immer unwahrscheinlicher, im nächsten Token einen neuen Worttype anzutreffen, denn es verbleiben fast nur noch niederfrequente Worttypes, die Sie noch ein erstes Mal sehen könnten.
Während Sie anfangs häufig neue, bisher ungesehene Worttypes vorfinden, werden Sie später also immer länger warten müssen, bis Sie das nächste Mal einem neuen Worttype begegnen. Die Anzahl der Tokens Ihres Teilkorpus wächst bei dieser Vorgehensweise natürlich gleichmäßig mit jedem Worttoken, das Sie betrachten – die Anzahl der Types jedoch wächst anfangs sehr schnell und später immer langsamer. Daher ist das Type-Token-Verhältnis anfangs hoch und wird später allmählich immer kleiner. Hiermit ist jedoch nur der grobe Verlauf der TTR-Kurve gemeint – zwischendurch kann der TTR-Wert punktuell durchaus ein wenig ansteigen.

Für das Beispiel von S. E6-2 lässt sich also nicht direkt entscheiden, in welchem der beiden Beispielkorpora die lexikalische Vielfalt größer ist: Wir haben für das größere Korpus (BRZ08) den kleineren TTR-Wert beobachtet und können ohne weitere Hilfsmittel nicht beurteilen, ob dies

ausschließlich Folge der Korpusgrößenabhängigkeit des TTR-Maßes ist oder ob sich dahinter zumindest teilweise auch eine unterschiedliche lexikalische Vielfalt verbirgt. Der Einfluss der Korpusgröße überlagert den Einfluss der lexikalischen Vielfalt, und um letztere messen zu können, benötigen wir daher eine Möglichkeit, den Einfluss der Korpusgröße sozusagen aus dem TTR-Maß herauszurechnen.

alternative Maße

Es wurden hierfür in der Literatur zahlreiche Vorschläge gemacht, eine Übersicht und kritische Diskussion findet sich bei Tweedie/Baayen (1998) und auch bei McCarthy (2005). Eine naheliegende, pragmatische Lösung, die offenbar zuerst von Johnson (1944) präsentiert wurde, ist die *mean segmental type-token ratio* (*MSTTR*), in der Korpuslinguistik auch unter der Bezeichnung *Standardisiertes Type-Token-Verhältnis* (Englisch: *standardized type-token ratio*, kurz: *STTR*) bekannt.

MSTTR

STTR

STTR-Berechnung

Um einen STTR-Wert zu berechnen, ermittelt man zunächst den normalen TTR-Wert für die ersten L Worttokens (als würden sie zusammen ein eigenes kleines Korpus bilden), danach dasselbe für die nächsten L Worttokens und so weiter, bis man am Ende des Korpus angelangt ist. Auf diese Weise wird das gesamte Korpus in viele Segmente derselben Größe L zerlegt, so dass die TTR-Werte der einzelnen Segmente miteinander vergleichbar sind. Der STTR-Wert für das gesamte Korpus ist nun einfach als der Durchschnitt der TTR-Werte aller Segmente definiert. Genau wie die einzelnen TTR-Werte liegt auch der STTR-Wert zwischen 0 und 100%. Als Segmentlänge wählt man typischerweise einen Wert zwischen 100 und 2.000, bei sehr großen Korpora sind aber auch deutlich größere Segmentlängen sinnvoll.

Bei der hier beschriebenen Vorgehensweise wird das Korpus so behandelt, als wäre es ein einziger großer Text, d. h. die Grenzen zwischen den Korpustexten werden ignoriert und Segmente können über diese Textgrenzen hinaus gebildet werden. Nach dem letzten Segment bleibt normalerweise restliches Textmaterial übrig, das aus weniger als L Tokens besteht und daher kein eigenes Segment bildet – dieses Material wird bei der Berechnung des STTR-Werts i. A. ignoriert.

Segmente und Textgrenzen

alternative STTR-Berechnung

Segmente über Textgrenzen hinaus zu bilden, ist durchaus fragwürdig, denn unterschiedliche Texte stellen natürlich unterschiedliche Sprachproduktionsakte dar und können sich in ihren Eigenschaften (insbesondere in ihrer lexikalischen Vielfalt) stark voneinander unterscheiden. Dieses Problem ließe sich umgehen, indem man einen STTR-Wert für jeden einzelnen Korpustext berechnet und anschließend aus allen STTR-Werten einen Mittelwert für das gesamte Korpus bildet. Der Nachteil hierbei ist aber, dass für jeden Text nach dem letzten Segment Textmaterial ignoriert werden muss und dass zudem Texte, die kürzer als ein Segment sind, gar nicht in die STTR-Berechnung einfließen. Daher behandeln wir in diesem Abschnitt bei der STTR-Berechnung das Korpus wie einen einzigen Text und lassen also Segmente, die über Textgrenzen hinausgehen, zu.

Fortsetzung des Beispiels von S. E6-2
Verwendet man als Segmentlänge $L = 2.000$, dann ist das Standardisierte Type-Token-Verhältnis in den beiden Beispielkorpora BRZ08 und NUN08 (in %):

BRZ08: 50,444 NUN08: 51,619

Zur Interpretation dieser Zahlen: In den Segmenten im Korpus BRZ08 wurde durchschnittlich ein TTR-Wert von 50,444% gemessen. Da die Segmente jeweils 2.000 Worttokens umfassen, wurden also pro Segment durchschnittlich ca. 1.009 Worttypes beobachtet. Für NUN08 ergibt sich analog, dass pro Segment durchschnittlich ca. 1.032 Worttypes beobachtet wurden.
Diese STTR-Ergebnisse legen nahe, dass der gegenüber BRZ08 beobachtete größere TTR-Wert beim Korpus NUN08 (vgl. das Beispiel auf S. E6-2) in erheblichem Ausmaß an dessen kleinerer Korpusgröße lag, denn die entsprechenden STTR-Werte liegen relativ nahe beieinander. Die lexikalische Vielfalt scheint in beiden Korpora sehr ähnlich zu sein, wenn auch für NUN08 noch immer etwas größer – ein Befund, der angesichts der ähnlichen Textsorte durchaus plausibel ist, aber nicht unbedingt zu erwarten war. Wir werden auf dieses Zwischenfazit zurückkommen.

STTR-Nachteil

Das STTR-Maß hängt – im Gegensatz zum TTR-Maß – nicht von der Korpusgröße ab. Dennoch ist es keine optimale Lösung, denn STTR-Werte können nur bei Verwendung derselben Segmentlänge L verglichen werden (wie im Beispiel oben). Dieser Nachteil ließe sich theoretisch umgehen, indem man eine einheitliche Segmentlänge vereinbart – genaugenommen wäre das STTR-Maß erst dann wirklich standardisiert. Leider aber gibt es keine in irgendeiner Hinsicht besonders naheliegende oder gar optimale Wahl für L. Bei einer größeren Segmentlänge hat das STTR-Maß eine größere Messempfindlichkeit, kann also die lexikalische Vielfalt verschiedener Korpora besser unterscheiden (vgl. McCarthy 2005, Kap. 2), gleichzeitig kann das STTR-Maß mit Segmentlänge L nicht auf Korpora oder Einzeltexte angewendet werden, die kürzer sind als L. Für möglichst breite Anwendbarkeit würde man L also möglichst klein wählen, angesichts ständig wachsender Korpora hingegen würde man gerne ein größeres L wählen, so dass eine optimale und dauerhafte Festlegung von L nicht möglich ist.

Messempfindlichkeit

Erst vor wenigen Jahren hat McCarthy (2005) ein Maß entwickelt, das diesen Konflikt offenbar behebt: Denn es ist nach bisherigem Forschungsstand für alle Korpusgrößen zuverlässig anwendbar und bietet gleichzeitig eine hohe Messempfindlichkeit. Mit diesen Eigenschaften hat sein *measure of textual lexical diversity* (kurz: *MTLD*) gute Aussichten, sich zukünftig in der Korpuslinguistik und angewandten Disziplinen als Standardmaß für lexikalische Vielfalt zu etablieren.

MTLD

Für die genaue Definition des MTLD-Maßes verweisen wir auf McCarthys o. g. Dissertation und beschränken uns hier auf eine sehr knappe Beschreibung der Grundidee. Wie beim STTR-Maß wird das Korpus auch hier in Segmente zerlegt, mit dem Unterschied, dass beim MTLD-Maß die Segmentlängen variieren. Die Länge eines Segments hängt von der lexikalischen Vielfalt in diesem Segment ab: je vielfältiger, desto länger ist das Segment. Dabei macht sich MTLD den ursprünglich problema-

Grundidee

tischen Umstand, dass der TTR-Wert bei einem wachsenden Korpus tendenziell allmählich fällt (vgl. den Kasten auf S. E6-3), sogar zunutze:

MTLD-Berechnung

Bei der Berechnung von MTLD-Werten geht man das Korpus Token für Token durch und berechnet nach jedem Token den TTR-Wert. Das Ende eines Segments ist erreicht, sobald der TTR-Wert einen fest vorgegebenen Wert t unterschreitet. Mit dem nächsten Token beginnt ein neues Segment, und Types und Tokens werden wieder von Null an gezählt. Mit anderen Worten: Während beim STTR-Maß TTR-Werte zu Segmenten derselben Segmentlänge berechnet werden, werden beim MTLD-Maß Segmente mit etwa demselben TTR-Wert gezählt. Zu dieser Segmentzahl wird noch ein Segmentbruchteil für das Textmaterial nach dem letzten Segment hinzugezählt – anders als beim STTR-Maß berücksichtigt MTLD also das gesamte Korpus. Aus der Anzahl der so gefundenen Segmente leitet McCarthy einen globalen MTLD-Wert ab (die konkrete Formel finden Sie in McCarthy 2005). Analog zu TTR und STTR deutet auch ein größerer MTLD-Wert auf eine größere lexikalische Vielfalt hin.

MTLD-Parameter

Auch beim MTLD-Maß gibt es einen Parameter, nämlich den vorgegebenen Wert t. Anders als bei dem Parameter L des STTR-Maßes lässt sich für t jedoch ein empirisch begründeter Standardwert festlegen, der auch bei immer größer werdenden Korpora unverändert bleiben kann. McCarthys bisherige Studien legen den Wert $t = 0{,}71$ nahe. Sofern sich dieser

MTLD ist standardisiert

(oder auch ein anderer) Standardwert durchsetzt, ist MTLD tatsächlich ein standardisiertes Maß für die lexikalische Vielfalt eines Korpus. Der

MTLD-Nachteil

einzige Nachteil des MTLD-Maßes ist, dass MTLD-Werte nicht intuitiv charakterisiert werden können (im Gegensatz etwa zu TTR-Werten, die man als »Anteil der verschiedenen Tokens an allen Tokens« beschreiben und interpretieren kann, vgl. S. E6-2). Diesen Preis muss man offenbar zahlen, wenn man in der Lage sein möchte, die lexikalische Vielfalt verschiedener Korpora zu messen und miteinander zu vergleichen.

Fortsetzung des Beispiels von S. E6-5
Für die beiden Beispielkorpora BRZ08 und NUN08 ergeben sich (mit $t = 0{,}71$) die folgenden MTLD-Werte:

BRZ08: 173,46 NUN08: 215,76

Im Vergleich zu den STTR-Analysen im Kasten auf S. E6-5 tritt der Unterschied zwischen NUN08 und BRZ08 hier deutlicher hervor. Dieser deutlichere Unterschied ist durchaus typisch für das MTLD-Maß und illustriert seine größere Messempfindlichkeit (s.o.).

Wir haben in diesem begleitenden Abschnitt drei verschiedene Maße für lexikalische Vielfalt vorgestellt und ihre jeweiligen Vorzüge und Schwächen kurz skizziert. Genau wie Wortfrequenzen liefern diese Maße aber zunächst nur Beobachtungen für das konkret untersuchte Korpus und sagen noch nichts über die zugrunde liegende Grundgesamtheit aus (vgl. Abschn. 3.2 im Buch). Bei den Wortfrequenzen gibt es etablierte

lexikalische Vielfalt in der Grundgesamtheit?

Methoden, um von den beobachteten Frequenzen Rückschlüsse auf die jeweilige Grundgesamtheit zu ziehen, insbesondere mithilfe von Konfidenzintervallen (s. Abschn. 6.4 und 6.6 im Buch). Auch für die Maße TTR, STTR und MTLD wären Konfidenzintervalle grundsätzlich geeignet, um Schlussfolgerungen vom Korpus auf die Grundgesamtheit kontrolliert zu handhaben. Es ist aber unklar, wie Konfidenzintervalle für diese drei Maße berechnet werden müssten, und nach unserer Kenntnis liegen hierzu bislang keine Forschungsarbeiten vor. Klar ist nur, dass Konfidenzintervalle für diese Maße anders berechnet werden müssen als für einzelne Wortfrequenzen, z. B. weil die Maße mit Sicherheit nicht binomialverteilt sind (vgl. Abschn. 6.4 im Buch), auch nicht näherungsweise.

Konfidenzintervalle

Dieser ergänzende Abschnitt verdeutlicht zwei Dinge, die über das Konzept der lexikalischen Vielfalt hinausgehen: (i) dass die Zipf-Verteilung einige nicht zu unterschätzende Konsequenzen auf die korpuslinguistische Arbeit hat und (ii) dass aussagekräftige korpuslinguistische Forschungsergebnisse nur dann möglich sind, wenn man sich zunächst mit den Eigenschaften der verwendeten Methoden und Daten auseinandersetzt.

übergeordnete Erkenntnisse

Literatur

Johnson, W. (1944): Studies in language behavior: I. A program of research. *Psychological Monographs*, 56, 1–15.

McCarthy, P. M. (2005): *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. PhD thesis, The University of Memphis. https://umdrive.memphis.edu/pmmccrth/public/Papers/MTLD%20dissertation.doc (07.06.2010).

Richards, B. (1987): Type/Token Ratios: what do they really tell us? *Journal of Child Language*, 14, 201–209.

Tweedie, F. J. / Baayen, R. H. (1998): How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–352.