Publié dans Revue Tranel (Travaux neuchâtelois de linguistique) 55, 45-57, 2011,
source qui doit être utilisée pour toute référence à ce travail

Vergleichbarkeit von sprachstatistischen Messungen

Uwe QUASTHOFF & Thomas ECKART
Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig

The quality of a measurement can be described by a specification of the process creating the result and by inspection of the result. In the case of inspection, an exact measurement of previously specified parameters ensures compatibility of different measurements performed by different researchers on similar objects. The comparison of different measured values requires an exact description of the measuring process.
Today, even simple measurements like the size of a corpus measured by the number of tokens or the average sentence length measured in words are not comparable. Therefore a more standardized way of corpus measurement is strongly needed.

1. Einleitung

Quantitative Messungen dienen der Beschreibung von Objekten durch (möglicherweise mit Maßeinheiten behaftete) Zahlen und ermöglichen dadurch den Vergleich verschiedener Objekte. Dabei werden diese Objekte zu Messobjekten. Zunächst wird die zu bestimmende Messgröße ausgewählt, dazu passend eine Messmethode. Das Messverfahren beschreibt schließlich das praktische Vorgehen zur Gewinnung der Messwerte (DIN 1319 1995).

An solche Messungen werden in der Regel einige Anforderungen gestellt:

- Der Messwert soll unabhängig sein von Messmethode und Messverfahren.
- Die Messung soll zu einem anderen Zeitpunkt am selben oder einem ähnlichen Messobjekt wiederholbar sein und dann das gleiche (oder ein ähnliches) Ergebnis liefern.
- Systematische Abweichungen müssen berücksichtigt (und eventuell korrigiert) werden.

In der Praxis besitzen Messergebnisse einen gewissen Messfehler, d.h. der zu messende Wert kann nur näherungsweise ermittelt werden. Die Größe des Messfehlers entscheidet darüber, ob zwei Objekte mit nahe beieinander liegenden Messwerten als gleich oder unterschiedlich groß betrachtet werden sollen.

Angewandt auf den Fall textueller Messobjekte bedeutet dies: Durch Messungen an Texten, Korpora oder Sprachen lassen sich charakteristische Messwerte erzeugen, welche die entsprechenden Objekte beschreiben. Mehrere solche Messwerte können möglicherweise dazu benutzt werden, um beispielsweise

- verschiedene Sprachen,
- Korpora einer Sprache, aber verschiedener Genres, oder
- Texte verschiedener Autoren

zu unterscheiden. Ebenso sollte es möglich sein, Ähnlichkeiten zwischen solchen Objekten auf der Basis vergleichbarer Messwerte zu identifizieren, etwa dass die niederländische Sprache dem Deutschen ähnlicher ist als das Finnische.

Nimmt die Zahl der zu vergleichenden Objekte allerdings zu, so wird es schwieriger, diese Objekte durch ihre Messwerte voneinander zu unterscheiden. Hier wird es wichtig, über eine große Anzahl von Messgrößen zu verfügen und die Messfehler möglichst klein zu halten, um bestehende Unterschiede nachweisen zu können.

2. Messungen an Korpora

Die folgenden Schwierigkeiten im Zusammenhang mit Messungen an Korpora wurden den Autoren bei den sprachstatistischen Analysen zu einem Häufigkeitswörterbuch (Quasthoff et al., 2011) besonders deutlich.

Betrachten wir Messgrößen einer Sprache etwas näher. Einige Messgrößen besitzen scheinbar einfache Messverfahren wie

- die durchschnittliche Wortlänge,
- die durchschnittliche Satzlänge (gemessen in der Anzahl von Zeichen oder Wörtern) und
- die Type-Token-Ratio

Andere Messverfahren erscheinen komplizierter, etwa für Messgrößen wie

- die durchschnittliche Silbenlänge und
- den Anstieg der Ausgleichsgeraden für die Darstellung des Zipfschen Gesetzes in doppelt-logarithmischen Koordinaten.

Die Ursache für Ungenauigkeiten in den Messwerten hat verschiedene Gründe:

1. Die Messung kann nicht direkt an ‚der Sprache' vorgenommen werden, sondern an einem für die Messung auszuwählenden (oder zusammenzustellenden) Korpus. Diese Zusammenstellung hat möglicherweise Einfluss auf die Messergebnisse. Aber nicht nur die Zusammensetzung eines Korpus, sondern auch allein seine Größe kann Auswirkungen auf die Größe eines Messwertes haben. In günstigen Fällen hängt ein

Messwert nicht von der Korpusgröße ab (z.B. durchschnittliche Satzlänge). Ebenso leicht zu handhaben ist eine lineare Abhängigkeit von der Korpusgröße wie bei der Anzahl der laufenden Wortformen. Komplizierter ist der Fall einer nichtlinearen oder unbekannten Abhängigkeit (Type-Token-Ratio). Beim Vergleich der Messwerte für verschiedene Korpora bzw. Sprachen sind hier die Vergleichbarkeit der zugrundeliegenden Daten und der gleiche Umfang der Datenmenge die wichtigsten Voraussetzungen.

2. Bei der Erstellung des Korpus werden Vorverarbeitungsschritte ausgeführt. Bevor die durchschnittliche Satzlänge bestimmt wird, werden möglicherweise nicht-wohlgeformte Sätze entfernt. Da durch Fehler in der Vorverarbeitung entstehende nicht-wohlgeformte Sätze (z.B. durch Anfügen des ersten Satzes an die Überschrift) häufig länger als üblich sind, hat die Vorverarbeitung Auswirkungen auf die Messung.
3. Die durchschnittliche Wortlänge hängt von der Wortdefinition ab. Die durchschnittliche Länge der durch Leerzeichen getrennten Zeichenketten unterscheidet sich von der durchschnittlichen Länge der tatsächlich vorkommenden Wörter (also z.B. ohne Satzzeichen, Zahlen usw.). Bei den möglicherweise kleinen Unterschieden zwischen verschiedenen Sprachen oder Genres sichert nur eine exakte Wortdefinition die Vergleichbarkeit der Messwerte. Ebenso hat die Wortdefinition Auswirkungen auf die Type-Token-Ratio, wie im Abschnitt 4.2 sichtbar wird.
4. Möglicherweise kann nicht die gewünschte Größe selbst gemessen werden, sondern es wird auf eine einfacher zu messende Hilfsgröße zurückgegriffen. Diese Hilfsgröße weicht von der ursprünglich zu messenden Größe leicht ab, der Messwert muss dann korrigiert werden. Als Beispiel dienen hier Silbenlänge und Silbenzahl. Die Bestimmung der durchschnittlichen Silbenlänge oder der durchschnittlichen Silbenzahl pro Wort erfordert die Bestimmung der Silbenzahl für jedes Wort. Anders als man vielleicht zunächst vermutet, ist zur Bestimmung der Silbenzahl nicht die korrekte Zerlegung in Silben nötig. Für ein zweisilbiges Wort muss nur erkannt werden, das darin genau eine Silbengrenze vorkommt, nicht aber, wo diese genau liegt. Da jede Silbe genau einen Silbengipfel enthält und dieser aus einem oder mehreren unmittelbar benachbarten Vokalen (im Folgenden Vokalgruppe genannt) besteht und die Silbe sonst keine weiteren Vokale enthält, entspricht die Anzahl der Silben genau der Anzahl der Silbengipfel. In den allermeisten Fällen stehen zwischen verschiedenen Silbengipfeln weitere Konsonanten, so dass die Anzahl der Silben gleich der Anzahl der Vokalgruppen ist. Es ist also leicht festzustellen, dass das Wort *Tischtuch* wegen der zwei isolierten Vokale aus zwei Silben besteht. Die Ermittlung der Silbengrenze ist auf diese

einfache Weise jedoch nicht möglich, da es hierfür mehrere Möglichkeiten gibt. Jedoch kann die tatsächliche Silbenzahl von der Zahl der Vokalgruppen abweichen, wenn z.B. zwei Silbengipfel ohne dazwischenliegenden Konsonanten aufeinander treffen, z.B. bei *beachten* oder *Teeei*. Wenn bekannt ist, bei welchem Anteil der Silbengrenzen dieser Effekt auftritt, kann dies genutzt werden, um den Messwert in die richtige Richtung zu korrigieren.

5. Eine Messung kann auch technisch kompliziert sein, wie am Beispiel des Zipfschen Gesetzes gezeigt werden soll. Das Zipfsche Gesetz (Zipf, 1935) sagt auch in seiner Verallgemeinerung durch Mandelbrot (Mandelbrot, 1953) folgenden Zusammenhang voraus: Ordnet man die Wortformen eines Korpus nach ihrer Häufigkeit und wählt die Position eines Wortes als seinen Rang, dann liefert der Zusammenhang zwischen dem Rang eines Wortes und seiner Häufigkeit bei doppeltlogarithmischer Darstellung näherungsweise eine Gerade (s. Abb. 1). Von Interesse ist nun der Anstieg dieser Gerade, der näherungsweise -1 beträgt, sich aber für verschiedene Korpora oder Sprachen leicht unterscheiden kann.

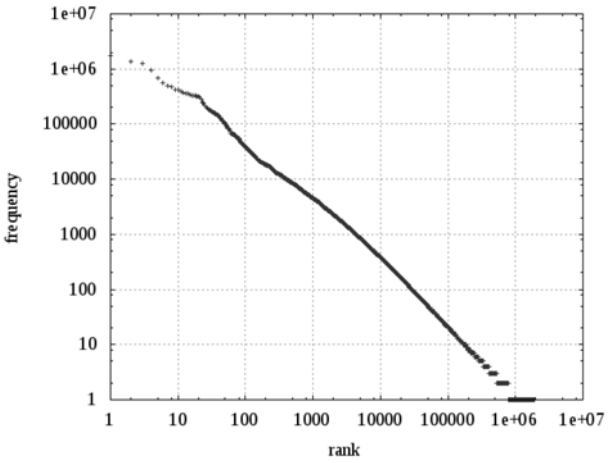

Abbildung 1: Das Zipfsche Gesetz für ein Korpus der deutschen Sprache

Die Schwierigkeiten bei der Messung haben zwei Gründe: Erstens sind die zu messenden Unterschiede nur gering, die Messung muss also mit hoher Exaktheit durchgeführt werden. Zweitens erfüllt ein Korpus typischerweise das Zipfsche Gesetz nur näherungsweise, d.h. der tatsächliche Zusammenhang zwischen logarithmierter Häufigkeit und logarithmiertem Rang ist nur näherungsweise linear, speziell für sehr kleine und große Ränge sind die Abweichungen oft erheblich. Aus linguistischen Gründen werden die von den allerhäufigsten Wörtern verursachten Abweichungen als nicht relevant betrachtet. Um sinnvoll vom Anstieg sprechen zu können, möchte man als zu berücksichtigenden Messbereich möglicherweise nur den mittleren Frequenzbereich untersuchen, in dem noch am ehesten ein linearer Zusammen-

hang vorliegt. Die Beispiele zeigen jedoch, dass sich die Auswahl des Messbereichs relativ stark auf den Messwert auswirkt, und Veränderungen des Messbereichs zu stärkeren Abweichungen führen können als sie zwischen verschiedenen Korpora oder Sprachen zu finden sind. Hier ist also eine genaue Festlegung des Messvorganges nötig, damit Veränderungen bei der Messung nicht größere Abweichungen erzeugen, als sie zwischen verschiedenen Messobjekten gemessen werden sollen.

3. Einheitliche Korpuserstellung

Grundlage jeder hier beschriebenen Messung ist ein Korpus. Um vergleichbare Messwerte zu erhalten, sind vergleichbare Korpora notwendig, die nach einem einheitlichen Verfahren erstellt wurden. Im Folgenden wird der Prozess der Korpuserstellung mit den Auswirkungen auf die nachfolgenden Messungen beschrieben.

3.1 Datenaufbereitung

Das Wortschatzprojekt der Universität Leipzig[1] sammelt kontinuierlich Textmaterial verschiedener Sprachen, Register und Genres für die Erstellung großer bis sehr großer Textkorpora und reichert diese mit diversen Metadaten an (wie Wortart-Annotationen, Grundforminformationen etc.). Mittlerweile liegen mehrere hundert Korpora in rund 100 verschiedenen Sprachen vor. Um auch mit dem kontinuierlichen Zustrom von Daten umgehen zu können (derzeit wird bei steigender Tendenz monatlich 30 Gigabyte neues Textmaterial gesammelt) und dessen zügige Weiterverarbeitung zu garantieren, wurde eine dedizierte Prozesskette zur Korpuserstellung entwickelt, die Gegenstand ständiger Verbesserung ist (Quasthoff & Eckart, 2009).

Haupteinsatzgebiet der erstellten Korpora ist die statistische und musterbasierte Verarbeitung natürlicher Sprache beziehungsweise die automatische Wissensextraktion. Konkret beinhaltet dies Arbeitsgebiete wie Named Entity Recognition, Textclustering und ähnliches. Die im Rahmen des Projektes erstellten Daten und Algorithmen werden u.a. auch über verschiedene REST und SOAP basierte Webservices angeboten (Büchler & Heyer, 2009).

Da eine große Breite an Eingabematerial verwendet wird, erschien ein möglichst standardisierter Workflow als eine essentielle Vorbedingung, um eine größtmögliche Vergleichbarkeit zwischen Korpora zu erhalten. Aufgrund der unterschiedlichen Verfügbarkeit von Werkzeugen variiert der Umfang (und auch die Qualität) der erstellten Metadaten. Ein Korpus

[1] Online erreichbar unter http://wortschatz.uni-leipzig.de/

besteht dabei mindestens aus Sätzen, Informationen über genutzte Quellen und diverse statistische Analysedaten wie Worthäufigkeiten und statistische Wortkookkurrenzen. Falls für die jeweilige Sprache verfügbar, werden zusätzliche Informationen wie Wortart, Grundformen oder semantische Kategorien hinzugefügt.

Abbildung 2 zeigt eine graphische Übersicht des gesamten Prozesses.

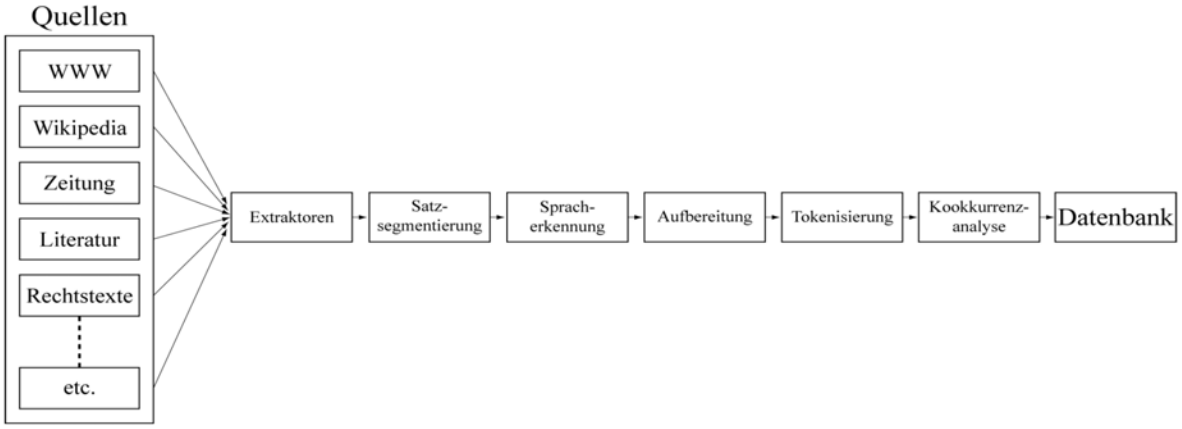

Abbildung 2: Übersicht des Aufbereitungsprozesses

3.2 Textquellen

Das Wortschatzprojekt nutzt diverse Quellen, um Textmaterial und ergänzende Metadaten zu sammeln und für die weitere Nutzung aufzubereiten. Erste Aktivitäten zur Textdatenbeschaffung begannen im Jahr 1995; seit 2000 stellt das World Wide Web die zentrale Ressource dar. Aus diesem Grund wurden verschiedene Textcrawler und Extraktoren für teilweise sehr spezifische Ressourcen erstellt, darunter der verteilte Webcrawler Findlinks (Biemann et al., 2007).

Derzeit werden hauptsächlich folgende Textquellen zur Korpuserstellung genutzt:

- Online verfügbare Zeitungsartikel,
- zufällig gecrawlte Texte aus dem WWW sowie
- die Onlineenzyklopädie Wikipedia.

Falls nötig werden weitere Quellen als Eingabematerial für neue Korpora verwendet. Unter anderem existieren mittlerweile Korpora basierend auf Filmuntertiteln, Chatroom-Texten, Twitter-Nachrichten oder weiterem öffentlich verfügbaren Textmaterial wie beispielsweise Texte des Projekts Gutenberg.

3.3 Vorverarbeitung

Der gesamte Vorverarbeitungsprozess arbeitet auf einem einheitlichen XML-basierten Eingabeformat, das durch die verschiedenen Textcrawler

bereitgestellt wird. In einem ersten Schritt der Vorverarbeitung werden die Texte durch einen musterbasierten Segmentierer in Sätze zerlegt. Zusätzliche Ressourcen wie Listen mit sprachspezifischen Satzendezeichen oder gängigen Abkürzungen verbessern die Qualität der Satzsegmentierung.

Um die gerade bei der Nutzung von heterogenen Textressourcen häufig anfallenden unerwünschten Sätze zu entfernen, werden verschiedene Prozeduren durchgeführt um das Material zu bereinigen. Im ersten Putzschritt werden eventuell vorkommende Duplikate entfernt, die für die spätere statistische Auswertung und Analyse problematisch sein können. Gerade bei der Nutzung generischer Webseiten ist dies ein unverzichtbarer Arbeitsschritt, um dem hohen Anteil an Standardklauseln oder Textbausteinen zu begegnen.

Anschließend werden die Sätze durch einen statistischen Sprachidentifizierer nach Sprachen aufgetrennt. Dabei werden die Wortfrequenzen jedes einzelnen Satzes mit vorhandenen Sprachprofilen (auf der Basis relativer Worthäufigkeiten hochfrequenter Terme) verglichen und die jeweils ähnlichste Sprache gewählt (Dunning 1994). Im letzten Arbeitsschritt werden reguläre Ausdrücke auf die verbliebenen Sätze angewendet um automatisch unerwünschte nicht wohlgeformte Sätze zu entfernen. Analog zur Satzsegmentierung existieren allgemeine Regeln (beispielsweise untypische Satzlängen oder Buchstaben/Sonderzeichen-Verhältnisse), aber auch sprach- und genrespezifische Muster.

3.4 Erstellung von Textdatenbanken

Anschließend werden die Sätze durch die Textanalyse-Software Medusa verarbeitet (Büchler, 2008). Arbeitsschritte umfassen hier die Wortidentifikation im Tokenisierungsprozess, Erstellung einer inversen Liste und Wortkookkurrenzanalyse mittels diverser statistischer Signifikanzmaße (Dunning, 1993). Abschließend werden Informationen zu zeichenbasierter (Levenshtein-Distanz) und semantischer Wortähnlichkeit (wie der Ähnlichkeit der Wortkookkurrenz-Profile) hinzugefügt sowie (falls für die jeweilige Sprache verfügbar) Grundformreduktion und Wortartidentifikation (sog. POS-Tagging) durchgeführt. Die erstellten Korpora werden üblicherweise in relationalen Datenbanken gespeichert und können durch diverse Softwarewerkzeuge analysiert und weiterverarbeitet werden.

4. Beispiele

Die folgenden Beispiele zeigen verschiedene Abhängigkeiten der Messwerte von der Wortdefinition, von der Vorverarbeitung sowie von der Korpusgröße bzw. Messbereichsauswahl. Diese Abweichungen zwischen verschiedenen Messobjekten einer Sprache sind möglicherweise größer als

die Abweichungen zwischen verschiedenen Sprachen für Messobjekte, die nach gleichen Kriterien erstellt wurden.

4.1 Einfluss der Wortdefinition auf die Wortanzahl

Für die Definition eines Wortes gibt es verschiedenste Ansätze (Fuhrhop 2008). Als einfache Vorgehensweise, die bereits für viele Anwendungen tauglich ist, lässt sich ein Wort als eine Folge von Zeichen definieren, die von Leerräumen oder Satzzeichen umgeben sind. Um saubere Wortlisten zu erhalten sind bei dieser Herangehensweise allerdings weitere Vorverarbeitungsschritte nötig, bei denen Nichtwörter entfernt werden. Zu diesen unerwünschten Zeichenketten können beispielsweise URLs, offensichtliche Schreibfehler oder auch Zahlen gehören. Im Rahmen einer automatischen Textverarbeitung und -aufbereitung wird üblicherweise ein musterbasierter Ansatz verfolgt, der diese Fehlerklassen über spezifische Merkmale identifiziert. Da diese Muster allerdings nicht sprachunabhängig sind und selten in gleich guter Qualität vorliegen, unterscheiden sich die erzielten Resultate teils deutlich. Als Konsequenz ergeben sich in Abhängigkeit von Wortdefinition sowie Umfang und Qualität der Vorverarbeitung deutlich unterschiedliche Anzahlen der identifizierten Wörter.

Eine Beispielanalyse zeigt diesen Zusammenhang für verschiedene Korpora der Sprachen Deutsch, Englisch und Französisch. Dabei wurden auf der Basis von Zeitungstexten jeweils drei Korpora mit 10.000, 100.000 und 1.000.000 Sätzen erstellt. Als Baseline diente die Implementation eines Standardtokenisierers, der abgesehen von Sonderbehandlung von Anführungsstrichen und zusätzlichen Mehrwortlexem- und Abkürzungslisten weitestgehend der oben skizzierten "naiven" Wortdefinition entspricht.

In Tabelle 1 wird die Anzahl der durch diesen Tokenisierer identifizierten Wortformen in Spalte 2 dargestellt. Im Rahmen eines Säuberungsprozesses wurden auf der Basis dieser Wortformenliste alle Wörter entfernt, die nicht ausschließlich aus einer festen Menge von Buchstaben, maximal zwei Ziffern und ausgewählten Sonderzeichen (wie dem Apostroph) bestehen (Spalte 3). Ein solcher Filter erwies sich bei der Erstellung eines deutschen Frequenzwörterbuches auf der Basis heterogener Textdaten (Quasthoff et al., 2011) als sinnvoll. Ein zweiter Filter (Spalte 4) entfernt zusätzlich alle Mehrwortlexeme und alle Zahlen beziehungsweise Datumsangaben. Für jeden Wert wird zusätzlich der prozentuale Anteil im Verhältnis zur ursprünglich bestimmten Wortmenge angegeben.

Korpus (Anzahl Sätze)	Anzahl Types (in %)	Filter 1: Anzahl Types ohne unerwünschte Sonderzeichen (in %)	Filter 2: Zusätzliche Säuberungsprozeduren (in %)
Deutsch (10.000)	39.334 (100%)	38.848 (98,8%)	33.530 (85,2%)
Deutsch (100.000)	191.041 (100%)	187.028 (97,9%)	159.184 (83,3%)
Deutsch (1.000.000)	830.641 (100%)	799.010 (96,2%)	683.651 (82,3%)
Englisch (10.000)	27.882 (100%)	27.215 (97,6%)	24.711 (88,6%)
Englisch (100.000)	109.734 (100%)	104.907 (95,6%)	90.922 (82,9%)
Englisch (1.000.000)	414.380 (100%)	377.404 (91,1%)	317.425 (76,6%)
Französisch (10.000)	31.898 (100%)	30.054 (94,2%)	28.798 (90,3%)
Französisch (100.000)	119.728 (100%)	109.750 (91,7%)	103.477 (86,4%)
Französisch (1.000.000)	422.476 (100%)	372.492 (88,2%)	342.483 (81,1%)

Tabelle 1: Wortanzahl für verschiedene Wortdefinitionen und Korpora

Abbildung 3 zeigt zusätzlich die Entwicklung der Wortanzahl für fünf verschiedene Korpusgrößen auf der Basis von deutschsprachigen Zeitungstexten für die gleichen Verarbeitungsprozeduren.

Es zeigt sich insgesamt sehr deutlich dass sich bereits bei diesen einfachen Veränderungen in der Vorverarbeitung und Aufbereitung signifikant unterschiedliche Anzahlen ergeben: Für einzelne Korpora ergibt sich ein Verlust an Wortformen von bis zu 24%. Dieser Unterschied stellt somit eine massive Fehlergröße für folgende Analysen dar, so dass die Nutzung unterschiedlicher Wortdefinitionen für vergleichende Arbeiten ein schwer korrigierbares Problem darstellt.

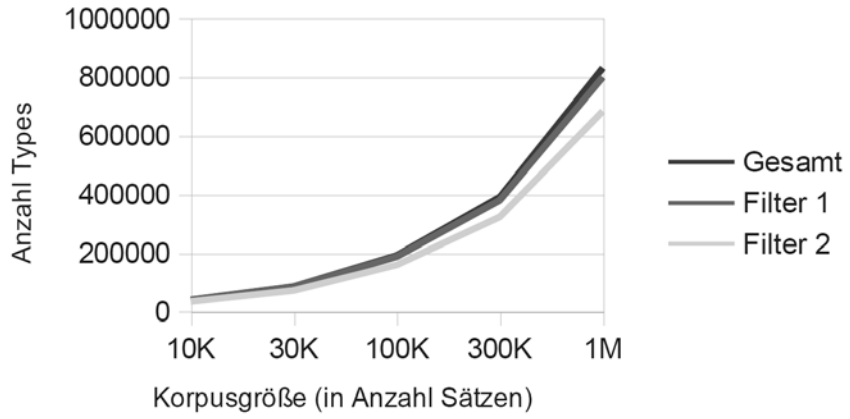

Abbildung 3: Anzahl Types für verschiedene Wortdefinitionen und Korpusgrößen

4.2 Einfluss der Wortdefinition auf die Type-Token-Ratio

Ein beliebtes Maß zur Analyse von Texten beziehungsweise Korpora ist die Type-Token-Ratio. Für die folgenden Angaben wurde jeweils als einfache Definition dieses Maßes TTR= |Anzahl Types|/|Anzahl Tokens| genutzt.

Tabelle 2 zeigt die Auswirkung der drei verschiedenen in Kapitel 4.1 beschriebenen Wortdefinitionen auf die sich ergebenden Werte. Deutlich erkennbar sind die weitestgehend systematischen Unterschiede durch die verschiedenen Mengen berücksichtigter Wortformen. Einerseits verkleinert sich das Type-Token-Ratio bei wachsender Korpusgröße. Andererseits geht die Veränderung bei einer strengeren Wortdefinition nicht immer in die gleiche Richtung. Filter 1 entfernt mehr häufige Wörter, so dass das Type-Token-Ratio steigt. Filter 2 hingegen sondert eher seltene Wörter aus, wodurch das Type-Token-Ratio wieder sinkt.

Korpus (Anzahl Sätze)	TTR für alle Wortformen	TTR mit Filter 1	TTR mit zusätzlichem Filter 2
Deutsch (10K)	0,187	0,236	0,218
Deutsch (100K)	0,09	0,113	0,103
Deutsch (1M)	0,039	0,048	0,044
Englisch (10K)	0,121	0,147	0,138
Englisch (100K	0,048	0,057	0,051
Englisch (1M)	0,018	0,020	0,018
Französisch (10K)	0,122	0,148	0,148
Französisch (100K)	0,046	0,054	0,052
Französisch (1M)	0,016	0,018	0,017

Tabelle 2: Type-Token-Ratio für verschiedene Wortdefinitionen und Korpora

4.3 Auswirkung der Vorverarbeitung auf die Satzlängenverteilung

Bei der automatischen Extraktion von Sätzen aus heterogenen Textquellen haben sich zur Minimierung unerwünschter Bestandteile (Satzfragmente, Markup) verschiedene Muster als sinnvoll herausgestellt. Nach Bereinigung der Satzliste ergibt sich meist eine Satzlängenverteilung wie in Abbildung 4 dargestellt. Dort wird der prozentuale Anteil von Sätzen mit einer bestimmten Zeichenanzahl am Beispiel eines deutschen Webkorpus dargestellt. Eine solche Verteilung lässt sich (selbstverständlich mit unterschiedlichen Ausprägungen) sprachübergreifend antreffen.

Ebenfalls sprachübergreifend konnte festgestellt werden, dass im Bereich der überdurchschnittlich langen Sätze der Anteil unerwünschter Zeichenketten mit steigender Länge stetig zunimmt. Zumindest für die automatische Korpuserstellung hat sich für viele europäische Sprachen eine maximal akzeptierte Satzlänge von ungefähr 250 Zeichen als sinnvoll herausgestellt, längere Sätze werden hier ignoriert. Bei der Unmöglichkeit, bei großen Korpora (mit bis zu mehreren hundert Millionen Sätzen), manuell wohlgeformte von nicht wohlgeformten Sätzen zu trennen, scheint diese Vorgabe ein nützlicher Vorverarbeitungsschritt zu sein.

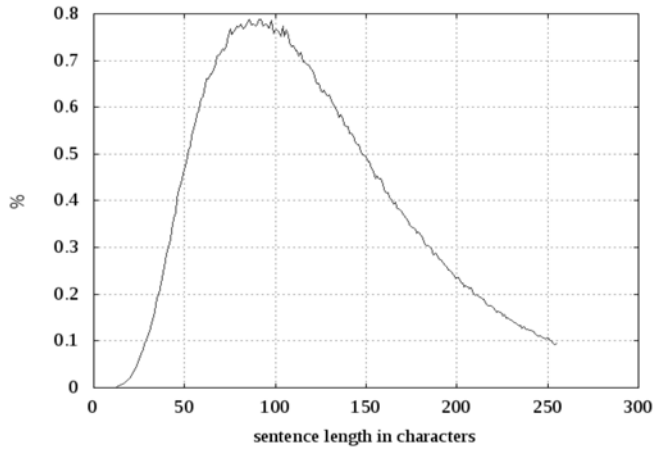

Abbildung 4: Typische Satzlängenverteilung bis 255 Zeichen

Die Konsequenzen aus diesem Eingriff in die Aufbereitung zeigen sich natürlich bereits bei einfachen Größen wie der durchschnittlichen Satzlänge. Tabelle 3 stellt die sich ergebende durchschnittliche Satzlänge für verschiedene maximal zulässige Maxima auf der Basis eines spanischen Zeitungskorpus dar.

Maximal zulässige Satzlänge (in Zeichen)	unbeschränkt	150	200	250	300	350	400
Durchschnittliche Satzlänge (in Zeichen)	134,38	84,07	105,33	122,28	132,51	134,36	134,38

Tabelle 3: Zusammenhang eines möglichen Vorverarbeitungsparameters auf die durchschnittliche Satzlänge eines spanischen Zeitungskorpus

4.4 Zipfsches Gesetz

Die folgende Tabelle 4 zeigt die unterschiedlichen Werte für den Anstieg[2] der Ausgleichsgeraden analog zu Abb. 1, wenn dieser für unterschiedlich große Korpora und verschiedene Rangbereiche (jeweils zwischen MinRank und MaxRank) berechnet wird. Diese großen Abweichungen bei gleicher Korpusgröße ergeben sich daher, dass die Rang-Häufigkeits-Kurve leicht konvex ist und deshalb der Anstieg der Ausgleichsgeraden etwas steiler ist, wenn die Messung für größere Rangwerte vorgenommen wird.

Rangbereich		Korpusgröße		
MinRank	MaxRank	100K	1M	10M
1	1000	-0.977	-0.977	-0.977
1	10.000	-0.999	-1.002	-1.002
1	100.000	-1.055	-1.072	-1.074

[2] Der Anstieg wurde im angegebenen Bereich mittels auf der logarithmischen Skala äquidistanter Rangwerte berechnet.

1	1.000.000	-1.046	-1.123	-1.165
10	1000	-1.053	-1.053	-1.053
10	10.000	-1.030	-1.033	-1.033
10	100.000	-1.076	-1.095	-1.098
10	1.000.000	-1.063	-1.142	-1.187
100	1000	-1.009	-1.013	-1.012
100	10.000	-1.011	-1.016	-1.017
100	100.000	-1.087	-1.114	-1.118
100	1.000.000	-1.066	-1.166	-1.223

Tabelle 4: Änderung des Anstiegs für verschiedene Rangbereiche und Korpusgrößen

Für die Vergleichbarkeit ist es also unbedingt notwendig, die entsprechende Messung für gleiche Rangbereiche vorzunehmen. Da im Bereich der häufigsten Wörter ohnehin größere Abweichungen vom Zipfschen Gesetz vorliegen, bietet sich eine Messung im Rangbereich von 10 bis 10.000 an, da dabei einerseits ein Rangbereich von einer gewissen Größe abgedeckt wird und andererseits die Abhängigkeit von der Korpusgröße nicht stark ausgeprägt ist.

5. Schlussfolgerungen

Die ausgewählten Beispiele zeigen, dass statistische Messwerte für Korpora möglicherweise nur schwer vergleichbar sind. Deshalb gehört zu einer Messung immer eine Beschreibung des Messvorgangs und speziell auch die Vorbereitung des Messobjekts, d.h. das Verfahren der Korpuserstellung. Die Vergleichbarkeit verschiedener Messwerte kann erreicht werden durch eine Standardisierung der einzelnen Arbeitsschritte für Vorbereitung und Messung, für die hier erste Vorschläge gemacht wurden. Eine solche Standardisierung kann letztlich allerdings auf verschiedene Arten erfolgen. Üblicherweise wird auf eine mehr oder weniger detaillierte erklärende Darstellung der gewählten Arbeitsschritte und deren Durchführung zurückgegriffen. Diese sollte sinnvollerweise auch die "selbstverständlichen" Details enthalten, da fehlende Angaben unter anderem Blickwinkel eine alternative Durchführung erlauben könnten. Darüber hinaus ist ergänzend ein Ansatz wünschenswert der im Sinne einer "Reproducible Research" (Buckheit & Donoho, 1995) die direkte Reproduktion einer Messung ermöglicht. In einem solchen Fall enthält die Dokumentation eines Messvorganges nicht nur Quellenangaben und beschreibende Texte, sondern die vollständige Messumgebung mit allen verwendeten Werkzeugen, Parametern und Daten.

Bibliografie

Biemann, C., Heyer, G., Quasthoff, U. & Richter, M. (2007): The Leipzig Corpora Collection - Monolingual corpora of standard size. In: Proceedings of Corpus Linguistic 2007, Birmingham, UK, 2007.

Büchler, M. (2008): Medusa: Performante Textstatistiken auf großen Textmengen - Kookkurrenzanalyse in Theorie und Anwendung, Vdm Verlag Dr. Müller, 2008.

Büchler, M. & Heyer, G. (2009): Leipzig Linguistic Services - A 4 Years Summary of Providing Linguistic Web Services. In: Proceeding of TMS 2009 conference, Augustusplatz 10/11, 04109 Leipzig, Germany, 2009.

Buckheit, J. & Donoho, D. (1995): WaveLab and Reproducible Research, In: Wavelets and Statistics, Springer-Verlag, 1995, 55-81.

DIN 1319 (1995): Grundlagen der Messtechnik, Deutsches Institut für Normung e.V., 1995.

Dunning, T. (1993): Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, Volume 19, number 1.

— (1994): Statistical Identification of Language. In: Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University, March 1994.

Fuhrhop, N. (2008): Das graphematische Wort (im Deutschen): Eine erste Annäherung. Zeitschrift für Sprachwissenschaft, Vol. 27, Walter de Gruyter, 2008.

Mandelbrot, Benoit B. (1953); An information theory of the statistical structure of language, New York, Academic Press, In: Jackson,W. (Ed.) Communication Theory (S. 503-512).

Quasthoff, U. & Eckart, T. (2009): Corpus Building Process of the Project "Deutscher Wortschatz". GSCL Workshop: Linguistic Processing Pipelines, Potsdam, Germany, 2009.

Quasthoff, U., Fiedler, S. & Hallsteinsdóttir, E. (2011): Häufigkeitswörterbuch DEU, Leipziger Universitätsverlag 2011.

Zipf, G. K. (1935): The Psycho-Biology of Language. An Introduction to Dynamic Philology, Boston.