Textdatenbanken

Sommersemester 2020

11. Vorlesung

- Text-Genres und Korpusstatistik, Teil 2 -

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

Mögliche Vergleichskriterien: Numerische Parameter, lexikalischer Vergleich, struktureller Vergleich

Numerischer Vergleich: Wir messen numerische Parameter für Texte und vergleichen diese für verschiedenen Korpora in derselben oder in verschiedenen Sprachen.

Lexikalischer Vergleich durch Differenzanalyse:

- Welche Wörter treten im Korpus A auffällig häufig verglichen mit Korpus B auf?
- Korpus B kann ein anderes Spezialkorpus oder ein generelles Referenzkorpus sein.

Struktureller Vergleich verschiedener Sprachen: Wir zählen das Auftreten von Strukturen und nehmen große / kleine Häufigkeiten als Existenz / Nichtexistenz.

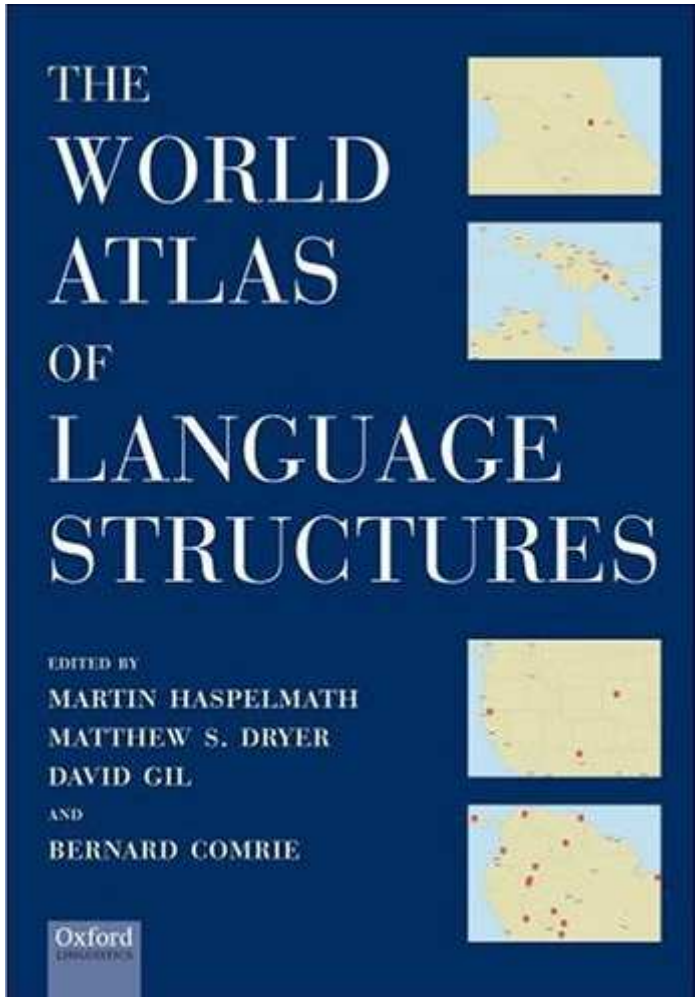
Struktureller Vergleich

Beispiel: Vergleich von unterschiedlichen Sprachen

The World Atlas of Language Structures - WALS

Online: *https://wals.info/*

- Fragen über jede Sprache wurden von Experten beantwortet
- Daraus wird ein Feature-Vektor erstellt
- Diese Vektoren werden geclustert.



WALS-Fragebogen (Ausschnitte)

Area	Chapter	Title	Values	Languages
Phonology	1	Consonant Inventories	1 - Small 2 - Moderately small 3 - Average 4 - Moderately large 5 - Large	563
Phonology	2	Vowel Quality Inventories	1 - Small (2-4) 2 - Average (5-6) 3 - Large (7-14)	564
Phonology	3	Consonant-Vowel Ratio	1 - Low 2 - Moderately low 3 - Average 4 - Moderately high 5 - High	564
Phonology	4	Voicing in Plosives and Fricatives	1 - No voicing constrast 2 - In plosives alone 3 - In fricatives alone 4 - In both plosives and fricatives	567
Phonology	5	Voicing and Gaps in Plosive Systems	1 - Other 2 - None missing in /p t k b d g/ 3 - Missing /p/ 4 - Missing /g/	567

Morphology	25	Locus of Marking: Whole-language Typology	1 - Head-marking 2 - Dependent-marking 3 - Double-marking 4 - Zero-marking 5 - Inconsistent or other	236
Morphology	26	Prefixing vs. Suffixing in Inflectional Morphology	1 - Little affixation 2 - Strongly suffixing 3 - Weakly suffixing 4 - Equal prefixing and suffixing 5 - Weakly prefixing 6 - Strong prefixing	894
Morphology	27	Reduplication	1 - Productive full and partial reduplication 2 - Full reduplication only 3 - No productive reduplication	368
Morphology	28	Case Syncretism	1 - No case marking 2 - Core cases only 3 - Core and non-core 4 - No syncretism	198
Morphology	29	Syncretism in Verbal Person/Number Marking	1 - No subject person/number marking 2 - Syncretic 3 - Not syncretic	198

Nominal Categories	30	Number of Genders	1 - None 2 - Two 3 - Three 4 - Four 5 - Five or more	257
Nominal Categories	31	Sex-based and Non-sex-based Gender Systems	1 - No gender 2 - Sex-based 3 - Non-sex-based	257
Nominal Categories	32	Systems of Gender Assignment	1 - No gender 2 - Semantic 3 - Semantic and formal	257
Nominal Categories	33	Coding of Nominal Plurality	1 - Plural prefix 2 - Plural suffix 3 - Plural stem change 4 - Plural tone 5 - Plural complete reduplication 6 - Mixed morphological plural 7 - Plural word 8 - Plural clitic 9 - No plural	958
Nominal Categories	34	Occurrence of Nominal Plurality	1 - No nominal plural 2 - Only human nouns, optional 3 - Only human nouns,	291

Korpusvergleich

Verfahren:

- Fragestellungen an die Datenbank, die sich
  - Als Script formulieren lassen und
  - Effektiv beantworten lassen
- Scripte arbeiten auf der Datenbank als
  - Select-Statement(s)
  - PHP-Scripte
  - Andere kleine Programme
- Die entspricht „Messungen am Korpus“

Resultate beim Korpusvergleich

Messungen am Korpus erlauben verschiedene Auswertungen:

- Untersuchung der Datensätze mit extremen (d.h. großen oder kleinen) Messwerten
- Verlauf der Messwerte über alle Messungen (Glockenkurve, Zipf-Kurve, ...)
- Vergleich der Messungen für verschiedene Korpora

Mögliche Erkenntnisse

- Erkenntnisse über statistische Regelmäßigkeiten der Sprache
- Auffinden einzelner, relevanter Parameter (z.B. Anstieg einer Geraden)
- Vergleich dieser Parameter für verschiedene Korpora / Sprachen
- Finden von Datensätzen mit extremalen Eigenschaften
- Auffinden von Lücken in der Vorverarbeitung, weil die die extremalen Datensätze nur Datenmüll enthalten

Character frequencies

character	frequency
"e"	603839
"n"	386658
"i"	344145
"d"	264977
"r"	254826
"s"	203013
"a"	199733
"t"	167670
"h"	148210
"u"	137395
"m"	103304

The most frequent 50 Words

The table shows the top-50 words of the corpus. Usually we see stopwords.

This list is a good candidate for a first stopword list for a language.

Usually a small, balanced corpus is enough to get a good list of high frequent words. But if the small corpus has some very prominent topic, this will be visible even in the top word lists.

Rank in Wordlist	Word
1	die
2	van
3	en
4	in
5	n
6	is
7	het

Longest Words in Top-1000 by length

The most frequent 1000 words contain a many stopwords and the most frequent content words. Assuming that stopwords are usually short, the list presented here shows some important content words.

In the case of very poor pre-processing, some non-words may appear in the list. But there will be much finer tests for poor pre-processing below.

Local Rank	Rank in Wordlist	Word	Length
1	859	Staatsanwaltschaft	18
2	766	Ministerpräsident	17
3	747	Geschäftsführer	15
4	707	internationalen	15
5	490	Bundesregierung	15
6	726	beispielsweise	14
7	904	Zusammenarbeit	14

Longest Words

The longest words of the corpus with minimum frequency 2 are shown. The words are seen at least twice, hence, there is some chance for not seeing misprinted words.

Surprisingly, there is no longest word which is much longer than the second one. This, again, argues for correct preprocessing.

Length	Frequency	Word
34	2	Pretoria-Witwatersrand-Vereeniging
32	3	Christus-die-Verlosser-katedraal
31	3	Hannover-Braunschweig-Göttingen
31	2	Dansgaard-Oescher-gebeurtenisse

Frequency of numbers I

Numbers can be expected to appear with regular frequencies. In this subsection we present the frequencies for the most frequent cardinal numbers with a fixed number of digits.

For very small numbers, the ordering by frequency is nearly the natural ordering with a slight preference for round lots.

Number	Frequency
1	293
2	201
3	151
5	118
4	103
6	101

Frequency of numbers II

In the case of four-digit numbers we usually get years. Here we see the temporal focus of the corpus.

Number	Frequency
2000	295
1999	252
2002	240
1995	239
1997	230

Most Frequent Abbreviations

If the tokenizer used for word tokenization knows abbreviations (ending with full stop; only those abbreviations may yield to problems during sentence segmentation), these are included in the word list as words ending in a full stop. The list shows the 50 most frequent words of this form.

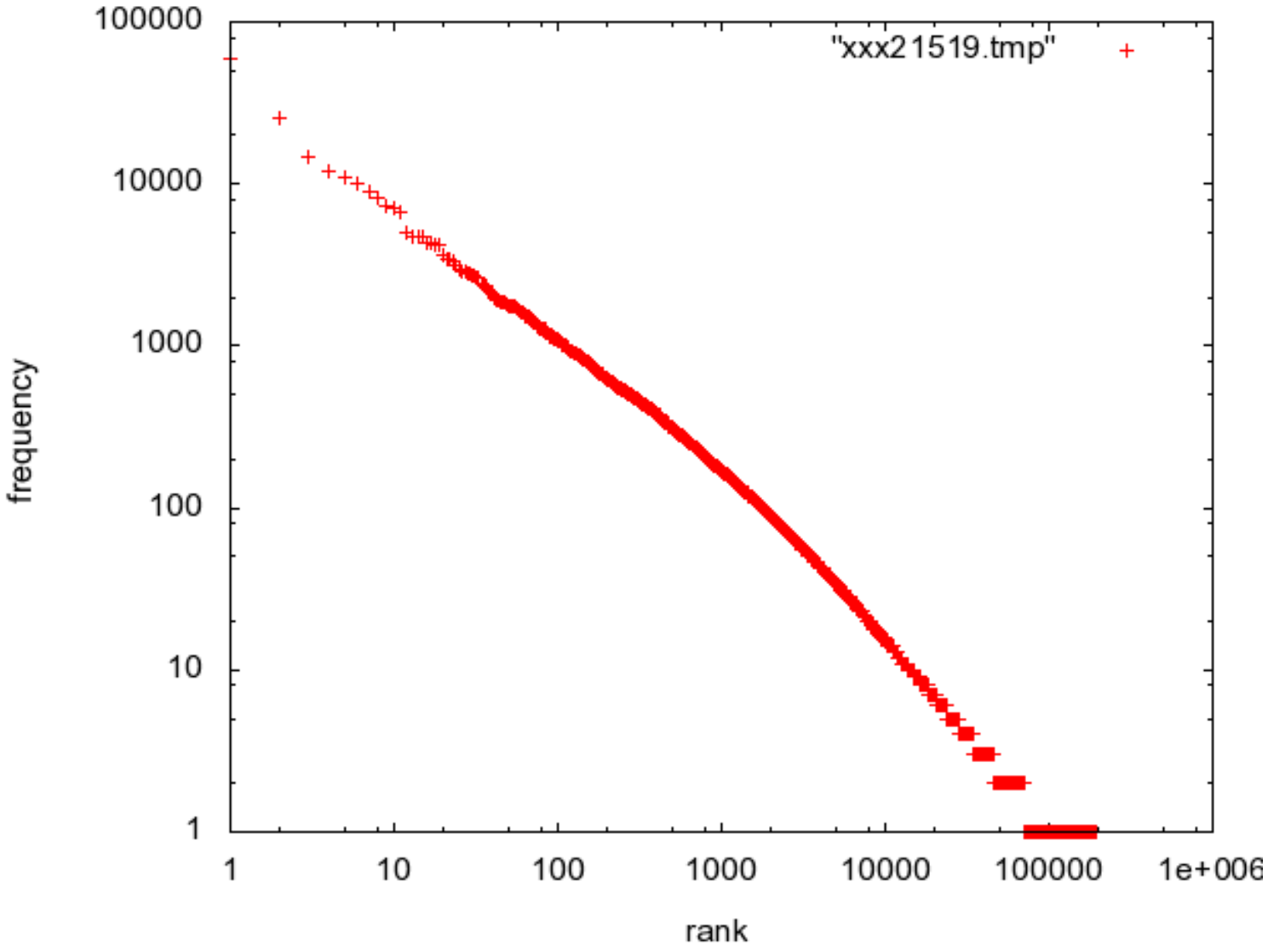
Rank	Abbreviation	Frequency
643	Mrd.	214
876	Mio.	159
1454	Dr.	98
1511	St.	95
2877	bzw.	50
3225	Mill.	44

Mainly noise: Words containing special characters

Rank in Wordlist	Frequency	Word
61170	1	6:1-(4:1-
81465	1	Existenzgründer(innen
92493	1	Himmelb(l)au
103472	1	Liberty(R
107066	1	MetaFramer(TM
109418	1	Märchen-Kini(ch
111495	1	Nord(ost)-Deutsche
127923	1	Staats(nahe)industrien
138849	1	Wa(h)re
140782	1	Whisk(e)y

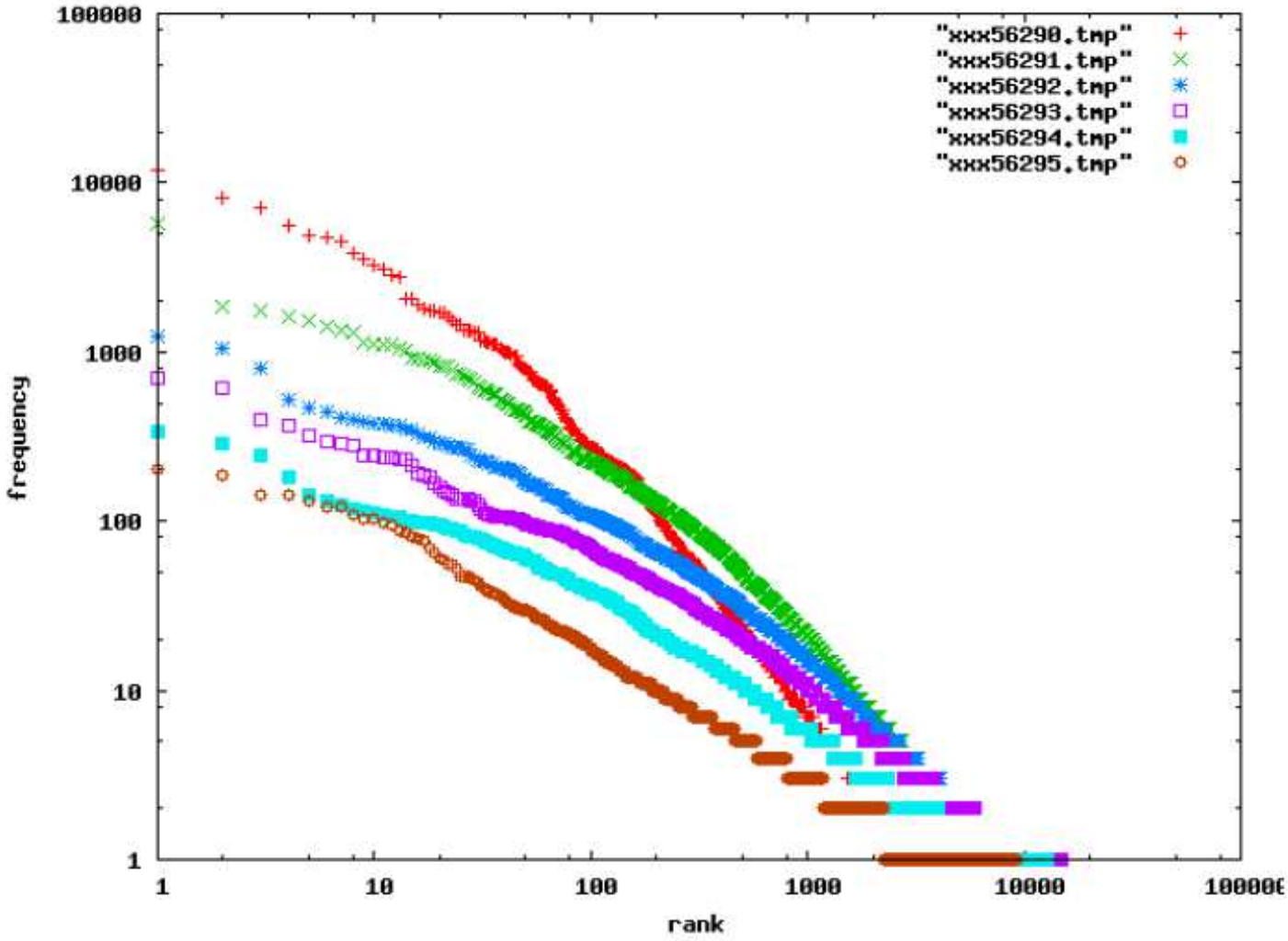
Rank in Wordlist	Frequency	Word
12521	9	Bild"-Zeitung
49521	2	Spiegel"-Bericht
50236	2	Tatort"-Kommissar
52253	2	Zeit"-
58779	1	12"-Delitzscher
59594	1	2000"-Aktion
59595	1	2000"-Programm
61177	1	7"-Gipfel
61876	1	ARD-"Tagesthemen
63357	1	All-Saints"-Christen

Zipf's law (Standard version)



Zipf's law for words of fixed lengths

The validity of Zipf's law for all words does not imply its validity for words of fixed length. The tables show the most frequent words for fixed word lengths 4, 6, 8, …, 14.



Most frequent word beginnings

The tables show the most frequent letter-N-grams at the beginning of words for N=1…5. Their frequency is count without multiplicity, otherwise the stopwords would dominate the tables.

Word prefixes are clearly visible. In the above example, ver- and ein- are prefixes, and Sch- is not. At the end of a prefix we typically have a wide variety of possible continuations. Hence a prefix of length k will be prominent in the table for N=k, but typically not in the table for N=k+1. The prominent entries Schw- and Schl- for N=4 tell us that Sch- is no prefix.

Top Character Trigrams

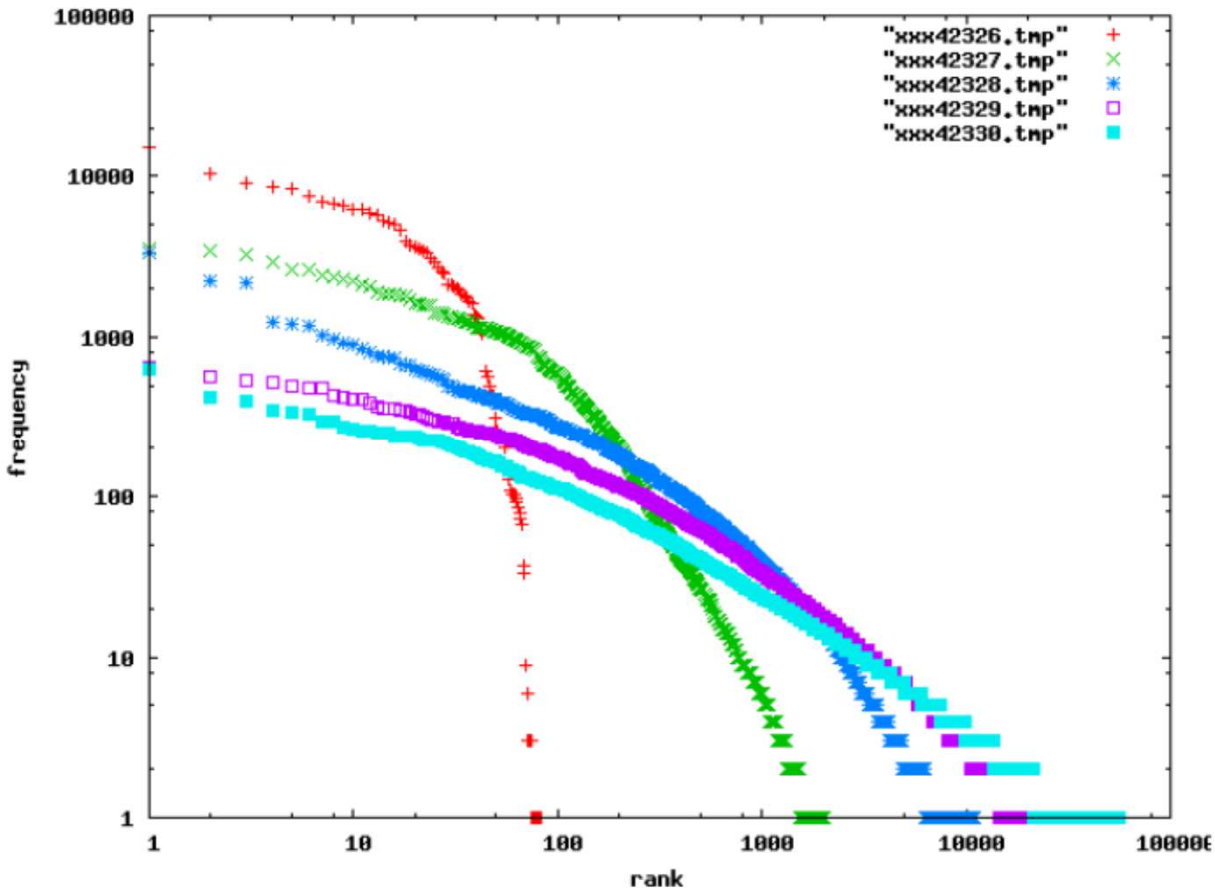
word rank	frequency	n-gram
1	3380	Sch-
2	2262	Ver-
3	2203	ver-
4	1242	ein-
5	1215	sch-

Top Character 4-Grams

word rank	frequency	n-gram
1	649	Bund-
2	565	über-
3	539	Schw-
4	526	Schl-
5	489	Land-

Most frequent word beginnings

Zipf's diagram is plotted with both axis in logarithmic scale, hence we expect nearly straight lines. The graphs look more typical for larger N. Especially for N=3 we find only a small number of trigrams resulting in a sharp decay.



Zipf's law for words with same first letter

Top words a-

rank	frequency	word
1	11247	auf
2	7384	als
3	7216	auch
4	7064	an
5	5940	aus

Top words b-

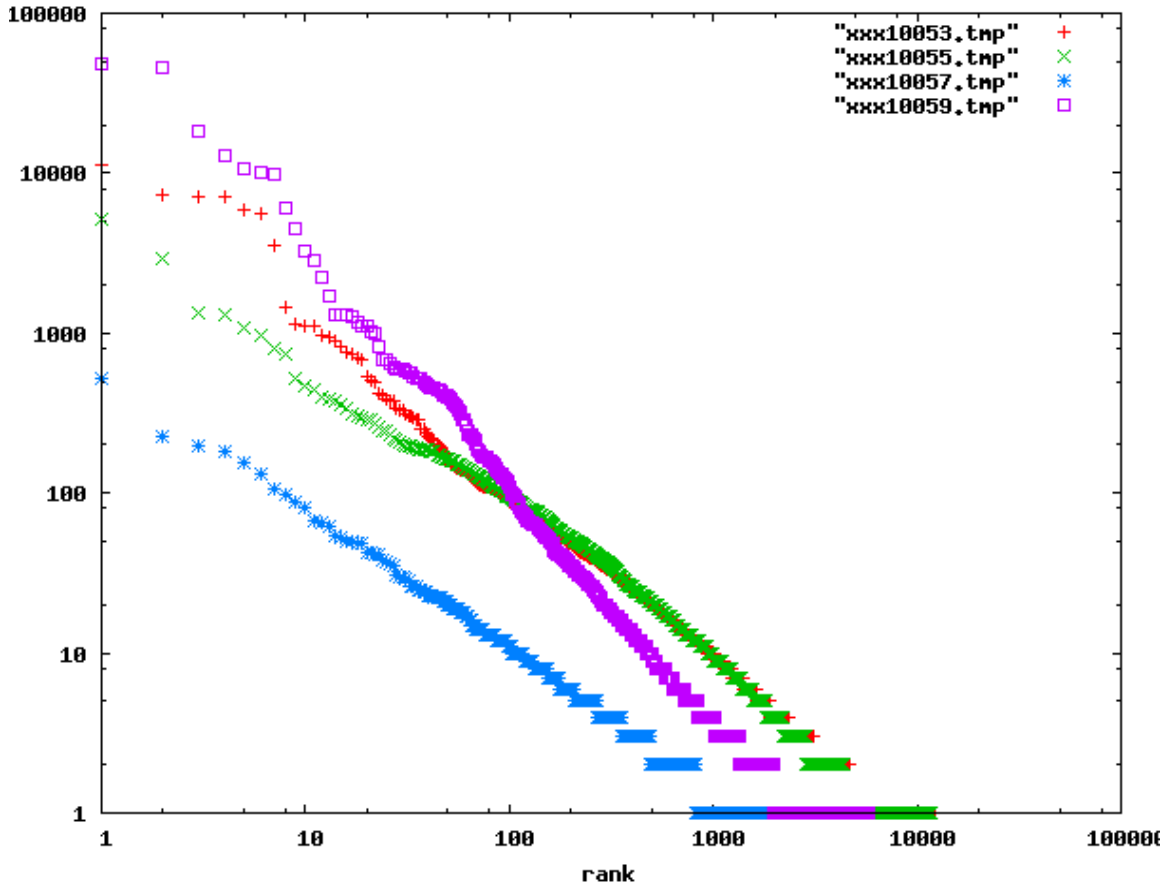
rank	frequency	word
1	5099	bei
2	2944	bis
3	1323	Berlin
4	1301	beim
5	1092	bereits

Top words c-

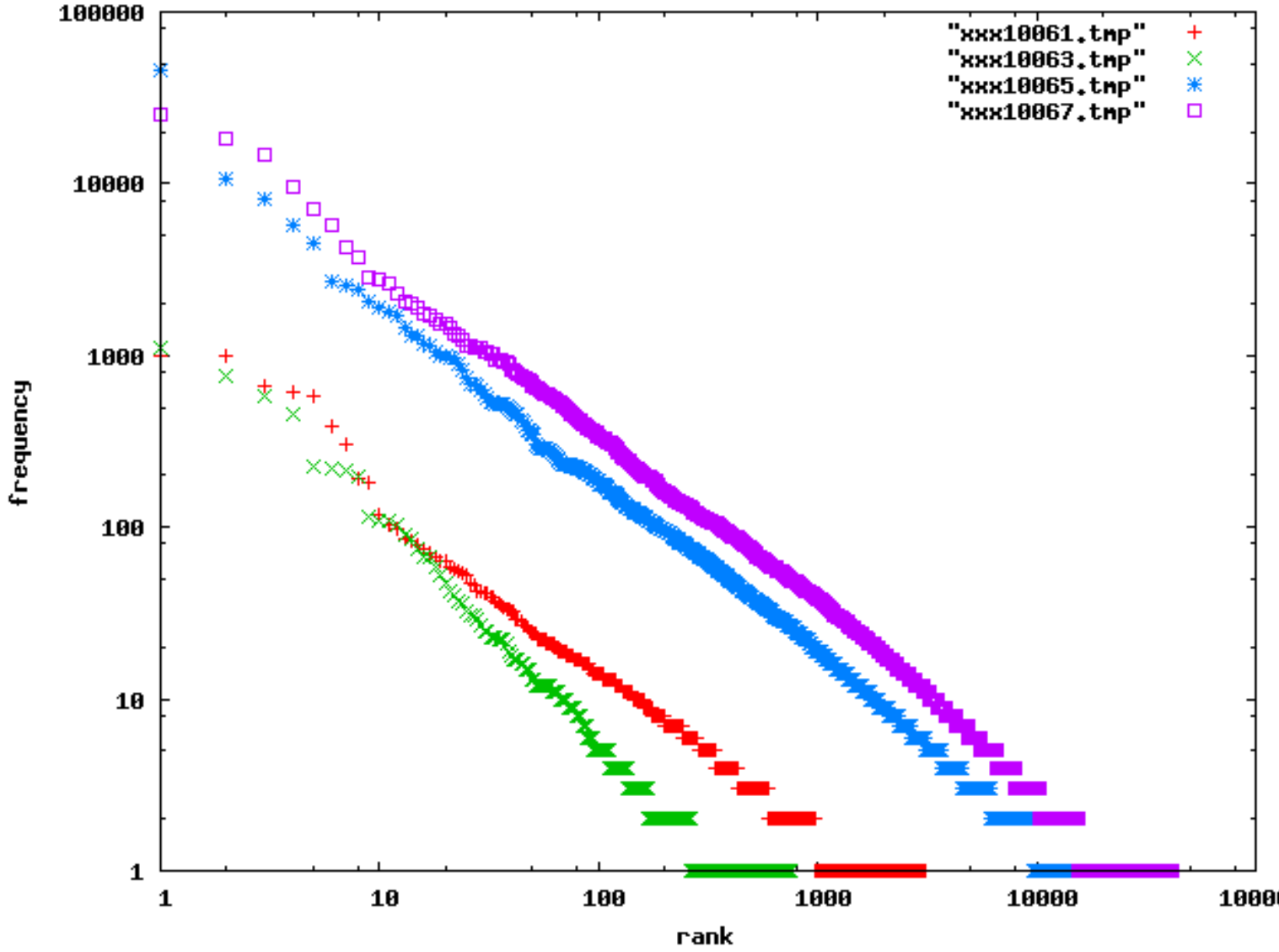
rank	frequency	word
1	515	CDU
2	224	Chance
3	198	Chef
4	183	China
5	155	Chancen

Top words d-

rank	frequency	word
1	48054	der
2	45544	die
3	18527	den
4	13006	das
5	10764	Die



Zipf's law for words with same last letter: a, b, e, n



Allgemeines zum Messen

Ziel einer Messung: Aussage über die unbekannte Größe eines Objekts. Beispiel: Mittlere Länge eines deutschen Wortes in Buchstaben.

Messwert: Am gemessenen Objekt hat die Messgröße den Wert x. Diese Messgröße kann fehlerbehaftet sein.

Genauigkeit der Messung:

Wie groß ist der Messfehler am gemessenen Objekt?

Wie groß ist die Abweichung des gemessenen Objekts vom „allgemeinen Fall“?

Schritte zur (linguistischen) Messung

- **Eindeutige Definition der Messaufgabe** und der Messgröße
- **Zusammenstellen der Randbedingungen** (Quellen, Vorverarbeitung, Wortdefinition, ...)
- **Wahl des Messverfahrens** (z.B. mittels Programm)
- **Festlegung des Messablaufs**: Einzelmessung oder Messreihe auch verschiedenen, vergleichbaren Objekten? Ferner können Messungen verschiedener Größen erforderlich sein, aus denen der Messwert der gesuchten Größe nach festgelegten mathematischen Beziehungen zu berechnen ist.
- **Durchführen der Messung** und Ermitteln des Messergebnisses
- **Berücksichtigung der Auswirkungen von Einflussgrößen**: Korrektur von systematischen Messabweichungen. (z.B. durch spezielles Genre)
- **Ermitteln des vollständigen Messergebnisses**: Ein vollständiges Messergebnis besteht aus dem Messwert (gegebenenfalls Mittelwert aus einer oder mehreren Messreihen oder dem berechneten Wert aufgrund anderer Messungen), ergänzt durch quantitative Aussagen zur Messunsicherheit.

Quelle: Wikipedia

Normierung der Messungen
Beispiel: Type - Token - Ratio

1. Definitionsprobleme

Wie groß ist das Type-Token Ratio im Satz

„Ich komme um 8 Uhr", sagte ich um 7 Uhr.

- Wieviele Tokens? 8, 10, oder 14?
- Wieviel Types? 5, 6, 8 oder 12?
- Type-Token Ratio: 5/8, 6/8, 8/10 oder 12/14?

2. Abhängigkeit von der Korpusgröße

Nach dem Zipfschen Gesetz wächst das Type-Token-Ratio logarithmisch mit der Korpusgröße.

Auch problematisch sind deshalb: durchschnittliche Wortlänge, durchschnittliche Satzlänge, Anzahl Tokens im Korpus, Textabdeckung, ...

Aber auch: Buchstabenhäufigkeiten, Anzahl Sätze, längster Satz, längstes Wort, ...

Messbarmachung / Operationalisierung

Mit welchen (messbaren) Indikatoren soll ein linguistisches Konstrukt gemessen werden?

Beispiele:

- **Vokal-Konsonanten-Verhältnis**: Verwende Vokalbuchstaben statt Vokalen usw.
- **Silbenzahl**: Bestimme Anzahl der Silbengipfel aus lokalen Vokalgruppen. Hier tritt ein systematischer Fehler auf, der sich abschätzen und korrigieren läßt
- **Variabilität der Satzstellung**: Je strenger die Satzstellung, desto mehr feste Aufeinanderfolgen von Wörtern. Also: Die Anzahl der Nachbarschaftskokkurrenzen steigt im Vergleich zu den Satzkokkurrenzen.
- **Flexion** mittels Stringähnlichkeit: Mehr regelmäßig flektierte Wörter sorgen für mehr Cluster stringähnlicher Wörter. Die meisten Unterschiede (z.B. Flexionsendungen) lassen sich in wenige Klassen aufteilen.
- **Affixe bzw. Silben** mit Hilfe von Buchstaben-N-Gramme