

Die Repräsentation und Auflösung von ambigen Wortbedeutungen in der Computerlinguistik

Korpus-Statistik

PD Dr. Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

16. Januar 2010

- Korpora stellen große Datenmengen für empirische Forschung zur Verfügung
- Welche Möglichkeiten gibt es für die **Exploration der Korpus-Daten**?
 - Belege suchen
 - über Belege generalisieren
- Beleg: Zeichen, Zeichenkette, Sequenz von Zeichenketten
- Muster als Beleg-Beschreibung
- **Statistik** für die Beleg-Analyse (quantitativ, aber auch qualitativ)

↔ **lexikalisches Wissen**

Quantitative Korpus-Untersuchungen

- Typen und Token
- absolute Häufigkeit
- relative Häufigkeit
- Rangplatz aufgrund von Häufigkeit
- Häufigkeiten von Sequenzen

Typen und Token

- **Token:** Anzahl von Wort-Instanzen in einem Korpus
→ Korpus-Größe

*Peters*₁ *Vater*₂ *ist*₃ *ein*₄ *Koch*₅ *.*₆
*Peters*₇ *Mutter*₈ *ist*₉ *Köchin*₁₀ *.*₁₁

- **Typ:** Anzahl von verschiedenen Wörtern in einem Korpus
→ Vokabular-Größe

*Peters*₁ *Vater*₂ *ist*₃ *ein*₄ *Koch*₅ *.*₆
*Peters*₆ *Mutter*₇ *ist*₇ *Köchin*₈ *.*₈

Typ-Token-Abbildung

- Definition von Token \rightsquigarrow Tokenisierung;
erbt Probleme der Tokenisierung, z.B. Satzzeichen, Zahlen,
Mehrwortausdrücke
- Abbildung von Token auf Typen \rightsquigarrow Normalisierung
(z.B. bei Klein-/Großschreibung, Lemmatisierung)
- Was wird nicht berücksichtigt? \rightsquigarrow Wortbedeutungen

- Lexikalische Statistik: Häufigkeitsverteilungen von Wörtern
- Frequenz: absolute Häufigkeit
- Wahrscheinlichkeit: relative Häufigkeit

- **Frequenzliste:** Typen im Korpus und ihre Frequenz
- Beispiel:

Typ	Frequenz
ein	1
ist	2
Koch	1
Köchin	1
Mutter	1
Peters	2
Vater	1

Frequenz-Rangliste

- **Frequenz-Rangliste:** Frequenzen werden nach Größe sortiert
- Beispiel:

Rang	Frequenz
1	2
2	2
3	1
4	1
5	1
6	1
7	1

Frequenzspektrum

- **Frequenzspektrum**: Häufigkeit von Frequenzen
- Frequenzspektrum kann aus Frequenz-Rangliste abgeleitet werden (und umgekehrt)
- Beispiel:

Frequenz	Häufigkeit der Frequenz
2	2
1	5

Beispiel: Ausschnitt aus dem *DeWaC* mit 448,675 Wörtern (Anfang)

Rang	Frequenz	Wort(beispiele)
1	23848	,
2	18851	.
3	11907	der
4	10973	die
5	10705	und
6	5880	in
7	4276	den
8	4063	"
9	3967	zu
10	3899	von

Beispiel: Ausschnitt aus dem *DeWaC* mit 448,675 Wörtern (Ende)

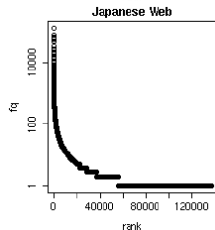
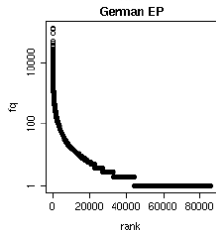
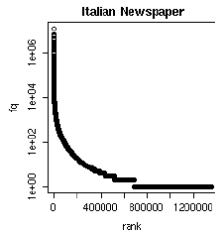
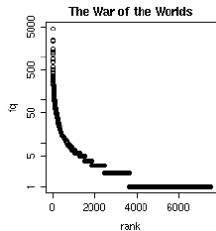
Rang	Freq.	Wort(beispiele)
3750-4609	10	zeitlich, wovon, Tempel, stirbt, Ordnungsmittel
4150-4610	9	samt, planen, normalerweise, kräftig, Jerusalem
4611-5244	8	EDEKA, Genuss, festgenommen, ehrenamtlich, dpa
5245-5981	7	liebt, Möhrenbrei, Kurzfassung, 700, artig
5982-6975	6	Sakristei, seufzte, Rhein, rote, Oh
6976-8442	5	Flower, effektive, Bio-Markt, betreten, CD-Rom
8443-10662	4	unscharf, Tunnel, regeln, Mabuse, BILD
10663-14501	3	Stiefvater, solidarisch, siedelten, Sex, abenteuerliche
14502-23304	2	zzgl., Wirtschaftsbosse, worum, seltsames, schälen
23305-60652	1	Zwickmühle, zweymal, zur., www.tui.com, Vortänzer

Frequenz-Verteilungen

- Es gibt typische Frequenz-Verteilungen für Korpora, die Korpus-übergreifend beobachtet werden können.
- Anfang der Frequenz-Liste:
 - Funktionswörter und Satzzeichen
 - Frequenz von Rang x_i deutlich größer als Frequenz von Rang x_{i+1}
 - im Beispiel: Summe der Frequenzen der ersten 10 Ränge entspricht 22% aller Token
- Ende der Frequenz-Liste:
 - Inhaltswörter, Komposita, Neologismen, Fehler, spezielle Ausdrücke wie Webseiten
 - Anzahl von Typen mit Frequenz x_i deutlich größer als Anzahl von Typen mit Frequenz x_{i+1} , z.B. 37347 vs. 8802 vs. 3838 etc.
 - im Beispiel: Wörter mit Frequenz 1 (Hapax Legomena) stellen 62% der Typen dar, Wörter mit Frequenz 1-10 repräsentieren 94%

Beispiel: Frequenz-Ranglisten

- Korpora: *The War of the Worlds*, italienische Zeitung *La Repubblica*, Teil des deutschen *EuroParl* Korpus', Korpus mit japanischen Webseiten
- Quelle: Baroni (2009)



- **Zipf's Law** (1949, 1965): Modell für die Vorhersage der Frequenz eines Wortes in Bezug auf den Rang

$$f(w) = \frac{C}{r(w)^a}$$

Wort-Frequenz $f(w)$, Wort-Rang $r(w)$, Korpus-abhängige Konstanten C und a

- $a = 1$: C ist Korpus-Frequenz des häufigsten Wortes
- Sagt raschen Abfall der Frequenz bei hohen Rängen und langes Plateau von Wörtern mit ähnlich niedrigen Frequenzen vorher.
- Modell ist auch gültig für andere Phänomene der natürlichen Sprache (z.B. Wortbedeutungen) sowie andere Phänomene (z.B. Stadtbevölkerung, Einkommen)
- Probleme in der Sprachverarbeitung: spärliche Daten und unvollständige und nicht-vergleichbare Sprachabdeckung
↪ schiefe Verteilungen; Fehlschätzungen von relativen Häufigkeiten

- Relative Häufigkeit: absolute Häufigkeit eines Ereignisses im Verhältnis zur Gesamtheit der Ereignisse
- Wahrscheinlichkeit p , mit der ein bestimmtes Ereignis x eintritt
- Summe aller Wahrscheinlichkeiten ergibt 1

- Wahrscheinlichkeitsverteilung:

$$0 \leq p(x) \leq 1$$
$$\sum_{x \in X} p(x) = 1$$

- Beispiele:
 - Wahrscheinlichkeit, dass man eine 6 würfelt
 - Wahrscheinlichkeit, dass auf *ins Gras* das Verb *beißen* folgt

- **Konkordanz**: Wort in seinem unmittelbaren Kontext;
vgl. **KWIC** (key word in context)
- **n -Gramm**: Wortsequenz mit Länge n , z.B. Bigramme, Trigramme
- **Kollokation**: Mehrwortausdruck mit (statistisch) starkem Zusammenhalt

- Analyse von Schlüsselwörtern
- Analyse von Wort-Frequenzen
- Vergleich von verschiedenen Gebräuchlichkeiten desselben Wortes über Kontext und Struktur
- Bestimmen von Wortbedeutungen über Kontext-Verallgemeinerungen
- Exploration von Kollokationen

Anfrage: [lemma='take' & pos='VB.*'],

Kontext: 5 Wörter

A Christmas Carol, Chapter 1

... nothing more remarkable in his **taking** a stroll at night , ...

A Christmas Carol, Chapter 1

... ' said the gentleman , **taking** up a pen , ` ...

A Christmas Carol, Chapter 1

... play at blindman's-buff . Scrooge **took** his melancholy dinner in his ...

A Christmas Carol, Chapter 1

... up that staircase , and **taken** it broadwise , with the ...

A Christmas Carol, Chapter 1

... secured against surprise , he **took** off his cravat ; put ...

A Christmas Carol, Chapter 1

... down before the fire to **take** his gruel . It was ...

A Christmas Carol, Chapter 1

... himself in a condition to **take** a chair ; and felt ...

A Christmas Carol, Chapter 1

... horror , when the phantom **taking** off the bandage round its ...

A Christmas Carol, Chapter 1

... ` Could n't I **take** ` em all at once ...

A Christmas Carol, Chapter 1

... these words , the spectre **took** its wrapper from the table ...

Anfrage: [lemma='take' & pos='VB.*'],

Kontext: 5 Wörter mit Wortart

A Christmas Carol, Chapter 1

... nothing/NN more/RBR remarkable/JJ in/IN his/PP\$ **taking/VBG** a/DT stroll/VBP at/IN night/NN ,/ , ...

A Christmas Carol, Chapter 1

... 'I" said/VBD the/DT gentleman/NN ,/ , **taking/VBG** up/RP a/DT pen/NN ,/ , 'I' ...

A Christmas Carol, Chapter 1

... play/VB at/IN blindman's-buff/NN ./SENT Scrooge/NN **took/VBD** his/PP\$ melancholy/JJ dinner/NN in/IN his/PP\$...

A Christmas Carol, Chapter 1

... up/IN that/DT staircase/NN ,/ , and/CC **taken/VBN** it/PP broadwise/RB ,/ , with/IN the/DT ...

A Christmas Carol, Chapter 1

... secured/VBN against/IN surprise/NN ,/ , he/PP **took/VBD** off/RP his/PP\$ cravat/NN ;/ ; put/VBN ...

A Christmas Carol, Chapter 1

... down/RP before/IN the/DT fire/NN to/TO **take/VB** his/PP\$ gruel/NN ./SENT It/PP was/VBD ...

A Christmas Carol, Chapter 1

... himself/PP in/IN a/DT condition/NN to/TO **take/VB** a/DT chair/NN ;/ ; and/CC felt/VBD ...

A Christmas Carol, Chapter 1

... horror/NN ,/ , when/WRB the/DT phantom/JJ **taking/VBG** off/RP the/DT bandage/NN round/VB its/PP\$...

A Christmas Carol, Chapter 1

... ./SENT `I' Could/MD n't/RB I/PP **take/VBP** `I' em/NN all/RB at/IN once/RB ...

A Christmas Carol, Chapter 1

... these/DT words/NNS ,/ , the/DT spectre/NN **took/VBD** its/PP\$ wrapper/NN from/IN the/DT table/NN ...

Anfrage: [lemma='take' & pos='VB.*'],
Kontext: 5 Lemmata mit Wortart

A Christmas Carol, Chapter 1

... nothing/NN more/RBR remarkable/JJ in/IN his/PP\$ take/VBG a/DT stroll/VBP at/IN night/NN ,/ , ...

A Christmas Carol, Chapter 1

... "I" say/VBD the/DT gentleman/NN ,/ , take/VBG up/RP a/DT pen/NN ,/ , `f` ...

A Christmas Carol, Chapter 1

... play/VB at/IN blindman's-buff/NN ./SENT Scrooge/NN take/VBD his/PP\$ melancholy/JJ dinner/NN in/IN his/PP\$...

A Christmas Carol, Chapter 1

... up/IN that/DT staircase/NN ,/ , and/CC take/VBN it/PP broadwise/RB ,/ , with/IN the/DT ...

A Christmas Carol, Chapter 1

... secure/VBN against/IN surprise/NN ,/ , he/PP take/VBD off/RP his/PP\$ cravat/NN ;/ ; put/VBN ...

A Christmas Carol, Chapter 1

... down/RP before/IN the/DT fire/NN to/TO take/VB his/PP\$ gruel/NN ./SENT it/PP be/VBD ...

A Christmas Carol, Chapter 1

... himself/PP in/IN a/DT condition/NN to/TO take/VB a/DT chair/NN ;/ ; and/CC feel/VBD ...

A Christmas Carol, Chapter 1

... horror/NN ,/ , when/WRB the/DT phantom/JJ take/VBG off/RP the/DT bandage/NN round/VB its/PP\$...

A Christmas Carol, Chapter 1

... ./SENT `f` Could/MD not/RB I/PP take/VBP `f` em/NN all/RB at/IN once/RB ...

A Christmas Carol, Chapter 1

... these/DT word/NNS ,/ , the/DT spectre/NN take/VBD its/PP\$ wrapper/NN from/IN the/DT table/NN ...

Anfrage: [lemma='take' & pos='VB.*'] [lemma='up' & pos='RP'] []1,3 [pos='NN'],

Kontext: 5 Wörter

A Christmas Carol, Chapter 1

... ' said the gentleman , **taking up a pen** , ' it is more ...

David Copperfield, Chapter 1

... of a room , to **take up the less space** . He walked as softly ...

David Copperfield, Chapter 4

... an impressive look , and **took up his book** . This was a good ...

David Copperfield, Chapter 4

... he rose and said , **taking up the cane** : ' Why , Jane ...

David Copperfield, Chapter 5

... , ' he said , **taking up a table-spoon** , ' is my favourite ...

David Copperfield, Chapter 14

... earnestly at me , and **taking up his pen** to note it down , ...

David Copperfield, Chapter 17

... , it will not be **taken up** . **The result** is destruction . The bolt ...

David Copperfield, Chapter 30

... ' said Mr. Omer , **taking up his glass** , ' because it 's ...

David Copperfield, Chapter 32

... his birth , - to **take up in a moment** with a miserable girl , ...

David Copperfield, Chapter 35

... from the time of my **taking up my residence** in Mr. Wickfield 's house ...

- Wortsequenz mit Länge n , z.B. Bigramme, Trigramme
- Basis für Sprachmodelle (*language models*) zur Vorhersage des nächsten Wortes, gegeben die vorhergehenden Wörter:
$$p(w_n) = p(w_n | w_1, \dots, w_{n-1})$$
- Einschränkungen von n :
 - Anzahl Parameter vs. spärliche Daten
 - Anzahl Parameter vs. Übertragbarkeit auf anderes Korpus
 - Normalisierung, z.B. Wortformen vs. Lemmata
- Beispiele:
 - Vorhersage eines Nomens aus der Menge { *See, Auto, Druck, Muschel* } mit der vorhergehenden Wort-Sequenz *schwamm in dem blauen*
 - Übersetzung des deutschen Wortes *scheinen* ins Englische *shine* vs. *seem* mit dem vorhergehenden Wort *Sonne* vs. *mir*

- Lexikalische Einheiten in einer bestimmten funktionalen Relation zueinander
- **habituell**: *Eis essen, Wein trinken, ganz schön*
- **idiomatisch**: *ins Gras beißen, aus und vorbei*
- Induktion von Kollokationen durch absolute und relative Häufigkeiten und statistische Maße

↪ **lexikalisches Wissen zu Mehrwortausdrücken**

Korpus-Statistik: Zusammenfassung

- Korpora stellen große Datenmengen für empirische Forschung zur Verfügung
- Welche Möglichkeiten gibt es für die [Exploration der Korpus-Daten](#)?
- [Statistik](#) für die Beleg-Analyse (quantitativ, aber auch qualitativ)

↪ [lexikalisches Wissen zu Wörtern und Mehrwortausdrücken](#)

↪ [lexikalisches Wissen zu Wortbedeutungen?](#)



George K. Zipf.

Human Behaviour and the Principle of Least-Effort.

Addison-Wesley, Cambridge, 1949.



Christopher D. Manning and Hinrich Schütze.

Foundations of Statistical Natural Language Processing.

MIT Press, Cambridge, MA, 1999.



Marco Baroni.

Distributions in Text.

In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook.*, volume 2 of *Handbooks of Linguistics and Communication Science.* Mouton de Gruyter, Berlin, 2009.