

# Die Repräsentation und Auflösung von ambigen Wortbedeutungen in der Computerlinguistik

## Distributionelle Semantik

PD Dr. Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung  
Universität Stuttgart

26. März 2010

# Distributionelle Semantik

- **Kontext** einer linguistischen Einheit enthält Indikatoren für Wortbedeutung
- Beispiel: ein Korpus enthält die Information, dass man einen Apfel kaufen, schälen und essen kann

- **Distributionelle Hypothese:**

*You shall know a word by the company it keeps.* (Firth, 1957)

*Each language can be described in terms of a distributional structure, i.e., in terms of the occurrence of parts relative to other parts.* (Harris, 1968)

- Problem: Korpus enthält kein **Weltwissen**;  
*inferential (i.e., how to use language distributionally) vs. referential (incorporating world knowledge) abilities* (Marconi, 1997), z.B.  
*Ananas-gelb; auftauen-Wasser*

# Kookkurrenz

- **Kontext** bezieht sich auf **Kookkurrenz**
- Kookkurrenz kann auf verschiedenen linguistischen Ebenen erfolgen

# Kookkurrenz

- **Kontext** bezieht sich auf **Kookkurrenz**
- Kookkurrenz kann auf verschiedenen linguistischen Ebenen erfolgen
  - Laut

# Kookkurrenz

- **Kontext** bezieht sich auf **Kookkurrenz**
- Kookkurrenz kann auf verschiedenen linguistischen Ebenen erfolgen
  - Laut
  - Zeichen:  
AP EL

# Kookkurrenz

- **Kontext** bezieht sich auf **Kookkurrenz**
- Kookkurrenz kann auf verschiedenen linguistischen Ebenen erfolgen
  - Laut
  - Zeichen:  
AP EL → APFEL

# Kookkurrenz

- **Kontext** bezieht sich auf **Kookkurrenz**
- Kookkurrenz kann auf verschiedenen linguistischen Ebenen erfolgen
  - Laut
  - Zeichen:  
AP EL → APFEL  
MÜT E

# Kookkurrenz

- **Kontext** bezieht sich auf **Kookkurrenz**
- Kookkurrenz kann auf verschiedenen linguistischen Ebenen erfolgen
  - Laut
  - Zeichen:  
AP EL → APFEL  
MÜT E → MÜTZE

# Kookkurrenz

- **Kontext** bezieht sich auf **Kookkurrenz**
- Kookkurrenz kann auf verschiedenen linguistischen Ebenen erfolgen
  - Laut
  - Zeichen:  
AP EL → APFEL  
MÜT E → MÜTZE
  - Wort:  
Oma        einen Kuchen

# Kookkurrenz

- **Kontext** bezieht sich auf **Kookkurrenz**
- Kookkurrenz kann auf verschiedenen linguistischen Ebenen erfolgen
  - Laut
  - Zeichen:  
AP EL → APFEL  
MÜT E → MÜTZE
  - Wort:  
Oma        einen Kuchen → Oma backt einen Kuchen

# Kookkurrenz

- Kookkurrenz wird zur semantischen Beschreibung von linguistischen Einheiten verwendet, d.h. die Summe von Kontext-Beschreibungen entspricht einer lexikalischen Beschreibung (idealisierte Vorstellung)
- Arten von Kookkurrenz:
  - Wörter im Satz, Paragraphen, Dokument
  - Wörter in einem festen Wortfenster
  - Syntax-basierte Wörter
  - Muster-basierte Kookkurrenz
  - Subkategorisierung
  - Kookkurrenz *n*ter Ordnung
  - etc.
- unterschiedliche Arten setzen unterschiedliche Korpus-Vorverarbeitung voraus

# Kookkurrenz

- Quelle von Kookkurrenz: Korpora
- Vorgehensweise: zählen
- Darstellung von Kookkurrenz:

# Kookkurrenz

- Quelle von Kookkurrenz: Korpora
- Vorgehensweise: zählen
- Darstellung von Kookkurrenz:
  - Matrix:

	grün	gelb	schälen	fallen	Baum
Apfel	80	1	311	22	105
Banane	13	56	83	2	8
Blatt	258	0	1	98	244

# Kookkurrenz

- Quelle von Kookkurrenz: Korpora
- Vorgehensweise: zählen
- Darstellung von Kookkurrenz:
  - Matrix:

	grün	gelb	schälen	fallen	Baum
Apfel	80	1	311	22	105
Banane	13	56	83	2	8
Blatt	258	0	1	98	244

- Vektor:

Apfel:  $\langle 80, 1, 311, 22, 105 \rangle$

Banane:  $\langle 13, 56, 83, 2, 8 \rangle$

Blatt:  $\langle 258, 0, 1, 98, 244 \rangle$

# Kookkurrenz

- Quelle von Kookkurrenz: Korpora
- Vorgehensweise: zählen
- Darstellung von Kookkurrenz:

- **Matrix:**

	grün	gelb	schälen	fallen	Baum
Apfel	80	1	311	22	105
Banane	13	56	83	2	8
Blatt	258	0	1	98	244

- **Vektor:**

Apfel:  $\langle 80, 1, 311, 22, 105 \rangle$

Banane:  $\langle 13, 56, 83, 2, 8 \rangle$

Blatt:  $\langle 258, 0, 1, 98, 244 \rangle$

- Zeilen = Interesse; Spalten = linguistisch relevante Dimensionen

# Kookkurrenz: Wörter

- Kookkurrenz mit Wörtern im Satz, Paragraphen, Dokument  
↪ *Bag-of-Words*
- Auswahl von Dimensionen (Wörtern):  
Inhaltswörter, z.B. Nomen; mittel- bis hochfrequent
- Variationen: Lemmatisierung, Filtern von Dimensionen etc.
- grobe semantische Charakterisierung

# Kookkurrenz: Wortfenster

- **Wortfenster**:  $x$  Wörter links und/oder rechts von Interesse
- Einschränkung der Korpus-Umgebung unabhängig von Satzgrenzen
- Variation von sehr kleinem bis sehr großem  $x$   
↪ Kollokationen ↪ *Bag-of-Words*

# Kookkurrenz: Wortfenster

- **Wortfenster:**  $x$  Wörter links und/oder rechts von Interesse
- Einschränkung der Korpus-Umgebung unabhängig von Satzgrenzen
- Variation von sehr kleinem bis sehr großem  $x$   
↔ Kollokationen ↔ *Bag-of-Words*
- Beispiel mit  $x = 2$  (l+r):

... weil Peter den *Apfel* essen wollte ...

... Hannah pflückte Peter den *Apfel* vom Baum ...

# Kookkurrenz: Wortfenster

- **Wortfenster:**  $x$  Wörter links und/oder rechts von Interesse
- Einschränkung der Korpus-Umgebung unabhängig von Satzgrenzen
- Variation von sehr kleinem bis sehr großem  $x$   
 $\rightsquigarrow$  Kollokationen  $\rightsquigarrow$  *Bag-of-Words*
- Beispiel mit  $x = 2$  (l+r):

*... weil Peter den **Apfel** essen wollte ...*

*... Hannah pflückte Peter den **Apfel** vom Baum ...*

	Baum	den	essen	Peter	vom	wollte
<b>Apfel</b>	1	2	1	2	1	1

# Kookkurrenz: Wortfenster

- **Wortfenster:**  $x$  Wörter links und/oder rechts von Interesse
- Einschränkung der Korpus-Umgebung unabhängig von Satzgrenzen
- Variation von sehr kleinem bis sehr großem  $x$   
 $\rightsquigarrow$  Kollokationen  $\rightsquigarrow$  *Bag-of-Words*
- Beispiel mit  $x = 2$  (l+r):

... weil Peter den *Apfel* essen wollte ...

... Hannah pflückte Peter den *Apfel* vom Baum ...

	Baum	den	essen	Peter	vom	wollte
<i>Apfel</i>	1	2	1	2	1	1

- Gewichtung von Entfernung möglich

# Kookkurrenz: Syntax-basierte Wörter

- Syntaktische Funktionen von Dimensionen werden einbezogen, z.B. *NPnom*, *NPakk*, *PP.von* etc.
- dadurch typischerweise konzentriert auf Inhaltswörter

# Kokkurrenz: Syntax-basierte Wörter

- Syntaktische Funktionen von Dimensionen werden einbezogen, z.B. *NP<sub>nom</sub>*, *NP<sub>akk</sub>*, *PP.von* etc.
- dadurch typischerweise konzentriert auf Inhaltswörter
- Beispiel:

... weil Peter den *Apfel* essen wollte ...

... Hannah pflückte Peter den *Apfel* vom Baum ...

... Hannah isst am liebsten *Äpfel* ...

... Der *Apfel* ist gerade vom Baum gefallen ...

# Kokkurrenz: Syntax-basierte Wörter

- Syntaktische Funktionen von Dimensionen werden einbezogen, z.B. *NPnom*, *NPakk*, *PP.von* etc.
- dadurch typischerweise konzentriert auf Inhaltswörter
- Beispiel:

... weil Peter den *Apfel* essen wollte ...

... Hannah pflückte Peter den *Apfel* vom Baum ...

... Hannah isst am liebsten *Äpfel* ...

... Der *Apfel* ist gerade vom Baum gefallen ...

	NPnom:Peter	NPnom:Hannah	NPdat:Peter	PP.von:Baum
<i>Apfel</i>	1	2	1	2
	V:essen	V:fallen	V:pflücken	
<i>Apfel</i>	2	1	1	

# Kookkurrenz: Syntax-basierte Wörter

- Syntaktische Funktionen von Dimensionen werden einbezogen, z.B. *NPnom*, *NPakk*, *PP.von* etc.
- dadurch typischerweise konzentriert auf Inhaltswörter
- Beispiel:

... weil Peter den *Apfel* essen wollte ...

... Hannah pflückte Peter den *Apfel* vom Baum ...

... Hannah isst am liebsten *Äpfel* ...

... Der *Apfel* ist gerade vom Baum gefallen ...

	NPnom:Peter	NPnom:Hannah	NPdat:Peter	PP.von:Baum
<i>Apfel</i>	1	2	1	2

	V:essen	V:fallen	V:pflücken
<i>Apfel</i>	2	1	1

	Peter	Hannah	Baum	essen	fallen	pflücken
<i>Apfel</i>	2	2	2	2	1	1

# Kookkurrenz: Muster

- Kookkurrenz ist beschränkt auf syntagmatische Muster
- quantitative Auswertung von Muster-Vorkommen liefert zusätzlich Information zu semantischen Relationen

- Beispiele:

... große *Tiere* wie z.B. *Tiger, Elefanten und Gorillas* ...

... weil das Kind *stolperte und deshalb fiel* ...

- Muster-Beispiele:

N: *NP* wie z.B. *NP<sub>1</sub>, NP<sub>2</sub>, ..., NP<sub>i-1</sub> und NP<sub>i</sub>* – Hyperonymie

V: *V<sub>1</sub> und deshalb V<sub>2</sub>* – Kausalität

# Kookkurrenz: Subkategorisierungsrahmen

- Alternationen von syntaktische Umgebungen geben Hinweis auf lexikalische Semantik (Pinker (1989), Levin (1993))
- hauptsächlich aber nicht ausschließlich bei Verben angewandt
- Beispiele:

... weil Peter an Gott *glaubt* ...

... Hannah *glaubt* Peter nicht ...

... das Kind *glaubt* noch an den Weihnachtsmann ...

... Sie *glaubt* nicht, dass ...

... denn er *glaubt*, sie sei ...

# Kokkurrenz: Subkategorisierungsrahmen

- Alternationen von syntaktische Umgebungen geben Hinweis auf lexikalische Semantik (Pinker (1989), Levin (1993))
- hauptsächlich aber nicht ausschließlich bei Verben angewandt
- Beispiele:

... weil Peter an Gott *glaubt* ...

... Hannah *glaubt* Peter nicht ...

... das Kind *glaubt* noch an den Weihnachtsmann ...

... Sie *glaubt* nicht, dass ...

... denn er *glaubt*, sie sei ...

	$\langle \text{NPnom, NPdat} \rangle$	$\langle \text{NPnom, PP.an} \rangle$	$\langle \text{NPnom, S.dass} \rangle$	$\langle \text{NPnom, S-2} \rangle$
<i>glauben</i>	1	2	1	1

# Zusammenfassung: Kookkurrenz

- **Kontext** einer linguistischen Einheit enthält Indikatoren für Wortbedeutung
- Kontext bezieht sich auf **Kookkurrenz**
- Kookkurrenz kann in verschiedenen Arten stattfinden;  
allgemeine bis tiefe semantische Information;  
flache bis tiefe Vorverarbeitung erforderlich
- Kookkurrenz wird in Matrizen oder Vektoren dargestellt;  
erlaubt mathematische fundierte Berechnungen

# Distributionelle Ähnlichkeit

- Summe von Kontext-Beschreibungen entspricht lexikalischer Beschreibung; Dimensionen = Eigenschaften
- Vergleich von Kontext-Beschreibungen vergleicht Semantik von lexikalischen Einträgen  $\rightsquigarrow$  distributionelle Ähnlichkeit  $\rightsquigarrow$  semantische Ähnlichkeit
- Stärke des distributionellen Zusammenhalts einzelner Kombinationen lexikalischer Einheiten weist auf Kollokationen hin
- Stärke des distributionellen Zusammenhalts innerhalb eines Musters weist auf semantische Relationen hin

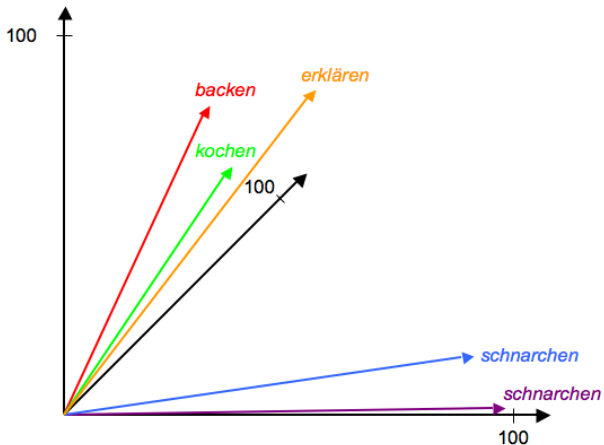
# Distributionelle Ähnlichkeit: Beispiel

## Subkategorisierungsrahmen von Verben

	$\langle \text{NP}_{\text{nom}} \rangle$	$\langle \text{NP}_{\text{nom}}, \text{NP}_{\text{akk}} \rangle$	$\langle \text{NP}_{\text{nom}}, \text{NP}_{\text{akk}}, \text{NP}_{\text{dat}} \rangle$
schlafen	98	1	1
kochen	35	50	15
backen	14	70	16
erklären	10	32	58
schnarchen	90	1	9

# Distributionelle Ähnlichkeit: Beispiel

## Vektor-Geometrie



# Distributionelle Ähnlichkeit: Distanzmaße

- Minkowski Metric /  $L_q$  Metric:

$$L_q(x, y) = \sqrt[q]{\sum_{i=1}^n (x_i - y_i)^q}$$

- Manhattan Distance:  $q = 1$

$$L_1(x, y) = \sum_{i=1}^n |x_i - y_i|$$

- Euclidean Distance:  $q = 2$

$$L_2(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

# Distributionelle Ähnlichkeit: Beispiel $L_2$

	$\langle \text{NP}_{\text{nom}} \rangle$	$\langle \text{NP}_{\text{nom}}, \text{NP}_{\text{akk}} \rangle$	$\langle \text{NP}_{\text{nom}}, \text{NP}_{\text{akk}}, \text{NP}_{\text{dat}} \rangle$
schlafen	98	1	1
kochen	35	50	15
backen	14	70	16
erklären	10	32	58
schnarchen	90	1	9

	schlafen	kochen	backen	erklären	schnarchen
schlafen	0	81	110	109	11
kochen	81	0	29	53	74
backen	110	29	0	57	103
erklären	109	53	57	0	99
schnarchen	11	74	103	99	0

Beispiel:

$$L_2(\text{kochen}, \text{backen}) = \sqrt{(35 - 14)^2 + (50 - 70)^2 + (15 - 16)^2}$$

$$= \sqrt{21^2 + 20^2 + 1^2} = \sqrt{842} = 29$$

# Distributionelle Ähnlichkeit: Distanzmaße

- **Kosinus**: misst Winkel zwischen zwei Vektoren

$$\text{Kosinus}(x, y) = \frac{\sum_{i=1}^n x_i * y_i}{\sqrt{\sum_{i=1}^n x_i^2} * \sqrt{\sum_{i=1}^n y_i^2}}$$

- Bereich: -1 (180 Grad) bis 1 (0 Grad)

# Distributionelle Ähnlichkeit: Beispiel Kosinus

	$\langle \text{NPnom} \rangle$	$\langle \text{NPnom}, \text{NPakk} \rangle$	$\langle \text{NPnom}, \text{NPakk}, \text{NPdat} \rangle$
schlafen	98	1	1
kochen	35	50	15
backen	14	70	16
erklären	10	32	58
schnarchen	90	1	9

	schlafen	kochen	backen	erklären	schnarchen
schlafen	1,00	0,57	0,20	0,16	0,99
kochen	0,57	1,00	0,92	0,67	0,59
backen	0,20	0,92	1,00	0,67	0,22
erklären	0,16	0,67	0,67	1,00	0,24
schnarchen	0,99	0,59	0,22	0,24	1,00

Beispiel:

$$\cos(\text{kochen}, \text{backen}) = \frac{35 \cdot 14 + 50 \cdot 70 + 15 \cdot 16}{\sqrt{35^2 + 50^2 + 15^2} \cdot \sqrt{14^2 + 70^2 + 16^2}} = \frac{4230}{\sqrt{3950} \cdot \sqrt{5352}} = 0.92$$

# Beispielansätze

- 1 Church & Hanks (1990): Wort-Assoziationen
- 2 Lin (1998): Thesaurus-Erstellung
- 3 Hearst (1992): Hyperonymie

# Wort-Assoziationen

- **Association ratio**: statistical measure to model **word association**
- Claims an objective measure to create association norms, in comparison to standard association norms in psycholinguistics
- Word association captures various semantic relationships:
  - idioms such as *bread and butter*
  - compounds such as *computer scientist*
  - semantic relations such as co-hyponymy, e.g., *man/woman*
  - phrasal verbs such as *refrain from*

# Mutual Information

- Association ratio is an instance of the information theoretic measure **Mutual Information**:

$$I(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

- Mutual information compares the probability of observing  $x$  and  $y$  together (the joint probability) with the probabilities of observing  $x$  and  $y$  independently (chance)
- Association ratio encodes linear precedence
- Corpora: total of approx. 60 million words from various versions of the AP corpus
- Window size for underlying frequencies: 5;  
ideal window size depends on type of semantic relationship

# Beispiele: Fenstergröße

Window size according to semantic relationship:

Relation	Word $x$	Word $y$	Separation	
			Mean	Variance
Fixed	bread	butter	2.00	0.00
	drink	drive	2.00	0.00
Compound	computer	scientist	1.12	0.10
	United	States	0.98	0.14
Semantic	man	woman	1.46	8.07
	man	women	-0.12	13.08
Lexical	refraining	from	1.11	0.20
	coming	from	0.83	2.89
	keeping	from	2.14	5.53

# Beispiele: Asymmetrie

Asymmetry in AP corpus:

$x$	$y$	$f(x, y)$	$f(y, x)$
doctors	nurses	99	10
man	woman	256	56
doctors	lawyers	29	19
bread	butter	15	1
save	life	129	11
save	money	187	11
save	from	176	18
supposed	to	1,188	25

# Beispiele: Assoziationen

Associations with *doctor* in AP corpus:

$x$	$y$	$I(x, y)$
honorary	doctor	11.3
doctors	dentists	11.3
doctors	nurses	10.7
doctors	treating	9.4
examined	doctor	9.0
doctors	treat	8.9
doctor	bills	8.7
doctor	visits	8.7
doctors	hospitals	8.6
nurses	doctors	8.4
	...	
doctor	with	0.96
a	doctor	0.95

# Beispiele: Partikelverben

Phrasal verbs in AP corpus:

$x$	$y$	$f(x)$	$f(y)$	$f(x, y)$	$I(x, y)$
set	up	13,046	64,601	2,713	7.3
set	off	13,046	20,693	463	6.2
set	out	13,046	47,956	301	4.4
set	on	13,046	258,170	162	1.1
set	in	13,046	739,932	795	1.8
set	about	13,046	82,319	16	-0.6

# Beispiele: Verb-Objekt-Paare

Verb-object pairs for *drink*:

$x$	$y$	$I(x, y)$
drink	martinis	12.6
drink	cup water	11.6
drink	champagne	10.9
drink	beverage	10.8
drink	cup coffee	10.6
drink	cognac	10.6
drink	beer	9.9
drink	toast	9.6
drink	alcohol	9.4

# Beispiele: Subkategorisierende Verben

Verbs subcategorising *telephone* as direct object:

$x$	$y$	$I(x, y)$
sit by	telephone	11.78
disconnect	telephone	9.48
answer	telephone	8.80
hang up	telephone	7.87
tap	telephone	7.69
pick up	telephone	5.63
return	telephone	5.01
be by	telephone	4.93
spot	telephone	4.43
repeat	telephone	4.39

# Automatische Thesaurus-Erstellung

- Automatic construction of a thesaurus
- Basis: similarity with respect to distributional patterns of words
- Similarity between two words is defined as amount of information contained in the commonality between the words divided by the amount of information in the individual descriptions of the words
- Corpus: dependency-parsed 64 million word newspaper data

# Dependenz-Tripel

- **Dependency triple**: two words  $w$  and  $w'$  and the grammatical relationship  $r$  between them
- Example: *I have a brown dog*
  - have subj I
  - have obj dog
  - dog adj-mod brown
  - dog det a
- $||w, r, w'||$ : frequency count of dependency triple  $(w, r, w')$
- with wild card  $*$ , frequency counts of dependency triples that match rest of pattern are summed up, e.g.,  $||cook, obj, *||$  is the total occurrences of cook-object relationships in the parsed corpus
- $||*, *, *||$ : total number of dependency triples in the parsed corpus

## Beispiele: Dependenz-Tripel

- **Description of a word  $w$ :** frequency counts of all dependency triples that match  $(w, *, *)$

- Example:

$||cell, subj\_of, absorb|| = 1$

$||cell, pobj\_of, in|| = 159$

$||cell, pobj\_of, inside|| = 16$

$||cell, nmod\_of, abnormality|| = 3$

$||cell, nmod\_of, anemia|| = 8$

$||cell, obj\_of, attack|| = 6$

$||cell, obj\_of, bludgeon|| = 1$

$||cell, obj\_of, call|| = 11$

$||cell, obj\_of, contain|| = 4$

$||cell, nmod, bacteria|| = 3$

$||cell, nmod, body|| = 2$

$||cell, nmod, bonemarrow|| = 2$

# Dependenz-Tripel

- An occurrence of a dependency triple  $(w, r, w')$  can be regarded as the co-occurrence of three events:
  - $A$ : a randomly selected word is  $w$
  - $B$ : a randomly selected dependency type is  $r$
  - $C$ : a randomly selected word is  $w'$
- Probability of  $A, B, C$  co-occurring if  $||w, r, w'||$  is unknown, assuming that  $A$  and  $C$  are conditionally independent:

$$P_{MLE}(A, B, C) = P_{MLE}(B) P_{MLE}(A|B) P_{MLE}(C|B) \text{ with}$$

- $P_{MLE}(B) = \frac{||*,r,*||}{||*,*,*||}$
  - $P_{MLE}(A|B) = \frac{||w,r,*||}{||*,r,*||}$
  - $P_{MLE}(C|B) = \frac{||*,r,w'||}{||*,r,*||}$
- Probability of  $A, B, C$  co-occurring if  $||w, r, w'||$  is known:

$$P_{MLE}(A, B, C) = \frac{||w,r,w'||}{||*,*,*||}$$

# Dependenz-Tripel und Ähnlichkeit

- **Commonality between two words:** dependency triples that appear in the descriptions of both words
- **Mutual Information between two words:**

$$I(w, r, w') = \frac{\log(P_{MLE}(A, B, C))}{\log(P_{MLE}(B) P_{MLE}(A|B) P_{MLE}(C|B))}$$
$$= \log \frac{||w, r, w'|| * ||*, r, *||}{||w, r, *|| * ||*, r, w'||}$$

- **Similarity  $sim(w_1, w_2)$  between two words  $w_1$  and  $w_2$ :**

$$\frac{\sum_{(r, w') \in T(w_1) \cap T(w_2)} I(w_1, r, w') + I(w_2, r, w')}{\sum_{(r, w) \in T(w_1)} I(w_1, r, w') + \sum_{(r, w') \in T(w_2)} I(w_2, r, w')}$$

with  $T(w)$  the set of pairs  $(r, w')$  such that  $I(w, r, w')$  is positive

# Beispiele: Ergebnis

## Examples of thesaurus entries:

- **brief (noun)**: affidavit 0.13, petition 0.05, memorandum 0.05, motion 0.05, lawsuit 0.05, deposition 0.05, slight 0.05, prospectus 0.04, document 0.04, paper 0.04, ...
- **brief (verb)**: tell 0.09, urge 0.07, ask 0.07, meet 0.06, appoint 0.06, elect 0.05, name 0.05, empower 0.05, summon 0.05, overrule 0.04, ...
- **brief (adjective)**: lengthy 0.13, short 0.12, recent 0.09, prolonged 0.09, long 0.09, extended 0.09, daylong 0.08, scheduled 0.08, stormy 0.07, planned 0.06, ...

# Automatische Induktion von Hyperonymie

- Automatic acquisition of hyponymy lexical relation from unrestricted text
- Basis: set of lexico-syntactic patterns
- Example:

$NP_0$  such as  $\{NP_1, NP_2, \dots, (and|or)\} NP_n$

$\rightsquigarrow$  for all  $NP_i, 1 \leq i \leq n, \text{hyponym}(NP_i, NP_0)$

# Relationen-Muster

- **Desiderata for patterns:**
  - occur frequently and in many text genres
  - (almost) always indicate the relation of interest
  - can be recognised with little or no pre-encoded knowledge
- **Discovery of patterns:**
  - ① Decide on a lexical relation
  - ② Gather a list of terms for which this relation holds
  - ③ Find places in the corpus where these expressions occur syntactically near one another and record the environment
  - ④ Find the commonalities among these environments and hypothesise in the form of patterns
  - ⑤ Gather more instances and go back to step 2

# Muster für Hyperonymie

## Patterns for hyponymy:

- *NP such as {NP, NP, ..., (and|or)} NP*  
The bow lute, such as the Bambara ndang, ...
- *such NP as {NP, } \* {(or|and)} NP*  
... works by such authors as Herrick, Goldsmith, and Shakespeare.
- *NP {, NP} \* {, } or other NP*  
Bruises, wounds, broken bones or other injuries ...
- *NP {, NP} \* {, } and other NP*  
... temples, treasuries, and other important civic buildings
- *NP {, } including {NP, } \* {or|and} NP*  
All common-law countries, including Canada and England ...
- *NP {, } especially {NP, } \* {or|and} NP*  
... most European countries, especially France, England, and Spain.

# Beispiele: Hyperonymie

## Examples of hyponymy:

Hypernym	Hyponyms
cereals	rice*, wheat*
countries	Cuba, Vietnam, France*
liqueurs	anisetten*, absinthe*
fabrics	acrylics*, nylon*, silk*
seabirds	penguins, albatross*
legumes	lentils*, beans*, nuts
fruit	olives*, grapes*
ideologies	liberalism, conservatism
industries	steel, iron, shoes
minerals	pyrite*, galena

Entries with \* indicate relations found in WordNet.

# Referenzen: Distributionelle Hypothese



John R. Firth.

*Papers in Linguistics 1934-51.*

Longmans, London, UK, 1957.



Zellig Harris.

Distributional Structure.

In Jerold J. Katz, editor, *The Philosophy of Linguistics*, Oxford Readings in Philosophy, pages 26–47. Oxford University Press, 1968.



Diego Marconi.

*Lexical Competence.*

MIT Press, Cambridge, MA, 1997.

# Referenzen: Syntax-Semantik-Interface



Beth Levin.




*English Verb Classes and Alternations.*  
The University of Chicago Press, 1993.



Steven Pinker.

*Learnability and Cognition: The Acquisition of Argument Structure.*  
MIT Press, Cambridge, MA, 1989.

# Referenzen: Distributionelle Ansätze

-  Kenneth W. Church and Patrick Hanks.  
Word Association Norms, Mutual Information, and Lexicography.  
*Computational Linguistics*, 16(1):22–29, 1990.
-  Marti Hearst.  
Automatic Acquisition of Hyponyms from Large Text Corpora.  
In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992.
-  Dekang Lin.  
Automatic Retrieval and Clustering of Similar Words.  
In *Proceedings of the 17th International Conference on Computational Linguistics*, Montreal, Canada, 1998.