Beileibe nichts unversucht lassen – NPI-Gewinnung aus partiell annotierten Korpora

Timm Lichte

SFB 441
Universität Tübingen

5. Februar 2006

Um was geht es?

- Gewinnung von NPIs aus partiell annotierten Korpora:
  - eingliedrige NPIs (*sonderlich*)
  - mehrgliedrige NPIs (*alle Tassen im Schrank haben*)
- Zweck: Eine Liste von **NPI-Kandidaten**, die der Linguist weiterverarbeiten kann.
- Idee: Negative Polarität ist ein kollokationelles Phänomen (van der Wouden(1997)).

Teil I

Eine kleine Einführung

Eine kleine Einführung: NPIs

- **Niemand** von uns war jemals im Jemen.
  * Jeder von uns war jemals im Jemen.
- Hast du noch alle Tassen im Schrank **?**
  * Ich schwöre dir, dass Peter alle Tassen im Schrank hatte.

$\Rightarrow$ NPI müssen von einem (geeigneten) Negativitätstrigger lizenziert werden.

Eine kleine Einführung: NPIs

- große Formenvielfalt
- nicht an eine bestimmte sytaktische oder semantische Kategorie geknüpft
- eingliedrig vs. mehrgliedrig
  (z.B.: *sonderlich* vs. *wahrhaben wollen*)
- nicht-polysem vs. polysem
  (z.B.: *wahrhaben wollen* vs. *einen Finger rühren*)
- Subklassen bezüglich der Stärke der Negativität (Zwarts(1997))
- Dokumentation deutscher NPIs relativ dünn:
  Kürschner(1983) nennt etwa 350 NPIs, während z.B. Hoeksema(2005) 760 NPIs des Niederländischen beschreibt.

Eine kleine Einführung: NPI-Lizenzierer

- große Formenvielfalt :
  - Negationspartikel und Negative Quantoren:
    (nicht, kein, niemand, niemals, wenig, höchstens, ...)
  - Negative Konjunktionen:
    (ohne dass, ob, bevor, ...)
  - Restriktoren von Superlativen und Allquantoren
  - Inhärent negative Verben:
    (bezweifeln, weigern, abstreiten, dementieren, ablehnen, ...)
  - Negierte Einstellungsverben:
    (glauben, vorstellen können, für möglich halten, ...)
  - Antezedent von Konditionalkonstruktionen
  - Fragen
  - Negative Prädikate:
    (unwahrscheinlich, unmöglich, ...)
  - zu-Komparative, als-Komparative
  - ...

Eine kleine Einführung: NPI-Lizenzierer

- große Formenvielfalt
- Hypothese: Alle NPI-Lizenzierer zeichnen sich logisch durch **fallenden Monotonizität** aus (Engl.: Downward Entailment, Downward Monotonicity).
  - **Few** congressmen eat vegetables.
    $\|\text{spinach}\| \subseteq \|\text{vegetables}\|$

    Few congressmen eat spinach.

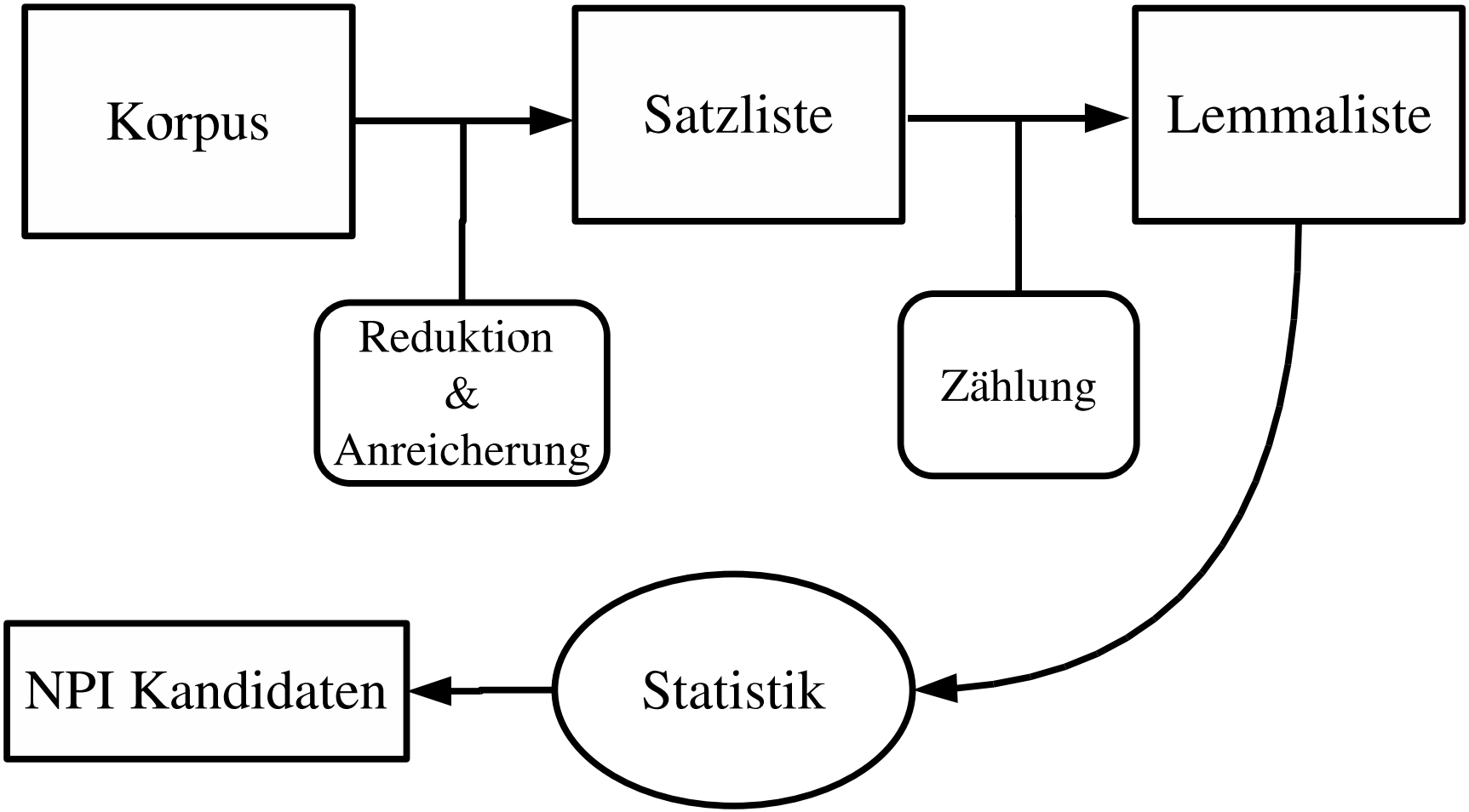
Eine kleine Einführung: NPI-Lizenzierer

- große Formenvielfalt
- Hypothese: Alle NPI-Lizenzierer zeichnen sich logisch durch **fallenden Monotonizität** aus (Engl.: Downward Entailment, Downward Monotonicity).
- Stufen der Negativität
  - **Fragen** triggern eine schwächere Form der Negation als **nicht**.

Teil II

Korpusextraktion der NPI-Kandidaten

Der Exktraktionsalgorithmus



Das Korpus: TüPP-D/Z

- *Tübinger partiell geparstes Korpus des Deutschen*(TüPP)
- basierend auf der Tageszeitung *die tageszeitung*(*taz*)
- Lemmatisierung, POS-Tags, Chunks, Teilsatzannotation, ...
- in toto: ca. 200 Millionen Wörter
- Wir nutzen einen Ausschnitt von ca. 2,7 Millionen Sätzen, das umfasst 4 von 12 Jahrgängen.

Die Satzliste

- Die Satzliste enthält ...
  - die Lemmaform der Wörter,
  - die Teilsatzgrenzen.
- NPI-Lizenzierer werden durch DEINT-Markierungen ersetzt oder lizenzierende Teilsätze werden markiert.
- Beispiel:

  CLstart1 können etwas wirklich gut sein allein aus das Grund, **CLstart2** weil es immer so sein **CLende2 DEINT CLende1**

  „Kann etwas wirklich gut sein allein aus dem Grund, weil es immer so ist? “

Die Lemmaliste

Jedes Lemma hat zwei Frequenzangaben:

- die Gesamtfrequenz des Auftretens
- die Frequenz des Auftretens innerhalb des Skopus einer DEINT-Markierung.

Übersee	174	27
...	..	..

- Lemmata mit einer Gesamtfrequenz ≤ 40 werden ignoriert! (641 035 $\longrightarrow$ 34 952)

Quantitative Evaluation

- Zuerst wird für jedes Lemma sein **Kontextverhältnis (KV)** berechnet:

$$KV = \frac{Auftreteninnegativen Kontexten}{Gesamtauftreten}$$

Übersee	174	27	0.16
...	..	..	...

N	Minimum	Maximum	Durchschnitt	Standardabweichung
34952	0.0	1.0	0.15	0.08

Quantitative Evaluation

- Zuerst wird für jedes Lemma sein **Kontextverhältnis (KV)** berechnet:
  $KV = \frac{Auftreteninnegativen Kontexten}{Gesamtauftreten}$
- Dann werden die Lemmata entsprechend ihres KV-Werts sortiert:

1	unversucht	50	50	1.00
..	...	..	..	...

Schließlich: Die Kandidatenliste

#	*Lemma*	*CR*
1	**unversucht**	1.00
2	***unterschätzender**	1.00
3	**umhin**	0.99
4	**nachstehen**	0.98
5	**lumpen**	0.98
6	**verhehlen**	0.97
7	**geheuer**	0.96
8	**beirren**	0.96
9	***Genaueres**	0.95
10	***wegdenken**	0.95
11	unähnlich	0.94
12	***allzuviel**	0.92
13	**sonderlich**	0.90
14	hinwegtäuschen	0.89
15	dagewesen	0.89
16	abneigen	0.89
17	behagen	0.85
18	**verdenken**	0.85
19	***missen**	0.84
20	**fruchten**	0.83
..	...	..

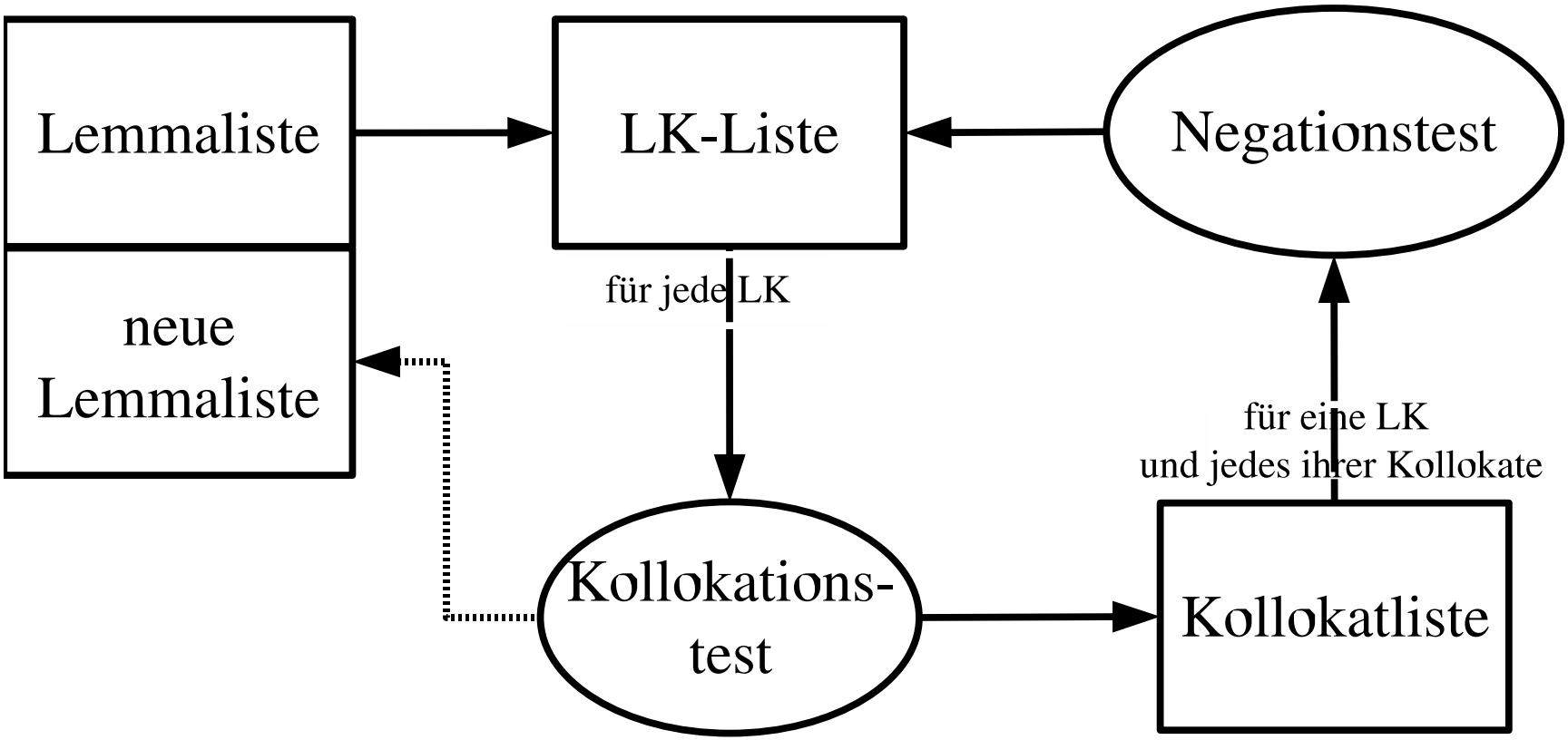
Schließlich: Die Kandidatenliste

Die Kandidatenliste sieht vielversprechend aus, aber ...

- Die NPI-Kandidaten sind eingliedrig, obwohl viele NPIs mehrgliedrig sind.
- Die Elemente mehrgliedriger NPIs können einen tiefen Ranglistenplatz haben und müssen deshalb 'disambiguiert' werden.

  (*alle Tassen im Schrank haben* $\longrightarrow$ *Tasse* auf Position 6934)

Erweiterung für mehrgliedrige NPIs



Kandidatenliste mit mehrgliedrigen Kandidaten

reichen hinten vorne
Zweifel lassen daran er daß
schlecht staunen Blitz solange löschen
aufgehen Rezept obwohl
bekannt zunächst über
geheuer ganz
ausstehen können
zuletzt verdanken
umhin VVIZU kommen zu
ganz geheuer
unterschätzender zu
Sorge Sie brauchen machen
Stellungnahme Redaktionsschluß bis
noch dementieren wollen bestätigen
Sache jedermanns
unversucht lassen
entbehren gewiß
jedermanns Sache
allzu lang Zeit vor
recht glauben so
übrig ander bleiben als VVIZU

nachstehen
lumpen lassen
darüber hinwegtäuschen der können daß
angehen es können daß
ändern daran auch
gar wissen ich
kommen umhin VVIZU
so gar schlimm
Wahl ander bleiben
einwenden VVIZU gegen
verkneifen können Sie
genug gehen weit
verstehen Aufregung
vermögen sagen VVIZU zu
erkennen woran Frau
müde betonen
erwähnen Wort mit
halt vor machen auch
verhehlen
...

Kandidatenliste mit mehrgliedrigen Kandidaten

- 'Disambiguierung' polysemer Kandidaten:

  Tasse $\longrightarrow$ Tasse Schrank (0.73, #232)

- 102 Elemente unserer Kandidatenliste ($\leq$ #1000) tauchen auch in Kürschners NPI-Sammlung auf.

- Wermutstropfen (ca. 80):

  notwendigerweise geben die auf Meinung wieder Seite erscheinend (1.00) $\rightarrow$ "Die auf dieser Seite erscheinenden Leserbriefe geben nicht notwendigerweise die Meinung der taz wieder.„

Teil III

Automatische Subklassifizierung

NPI-Klassifikation nach Zwarts(1997)

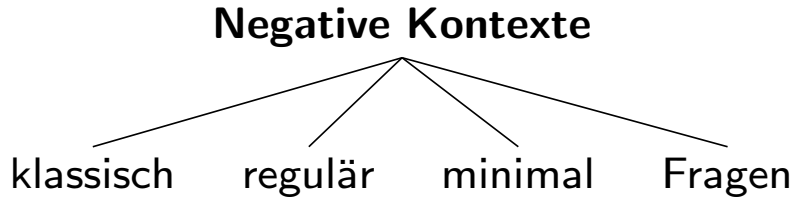
Negative Kontexte

klassisch	regulär	minimal	Fragen
nicht,	nichts	höchstens	?
ohne dass	niemand	kaum	
nicht glauben, dass	kein	wenn	
anti-morph	anti-additiv	monoton fallend	non-verdidikal

NPI	*Negation*		
	klassisch	regulär	minimal
superstark	+	-	-
stark	+	+	-
schwach	+	+	+

Wie geht das automatisch?

- Die Frequenzdaten werden feiner aufgeschlüsselt.



sonderlich (443)	357	39	1	2
Ahnung haben (548)	49	387	21	9
jemals (939)	150	101	53	213
Tasse Schrank (22)	4	2	5	5
brauchen fürchten zu (64)	44	12	1	0

Wie geht das automatisch?

- Die Frequenzdaten werden feiner aufgeschlüsselt.
- Eine simple statistische Methode wird angewandt:

Wie geht das automatisch?

- Die Frequenzdaten werden feiner aufgeschlüsselt.
- Eine simple statistische Methode wird angewandt:
  1. E_k := erwarteter Frequenzwert bezügl. einer Negationsklasse k
  2. O_k := beobachteter Frequenzwert bezügl. einer Negationsklasse k
  3. Berrechnung der Differenz zwischen E_k und O_k
  4. Standardisierung der Differenz
     (anhand der Gesamtfrequenz eines Kandidaten und anhand der Diffenrenzwerte anderer Lemmata (z-Wert-Statistik))

Wie geht das automatisch?

- Die Frequenzdaten werden feiner aufgeschlüsselt.
- Eine simple statistische Methode wird angewandt:

Negative Kontexte

	klassisch	regulär	minimal	Fragen
sonderlich	+++	o	–	o
(443)	357	39	1	2
Ahnung haben	—	+++	o	o
(548)	49	387	21	9
jemals	–	o	o	+++
(939)	150	101	53	213
Tasse Schrank	—	o	++	++
(22)	4	2	5	5
brauchen fürchten zu	++	o	-	-
(64)	44	12	1	0
Sache jedermanns	+++	–	–	o
(38)	37	0	0	1

Schlussbemerkungen

- Welche NPIs sind unerreichbar?
- Expandierung erst für 7000 Lemmata
- Viele Detailverbesserungen in Arbeit
- Kombination mit Sprecherurteilen?
- ...

Jetzt ist Jan-Philipp dran ...