The Retrieval and Classification of NPIs

Timm Lichte
Universität Tübingen

Hauptseminar (Frank Richter):
Negative Polarity Items: Semantics, Computational Linguistics, and Corpus Linguistics

20/25/27.06.2007

Aim of this work

- **"Retrieval":** to support the collection of German NPIs (CoDII-NPI.de).
- **"Classification":** Can we track down the mirror image of theoretical classifications (cf. Zwarts) in frequency profiles?
- **Validation?:** No!

Main references:
(1) Timm Lichte (2005): Korpusbasierte Akquirierung negativ-polärer Elemente. Master's Thesis, Eberhard-Karls-Universität Tübingen, Seminar für Sprachwissenschaft.

(2) Timm Lichte and Jan-Philipp Soehn (to appear). The Retrieval and Classification of Negative Polarity Items using Statistical Profiles. In: Sam Featherston and Wolfgang Sternefeld (Eds.): Roots: Linguistics in Search of Its Evidential Base. Mouton de Gruyter.

Basic idea of this work

distributional restriction $\Rightarrow$ significance of frequency data

- negative polarity is a **collocational phenomenon** (van der Wouden, 1997)
  $\rightsquigarrow$ arbitrary restriction on distribution!
- much work on collocation retrieval using **frequency statistics**
  $\rightsquigarrow$ distribution $\approx$ set of frequencies
- application of collocation retrieval techniques for the retrieval of NPI-candidates

Outline

1. First step: Collect the bits of theory
2. The instruments: Corpus and statistics
3. Method and results: Extraction of single lemmata and lemma chains
4. Method and results: Quantitative classification of NPIs

$\Longrightarrow$ To what extent can we apply theoretical insights???

Part I

Theoretical Background

NPIs - Basics

Negative Polarity Items (NPIs) show negative polarity.

Negative polarity

- Distributional restriction in favour of contexts that exhibit a certain semantic property: **"negativity"**.
- "Negativity" is induced by so-called **licensers**.
- Negative polarity is not predictable.
- NPIs are collocations, consisting of an lexical item and the class of licensers.

(1) a. **Niemand** von uns war jemals in Turin.
b.#Jeder von uns war jemals in Turin.

NPIs - Examples

- **adverbials**: *jemals, beileibe, sonderlich*
- **nouns**: *Deut, Menschenseele*
- **verbs**: *brauchen, ausstehen können, wahrhaben wollen*
- **multiword expressions**: *alle Tassen im Schrank haben, einen Finger rühren*

$\Longrightarrow$ NPIs can be found in every part-of-speech!

$\Longrightarrow$ The inventory of German NPIs is not properly documented, yet. There exists only a rudimentary collection of NPIs in Kürschner(1983).

Shape of licensers

- **n-words** (negative particles, negative quantifiers):
  niemand, niemals, nichts,...
- **antecedent of conditionals**: *Falls sie jemals...*
- **questions**: *Hat er jemals...?*
- **restrictor of universal quantifiers and superlatives**:
  Er war der erste, der jemals am Nordpol war.
  Jeder, der jemals...
- **non-affirmative verbs**: *bezweifeln, verbieten,...*
- **neg-raising verbs**: *Ich glaube nicht, dass er jemals ...*
- **negative conjunctions**: *ohne dass*
- **comparative than-sentences**:
  Die Lage ist wohl dramatischer, als er es wahrhaben wollte.
- **too-comparatives**: *Es gibt zu viele Opfer, als daß die Region jemals zur Ruhe kommen könnte.*
- **negative predicates**: *unwahrscheinlich*
- **other**: *endlich, nur*

Licensing property?

downward entailment (DE) (Fauconnier,1975; Ladusaw,1980)

$$X \subseteq Y \Rightarrow f(Y) \subseteq f(X)$$

Does not catch: e.g. questions, *bedauern, nur* (Strawson-DE?)

non-veridicality (Giannakidou,1997)

$$f(p) \not\Rightarrow p$$

\+ negative implicatures (Linebarger,1980,1987)

Does not catch: e.g. *überrascht sein, bedauern* (veridical!)

$\Longrightarrow$ The crucial property of licensers is somewhat unclear!

Licensing property?

downward entailment (DE) (Fauconnier,1975; Ladusaw,1980)

$$X \subseteq Y \Rightarrow f(Y) \subseteq f(X)$$

Does not catch: e.g. questions, *bedauern, nur* (Strawson-DE?)

non-veridicality (Giannakidou,1997)

$$f(p) \not\Rightarrow p$$

+ negative implicatures (Linebarger,1980,1987)

Does not catch: e.g. *überrascht sein, bedauern* (veridical!)

$\Longrightarrow$ The crucial property of licensers is somewhat unclear!

The scope of licensers (1)

(2) a. Phil would **not** give me anything.
b. *Anything Phil would **not** give me.

Scope of licensers $\subseteq$ Scope of negation

- Ladusaw(1980): c-command on S-structure
- Linebarger(1987): immediate scope constraint
  - no interaction with propositional operators on LF

  (3) #Sam didn't read every child any stories.

- Hoeksema(2000): licensing scope is semantic scope of negation?
  - Does not hold for **minimizer** (*ein Deut*, *eine Menschenseele*), for they take wide scope in pre-negative position.
  - **BUT:** all other NPIs are licensed under semantic scope.

The scope of licensers (1)

(2) a. Phil would **not** give me anything.
b. *Anything Phil would **not** give me.

Scope of licensers $\subseteq$ Scope of negation

- Ladusaw(1980): c-command on S-structure
- Linebarger(1987): immediate scope constraint
  - no interaction with propositional operators on LF

  (3) #Sam didn't read every child any stories.

- Hoeksema(2000): licensing scope is semantic scope of negation?
  - Does not hold for **minimizer** (*ein Deut*, *eine Menschenseele*), for they take wide scope in pre-negative position.
  - **BUT:** all other NPIs are licensed under semantic scope.

The scope of licensers (2)

Scope of licensers $\subseteq$ Scope of negation

- Minimizer
- Contrastive negation:

(4) ***Nicht** Peter hat der Caritas <u>eine müde Mark</u> gespendet (, sondern Georg)

- Caveat: Negative particles with narrow scope:

(5) a. **Not** long ago there was rain falling.
$\not\Rightarrow$ ¬(long ago there was rain falling)
b. ein **nicht** sonderlich interessierter Student

Zwarts classification of licensers and NPIs

Grades of negativity:

AM	AA	DE	DEINT
anti-morphic	anti-additive	downward-entailing	DE + interrogatives
classic	regular	minimal	
nicht, keinesfalls	*niemand, kein*	*wenige, kaum*	questions

Classification of NPIs:

NPI	Negation			
	classic	regular	minimal	
weak	+	+	+	*jemals*
strong	+	+	-	*sonderlich*
superstrong	+	-	-	???

Where licensers play a role here

shape of licensers licensing property the set of licensers	$\longrightarrow$	annotation of the licensers in the corpus
scope of licensers	$\longrightarrow$	compilation of frequencies
strength of licensers	$\longrightarrow$	NPI-classification

Part II

The instruments: Corpus and statistics

The corpus - TüPP-D/Z

- **T**übingen **P**artially **P**arsed corpus - **D**eutsch/**Z**eitung
- based on "die tageszeitung"
- altogether: ca. 200 Mio. words, volume 1986-1999
- we use: volume 1990-1998, 5.8 Mio. sentences
- XML format

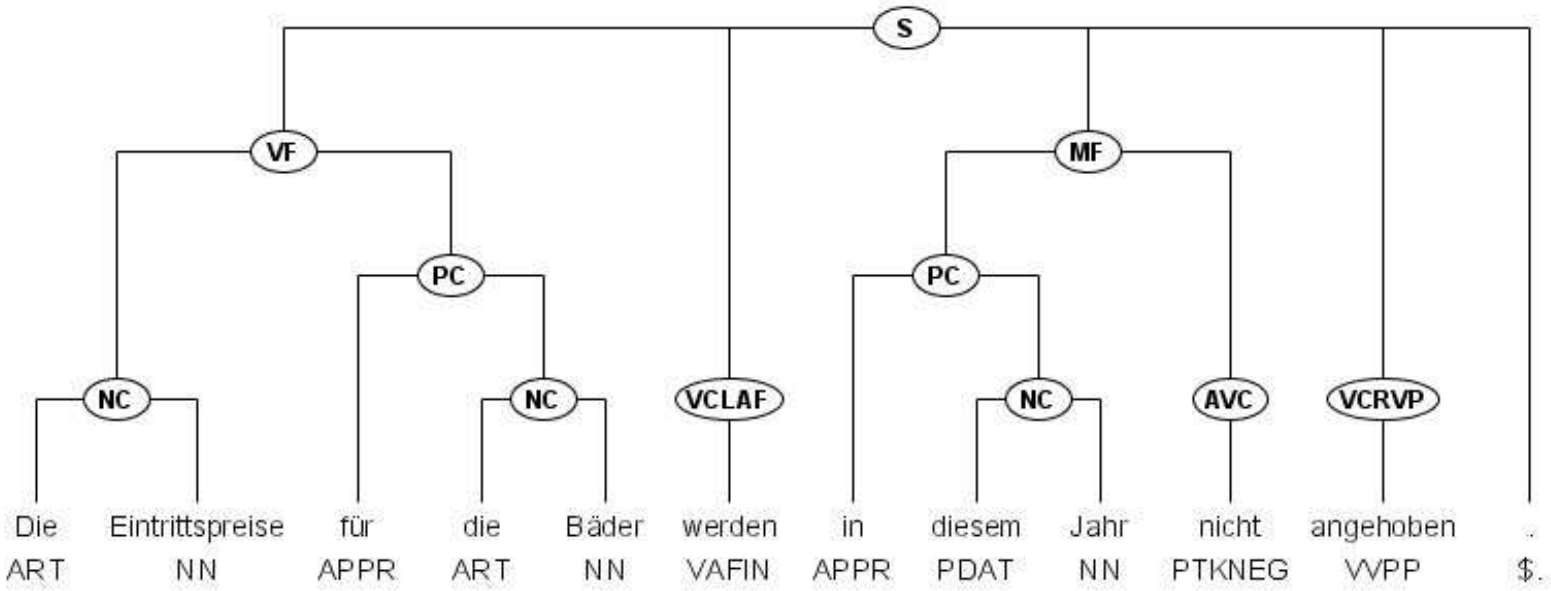
Why do we use a partially parsed corpus?
$\rightsquigarrow$ Due to the sparse data problem!

TüPP-D/Z - Annotation

- lemmatization
- part-of-speech
- chunks
- clauses
- topological fields

$\Longrightarrow$ Except topological fields, we can use all of this!

TüPP-D/Z - Example



TüPP-D/Z - Errors in annotation

... are unavoidable.

We apply a filter for defective clause annotation:

(6) <**cl**> Dem Papier ist zu entnehmen </**cl**>, **dass** gestern nach internen Eil-Verhandlungen auf höchster Ebene der Berliner Ramsch-Konzern 'Rudis-Reste-Rampe' und die koreanische 'Han-Jin-Shipping-Reederei' für die vollständige Finanzierung der Kosten in Höhe von 1,8 Milliarden Mark gewonnen werden konnten. [taz, 01.04.1998]

$\Longrightarrow$ Ca. 39% of the sentences get lost (Lichte,2005).
$\Longrightarrow$ The statistics become more expressive.

Statistics - Frequencies and contingency tables

The retrieval of NPIs $\approx$ The retrieval of collocations

"Collocations are identified by **the frequency of word co-occurrences** in corpora. Basically, word n-grams (mostly bi-grams) are collected from varying **spans**." (Krenn(1999),p.28)

Contingency tables:

	w_2	$\neg w_2$	
w_1	O_{11}	O_{12}	O_{1+}
$\neg w_1$	O_{21}	O_{22}	O_{2+}
	O_{+1}	O_{+2}	N

$\Longrightarrow$ Which bi-grams/contingency tables are significant?

Statistics - Frequencies and contingency tables

The retrieval of NPIs $\approx$ The retrieval of collocations

"Collocations are identified by **the frequency of word co-occurrences** in corpora. Basically, word n-grams (mostly bi-grams) are collected from varying **spans**." (Krenn(1999),p.28)

Contingency tables:

	w_2	$\neg w_2$	
w_1	O_{11}	O_{12}	O_{1+}
$\neg w_1$	O_{21}	O_{22}	O_{2+}
	O_{+1}	O_{+2}	N

$\Longrightarrow$ Which bi-grams/contingency tables are significant?

Statistics - Frequencies and contingency tables

The retrieval of NPIs $\approx$ The retrieval of collocations

"Collocations are identified by **the frequency of word co-occurrences** in corpora. Basically, word n-grams (mostly bi-grams) are collected from varying **spans**." (Krenn(1999),p.28)

Contingency tables:

	NEG	$\neg$*NEG*	
sonderlich	879	103	982
$\neg$ *sonderlich*	1456455	4309860	5766315
	1457334	4309963	5767297

$\Longrightarrow$ Which bi-grams/contingency tables are significant?

Statistics - Association measures

no significance statements ($\rightsquigarrow$ ranking)

Mutual Information (MI) / Context Ratio (CR)

$$CR(w_1) = \frac{O_{11}}{O_{1+}}$$

significance statements

X^2-Test
Log-likelihood Quotient (G^2)

$\Longrightarrow$ NPIs are expected to have a salient association measure wrt. triggers of negativity.

Statistics - Evaluation of association measures

Which association measure yields the best list of NPI-candidates?

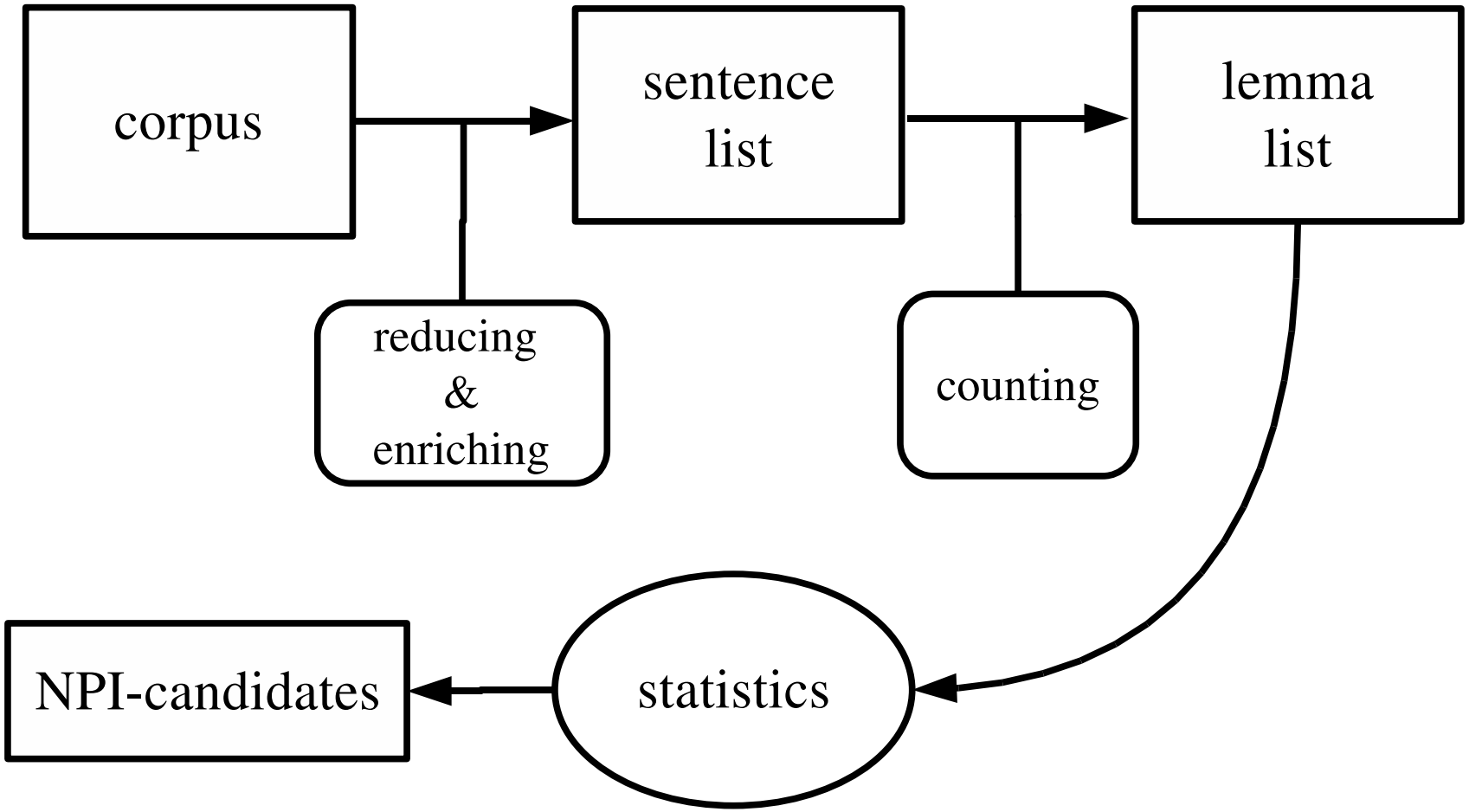
- precision & recall
- uninterpolated average precision (uap)
- Spearmen's rank correlation coefficient (ρ_s)

$\Longrightarrow$ **Context ratio** seems to be the most appropriate AM.

Part III

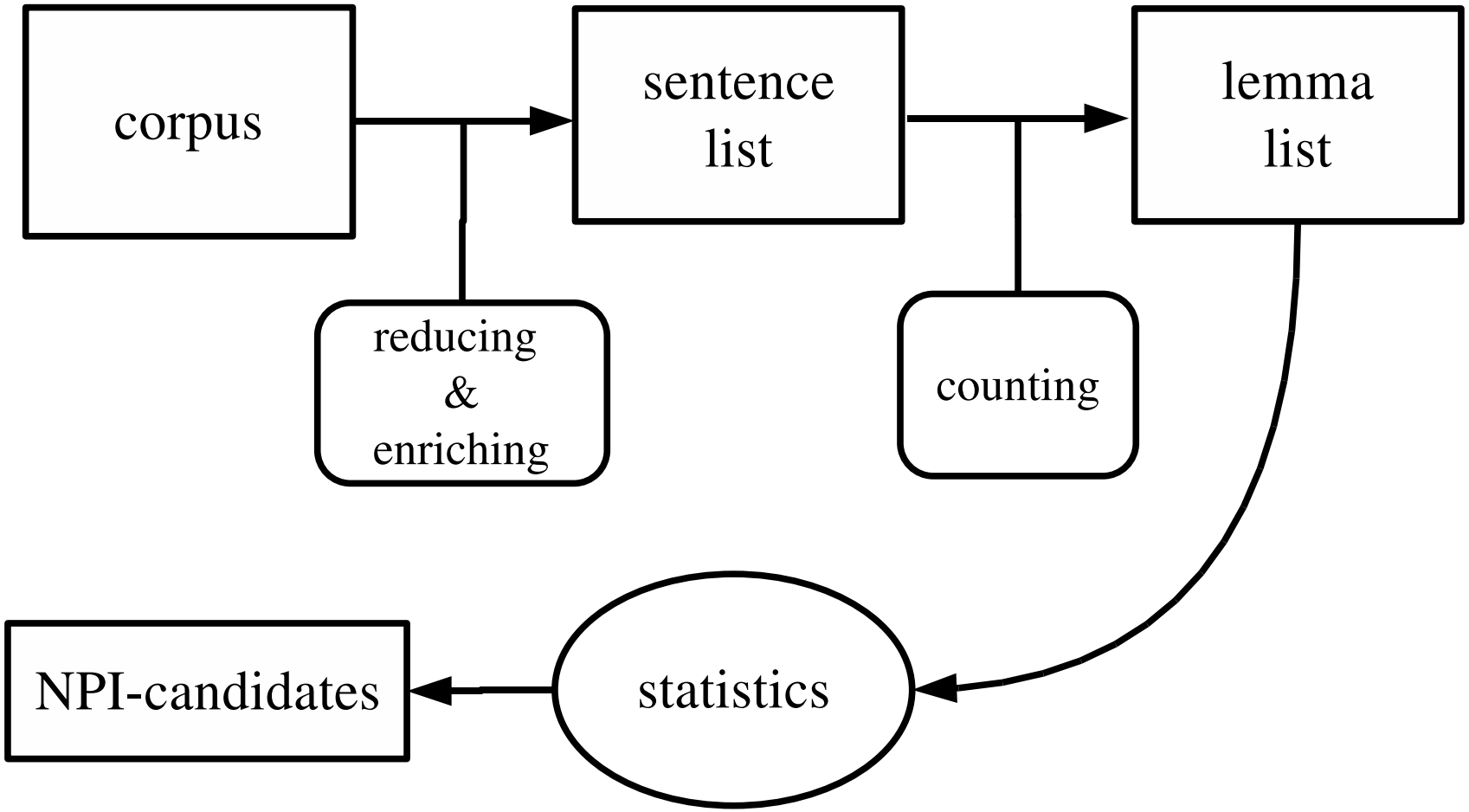
Extraction of single lemmata

Outline



- **Step 1:** Reduction and enrichment
- **Step 2:** Counting
- **Step 3:** Evaluation of frequency data

Outline



- **Step 1:** Reduction and enrichment
- **Step 2:** Counting
- **Step 3:** Evaluation of frequency data

Reduction

The original corpus contains more information than needed:

- XML format $\longrightarrow$ plain text format
- inflected words $\longrightarrow$ lemmatized words
- Clause boundaries are preserved.

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

("Can something be really good, only because it is as always?")

Enrichment

... concurrently the corpus is enriched with **negation markers (NEG)**

- **mode-1:** the licenser is replaced with a NEG marker.
- **mode-2:** if a licenser has licensing scope only over a clausal complement, then the clausal complement receives a NEG marker.

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

$\Longrightarrow$ What is the set of licensers?

Enrichment - The set of annotated licensers

The extraction mechanism will implement
(1) evident and (2) reliably identifiable licensers!

negative quantifier	*niemand, nichts, nix, kein, ...*	m-1
	nie, niemals, nirgendwo	m-1
	kaum, selten, wenig	m-1
negative conjunction	*ohne(KOUI)-zu, ohne dass*	m-1
	bevor, ob, weder	m-1
question mark	*?*	m-1
too-comparative	*zu* + adverb + *um*-clause	m-2
	zu + adverb + *um*-clause	m-2
universal quantifier	*alle/jeder* + relative clause	m-2
superlative	adjective (AJAC) + ending + rel. cl.	m-2
negative predicates	*unwahrscheinlich*	m-2
non-affirmative verbs	*bezweifeln, weigern, ablehnen, ...*	m-2
neg-raising verbs	*glauben, vorstellen können, ...*	m-2

Enrichment - Not included licensers

- **not reliably identifiable:**
  - extraposed relative clauses
  - *als*-comparative (comparative adjective + *als*-clause)
  - *so(fern) er denn jemals ....*
  - subjunctive clauses:
    Ich finde, er hätte einen Finger rühren können.
  - opaque conditionals: *You say anything, and I'll kill you.*
- **not evident:**
  - *endlich, nur*
- **other reasons:**
  - indirect questions: *sich fragen*

$\Longrightarrow$ Risk: licensing without NEG marking
CR of NPIs $\downarrow$, recall $\downarrow$

Enrichment - Caveats

- **narrow scope:**
  - NP-internal negation
  - Contrastive negation
- **not licensing negation:**
  - echo readings/predicates: *Es stimmt nicht, dass ...*
  - veridical predicates: *Peter weiss nicht, dass ...*
  - double negation
  - anti-multiplicative items: *nicht alle, nicht jeder*

$\Longrightarrow$ Risk: NEG marking without licensing
average CR $\uparrow$, precision $\downarrow$

Counting

For each lemma, two frequencies relative to the corpus are compiled:

- its overall frequency N
- its frequency in the scope of a licenser N_{lic} .

N and N_{lic} suffice to build the contingency table!

$\Longrightarrow$ What is **the scope of a licenser**?

Counting - The scope of licensers

Scope of NEG markings: the immediately embedding clause

(7) [$_A$ ... [$_{cl}$ ... **NEG** ... [$_B$ ...]] ...]

Hence, we ignore:

- immediate scope constraint
- precedence constraint for minimizer
- ...

$\Longrightarrow$ Risk of a simplification of licensing scope:
average CR $\uparrow$, precision $\downarrow$

Counting - The scope of licensers

Scope of NEG markings: the immediately embedding clause

(7) [$_A$ ... [$_{cl}$ ... **NEG** ... [$_B$ ...]] ...]

Hence, we ignore:

- immediate scope constraint
- precedence constraint for minimizer
- ...

$\Longrightarrow$ Risk of a simplification of licensing scope:
average CR $\uparrow$, precision $\downarrow$

Counting - The lemma list

Übersee	174	27
Betätigung	134	25
Hans-Peter	316	24
...	..	..

Number of lemmata: 1 058 462 (L&S,t.a.) / 641 035 (Lichte,2005)

- A ranking of lemmata is compiled based on an association measure.
- NPIs are expected to have a high ranking position!

Evaluation - Which association measure?

How to compare the rankings of CR, X^2, and G^2 quantitatively?

(Lichte,2005)

(1) precision & recall:

(2) uninterpolated average precision (uap):

Evaluation - Which association measure?

How to compare the rankings of CR, X^2, and G^2 quantitatively?

(Lichte,2005)

(1) precision & recall:

$$\textit{Precision} = \frac{\text{number of NPIs in } K}{\text{size of } K}$$

$$\textit{Recall} = \frac{\text{number of NPIs in } K}{\text{total number of NPIs}}$$

Standard precision and recall can only be computed if all true collocations have been identified in the data set.
(Villada-Moirón(2005),S.65)

(2) uninterpolated average precision (uap):

Evaluation - Which association measure?

How to compare the rankings of CR, X^2, and G^2 quantitatively?

(Lichte,2005)

(1) precision & recall:

(2) uninterpolated average precision (uap):
Given a ranked candidate list K with size n, and given a function $tp(i)$ for a rank position i with a true positive, then

$$tp(i) = \frac{\text{number of higher true positives} + 1}{i}$$

$$uap = \frac{\sum_{i=1,\ldots,n} tp(i)}{n}$$

Evaluation - Which association measure?

How to compare the rankings of CR, X^2, and G^2 quantitatively?

(Lichte,2005)

(1) precision & recall:

(2) uninterpolated average precision (uap):
Given a ranked candidate list K with size n, and given a function $tp(i)$ for a rank position i with a true positive, then

$$tp(i) = \frac{\text{number of higher true positives} + 1}{i}$$

$$uap = \frac{\sum_{i=1,\ldots,n} tp(i)}{n}$$

$\Longrightarrow$ Given a fairly large candidate list (e.g. $n = 1000$) and a list of known NPIs, **uap is more revealing**.

Evaluation - List of known NPIs and first run

List of known NPIs:

allzu, beileibe, fackeln, überhaupt, brauchen, Hehl, ausstehen, scheren, lumpen, sonderlich, verhehlen, umhin, umhinkommen, wahrhaben, Menschenseele, jemals

First run: (Lichte,2005)

CR	X^2	G^2
0.0625	0.0003644	0.0002867

First positions of the ranking with X^2 and G^2:

G^2	*mehr, können, sein, es, daß, gar, sie, die, noch, aber*
X^2	*mehr, können, gar, sein, daß, es, die, sie, noch, sondern*

Evaluation - List of known NPIs and first run

List of known NPIs:

allzu, beileibe, fackeln, überhaupt, brauchen, Hehl, ausstehen, scheren, lumpen, sonderlich, verhehlen, umhin, umhinkommen, wahrhaben, Menschenseele, jemals

First run: (Lichte,2005)

CR	X^2	G^2
0.0625	0.0003644	0.0002867

First positions of the ranking with X^2 and G^2:

G^2	*mehr, können, sein, es, daß, gar, sie, die, noch, aber*
X^2	*mehr, können, gar, sein, daß, es, die, sie, noch, sondern*

Evaluation - List of known NPIs and first run

List of known NPIs:

allzu, beileibe, fackeln, überhaupt, brauchen, Hehl, ausstehen, scheren, lumpen, sonderlich, verhehlen, umhin, umhinkommen, wahrhaben, Menschenseele, jemals

First run: (Lichte,2005)

CR	X^2	G^2
0.0625	0.0003644	0.0002867

First positions of the ranking with X^2 and G^2:

G^2	*mehr, können, sein, es, daß, gar, sie, die, noch, aber*
X^2	*mehr, können, gar, sein, daß, es, die, sie, noch, sondern*

Evaluation - Modification of X^2 and G^2

Since X^2 and G^2 seem to be biased in favour of lemmata of high frequency, they are **divided by the overall frequency** of a lemma.

CR	$X^{2}*$	$G^{2}*$
0.0625	0.0625	0.0625

Note that CR, $X^{2}*$ and $G^{2}*$ are biased in favour of low frequencies!

Evaluation - The frequency cut-off as a second variable (1)

Frequency		*CR*	$X^{2}*$	$G^{2}*$
≥ 0	(641035)	0.0625 [1]	0.0625 [1]	0.0625 [1]
≥ 5	(125505)	0.3437 [12]	0.3438 [12]	0.3437 [12]
≥ 10	(81060)	0.7734 [13]	0.7742 [13]	0.7734 [13]
≥ 20	(53475)	1.7103 [13]	1.7113 [13]	1.7103 [13]
≥ 30	(41835)	2.0777 [13]	2.0796 [13]	2.0777 [13]
≥ 40	(34957)	2.3098 [14]	2.3126 [14]	2.3098 [14]
≥ 60	(26923)	3.8596 [13]	3.8596 [13]	3.8596 [13]

(Lichte,2005)

- The AMs behave the same across frequency cut-offs: we choose CR for its simplicity.
- The higher the frequency cut-off, the higher the AM-score. Concurrently, however, the number of lemmata and known NPIs is decreasing.

$\Longrightarrow$ In (Lichte,2005), I chose CR together with ≥ 40.

Evaluation - The frequency cut-off as a second variable (1)

Frequency		*CR*	$X^{2}*$	$G^{2}*$
≥ 0	(641035)	0.0625 [1]	0.0625 [1]	0.0625 [1]
≥ 5	(125505)	0.3437 [12]	0.3438 [12]	0.3437 [12]
≥ 10	(81060)	0.7734 [13]	0.7742 [13]	0.7734 [13]
≥ 20	(53475)	1.7103 [13]	1.7113 [13]	1.7103 [13]
≥ 30	(41835)	2.0777 [13]	2.0796 [13]	2.0777 [13]
≥ 40	(34957)	2.3098 [14]	2.3126 [14]	2.3098 [14]
≥ 60	(26923)	3.8596 [13]	3.8596 [13]	3.8596 [13]

(Lichte,2005)

- The AMs behave the same across frequency cut-offs: we choose CR for its simplicity.
- The higher the frequency cut-off, the higher the AM-score. Concurrently, however, the number of lemmata and known NPIs is decreasing.

$\Longrightarrow$ In (Lichte,2005), I chose CR together with ≥ 40.

Evaluation - The frequency cut-off as a second variable (1)

Frequency		*CR*	$X^{2}*$	$G^{2}*$
≥ 0	(641035)	0.0625 [1]	0.0625 [1]	0.0625 [1]
≥ 5	(125505)	0.3437 [12]	0.3438 [12]	0.3437 [12]
≥ 10	(81060)	0.7734 [13]	0.7742 [13]	0.7734 [13]
≥ 20	(53475)	1.7103 [13]	1.7113 [13]	1.7103 [13]
≥ 30	(41835)	2.0777 [13]	2.0796 [13]	2.0777 [13]
≥ 40	(34957)	2.3098 [14]	2.3126 [14]	2.3098 [14]
≥ 60	(26923)	3.8596 [13]	3.8596 [13]	3.8596 [13]

(Lichte,2005)

- The AMs behave the same across frequency cut-offs: we choose CR for its simplicity.
- The higher the frequency cut-off, the higher the AM-score. Concurrently, however, the number of lemmata and known NPIs is decreasing.

$\Longrightarrow$ In (Lichte,2005), I chose CR together with ≥ 40.

Evaluation - The frequency cut-off as a second variable (1)

Frequency		*CR*	$X^{2}*$	$G^{2}*$
≥ 0	(641035)	0.0625 [1]	0.0625 [1]	0.0625 [1]
≥ 5	(125505)	0.3437 [12]	0.3438 [12]	0.3437 [12]
≥ 10	(81060)	0.7734 [13]	0.7742 [13]	0.7734 [13]
≥ 20	(53475)	1.7103 [13]	1.7113 [13]	1.7103 [13]
≥ 30	(41835)	2.0777 [13]	2.0796 [13]	2.0777 [13]
≥ 40	(34957)	2.3098 [14]	2.3126 [14]	2.3098 [14]
≥ 60	(26923)	3.8596 [13]	3.8596 [13]	3.8596 [13]

(Lichte,2005)

- The AMs behave the same across frequency cut-offs: we choose CR for its simplicity.
- The higher the frequency cut-off, the higher the AM-score. Concurrently, however, the number of lemmata and known NPIs is decreasing.

$\Longrightarrow$ In (Lichte,2005), I chose CR together with ≥ 40.

Evaluation - The frequency cut-off as a second variable (1)

Frequency		*CR*	$X^{2}*$	$G^{2}*$
≥ 0	(641035)	0.0625 [1]	0.0625 [1]	0.0625 [1]
≥ 5	(125505)	0.3437 [12]	0.3438 [12]	0.3437 [12]
≥ 10	(81060)	0.7734 [13]	0.7742 [13]	0.7734 [13]
≥ 20	(53475)	1.7103 [13]	1.7113 [13]	1.7103 [13]
≥ 30	(41835)	2.0777 [13]	2.0796 [13]	2.0777 [13]
≥ 40	(34957)	2.3098 [14]	2.3126 [14]	2.3098 [14]
≥ 60	(26923)	3.8596 [13]	3.8596 [13]	3.8596 [13]

(Lichte,2005)

- The AMs behave the same across frequency cut-offs: we choose CR for its simplicity.
- The higher the frequency cut-off, the higher the AM-score. Concurrently, however, the number of lemmata and known NPIs is decreasing.

$\Longrightarrow$ In (Lichte,2005), I chose CR together with ≥ 40.

Evaluation - The frequency cut-off as a second variable (2)

Rank	≥ 10	≥ 40	≥ 60
1	**verdenken**	**verdenken**	**umhin**
2	**unversucht**	**unversucht**	**nachstehen**
3	**unterschätzender***	**unterschätzender***	**lumpen**
4	entblöden	**umhin**	langgehen
5	**lockerlassen**	**nachstehen**	**verhehlen**
6	B.-Agent	**lumpen**	**beirren**
7	Bombay-Kino	langgehen	**geheuer**
8	handgekneten	**verhehlen**	unähnlich
9	**brauchen**	**beirren**	**wegdenken***
10	Bewusstseinsstadion	Genaues	**allzuviel***
11	durchwandeln	**geheuer**	**sonderlich**
12	verhohlen	unähnlich	abneigen
13	auseinanderzwingen	**wegdenken**	behagen
14	romantisch-harmlos	**allzuviel**	hinwegtäuschen
15	verstrubbeln	**sonderlich**	dagewesen
16	**umhinkommen**	abneigen	hingehören
17	Generationsstudie	behagen	draufstehen
18	Geschichten-Collage	hinwegtäuschen	**zimperlich***
19	niet	dagewesen	**missen***
20	abbleiben	hingehören	**antasten***

(Lichte,2005)

Evaluation - The final candidate list (1)

Rank	*Lemma*	*CR*
1	**verdenken**	1.00
2	**unversucht**	1.00
3	**unterschätzender***	1.00
4	**umhin**	0.99
5	**nachstehen**	0.99
6	**lumpen**	0.99
7	langgehen	0.99
8	**verhehlen**	0.97
9	**beirren**	0.97
10	Genaues	0.97
11	**geheuer**	0.96
12	unähnlich	0.96
13	**wegdenken***	0.95
14	**allzuviel***	0.93
15	**sonderlich**	0.92
16	abneigen	0.91
17	behagen	0.90
18	hinwegtäuschen	0.89
19	dagewesen	0.89
20	hingehören	0.88

Rank	*Lemma*	*CR*
21	Schöneres	0.88
22	draufstehen	0.87
23	**zimperlich***	0.85
24	**missen***	0.84
25	**antasten***	0.84
26	fruchten	0.83
27	**jedermanns***	0.83
28	hinauskommen	0.82
29	Ungewöhnliches	0.82
30	anbelangen	0.81
31	anbetreffen	0.80
32	**wahrhaben**	0.79
33	gar	0.79
34	**hinnehmbar***	0.79
35	nützen	0.78
36	**anhaben***	0.78
37	**Hehl**	0.78
38	durchsetzbar	0.77
39	Seltenheit	0.76
40	einwenden	0.75

(Lichte,2005)

Evaluation - The final candidate list (2)

Groups of candidates:

- **non-polysemous candidates:**
- **polysemous candidates:**
- **pseudo-NPIs:** (Hoeksema,1997)
- **statistical noise:**

Evaluation - The final candidate list (2)

Groups of candidates:

- **non-polysemous candidates:**
  sonderlich (particularly), nachstehen (to rank behind), Hehl (secret)

  (8) Die Spieler wollten dem natürlich in **nichts** nachstehen, weshalb am Sonntag abend mit offensichtlich großem Einsatz rustikal um jeden Ball gekämpft wurde. [taz, 03.03.1998]

  (9) Die Brand New Heavies machten **nie** einen Hehl daraus, daß im Soul auch Hits im herkömmlichen Sinne geschrieben werden könnten. [taz, 16.04.1998]

- **polysemous candidates:**
- **pseudo-NPIs:** (Hoeksema,1997)
- **statistical noise:**

Evaluation - The final candidate list (2)

Groups of candidates:

- **non-polysemous candidates:**
- **polysemous candidates:**
  wegdenken (ignore), brauchen (to need)

  (8) Sentimentales und Kitsch sind aus der privaten Geschichtsschreibung **kaum** wegzudenken. [taz, 09.04.1998]

  (9) Hedge-Fonds brauchen [...] **keinerlei** Mindestvorschriften zu beachten und sind keiner Aufsichtsbehörde Rechenschaft schuldig. [taz, 13.11.1998]

- **pseudo-NPIs:** (Hoeksema,1997)
- **statistical noise:**

Evaluation - The final candidate list (2)

Groups of candidates:

- **non-polysemous candidates:**
- **polysemous candidates:**
- **pseudo-NPIs:** (Hoeksema,1997)
  unähnlich (unalike), hinwegtäuschen (mislead)

  (8) Doch selbst gute Bilanzen können **nicht**
  darüber hinwegtäuschen, daß der Großteil deutscher
  Studenten Identifikationsprobleme mit seiner Uni hat.
  [taz, 03.04.1998]

- **statistical noise:**

Evaluation - The final candidate list (2)

Groups of candidates:

- **non-polysemous candidates:**
- **polysemous candidates:**
- **pseudo-NPIs:** (Hoeksema,1997)
- **statistical noise:**
  Genaues, Schöneres, langgehen
  The linguist cannot think of a correlating NPI ...

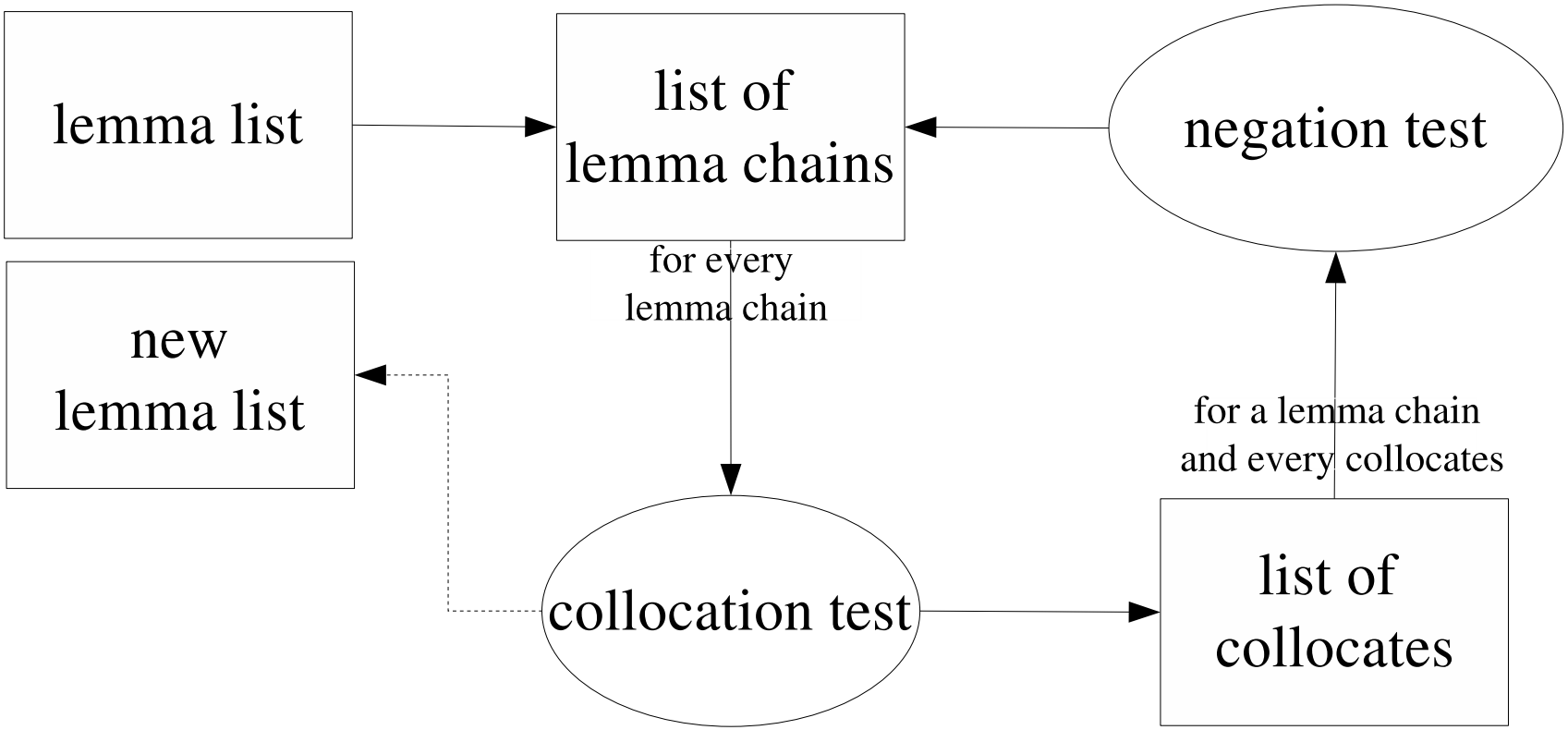
Part IV

Extraction of lemma chains

Reasons for extracting lemma chains

- completion of multiword NPIs
- disambiguation of polysmous NPIs
- detection of multiword NPIs consisting of inconspicuous parts
- statistical noise or hidden NPI?

Outline



Version of (L&S,t.a.):

- size of lemma list: ca. 65 000 entries
- frequency cut-off: ≥ 30

Collocation test and negation test

Collocation test:

- Identify lemmata that significantly co-occur!
- bi-grams with varying components
  $\rightsquigarrow G^2 \geq 250$
- frequency threshold for co-occurrence: ≥ 10
- span of co-occurrence: the immediate clause as annotated

Negation test:

- For a lemma (chain) *lc* and every of its collocates *col*:
  CR of *lc* $\leq$ CR of *lc* + *col* ?

Collocation test and negation test

Collocation test:

- Identify lemmata that significantly co-occur!
- bi-grams with varying components
  $\rightsquigarrow G^2 \geq 250$
- frequency threshold for co-occurrence: ≥ 10
- span of co-occurrence: the immediate clause as annotated

Negation test:

- For a lemma (chain) *lc* and every of its collocates *col*:
  CR of *lc* $\leq$ CR of *lc* + *col* ?

The new candidate list

erreichen Stellungnahme zu für gestern	37	37
Ofen locken hervor	20	20
müde betonen zu	55	55
weniger binnen Woche	47	47
Bericht dementieren bestätigen	26	26
reißen Serie ab	33	33
fantastisch unrealistisch wirken besonders	21	21
ändern Tatsache an	55	55
verstehen Welt mehr die	158	158
weniger binnen Tag	67	67
Rede mehr sein davon	127	127
versuchen erst gar	143	143
anhalten Grenzpolizei Mexico New	19	19
notwendigerweise Meinung Seite erscheinenend geben wieder auf die	142	142
Kinoheld fahren Tag sehen Herz heiter Himmel zum	33	33
unversucht lassen	89	89
unterschätzend ein zu	82	82
ansatzweise einmal	50	50

The new candidate list - Completion

Completed non-polysemous NPIs:

(8) `unversucht lassen` (#16, 1.00)
etw. unversucht lassen (to leave sth. undone)

(9) `daraus Hehl machen` (#449, 0.79)
einen Hehl machen aus etw. (to make a secret of sth.)

(10) `scheren Sie um` (#646, 0.66)
sich [nicht] scheren um etw. (to give a damn about sth.)

The new candidate list - Disambiguation

Disambiguated polysemous NPIs:

(11) `wegdenken aus sein` (#143, 0.97)
aus etw. wegzudenken sein (sth. without sth. is [not] to be thought of)

(12) a. `scheuen Vergleich brauchen` (#35, 1.00)
b. `Sorge Sie brauchen machen` (#92, 0.99)
c. `wundern brauchen zu` (#97, 0.99)
d. `Angst haben brauchen zu` (#159, 0.96)
e. `fürchten zu brauchen` (#204, 0.94)

$\rightsquigarrow$ *brauchen* + infinitival complement

The new candidate list - Detection

Detected multiword NPIs:

(13) `Staunen` (#12678)
`Staunen heraus` (#55, 1.00)
aus dem Staunen herauskommen (to stop wondering)

(14) `Tasse` (#15221)
`Tasse Schrank` (#433, 0.8)
alle Tassen im Schrank haben (to have lost one's marbles)

(15) `reichen` (#690)
`reichen hinten vorne` (#2, 1.00)
hinten und vorne reichen (to be adequate)

The new candidate list - Pseudo-NPIs

Completed/detected pseudo-NPIs:

(16) `hinwegtäuschen können darüber` (#156, 0.96)
über etw. hinwegtäuschen können
(to be able to hide the fact that . . .)

(17) a. `hinter Ofen hervor locken`
hinter dem Ofen hervorlocken (to get someone excited about sth.)

b. `entbehren gewiß Komik`
einer gewissen Komik entbehren (to be lacking in humor)

c. `Redaktionsschluss fest stehen noch bei`
bei Redaktionsschluss feststehen (be available at press date)

The new candidate list - Revising noise

Is it really noise?

(18) `Genaues`
`Genaues wissen` (#174, 0.96)
etwas Genaues wissen (to know sth. in detail)

The new candidate list - Garbage

Garbage due to recurring strings in the newspaper:

(19) `notwendigerweise Meinung Seite erscheinend geben wieder auf die`

Die auf dieser Seite erscheinenden Leserbriefe geben nicht notwendigerweise die Meinung der taz wieder.
("The reader's letters on this page don't necessarily reflect the opinion of the taz.")

$\rightsquigarrow$ Easily identifiable due to noticeable length and CR of 1!
$\rightsquigarrow$ ca. 10 such lemma chains

The new candidate list - Size of shortlist?

- The shortlist should contain **statistically significant** lemma chains.
- significance statements by using **z-scores**:

$$z = \frac{O - \mu_{CR}}{sd_{CR}}$$

  confidence level: $z = 2.58$ ($p \leq 0.01$)
- According to their z-score, 2000 lemma chains show statistically significant CRs.
- Recall wrt. Kürschners collection: 112 of 344 items.

The new candidate list - Size of shortlist?

- The shortlist should contain **statistically significant** lemma chains.
- significance statements by using **z-scores**:

$$z = \frac{O - \mu_{CR}}{sd_{CR}}$$

confidence level: $z = 2.58$ $(p \leq 0.01)$

- According to their z-score, 2000 lemma chains show statistically significant CRs.
- Recall wrt. Kürschners collection: 112 of 344 items.

The new candidate list - Size of shortlist?

- The shortlist should contain **statistically significant** lemma chains.
- significance statements by using **z-scores**:

$$z = \frac{O - \mu_{CR}}{sd_{CR}}$$

  confidence level: $z = 2.58$ ($p \leq 0.01$)
- According to their z-score, 2000 lemma chains show statistically significant CRs.
- Recall wrt. Kürschners collection: 112 of 344 items.

The new candidate list - Size of shortlist?

- The shortlist should contain **statistically significant** lemma chains.
- significance statements by using **z-scores**:

$$z = \frac{O - \mu_{CR}}{sd_{CR}}$$

  confidence level: $z = 2.58$ ($p \leq 0.01$)
- According to their z-score, 2000 lemma chains show statistically significant CRs.
- Recall wrt. Kürschners collection: 112 of 344 items.

Part V

Quantitative classification of NPIs

Outline

- input: NPIs
- refinement of the distributional patterns according to **the grades of negativity**:

AM	AA	DE	DEINT
anti-morphic	anti-additive	downward-entailing	DE + interrogatives
classic	regular	minimal	
nicht, keinesfalls	*niemand, kein*	*wenige, kaum*	questions

- lemma list with additional columns:

Lemma chain	**Total**	**Classic**	**Regular**	**Minimal**
sonderlich	878	782	92	4
brauchen VVIZU	2359	1660	625	74
jemals	1077	314	202	561
Tasse Schrank	28	10	2	16
jedermanns Sache	66	64	0	2
Menschenseele	28	4	22	2
sonst ja gönnen	27	0	27	0

Deviance ratio

(20) **Deviance ratio (DR):** Given a subclass of negative contexts sc, the observed frequency O_{sc} and expected frequency E_{sc} of an NPI in sc, its total frequency in negative contexts N_{neg}, and furthermore the fraction of sc with respect to the overall frequency of negative contexts R_{sc}, then we calculate:

$$DR = \frac{O_{sc} - E_{sc}}{N_{neg}} \qquad E_{sc} = R_{sc} * N_{neg}$$

$p > 0.05$	(not significant, $z <$ -/+1.96)	$\longrightarrow$ o
$p \leq 0.05$	(significant, $z \geq$ -/+1.96)	$\longrightarrow$ $-/+$
$p \leq 0.01$	(significant, $z \geq$ -/+2.58)	$\longrightarrow$ $--/++$

Deviance ratio

(20) **Deviance ratio (DR):** Given a subclass of negative contexts sc, the observed frequency O_{sc} and expected frequency E_{sc} of an NPI in sc, its total frequency in negative contexts N_{neg}, and furthermore the fraction of sc with respect to the overall frequency of negative contexts R_{sc}, then we calculate:

$$DR = \frac{O_{sc} - E_{sc}}{N_{neg}} \qquad E_{sc} = R_{sc} * N_{neg}$$

$p > 0.05$	(not significant, $z <$ -/+1.96)	$\longrightarrow$	o
$p \leq 0.05$	(significant, $z \geq$ -/+1.96)	$\longrightarrow$	$-/+$
$p \leq 0.01$	(significant, $z \geq$ -/+2.58)	$\longrightarrow$	$--/++$

Results

NPI	*Negation*		
	classic	regular	minimal
weak	x	x	x
strong	x	x	–
superstrong	x	–	–

Lemma chain	**Classic**	**Regular**	**Minimal**
sonderlich (878)	++ (782)	o (92)	−− (4)
brauchen VVIZU (2359)	o (1660)	o (625)	– (74)
jemals (1077)	– (314)	o (202)	+ (561)
Tasse Schrank (28)	o (10)	o (2)	++ (16)
jedermanns Sache (66)	++ (64)	– (0)	– (2)
Menschenseele (28)	−− (4)	++ (22)	– (2)
sonst ja gönnen (27)	−− (0)	++ (27)	−− (0)