Negative Polarity Items
Corpus Linguistics, Semantics, and Psycholinguistics

Frank Richter, Janina Radó, Manfred Sailer

Universität Tübingen and Universität Göttingen

ESSLLI 2008, Hamburg

Quantitative Methods in NPI Research

Timm Lichte: The Retrieval and Classification of NPIs
Lichte (2005a, 2005b), Lichte & Soehn (2007)

The idea and its extensions:

- Use partially parsed corpora and statistical analysis to extract NPI candidates from corpora. (note: no validation!)
- Extend the basic method from the extraction of single-word expressions to multiword expressions.
- Extend the method to the classification of NPIs according to their licensing requirements (superstrong, strong, weak).

Extraction of NPI Candidates: TüPP-D/Z

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

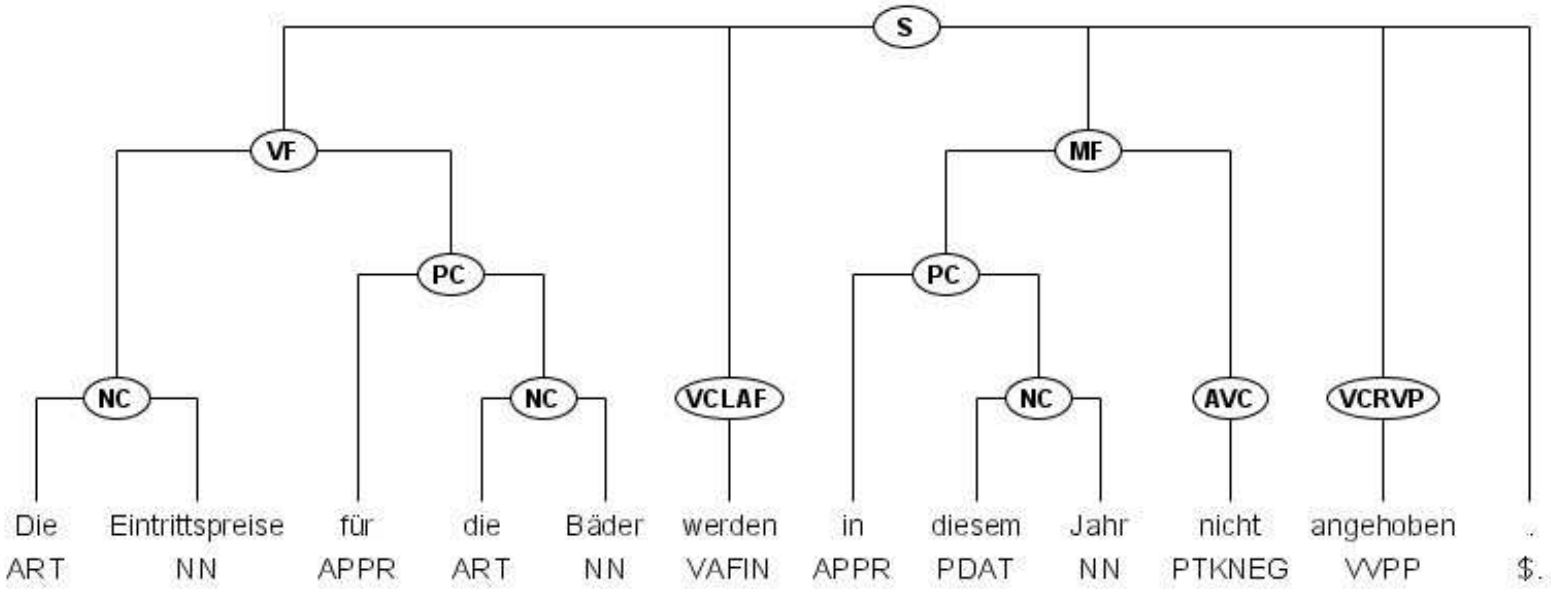
Why use a partially parsed corpus?

Extraction of NPI Candidates: TüPP-D/Z

- lemmatization
- part-of-speech tags
- chunks
- clause boundaries
- topological fields

$\Longrightarrow$ Except topological fields, all of this information can be used.

Extraction of NPI Candidates: TüPP-D/Z



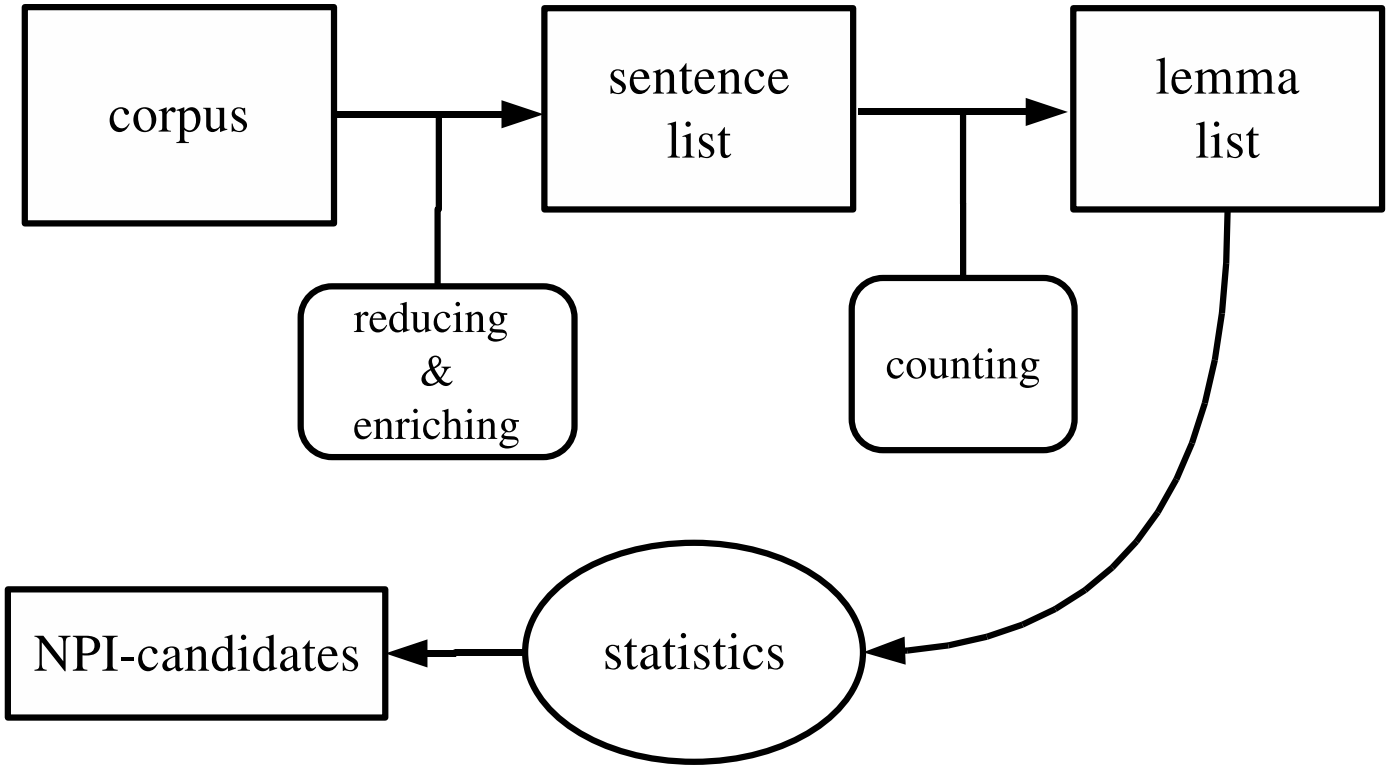
Extraction of NPI Candidates: TüPP-D/Z

Lichte applies a filter for defective clause annotation:

(1) <**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 are discarded (Lichte 2005).
$\Longrightarrow$ The statistics become more expressive.

Extraction of NPI Candidates: The Procedure



- **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 $\rightarrow$ plain text format
- inflected words $\rightarrow$ 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 has always been the case?”)

Enrichment 1

At the same time, 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**

Question: What is the set of licensers?

Enrichment 2

The extraction mechanism considers
(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

The Result: A List of NPI Candidates

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

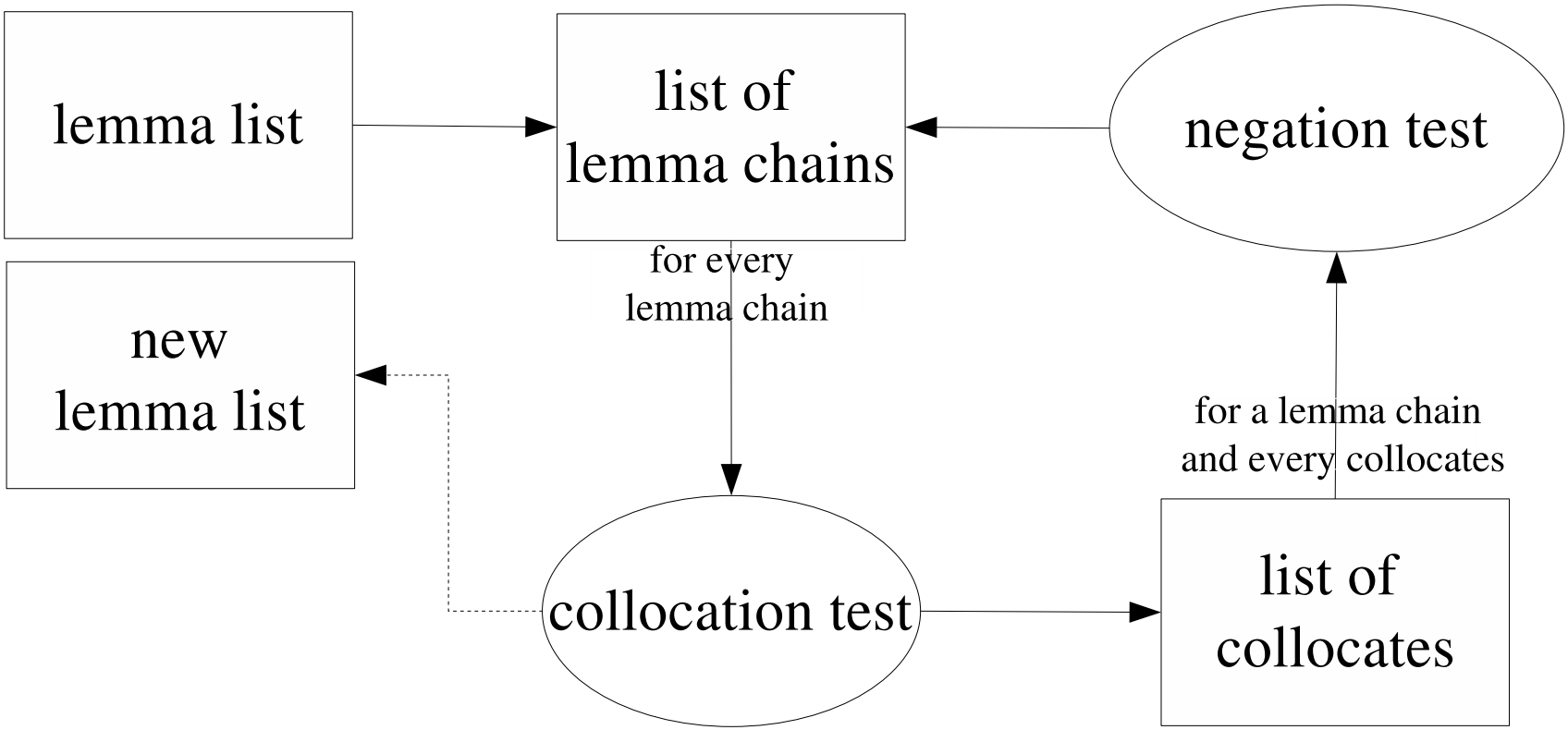
(Lichte 2005)

Extension 1: Extraction of Multiword NPIs

Lichte's first extension of the basic NPI extraction mechanism aims at

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

Extension 1: Schematic Overview



(Lichte & Soehn 2007):

- size of lemma list: about 65 000 entries
- frequency cut-off: $\geq$ 30

Extension 2: Classification of NPIs

- Input: NPIs
- Refinement of the distributional patterns according to **3 degrees of negativity**:

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

Extension 2: Classification of NPIs: Results

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

Lemma chain	**Classical**	**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)

Association Measures

Mutual Information (MI) / Context Ratio (CR)

$$CR(w) = \frac{w\&NEG}{w}$$

$\Longrightarrow$ ranking of candidates

Association Measures

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

Contingency tables:

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

Association Measures

X^2-test:

$$X^2 = \Sigma_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

where $O_{i,j}$ is the cell in row i and column j of the contingency table, and $E_{i,j}$ is the expected value for $O_{i,j}$, and

$$E_{i,j} = \frac{O_{i+} * O_{+j}}{N}$$

Association Measures

Log-likelihood Quotient (G^2)

$$G^2 = 2\Sigma_{i,j} O_{i,j} log_2 \frac{O_{i,j}}{E_{i,j}}$$

Evaluation - Which Association Measure?

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, p. 65)

Evaluation - Which Association Measure?

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

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

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

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* ?

Acknowledgments

Most of the slides in this section go back to Timm Lichte's slide presentation of his work. Timm Lichte's original slides are available from

`www.sfs.uni-tuebingen.de/~fr/teaching/ss07/npi/presentations/timms-folien.pdf`

Thanks are due to him for sharing the latex sources of his slides with us.