Pattern-based Extraction of Negative Polarity Items from Dependency-parsed Text

Fabienne Fritzinger[1], Frank Richter[2], Marion Weller[1]

[1]Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart

[2]Universität Tübingen
Seminar für Sprachwissenschaft
Abteilung Computerlinguistik
Wilhelmstr. 19
D 72074 Tübingen

Background

Largest collection of German NPIs:
Subcollection CoDII-NPI.DE in the **CO**LLECTION OF **D**ISTRIBUTIONALLY **I**DIOSYNCRATIC **I**TEMS:

CoDII-NPI.DE: 165 entries (Trawiński et al, 2008)

Previous extraction methods (Lichte and Soehn, 2007)

- primarily targets single-word NPIs
- findings included in CoDII-NPI.DE

Negative Polarity Items

(1) a. Kim never saw any student in the hallway.
b. *Kim saw any student in the hallway.

Negative Polarity Items

(1) a. Kim never saw any student in the hallway.
b. *Kim saw any student in the hallway.

(2) a. Few lecturers saw any student in the hallway.
b. *Some lecturers saw any student in the hallway.

Negative Polarity Items

(1) a. Kim never saw any student in the hallway.
b. *Kim saw any student in the hallway.

(2) a. Few lecturers saw any student in the hallway.
b. *Some lecturers saw any student in the hallway.

(3) a. Nobody had the slightest inkling about where to go.
b. *Few visitors had the slightest inkling about where to go.

Negative Polarity Items

(1) a. Kim never saw any student in the hallway.
b. *Kim saw any student in the hallway.

(2) a. Few lecturers saw any student in the hallway.
b. *Some lecturers saw any student in the hallway.

(3) a. Nobody had the slightest inkling about where to go.
b. *Few visitors had the slightest inkling about where to go.

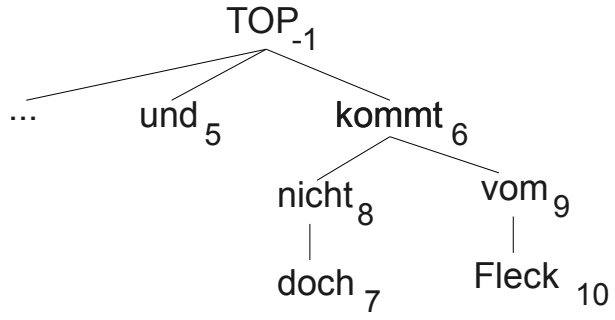
(4) a. Sie erweisen sich keinen Deut besser als ihre Vorgänger.
(*They are not one whit better than their predecessors.*)
b. *Sie erweisen sich einen Deut besser als ihre Vorgänger.

Selection of German NPIs

- **Adverbs**: jemals (*ever*), beileibe (*by no means*)
- **Adjectives**: geheuer (*mysterious/scary*), gefeit (*immune*)
- **Nouns**: Deut (*farthing*), Sterbenswörtchen (*dying word*)
- **Verbs**:
  ausstehen können (*can stand*),
  wahrhaben wollen (*want to see the truth*)
- **Idiomatic expressions**:
  alle Tassen im Schrank haben (*to play with a full deck*),
  einen Finger rühren (*to lift a finger*),
  ein Hehl aus etw. machen (*to conceal sth.*)

Extraction of candidate PNV-triples

FSPAR (Schiehlen, 2003)

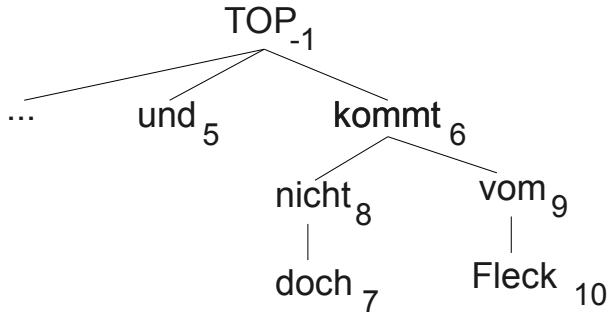


	word form	pos tag	lemma	morph. features	gover-nor	gramm. function	*engl.*
5	und	KON	und	|	-1	KON	and
6	kommt	VVFIN	kommen	|	-1	TOP	comes
7	doch	ADV	doch	|	8	ADJ	yet
8	nicht	PTKNEG	nicht	|	6	ADJ	not
9	vom	APPRART	von	Dat:M:Sg	6	PP	from the
10	Fleck	NN	Fleck	Dat:M:Sg	9	PCMP	spot

and yet, he is not moving forward.

Extraction of candidate PNV-triples

FSPAR (Schiehlen, 2003)



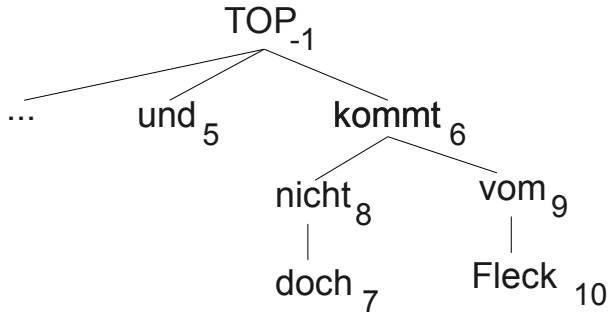
	word form	pos tag	lemma	morph. features	gover-nor	gramm. function	*engl.*
5	und	KON	und	|	-1	KON	and
6	kommt	VVFIN	kommen	|	-1	TOP	comes
7	doch	ADV	doch	|	8	ADJ	yet
8	nicht	PTKNEG	nicht	|	6	ADJ	not
9	vom	APPRART	von	Dat:M:Sg	6	PP	from the
10	Fleck	NN	Fleck	Dat:M:Sg	9	PCMP	spot

and yet, he is not moving forward.

kommen

Extraction of candidate PNV-triples

FSPAR (Schiehlen, 2003)



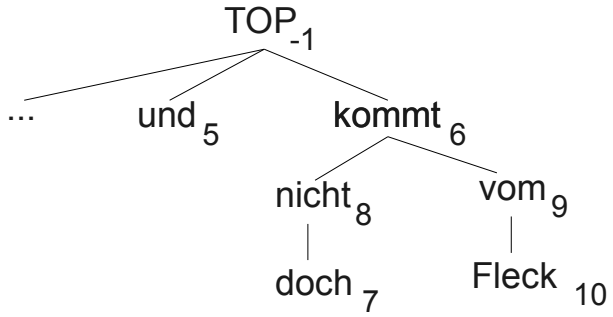
	word form	pos tag	lemma	morph. features	gover-nor	gramm. function	*engl.*
5	und	KON	und	|	-1	KON	and
6	kommt	VVFIN	kommen	|	-1	TOP	comes
7	doch	ADV	doch	|	8	ADJ	yet
8	nicht	PTKNEG	nicht	|	6	ADJ	not
9	vom	APPRART	von	Dat:M:Sg	6	PP	from the
10	Fleck	NN	Fleck	Dat:M:Sg	9	PCMP	spot

and yet, he is not moving forward.

von kommen

Extraction of candidate PNV-triples

FSPAR (Schiehlen, 2003)



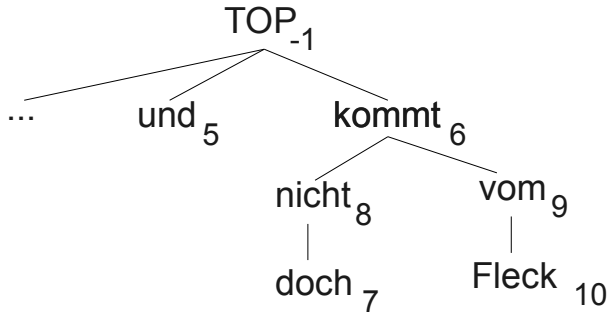
	word form	pos tag	lemma	morph. features	gover-nor	gramm. function	*engl.*
5	und	KON	und	|	-1	KON	and
6	kommt	VVFIN	kommen	|	-1	TOP	comes
7	doch	ADV	doch	|	8	ADJ	yet
8	nicht	PTKNEG	nicht	|	6	ADJ	not
9	vom	APPRART	von	Dat:M:Sg	6	PP	from the
10	Fleck	NN	Fleck	Dat:M:Sg	9	PCMP	spot

and yet, he is not moving forward.

von Fleck kommen

Extraction of candidate PNV-triples

FSPAR (Schiehlen, 2003)



	word form	pos tag	lemma	morph. features	gover-nor	gramm. function	*engl.*
5	und	KON	und	|	-1	KON	and
6	kommt	VVFIN	kommen	|	-1	TOP	comes
7	doch	ADV	doch	|	8	ADJ	yet
8	nicht	PTKNEG	nicht	|	6	ADJ	not
9	vom	APPRART	von	Dat:M:Sg	6	PP	from the
10	Fleck	NN	Fleck	Dat:M:Sg	9	PCMP	spot

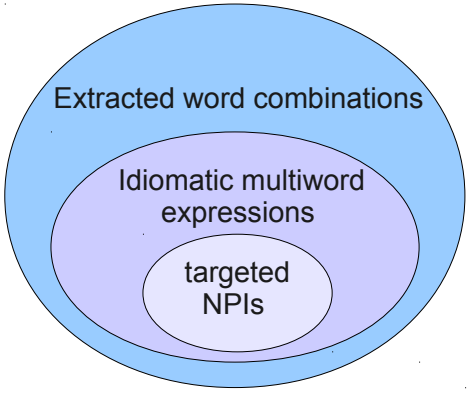
and yet, he is not moving forward.

nicht von Fleck kommen

Data

Corpus collection: 269 million tokens
(Europarl: 35 mio, newspaper text: 234 mio).

pattern	NPI
NV	Hehl machen
ANV	blass Schimmer haben
PNV	über Herz bringen
NPNV	Blatt vor Mund nehmen
PANV	mit recht Ding zugehen



	NV	ANV	PNV	NPNV	PANV
types	2 069 393	1 143 104	6 337 849	3 033 148	2 475 122
tokens	5 194 941	1 442 865	11 420 865	3 388 758	2 906 645

Modelling Negative Contexts

strong negation *with verbs:* nicht, *with nouns:* kein

Modelling Negative Contexts

strong negation	*with verbs:* nicht, *with nouns:* kein
nouns	niemand (*nobody*), nichts (*nothing*)

Modelling Negative Contexts

strong negation	*with verbs:* nicht, *with nouns:* kein
nouns	niemand (*nobody*), nichts (*nothing*)
adverbs	kaum (*hardly*), nirgendwo (*nowhere*) keinesfalls (*by no means*), wenig (*few*)

Modelling Negative Contexts

strong negation	*with verbs:* nicht, *with nouns:* kein
nouns	niemand (*nobody*), nichts (*nothing*)
adverbs	kaum (*hardly*), nirgendwo (*nowhere*) keinesfalls (*by no means*), wenig (*few*)
inherently negative verbs	bezweifeln (*to doubt*), verhindern (*to impede*) weigern (*to refuse*), abstreiten (*to deny*)

Statistical Processing

NPI candidate	contexts	
aus Kopf gehen	NEG: 47	NONEG: 0
Wald vor Baum sehen	NEG: 46	NONEG: 4
zu Schaden kommen	NEG: 246	NONEG: 199

Statistical Processing

NPI candidate	contexts	
aus Kopf gehen	NEG: 47	NONEG: 0
Wald vor Baum sehen	NEG: 46	NONEG: 4
zu Schaden kommen	NEG: 246	NONEG: 199

Statistical association measures

Assuming that NPIs are highly associated with their negative context, we compute association scores for pairs of MWEs and their respective context label (NEG or NONEG).

NPI candidate	context label	log-likelihood score
Hehl machen	NEG	4.3197549e+03
Wahl annehmen	NEG	2.3230087e+01

Candidates are then sorted according to their association scores.

The 500 highest scoring candidates of each pattern with a NEG-label were manually annotated.

Results

	NPNV-pattern with negative context (NEG)	fr.	position POIS	LL	f
+	**Blatt vor Mund nehmen**	139	1	1	50
-	Angabe über Höhe machen	78	2	2	160
-	Richtlinie in Recht umsetzen	61	3	3	262
-	Ziel aus Auge verlieren	116	4	4	76
+	**Wald vor Baum sehen**	50	5	7	367
-	Angabe über Kaufpreis machen	42	6	6	466
(+)	**Hehl** aus Sympathie **machen**	38	7	8	561
(+)	**Hehl** aus Enttäuschung **machen**	37	8	9	594
-	Arbeit für Stunde niederlegen	37	9	11	573
(+)	Gefahr **von Hand weisen**	36	10	10	736
-	Stein in Weg legen	84	11	13	142
-	Zugang zu Trinkwasser haben	29	12	12	896
-	Änderungsantrag aus Grund akzeptieren	36	13	16	612
+	**Mördergrube aus Herz machen**	28	14	14	814
(+)	**Hehl** aus Abneigung **machen**	28	15	17	868

Linguistic Processing

Hypothesis: Idiomatic multiword expressions are morpho-syntactically fixed.

Linguistically motivated scores, monolingual and multilingual:

#NEG	the percentage of negative contexts
FIX	degree of morpho-syntactic fixedness
TE	degree of diversity when translated
PDA	percentage of trivial translations

the FIX-score is based on the percentage of *number* and *article* setting and the #NEG-score.

Linguistic Processing

Hypothesis: Idiomatic multiword expressions are morpho-syntactically fixed.

Linguistically motivated scores, monolingual and multilingual:

#NEG	the percentage of negative contexts
FIX	degree of morpho-syntactic fixedness
TE	degree of diversity when translated
PDA	percentage of trivial translations

the FIX-score is based on the percentage of *number* and *article* setting and the #NEG-score.

Problem:

NPIs tend to be lowfrequent, but the TE, PDA and FIX-scores work better with high-frequent data.

Conclusions and Future Work

Our results

- retrieved 142 NPIs (all patterns), out of which
  - 28 are in CoDII, and
  - 114 are new !

NPIs in top 500	NV	ANV	PNV	NPNV	PANV
poisson	29	77	31	5	4

Future work

- more fine-grained distinction of negative context in extraction
- more data: benefits for both the statistical and linguistic processing
- psycholinguistic experiments on NPIs (Radó, Richter, Sailer)

References

- UCS-toolkit: www.collocations.de
- Stefan Evert, 2004: *The statistics of word cooccurrences: word pairs and collocations.* Phd thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart
- Michael Schiehlen, 2003: *A cascaded finite-state parser for German.* In *Proceedings of the 10th* EACL, Budapest, Hungary
- Timm Lichte and Jan-Philipp Soehn, 2007: *The Retrieval and Classification of Negative Polarity Items using Statistical Profiles.* In Sam Featherston and Wolfgang Sternefeld, editors, *Roots; Linguistics in Search of its Evidential Base*, pages 249-266. Mouton de Gruyter, Berlin
- Beata Trawiński, Jan-Philipp Soehn, Manfred Sailer, Frank Richter, 2008: *A Multilingual Electronic Database of Distributionally Idiosyncratic Items.* In Elisenda Bernal and Janet DeCesaris, editors, *proceedings of the* XIII *Euralex International Congress*, Barcelona, Spain