

A Collocational Theory of Polarity Item Licensing

Manfred Sailer
 Seminar für Englische Philologie
 Universität Göttingen
 manfred.sailer@phil.uni-goettingen.de

Polarity Meets Psycholinguistics
 University of Potsdam
 December 8, 2005

1 Introduction

- Based on joint work with Timm Lichte, Frank Richter Jan-Philipp Soehn and Beata Trawiński, carried out in the project *Distributional Idiosyncrasies* of the SFB 441 *Linguistic Data Structures*, Tübingen and at the Seminar für Englische Philologie, Göttingen.
- Corpora as a source for empirical research on polarity items.
- Overview
 1. Introduction
 2. Study 1: Corpus data on some NPIs with additional collocational restrictions
 3. Study 2: Extraction of lemmata with high occurrence ratios in NPI licensing contexts
 4. Collocational module
 5. Relation to other models
 6. Conclusion

1.1 The Collocational Perspective: van der Wouden

What are collocations?

“Collocation is the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening.”
 (Sinclair, 1991, p.170)

“Collocation has long been the name given to the relationship a lexical item has with items that appear with greater than random probability in its (textual) context.” (Hoey, 1991, p.7)

“The defining feature of a lexical item, by which such an item is recognized, is its pattern of co-occurrence with other items, that is its *collocational* behaviour. A lexical item is recognized as different from other lexical items because its total pattern of collocation is unique.”
 (Butler, 1985, p.130)

van der Wouden (1992, 1997) generalizes from the notion of collocations from a lexeme-lexeme relation to more abstract patterns:

“Collocation: idiosyncratic restriction on the combinability of lexical items.” (van der Wouden, 1997, p.5)

	<i>nicht</i> (not)		<i>nicht mehr</i> (no more/longer)		<i>ohne</i> (without)		<i>nie, etc</i> (n-word)		no Neg		Σ	
	#	%	#	%	#	%	#	%	#	%	#	%
<i>lange</i>	546	84.129	11	1.694	47	7.242	12	1.848	3	0.462	619	95.378
other mod	1	0.154	-	0	-	0	2	0.308	-	0	3	0.308
<i>viel</i>	-	0	-	0	1	0.154	-	0	-	0	1	0.154
no mod	16	2.465	1	0.154	5	0.770	1	0.154	4	0.616	27	4.160
Σ	563	86.661	12	1.849	53	8.166	15	2.310	7	1.077	650	100

Table 1: Distribution of *fackeln* in the publically available corpora of the IDS (March 2003, from Sailer (2004))

“If all restrictions on the distribution of lexical items are taken to be collocational, uniform listing of collocational combinations in the lexicon is impossible: certain open classes of lexical items (e.g. negative polarity items) are restricted to the neighbourhoods of member of another open class (viz. monotone decreasing operators).” (van der Wouden, 1997, p.50)

1.2 Primary and Secondary Licensing

Hoeksema (2001) discusses Polarity Items with further restrictions on their possible occurrence contexts.

- (1) a. I wouldn’t talk to him for love or money.
 b. John did not date her for love or money, *(but at the behest of his father.)

Hoeksema distinguishes:

- *primary licensing*: a negative (non-negative) context is required.
- *secondary licensing*: there is a further context restriction. (in (1): nonveridical/irrealis context)

2 Study 1: Negation Sensitivity among Other Abstract Collocational Restrictions

2.1 Data on *fackeln*

The German verb *fackeln* has two meanings: (i), *to burn* occurring mainly with particles such as *ab-* (*flare off*); (ii), *to dither, to shilly-shally, to dilly-dally*. We will focus on this second meaning. The corpus data stem from the publically available corpora of the *Institut für Deutsche Sprache* (IDS), Mannheim (www.ids-mannheim.de/~cosmas),

- (2) *lange + nicht* (84.1%)

Der neue Geschäftsführer fackelte nicht lange.
 the new manager dithered not long ‘The new manager didn’t dither very long.’

2.1.1 Corpus Data on *fackeln*

- (3) a. *lange + ohne* (7.2%):

Ohne lange zu fackeln, rannte der Helfer in die Wohnung.
 without long to dither ran the helper into the apartment
 ‘The helper ran into the apartment without dithering very long.’

- b. *lange + n-word* (1.8%):

Beim Kampf gegen die Zeit durfte keiner lange fackeln.
 in the fight against the time was allowed no-one long dither
 ‘In the fight against time no-one was allowed to dither very long.’

- c. other modifier + n-word (0.3%):
 Es gebe zwei Punkte, in denen die Politesse keine Minute fackle,
 there are two points at which a traffic warden no minute dithers
 ‘There were two points at which a traffic warden wouldn’t dither a minute.’
- d. *lange* + no negation (0.5%)
 Harsche SVP-Kritik erntete die Regierung dafür, dass sie so lange fackle
 harsh SVP criticism reaped the government that for that it so long dithered
 ‘The government reaped harsh SVP criticism for dithering so long.’
- e. no modifier + *ohne* (0.8%):
 Ohne zu fackeln, zückte einer von ihnen ein Messer.
 without to dither drew one of them a knife
 ‘One of them drew a knife without dithering.’

The seven cases without an overt negation still fall into the class of NPI licensing environments:

- (4) 3 cases with *lange*, but no negative item:
- a. Question:
Warum nach Sprüchen suchen und lang fackeln, laß’ den Blauen doch einfach
 why for excuses search an long dither let the 100 mark bill simply
 rüberwackeln.
 over.wobble
- b. *too*-clause:
 Gelandet ist er aber in Mexiko, weil die US-Behörden mit der Aufenthaltsgenehmigung zu
 lange fackelten.
 ‘But he landed in Mexico because US authorities hesitated too long with his residence permit’
- c. licensing not clear :
 Harsche SVP-Kritik erntet die Regierung dafür, dass sie so lange fackle und noch ein zeitaufwendiges
 Vernehmlassungsverfahren inszeniere.
 ‘The government reaped harsh SVP criticism for dithering so long ...’
- (5) 4 cases with neither a durative modifier nor a negative item:
- a. L97/OKT.00130, Berliner Morgenpost, 19.10.1997, Ressort: FEUILLETON, S. 32; Wenn
 Scherze schmerzen Züngelt, strippt und schreit im Charlottchen: Heike Schneider in ‘Geklont’:
 Sie fackelt — ganz schlechtes Timing!
 she dithered very bad timing
- b. n91/JUL.06225 Salzburger Nachrichten, 30.07.1991, S. ; beitrage für das mozartjahr mozart/sänger/reinhardt
fackel.
- c. N92/AUG.30688 Salzburger Nachrichten, 22.08.1992; Wien hat keine Jugoslawien-Politik:
 Zur gleichen Zeit aber fackeln Politiker wild herum:
 at the same time but dither politicians wildly around
 die einen verdammen die EG, die anderen retten die Neutralität, und einer will gerade den
 Anschluss Österreichs an die NATO, als ob die NATO noch ein Zukunftsmodell wäre.
- d. O97/JUN.60636 Neue Kronen-Zeitung, 09.06.1997, S. 32; *:
 Günther König ließ sich nicht lumpen: Als sich beim Bischofshofener SK einiges wieder ein-
 gerenkt hatte, ...
 kehrte der Tormann mit anderen Routiniers zurück — und fackelte nun auch
 returned the goal keeper with other experienced players back and dithered now also
 gegen St. Martin/L.
 against St. Martin/L.

For these cases, it is not clear whether we are dealing with the same use of *fackeln* at all. Thus, we might consider them as “non-classifiable”. This leaves us with a single case of a bone fide occurrence of *fackeln* without negation, namely sentence (4-b).

	<i>nicht</i>	<i>nicht mehr</i>	<i>ohne</i>	<i>nie, etc</i>	schwache Neg	kein Neg
<i>lange</i>	!	!	!	!	!	?*
anderer DM	!	!	!	!	!	?*
<i>viel</i>	!	!	!	!	?*	*
kein DM	?*	?	?*	?	*	*

Table 2: Grammaticality judgements on *fackeln* (taken from Sailer (2004))

collocates		numbers	judgments
NPI-LIC	<i>wollen (want)</i>	96 96%	ok
NPI-LIC	<i>möchten (would like)</i>	2 2%	ok
NPI-LIC	<i>können (can)</i>	1 1%	??
NPI-LIC	<i>müssen (must)</i>	1 1%	*
no NPI-LIC		0 0%	*
	no modal	0 0%	*
Σ		100 100%	

Table 3: Corpus data (IDS corpora, June 2002) and introspective data (Sailer and Richter, 2002b)

2.1.2 Introspective Data on *fackeln*

The verb *fackeln* may occur with a wide range of durational modifiers and in many contexts that license NPIs. If one of these requirements is not met, grammaticality decreases significantly.

- (6) a. other mod + n-word:
 Keiner hat einen Moment gefackelt, bevor er gesprungen ist,
 nobody has a moment dithered before he jumped has
 ‘Nobody dithered a moment before jumping.’
- b. no mod + n-word:
 ? Keiner hat gefackelt, bevor er gesprungen ist.
 Nobody has dithered before ... ‘Nobody dithered before ...’
- c. other mod + no NPI licenser:
 *? Peter hat einen Moment gefackelt, bevor ...
 Peter has a moment dithered before ...
 ‘Peter dithered a moment before ...’
- d. no mod + no NPI licenser:
 * Peter hat gefackelt, bevor ...
 Peter has dithered before ... ‘Peter dithered before ...’

Generalization The semantic contribution of *fackeln* must occur in the scope of a durative operator and the two must be in the scope of an NPI licensing operator.

2.2 Other NPIs with Secondary Licensing: *wahrhaben*

- (7) Canonical example (NPI-LIC + *want*; 96%):
 Man wolle die mögliche Diagnose Herzinfarkt nicht wahrhaben, sagte S.
 one wants the possible diagnosis heart attack not true.have said S.
 ‘People didn’t want to face the reality of the possible diagnosis heart attack, said S.’
- (8) More examples from the corpus:
 a. comparative + *would like*:
 ... 1995 sei „mehr passiert“, als manche wahrhaben möchten.
 1995 has more happened than some true.have would like

‘More happened in 1995 than some people would like to admit.’

- b. NPI-LIC + *can* (1%):
... in einem Mietshaus, dessen Bewohner nicht wahrhaben können, was ...
in a block of rental flats whose inhabitants not true.have can what
‘... in a block of rental flats whose inhabitants cannot grasp what ...’
- c. NPI-LIC + *must* (1%):
(*)(er) muß wahrhaben, daß er sich mit seiner Schwägerin ... nicht nur in Sachen
(he) must true.have that he himself with his sister-in-law not only in matters
Finanzen gut versteht.
of finances well gets along
‘he must admit that he gets along well with his sister-in-law, not only in financial matters.’

More introspective data

- (9) comparative + *want*:

Der Wald stirbt schneller als die regierungsamtlichen Zahlen es wahrhaben wollen.
the forest dies faster than the government official numbers it true.have want

‘The forest is dying faster than the government’s official figures indicate.’

Generalization The verb *wahrhaben* must fall within the scope of a semantic operator of volition, as introduced by certain uses of *wollen*, *möchten* or *Wille*, which, in turn, must fall within the scope of an NPI licensing operator.

2.3 Secondary Licensing: Selection? Construction? — Collocation!

Assumptions about selection

- the selector has access to a restricted part of the properties of the selected item.
- a head selects its complements.
- an adjunct selects the head it combines with.

wahrhaben: The verb is the complement of the obligatory modal.

fackeln:

- Option 1: the obligatory adjunct is selected by *fackeln*:
 - in all grammatical examples, *fackeln* can be replaced with *zögern* (*hesitate*).
 - for *zögern* there is no reason to assume that a durative modifier is selected.
- Option 2: independent principles enforce the presence of the adjunct:
 - This has been claimed for “semantically weak” verbs:

(10) Die Vorstellung dauerte *(drei Stunden)
the performance took three hours
 - Given the synonymy of *fackeln* and *zögern*, why should *fackeln* be a semantically weak verb?

Collocation Properties of our data:

- The restrictions are collocational in nature.
 - The restrictions are grammaticalized, i.e., their adherence is obligatory and can be subjected to grammaticality judgments.
- ⇒ The collocational behavior of these words should follow from a speaker’s linguistic knowledge.

Primary Licensing If the secondary licensing requires the item to be in the scope of a particular semantic operator, why should primary licensing be fundamentally different?

3 Study 2: Automatic Acquisition of NPIs Based on their Collocational Behavior

For usual collocations quantitative methods have been investigated to extract collocation pairs.

- Sinclair (1991): largely qualitative analysis of co-occurrence patterns of lexical items (see also Moon (1998)). Extracted co-occurrence partners are based on linearly defined spans.
- Manning and Schütze (1999): describe standard quantitative methods to extract different types of collocation pairs.
- Krenn (1999): compares the extraction of collocations from annotated and unannotated corpora. Extracted co-occurrence partners are based on linguistically defined spans (chunk, clause).
- Evert (2004): shows that the extraction of collocation pairs can be improved if partially annotated corpora are used.
- Villada and Bouma (2002): Evaluation of the result is problematic for all these studies. There are no clear linguistic tests for collocations, there are no exhaustive lists, and speaker intuition are not reliable!

The described collocation extraction methods deliver word-word co-occurrence patterns. Sailer (2004) argues that a collocation-based extraction of NPIs requires a different approach.

- The collocation partners of the NPI (i.e. the licensors) are known. However, there is no corpus which contains annotation of licensors and their licensing strength.
- In a licensing constellation the NPI must be in the scope of the licensor. Semantic scope is not available in the bare corpus data, nor can it be easily computed from surface syntax (Pinkal, 1996; Hoeksema, 2000).

Lichte (2005b) addresses problems that Hoeksema (1997) lists for automatic NPI extraction:

- Many NPIs are *polysemous*, i.e. the same form has both an NPI and a non-NPI use.
- *Sparse data problem*: many NPIs will not occur in the corpus (or not frequently enough).
- *Licensing*: we might not be able to detect all licensing environments.

There are some recent studies attempts to overcome these problems (Lichte and Sailer, 2004; Lichte, 2005a,b). In the following, I will focus on Lichte (2005b) and Sailer and Lichte (2004).

Rank	Lemma	ratio
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

Table 4: Lemmata with highest occurrence ratios in the scope of DEINT marking. Items which also occur in Kürschner (1983) are given in bold face; those that are considered additional promising NPI candidates are marked with an asterisk (Lichte, 2005b, p. 75)

3.1 Data

Corpus

- the years 1995–1998 of the TüPP-D/Z corpus (*Tübingen Partially Parsed Corpus of Written German*; see Ule and Müller 2004 and <http://www.sfs.uni-tuebingen.de/tupp>).
- TüPP-D/Z is based on the electronic version of the German newspaper *die tageszeitung (taz)*.
- It contains lemmatization, part-of-speech tagging, chunking and clause brackets.
- The used part of the corpus consists of 4,5 mio sentences.

Preprocessing

- Clearly wrongly annotated sentences are filtered out by a simple heuristics (about 39% of the data; remaining 2,8 mio sentences)
- Replacing (easily identifiable) licensors by a simple licensor tag (DEINT — downward entailing/interrogative).

3.2 Simple Lexeme Study

Extraction

- Scope of a DEINT marking (Lichte, 2005b, p. 63): A DEINT marking has scope over the lemmata in the same clause and over lemmata that occur in clauses embedded in this clause.
- For each lemma: extract the relative occurrence frequency in the scope of a DEINT marking.
- Set a threshold for the minimal occurrence frequency of a lemma (in the study: ≤ 40)

Results

- The highest ranked NPI candidates are given in Table 4.
- Lichte classifies about 50% of the highest ranked items as promising NPI candidates.

Rank	Lemma combination	corresponding expression	ratio
1	unversucht lassen	etw. unversucht lassen	1.00
2	umhin zu kommen	umhinkommen	1.00
3	geheuer ganz	ganz geheuer	1.00
4	entbehren gewiss	einer gewissen <Eigenschaft> entbehren	1.00
5	müde werden betonen	müde werden zu betonen, dass ...*	1.00
6	jedermanns Sache	jedermanns Sache sein*	1.00
7	aufgehen Rezept obwohl	das/ein Rezept geht auf	1.00
8	daran lassen der Zweifel	Zweifel daran lassen	0.99
9	lumpen lassen	sich lumpen lassen	0.99
10	verkneifen können Sie	sich verkneifen können*	0.98
:	:	:	:
57	Tasse Schrank	alle Tassen im Schrank haben	0.86

Table 5: Lemma combinations with highest occurrence ratios in the scope of DEINT marking. (Lichte, 2005b, p. 85)

Discussion

- The results are promising, but:
- some NPI candidates seem to be genre specific
- polysemous NPIs have low ratio
- complex NPIs are not found (*alle Tassen im Schrank haben*)

3.3 Complex Lexeme Study

Extraction

- To find complex NPI candidates and to disambiguate different readings of polysemous words, Lichte first determines lexeme collocators for the lemmata and, then, calculates the occurrence ratio of these bi-grams in the scope of DEINT marking.
- This is iterated until the ratio has reached a maximum.

Results

- Very promising (see Table 5)
Clear improvement for polysemous and complex NPIs

Discussion

- Sparse data: the frequency of the lemma combinations is relatively low
- Time consuming! (only 1000 high frequent lemmata are used; heuristically determined collocation threshold for lemma combinations applied)
- Some structural factors for disambiguation of polysemous lemmata are not available (*brauchen*)

3.4 Subclasses of NPIs

In unpublished work (Sailer and Lichte, 2004) we tried to derive *distributional profiles* for NPI subclasses.

	PTKNEG → AM	AM → AA	AA → DE	DE → DEINT
max	24.2	23.9	14.4	20.9
min	-0.72	-1.56	-2.1	-1.8
middle	0.56	4.3	7.9	5.3
sd	0.8	2.8	3.9	3.0
verkneifen	-0.71	-1.56	0.33	-1.33
hinnehmbar	-0.71	-0.96	-2.01	-0.68
zimperlich	1.80	-0.85	-1.04	-0.47
hinwegtäuschen	0.68	-0.77	-1.49	-1.40
beileibe	-0.71	5.42	-2.06	-1.76
sondern	1.32	1.4	-1.82	-1.63
sonderlich	1.87	1.72	-1.66	-1.60
verhehlen	3.0	0.54	-0.93	-1.26
scheren	5.01	3.32	4.27	0.99
unerheblich	0.45	-0.58	-0.42	2.10
Seltenheit	-0.71	23.91	-2.10	-1.76
Ahnung	0.31	18.13	-0.26	-1.05
Hehl	0.21	19.70	-1.69	-1.53
Handhabe	1.75	16.59	-1.06	-1.13

Table 6: Overview over the z-values for the growth ratio

3.4.1 Method

- Only 1 year of TüPP-D/Z was used and clause bracketing was ignored.
- We used a more elaborated licensor annotation in the preprocessing step (PTKNEG, AM, AA, DE, DEINT).
- From the middle and the standard deviation of the growth ratios we calculate how much the growth of the ratio of an item differs from the expected growth rate. Focusing on NPI candidates only, we investigated the following hypotheses:
 - H1: For a superstrong NPI we expect a low z-value for AM→AA, AA→DE and DE→DEINT.
 - H2: For a strong NPI only the z-value for AA→DE and DE→DEINT should be low.

3.4.2 Results

The results are summarized in Table 6. The first line indicates the maximal z-values, the second the lowest z-values for each context in our sample. The minimal values in Table 6 give us an indication what a “low” z-value means for a given context widening.

H1: Potential instances of superstrong NPIs in our sample are the first four lemmata in Table 6. Introspectively, however, they all are possible in other contexts as well.

- (11) Bei so einem Flopp kann sich wohl kaum jemand die Schadenfreude verkneifen.

H2: Candidates for strong NPIs have low values for the last two columns of Table 6. This is confirmed for all the candidate NPIs of the study by introspective judgments:

- (12) a. Die Aufführung war beileibe **nicht** so gut wie erwartet.
the performance was definitely not as good as expected
b. Die Aufführung war beileibe **kein** so großer Erfolg wie erwartet.
the performance was definitely no such big success as expected
c. *Die Aufführung war beileibe **kaum** (hardly) so gut wie erwartet.
d. *War die Aufführung beileibe so ein Erfolg wie erwartet?

3.4.3 Discussion

- The hypotheses were only partially confirmed.
- There seems to be a way to get a more detailed picture from the corpus data
- Using the clause structure will improve the results.
- Evaluation of subclasses is very problematic!

4 A Collocation Module

- If the co-occurrence data can be fruitfully exploited to determine the NPI-hood of a lexical item, we might hope that a purely representational account of the NPI licensing as part of a general collocation module is possible.
- Sailer (2003) presents a very general collocational module to treat (decomposable) idiomatic expressions within *Head-Driven Phrase Structure Grammar* (HPSG, Pollard and Sag (1994)).
- Empirical restriction 1: Only lexical items can have collocational restrictions.
- Empirical restriction 2: The required information must be present in the linguistic representation of the sentence which contains a collocationally restricted item.
- Information present: phonology, synt. category and structure, sem. representation (*logical form*).
- Information not present in current versions of HPSG: entailment properties¹, meaning

- (13) Informal example for *fackeln*:

phonological representation: /fak@l@n/
syntactic category: V
valence: subject: NP
logical form: $\lambda x.\text{hesitate}'(e, x)$
collocational requirement: must occur in a sentence with a logical form of the shape:
...NPI-LIC [...DurMod [...hesitate'(e, x) ...]...] ...

4.1 Collocator is a Lemma

Two directions to refine a collocation module for HPSG:

- Structural conditions on the relation between the collocates (Soehn, 2003; Soehn and Sailer, 2003; Sailer and Trawinski, ta)
 - Individual barriers for the “licensing domain” (Richter and Sailer, 1999; Soehn, 2004, 2006)
- (14) Typical instances of collocations:
- Idiosyncratic relation between lexemes.
 - head requires complement: wage war; **Angst einjagen**
 - complement requires head: **make headway**, Tacheles reden
 - head requires adjunct: **groß rauskommen**
 - sich **freuen** wie ein Schneekönig

Sailer and Trawinski (ta) studied a corpus of 450 German expressions with bound words. There seems to be no instance of a bound word which involves a relation between two lexemes which are further than two dependency steps.

¹Tonhäuser (2000) proposes encoding entailment properties as diacritics on parts of the semantic representation.

4.2 Collocator is an Operator

- Secondary Licensing: The data on *fackeln* and *wahrhaben* can be accounted for if we can enforce the presence of a certain type of operator in the semantic representation (–logical form).
- Primary Licensing: Is the information about the NPI licensing properties also available in the linguistic structure as assumed in HPSG?

4.2.1 Representation of Licensing Strength

(15) Representing entailment strength by lexical decomposition (Sailer and Richter, 2002b):

entailment classification	example	If representation
antimorphic	<i>nicht</i> (<i>not</i>)	$\neg[\dots\phi\dots]$
anti-additive	<i>niemand</i> (<i>nobody</i>)	$\neg\exists[\dots\phi\dots]$
downward entailing	<i>wenige</i> (<i>few</i>)	$\neg\text{many}(\dots)[\dots\phi\dots]$

An NPI expresses in its collocational requirements:

1. syntactic domain within which the licenser must be found (same clause; within the next higher finite clause, ...)
2. a characterization of the licensing operator (negation, durative modifier, ...)
3. accessibility condition of the operator (immediate scope; at most one intervening quantifier; ...).
 - Directly reflects the intuition that NPIs are licensed by “negation”.
 - Since syntactic structure is also present in the linguistic representation, constraints on the syntactic domain can be combined with constraints on the semantic representation in a simple way.
 - *Antimultiplicative* licensors (*not every*) are not a natural class in this account — and they fail to be a natural class for NPI licensing (van der Wouden, 1997).
 - Uniform account of Hoeksema’s primary and secondary licensing.
 - **Problem:** It is not clear whether such decomposition is possible in all cases, let alone motivated! (questions, implications, ...).

4.2.2 Constraints

1. Given the parallelism between collocations and NPIs we also expect to find “grammatical constructions” (Fillmore et al., 1988) (idiomatic patterns, Phraseoschablonen) which are NPIs:

- (16)
- a. **nicht aus X-dorf** sein/ kommen
not from X-village be/ come
‘not to be X’
 - b. **nicht aus Dummsdorf** sein/ kommen
‘not to be stupid’
 - c. **nicht aus Gebersdorf** sein/ kommen
‘not to be from the village of giving people’, i.e. ‘be scrooge’

2. Are the syntactic domains identical to those needed in syntax? (Progovac, 1988; Soehn, 2006)
3. Which other operators can occur in the collocational restrictions of lexical items? Hoeksema (1996) suggests that some items tend to occur in the scope of a universal quantifier (Dutch: *rozegeur en maneschijn* [NPI], *koek en ei* [not an NPI])

4. The data looked at so far suggest that at most two quantifiers can intervene between the negation and the NPI (Immediate Scope Constraint plus indirect licensing of (Linebarger, 1980). While this is an interesting parallel to the restriction on the distance between lexeme collocates on the dependence structure, neither restriction follows from the model.
5. Planned: collection of polarity items, analogous to the collection of bound words (CoDIL-BW, <http://sfs.uni-tuebinge.de/a5/codii>, Sailer and Trawinski (ta)) to test generalizations.

5 Relation to other Models

5.1 Syntactic Accounts

5.1.1 Immediate Scope Constraint: Linebarger

- Linebarger (1980) argues that no quantifier may intervene between an NPI and the licenser, and she assumes “indirect” licensing in the case of non-negative licensors, i.e. there must be an implicature within which the NPI is in the immediate scope of negation.
- The representational theory can capture licensing by negation directly. For indirect licensing, lexical decomposition is used. Lexical decomposition would lead to a more homogeneous overall framework.

5.1.2 Binding Theory: Progovac

The collocational approach has similarities to the binding theoretical approach of Progovac (1988, 1994):

1. Being genuinely “syntactic”, both accounts are flexible enough to capture a high degree of idiosyncrasy in the data.
2. Just like in a binding approach, we specify (i) the syntactic domain, (ii) the type of binder, and (iii) possible interveners.
3. Since the context elements that we require are part of the semantic representation, we avoid two problems of Progovac:
 - we do not stipulate phonologically empty syntactic operators which do not have a meaning either.
 - For Progovac only NEG at INFL (or an Operator in Comp) can bind an NPI. This deprives her from accounting for licensing by other operators. Lexical decomposition, but not the orientation towards a particular syntactic position makes our approach more flexible.

5.2 Entailment-based Accounts

1. “Text book theory” of NPI licensing (Fauconnier, 1975, 1978; Ladusaw, 1980).
2. If we can provide and motivate a lexical decomposition for the subclasses of downward entailing contexts (Zwarts, 1995, 1997), the collocational approach is equally powerful as the entailment-based approaches.
3. The collocational approach avoids the antimultiplicativity gap and can integrate syntactic constraints more easily. For a problem with Neg Raising predicates see van der Wouden (1995).

5.3 Pragmatic Accounts

1. Krifka (1995) proposes a combination of a semantic and a pragmatic account. A sentence with an unlicensed NPI is bad “because it expresses a sentence in which what is said systematically contradicts what is implicated”.

2. Krifka's and our approach can account for the high degree of lexical idiosyncrasy: For Krifka, the NPI lexically specifies a set of alternatives. These interact with a general assertion operator to guarantee well-formedness in licensed structures and oddness in others.
3. To account for the Linebarger facts, Krifka assumes that insertion of the assert operator can be triggered by quantifiers. We would derive the Linebarger effect by characterizing operators which may trigger existential closure as potential interveners.
4. Difference in the empirical domain:
Krifka (1995) explicitly excludes NPIs which require a clausemate negation as licenser (such as English *one bit*, Zwarts (1997)):
 - (17) a. Jan was **not one bit** happy about this proposal.
 - b. ***Nobody** was one bit happy about this proposal.
 - c. ***Few** people were one bit happy about this proposal.

Krifka treats these NPIs as *negative concord items* (such as Serbian/Croatian *ni-* n-words). However:

- A negative concord item can occur on its own in a short answer; a superstrong NPI cannot:

- (18) a. Kogo widziałeś? Nikogo. (Polish)
 Who have you seen? Nobody.GEN/ACC.
- b. How happy was Jan about the proposal? Not at all/ *(Not) One bit.

Richter and Sailer (2004) argues (contrary to Błaszczak (1999)) that Polish n-words are inherently negative and, thus, are not NPIs (whereas *one bit* would count as a genuine NPI)

- Superstrong NPIs are fine in Neg Raising context; negative concord items typically are not.

- (19) a. Sorry to say, but I don't **think** you understand one bit.
- b. *Jan **nie sądzi**, żeby Marysia nikogo lubiła. (Polish)
 Jan not believes that Mary nobody liked (Przepiórkowski and Kupś, 1997)

5. It is not clear whether all NPIs really induce scalar implicatures:

- Should the contrast between *zögern* and *fackeln* be described in terms of different implicatures?
- Hoeksema (2002) argues that for some minimal elements in Dutch the typical use is not scalar:

- (20) a. **Geen hond** gelooft dat jij onschuldig bent. (Dutch, non-scalar)
 no dog believes that you innocent are
- b. Er kon **geen** ?*hond/ muis meer door.
 there could no dog/ mouse get through.

6. Hoeksema (1996) also wants to relate the occurrence pattern of an NPI to its pragmatic impact ("understatement", ...). If our empirical studies are successful, we can derive the collocational restrictions from the actual language use. If certain collocational patterns correlate with certain pragmatic properties, it might be a "Huhn-oder-Ei" question which information is primary and should be explicitly encoded in the lexical entry of the item.

6 Conclusion

6.1 Planned Continuations

1. Collection of Polarity Items for German (Tübingen) and English (Göttingen).
2. Automatic extraction of NPIs for English (Göttingen).
3. Representational characterization of NPI licensing contexts.

6.2 Open Questions

- Help from psycholinguistic data for a systematic evaluation of the quality of the extracted NPI candidates? (mere statistical co-occurrence vs. real linguistic collocation, Hausmann (2004))

- Can processing models clarify licensing domain extension:

- (21) "Parasitic Licensing" (Horn, 1978, p. 146):
 - a. *A/any bomb will explode until midnight.
 - b. *I don't claim that a bomb will explode until midnight.
 - c. I don't claim that any bombs will explode until midnight.
- (22) Bound variables in the embedded clause (van der Wouden, 1997, p. 166ff.) (*bijster* (rather) is a weak NPI in Dutch):
 - a. *Piet_i vindt dat hij_i bijster mooi zingt.
 Piet finds that he rather beautiful sings
 - b. Piet vindt dat niemand bijster mooi zingt.
 Piet finds that no one rather beautiful sings
 - c. ?Niemand vindt dat Piet bijster mooi zingt.
 - d. Niemand vindt zichzelf bijster mooi zingen.
 No one finds himself rather beautiful sing
 'Nobody considers himself to sing rather beautifully.'

- Locality restriction: Why at most two intervening items of the relevant domain? Is this a processing constraint?

References

- Błaszczak, Joanna (1999). *Investigations into the Interaction between the Indefinites and Negation*. Ph. D. thesis, Humboldt Universität Berlin. Published in: Akademieverlag, Berlin, *Studia grammatica* 51, 2001.
- Butler, Christopher S. (1985). *Systemic Linguistics: Theory and Application*. Batsford, London.
- Evert, Stefan (2004). *The Statistics of Word Cooccurrences, Word Pairs and Collocations*. Ph. D. thesis, Universität Stuttgart.
- Fauconnier, Giles (1975). Polarity and the Scale Principle. In *Papers from the Eleventh Regional Meeting of the Chicago Linguistic Society*, pp. 188–199.
- Fauconnier, Giles (1978). Implication Reversal in Natural Language. In F. Guenther and S. Schmidt (Eds.), *Formal Semantics and Pragmatics for Natural Languages*, pp. 289–302. Reidel, Dordrecht.
- Fillmore, Charles, Kay, Paul, and O'Connor, M. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone*. *Language* 64, 501–538.
- Hausmann, Franz-Joseph (2004). Was sind eigentlich Kollokationen? In K. Steyer (Ed.), *Tagungsband der 39. Jahrestagung des IDS*, pp. 309–334. IDS Mannheim, Mannheim.
- Hoeksema, Jack (1996). Zeg het met rozen: een vergelijking van drie idiomatische uitdrukkingen. *TABU Bulletin voor Taahwetenschap* 26(3), 129–149.
- Hoeksema, Jack (1997). Corpus Study of Negative Polarity Items. Html version of a paper which appeared in the *IV-V Jornades de corpus linguistics 1996-1997*, Universitat Pompeu Fabre, Barcelona. URL: <http://odur.let.rug.nl/~hoeksema/docs/barcelona.html>.
- Hoeksema, Jack (2000). Negative Polarity Items: Triggering, Scope and C-Command. In L. Horn and Y. Kato (Eds.), *Negation and Polarity*, pp. 115–146. Oxford University Press, Oxford.

- Hoeksema, Jack (2001). Polarity Items in Context: Primary and Secondary Licensing. Handout to a talk given in the *Workshop on Negation*, Paris, October 2, 2001.
- Hoeksema, Jack (2002). Meaning and Use of Polarity Items. Handout zu einem Vortrag an der Universität Tübingen, 18.11.2002.
- Hoey, Michael (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.
- Horn, Laurence R. (1978). Remarks on Neg-Raising. In P. Cole (Ed.), *Pragmatics*, Volume 9 of *Syntax and Semantics*, pp. 129–220. New York, San Francisco, London: Academic Press.
- Krenn, Brigitte (1999). *The Usual Suspects. Data-Oriented Models for Identification and Representation of Lexical Collocations*, Volume 7 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. Saarbrücken: DFKI and Universität des Saarlandes.
- Krifka, Manfred (1995). The Semantics and Pragmatics of Weak and Strong Polarity Item. *Linguistic Analysis* 25, 209–257.
- Kürschner, Wilfried (1983). *Studien zur Negation im Deutschen*. Gunter Narr, Tübingen.
- Ladusaw, William (1980). *Polarity Sensitivity as Inherent Scope relations*. Garland Press, New York.
- Lichte, Timm (2005a). Corpus-based Acquisition of Complex Negative Polarity Items. In J. Gervain (Ed.), *Proceedings of the Tenth ESSLLI Student Session*, Edinburgh. Heriot-Watt University.
- Lichte, Timm (2005b, October). Korpusbasierte Acquirierung negativ-polärer Elemente. Master's thesis, Seminar für Sprachwissenschaft, University of Tübingen.
- Lichte, Timm and Sailer, Manfred (2004). Extracting Negative Polarity Items from a Partially Parsed Corpus. In S. Kübler, J. Nivre, E. Hinrichs, and H. Wunsch (Eds.), *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004)*, University of Tübingen. Seminar für Sprachwissenschaft.
- Linebarger, Marcia Christine (1980). *The Grammar of Negative Polarity*. Ph. D. thesis, MIT, cited after the reproduction by the Indiana University Linguistics Club, Indiana, 1981.
- Manning, Christopher D. and Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Moon, Rosamund (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.
- Pinkal, Manfred (1996). Radical Underspecification. In P. Dekker and M. Stokhof (Eds.), *Proceedings of the Tenth Amsterdam Colloquium*, pp. 587 – 606. ILLC/Department of Philosophy, University of Amsterdam.
- Pollard, Carl and Sag, Ivan A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago and London: University of Chicago Press.
- Progovac, Ljiljana (1988). *A Binding Approach to Polarity Sensitivity*. Ph. D. thesis, University of Southern California.
- Progovac, Ljiljana (1994). *Negative and Positive Polarity*. Cambridge University Press.
- Przepiórkowski, Adam and Kupś, Anna (1997). Unbounded Negative Concord in Polish: A Lexicalist HPSG Approach. In J. Landsbergen, J. Odiijk, K. van Deemter, and G. V. van Zanten (Eds.), *Computational Linguistics in the Netherlands 1996. Papers from the Seventh CLIN Meeting*, pp. 129–144.
- Richter, Frank and Sailer, Manfred (1999). LF Conditions on Expressions of Ty2: An HPSG Analysis of Negative Concord in Polish. In R. D. Borsley and A. Przepiórkowski (Eds.), *Slavic in HPSG*, pp. 247–282. Stanford: CSLI Publications.
- Richter, Frank and Sailer, Manfred (2004). Basic Concepts of Lexical Resource Semantics. In *ESSLLI 2003 – Course Material I*, Volume 5 of *Collegium Logicum*. Kurt Gödel Society Wien.
- Sailer, Manfred (2003). Combinatorial Semantics and Idiomatic Expressions in Head-Driven Phrase Structure Grammar. Phil. Dissertation (2000). Arbeitspapiere des SFB 340. 161, Universität Tübingen.
- Sailer, Manfred (2004). Distributionsidiosynkrasien: Korpuslinguistische Erfassung und grammatiktheoretische Deutung. In K. Steyer (Ed.), *Wortverbindungen – mehr oder weniger fest*, Institut für Deutsche Sprache, Jahrbuch 2003, Berlin, New York, pp. 194–221. de Gruyter.
- Sailer, Manfred and Lichte, Timm (2004). Distributional Profiles of Polarity Items. unpublished manuscript, University of Tübingen.
- Sailer, Manfred and Richter, Frank (2002a). Collocations and the Representation of Polarity. In G. Alberti, K. Balogh, and P. Dekker (Eds.), *Proceedings of the Seventh Symposium on Logic and Language*, Pécs, pp. 129–138.
- Sailer, Manfred and Richter, Frank (2002b). Not for Love or Money: Collocations! In G. Jäger, P. Monachesi, G. Penn, and S. Wintner (Eds.), *Proceedings of Formal Grammar 2002*, pp. 149–160.
- Sailer, Manfred and Trawinski, Beata (t.a.). Die Sammlung unikalener Wörter des Deutschen. Aufbauprinzipien und erste Auswertung. In A. Häcki-Buhofer (Ed.), *EUROPHRAS 2003*, pp. ??
- Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Soehn, Jan-Philipp (2003). Von Geisterhand zu Potte gekommen. Eine HPSG-Analyse von PPs mit unikalener Komponente. Magisterarbeit, Universität Tübingen, Seminar für Sprachwissenschaft.
- Soehn, Jan-Philipp (2004). License to COLL. In S. Müller (Ed.), *Proceedings of the HPSG-2004 Conference, Center for Computational Linguistics, Katholieke Universiteit Leuven*, pp. 261–273. Stanford: CSLI Publications. <http://csli-publications.stanford.edu/HPSG/5/>.
- Soehn, Jan-Philipp (2006). *Über Bärendienste und erstaunte Bauklötze. Idiome ohne freie Lesart in der HPSG*. Ph. D. thesis, Friedrich-Schiller-Universität Jena.
- Soehn, Jan-Philipp and Sailer, Manfred (2003). At First Blush on Tenterhooks. About Selectional Restrictions Imposed by Nonheads. In G. Jäger, P. Monachesi, G. Penn, and S. Wintner (Eds.), *Proceedings of Formal Grammar 2003*, pp. 149–161.
- Tonhauser, Judith (2000). An Approach to Polarity Sensitivity and Negative Concord by Lexical Underspecification. Handout to a talk given at *7th International Conference on HPSG*, University of Berkeley, 22–23 July 2000.
- Ule, Tylman and Müller, Frank Henrik (2004). KaRoPars: Ein System zur linguistischen Annotation großer Text-Korpora des Deutschen. In A. Mehler and H. Lobin (Eds.), *Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*. Opladen: Westdeutscher Verlag. to appear.
- Wouden, Ton van der (1992). Bepervingen op het optreden van lexicale elementen. *De Nieuwe Taalgids* 85(6), 513–538.
- Wouden, Ton van der (1995). A Problem with the Semantics of Negative Raising Predicates. In S. Fischer and M. Trautwein (Eds.), *Proceedings of Accolade*, Amsterdam, pp. 169–183.
- Wouden, Ton van der (1997). *Negative Contexts. Collocation, Polarity and Multiple Negation*. London: Routledge.
- Villada, Begoña and Bouma, Gosse (2002). A Corpus-based Approach to the Acquisition of Collocational Prepositional Phrases. In *Proceedings of EURALEX 2002, Copenhagen, Denmark*.
- Zwarts, Frans (1995). Nonveridical Contexts. *Linguistic Analysis* 25, 286–312.
- Zwarts, Frans (1997). Three Types of Polarity. In F. Hamm and E. W. Hinrichs (Eds.), *Plurality and Quantification*, pp. 177–237. Kluwer Academic Publishers, Dordrecht.