Frank Richter, Fabienne Fritzinger, Marion Weller

Who Can See the Forest for the Trees? Extracting Multiword Negative Polarity Items from Dependency-Parsed Text

1 Introduction

Ever since the groundbreaking work by Fauconnier (1975) and Ladusaw (1980), research on negative polarity items (NPIs) has been dominated by two fundamental assumptions about the licensing contexts of NPIs and their inherent semantic-pragmatic properties. The contexts in which NPIs may occur felicitously are said to have the semantic property of being *downward entailing* (which we will briefly explain below), and the elements themselves are often said to be located at the end of a pragmatically motivated scale, typically signalling a minimal amount, a smallest size, or similar concept. While the Ladusaw-Fauconnier theory has been substantially refined over time, and while there are very diverse variations on how the technical details of the theory are spelled out, its core insights are currently widely accepted and remain a point of reference for practically any 'formal' theory of NPIs. Some theories are syntactic in nature and formulate the relevant scope constraints relative to (possibly quite abstract) syntactic configurations, others are semantic and define hierarchies of negations of varying strength, and yet another group of theories is predominantly pragmatic, relying heavily on scalar implicatures, domain widening, and related concepts. There are, of course, also approaches in which syntax, semantics, and pragmatics all play a role. Overall, the number of papers and books that have been published on the subject of NPIs over the last 40 years is nothing short of intimidating.[1]

Given the sheer volume of the NPI literature, it is all the more surprising and striking that much of the discussion revolves around a very small set of items. Especially some of the most sophisticated and influential papers, such as Kadmon and Landman (1993), Krifka (1995), and Chierchia (2006), discuss hardly more than a handful of items, and some studies almost exclusively focus on one, *viz.* English *any*, which can be regarded as *the* classical example for a minimizer, with its variants *anything, anyone, anybody, anywhere*, etc. Since with *any* one of the most prominent items of interest is a minimizer, investigations into the significance of this particular property for the entire class of NPIs have turned into a dominating topic and occasionally even push aside the observation that being a minimizer is not a necessary (nor a sufficient) property of NPIs. As a result of its narrow empirical focus, the tendency to build a very comprehensive theory on an extremely small, carefully chosen but deeply researched set of examples is characteristic for large parts of the literature on NPIs. This might mean that only a fraction of the properties and behavior of NPIs are treated in current theories.

[1] To get a first impression of the amount of work published on the topic, the electronic NPI bibliography at www.sfb441.uni-tuebingen.de/a5/pib/XML2HTML/list.html is a good starting point. It lists well over 130 articles and books.

A more comprehensive overview of the landscape of empirical phenomena beyond the typical core examples of semantic and pragmatic studies of NPIs can be obtained by turning to research which approaches NPIs from a different angle. For German, Kürschner (1983) contains a collection of 344 (single-word and multiword) items which show strong affinity to negation and negative environments. Unfortunately, Kürschner's collection does not attempt a syntactic or semantic analysis within one of the major formal linguistic frameworks, and its data do not receive the kind of theoretical classification that would make them readily accessible to proponents of the Ladusaw-Fauconnier school. The most serious shortcoming in this respect might be the omission of a theoretically motivated distinction between items that strictly require licensing negation and those which do not grammaticalize this requirement and show only a preference for negative environments.

Another rich source for a broader picture of the empirical facts is provided in the work of a Dutch group of linguists around Jack Hoeksema, Ton van der Wouden, and Frans Zwarts. In contrast to Kürschner's strongly data-driven compilation, Jack Hoeksema's comprehensive studies of (predominantly Dutch) NPIs combines theoretical concepts from the Ladusaw-Fauconnier tradition with extensive synchronic and diachronic corpus studies. Hoeksema (1997) is an early example of the potential of using electronic corpora in this area and gives a first glimpse of the highly differentiated and exciting landscape of polarity phenomena that emerges when corpus data is systematically researched and investigated with the tools of formal NPI research.

Kürschner's collection and the comprehensive work of the above group of Dutch linguists on a wide range of polarity phenomena inspired the creation of a collection of theoretically classified and richly documented German NPIs as a subcollection of the electronic *Collection of Distributionally Idiosyncratic Items* (CoDII, Trawiński et al. (2008); Richter et al. (2010)).[2] The NPI collection in CoDII can be considered a thorough inventory of our current knowledge of the extent of German NPIs. Its empirical coverage subsumes all available sources, i.e. it lists all German NPIs mentioned in all of the literature that was surveyed in its creation, including the true NPIs in Kürschner's list. But CoDII not only collects items discussed elsewhere and reports their usage in systematically selected licensing environments with naturally occurring examples from corpora. Its compilation was also supported by search results from the implementation of the first semi-automatic extraction procedure for NPIs from corpora (Lichte, 2005a,b)[3].

The collection of German NPIs in CoDII and the NPI extraction procedure of Lichte and Soehn (2007) form the starting point of the present study. The motivation behind the new NPI extraction procedure which we will present is to prepare a wider empirical base for a future, more comprehensive theory of NPIs, to sharpen our understanding of the syntactic and semantic diversity of NPIs, and to provide the necessary material for psycholinguistic studies and the use of NPIs in language processing tasks. Our

[2] CoDII was compiled in a project of the former *Sonderforschungsbereich 441* and is available at www.sfb441.uni-tuebingen.de/a5/codii/

[3] Subsequently refined in Lichte and Soehn (2007)

immediate objective is to demonstrate that our method can significantly extend the set of known NPIs in German (as represented by the 165 entries in CoDII, the largest collection available today). In the absence of a complete repository of German NPIs that could serve as a gold standard, we will measure the success of our method by the number of items we can add to the CoDII collection.

Due to the striking frequency of multiword NPIs in CoDII, and based on the assumption that there is an affinity between the properties of NPIs and at least some classes of idiomatic expressions, our new method targets multiword NPI candidates. We adapt an extraction pipeline that was previously successfully applied in the identification of multiword expressions (MWEs, (Fritzinger and Heid, 2009)) using statistical association measures and two linguistically motivated scores, the degree of morpho-syntactic fixedness (Weller and Heid, 2010) and semantic opacity (Fritzinger, 2009) of an expression. The significant difference between our method and the basic form of the earlier algorithm by Lichte and Soehn is our focus on MWEs. Lichte and Soehn primarily search for single-word NPIs and capture multiword NPIs only indirectly in an extension to their basic extraction mechanism by building lemma chains of length $n+1$ from lemma chains of length n and checking if extending a lemma chain makes it a better NPI candidate.

Section 2 gives a very brief overview of NPIs and their licensing contexts. In Section 3 we characterize our corpora and our extraction method for MWE candidates, before we say more about how we model NPI licensing contexts in Section 4. In Section 5 we present optimizations to the statistical processing of NPI candidates that we apply to achieve a higher ratio of NPIs at the top of our candidate lists, and we propose some linguistic measures for the identification of idiomatic candidate expressions. Section 6 discusses the results. We conclude with a short outlook on future work in Section 7. The appendix lists the NPIs that our extraction method found.

2 npis and npi Licensing

NPIs are defined as single words or multiword expressions which require the presence of an appropriately 'negative' element in their utterance context. The negative element is said to *license* the NPI, and without a proper licenser the presence of an NPI results in ungrammaticality. Examples of extensively researched NPIs from English are the determiner *any*, the adverb *ever*, and the MWEs *red cent* and *to lift a finger*; good licensers are the sentential negation adverb *not* or negative quantifiers such as *no students*. In (1a/b)–(4a/b) we see sentence pairs with the NPI *any* which is licensed here in the scope of four different lexical licensers ((a)-sentences; licensers are underlined). The sentences become ungrammatical when the licenser is omitted or replaced by an element without the necessary licensing properties ((b)-sentences). The (c) and (d) sentences are parallel German counterparts to the standard English examples in (a) and (b), with the verb *scheren* ('to care') as NPI.

(1) a. Pat did not see **any** student in the hallway this morning.
b. *Pat saw **any** student in the hallway this morning.

c. Peter **schert** sich nicht um Lokalpolitik.
Peter cares REFL not about local politics
'Peter does not care about local politics.'

d. *Peter **schert** sich um Lokalpolitik.
Peter cares REFL about local politics
'Peter cares about local politics.'

(2) a. Nobody saw **any** student in the hallway this morning.

b. *Everybody saw **any** student in the hallway this morning.

c. Niemand **schert** sich um Lokalpolitik.
nobody cares REFL about local politics
'Nobody cares about local politics.'

d. *Jeder **schert** sich um Lokalpolitik.
everybody cares REFL about local politics
'Everybody cares about local politics.'

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

b. *Kim saw **any** student in the hallway this morning.

c. Peter **schert** sich niemals um Lokalpolitik.
Peter cares REFL never about local politics
'Peter never cares about local politics.'

d. *Peter **schert** sich um Lokalpolitik.
Peter cares REFL about local politics
'Peter cares about local politics.'

(4) a. Few lecturers saw **any** student in the hallway this morning.

b. *Some lecturers saw **any** student in the hallway this morning.

c. Wenige Bundespolitiker **scheren** sich um Lokalpolitik.
few federal politicians care REFL about local politics
'Few federal politicians care about local politics.'

d. *Einige Bundespolitiker **scheren** sich um Lokalpolitik.
some federal politicians care REFL about local politics
'Some federal politicians care about local politics.'

The question of how to characterize the necessary negativity more accurately and which structural, logical or pragmatic relationship must hold between an NPI and its licenser or licensing environment has been subject to intense debate in theoretical linguistics, and is far from being settled. According to the dominant view, the contextually necessary negativity can best be semantically characterized in terms of the entailment behavior of the licensing environment, and the entailment behavior is triggered by an operator that must stand in a certain structural relation to the licensed NPI. NPIs are licensed in the semantic scope of the relevant operators, and are ungrammatical in their absence (see Zwarts (1997) and van der Wouden (1997) for details). Note that a component of a larger constituent can be responsible for the licensing behavior of that constituent. The NP quantifier *few lecturers* in (4a) is a licenser due to the logical behavior of its determiner, *few*, as can be verified by the ungrammaticality of (4b), where *few* has been replaced by the determiner *some*.

To keep our terminology simple, in the remainder of this paper we will call all relevant licensing environments *negative*. It is important to keep in mind that, despite this naming convention, other operators whose negativity is much less apparent than in the case of sentential negation and negative quantifiers can also license NPIs. Examples of weaker forms of negation are the quantifier *few lecturers* (see (4a)) and questions, which are perfectly valid licensers for many NPIs. Most licensing environments are logically *downward entailing*, which means that they allow inferences from supersets to subsets. For example, the downward entailing operator *few doctors* is responsible for the valid inference from the truth of *few doctors recommended showers* to *few doctors recommended cold showers*. Questions are sometimes subsumed under a weaker class of negativity, called *nonveridicality* (Zwarts, 1995). Nonveridical operators prohibit inferring the truth of a proposition from it being uttered: *Did Peter come late?* does not entail that Peter came late.

The examples in (5)–(8) illustrate multiword NPIs and highlight additional factors that need to be taken into consideration when searching for them in corpora and when checking if a candidate expression is indeed an NPI. English examples are followed by their German translations into corresponding constructions with NPIs. All explanations below about the English examples also apply, *mutatis mutandis*, to their German counterparts.

(5) a. John didn't **drink a drop** (of alcohol) last night.
b. #John **drank a drop** (of alcohol) last night.
c. #Few students **drank a drop** (of alcohol) last night.
d. Hans hat letzte Nacht keinen **Tropfen** (Alkohol) **getrunken**.
e. #Hans hat letzte Nacht **einen Tropfen** (Alkohol) **getrunken**.
f. #Wenige Studenten haben letzte Nacht **einen Tropfen** (Alkohol) **getrunken**.

(6) a. Nobody had **the slightest inkling** about where to go.
b. *Few visitors had **the slightest inkling** about where to go.
c. Niemand hatte **die leiseste Vorstellung**, wohin man gehen sollte.
d. *Wenige Besucher hatten **die leiseste Vorstellung**, wohin sie gehen sollten.

(7) a. Thomas isn't **much of a** soccer player.

b. *Miroslav is **much of a** soccer player.

c. Thomas ist **beileibe** kein Fußballer.

d. *Miroslav ist **beileibe** ein Fußballer.

(8) a. This sentence will not parse **in a million years**.

b. #This sentence will parse **in a million years**.

c. Dieser Satz lässt sich **im Lebtag** nicht parsen.

d. *Dieser Satz lässt sich **im Lebtag** parsen.

A comparison between (5a) and (5b) shows that with multiword NPIs it becomes important to distinguish between different readings of candidate expressions. In (5a) the idiomatic reading (John did not drink any alcohol at all) is very prominent, whereas (5b), due to the absence of an appropriate licenser for the idiomatic NPI *drink a drop*, does not have the idiomatic reading. However, there is a literal meaning (the amount that John drank was one drop), which is in principle available. In (5) and (8) we indicate unavailable idiomatic NPI readings and available literal readings with '#'. In cases in which there is no literal meaning ((6) and (7)) this complication does not arise. The examples also demonstrate that (simple and complex) NPIs can be of almost any syntactic category. The present selection of multiword NPIs also provides examples for the class of minimizers (and maximizers, (8a)), which play such a prominent role in current pragmatic theories of NPIs. Finally, (5c) shows that the licensing requirements of NPIs may differ: (4a) confirmed that quantifiers with the determiner *few* license *any*, but *drink a drop* is not licensed by a corresponding quantifier in (5c), it needs a *stronger* type of negation such as sentential negation to be satisfied (5a). Throughout this study, we will ignore the observation that licensing requirements can be of different strength.[4]

For the present research, we follow an idea applied in the NPI extraction algorithm by Lichte and Soehn (2007) and exploit the fact that a finite set of particular lexemes (determiners, adverbs, subordinating conjunctions, a small number of verbs) and an equally small set of syntactic structures (antecedents of conditionals, questions, comparative constructions) are good indicators of licensing environments.[5] Although they do not cover all possible licensing environments, and although there can be additional syntactic or semantic properties present in a clause which prevent NPIs from being licensed in certain positions, we assume that we can detect enough licensing environments sufficiently well to obtain useful NPI candidate lists when using our heuristics in large corpora.

[4] See Lichte and Soehn (2007) for an attempt to use a hierarchy of three types of negation strengths in extracting NPIs from corpora.

[5] Details about this choice of licensing contexts (which is derived from the data-driven classification of NPIs in CoDII) will be discussed in Section 4, along with an explanation of why there is an unavoidable residue of licensing environments which cannot be detected with our type of heuristics. In Section 5.3 we illustrate concrete limitations of our extraction pipeline with problematic data from our corpus.

3 Preliminaries: Extraction Methodology

Finding multiword NPIs in corpora is not an easy task: NPIs are known to be rare, and many members of the subclass of multiword NPIs are probably even rarer. Lichte and Soehn (2007) report that they found 28 occurrences of the single-word NPI *Menschenseele* ('living soul'), and the same number of occurrences of the complex NPI *alle Tassen im Schrank haben* ('to have lost one's marbles'). In EUROPARL (see below) we found 36 occurrences of *ein Hehl aus etwas machen* ('to hide sth.'), 18 occurrences of *ein Blatt vor den Mund nehmen* ('to mince words'), and 6 occurrences of *einen Finger rühren* ('to lift a finger'). It is evident that in order to retrieve enough occurrences of an expression to apply statistical methods which can meaningfully support its association with negative environments, we would thus like to use as large a corpus as possible. Easily available unannotated text would fulfill this desideratum.

At the same time we are faced with a second requirement. Detecting negative environments and determining that several words form a multiword expression presupposes linguistic knowledge. For that reason these two tasks can be most easily accomplished with text that is already linguistically annotated and provides a syntactic basis for recognizing plausible multiword expression candidates and at least some indication of the scope of relevant semantic operators.[6] If we hypothesized naively, i.e. without paying attention to structure, that any collection of words in a sentence could be an MWE candidate, we would quickly run into an intractable combinatorial explosion of candidates. Syntactic knowledge about which groups of words form meaningful syntactic units is particularly relevant for languages with discontinuous constituents such as German. In short, we can benefit from annotation. Annotated corpora are, however, limited in size, and decrease the size of the available data base compared to plain text.

Our method tries to strike a balance between the conflicting needs of working with a large resource and being able to refer to structural linguistic knowledge. As outlined in Section 3.1, we start from unannotated corpora, and we obtain the necessary annotation by relying on a robust broad coverage dependency parser with rich lexical information. Our next step then, described in Section 3.2, is to extract certain complex syntactic patterns that we consider promising structural skeletons of multiword NPIs. The MWE candidates that we extract in this preprocessing step will later provide the foundation to finding those complex expressions that are statistically associated with NPI licensing contexts.

3.1 Data

NPIs are infrequent in text corpora and, presumably, in everyday language. In order to avoid problems in the statistical analysis arising from sparse data, we need a very large text corpus to start from.[7] An overview of our corpus collection is given in Table 1.

[6] Lichte and Soehn's work was based on a treebank newspaper corpus. Their heuristics for determining the scope of semantic operators was using the syntactic structure provided by the tree annotation.

[7] Lichte (2005b) reports that he achieves the best UAP index (see Section 5.2) for his candidate lists with a minimal frequency of 60 for the items considered in the candidate list. Since this threshold

It contains about 269 million words (tokens), comprising text from several German newspapers, and the proceedings of the European parliament debates, Europarl (Koehn, 2005).

source	size (tokens)	text type	years
Europarl	35 million	debates	1996-2006
Frankfurter Allg. Zeitung	70 million	news	1997-1998
Frankfurter Rundschau	40 million	news	1992-1993
Handelsblatt	36 million	news	1986/1988
Stuttgarter Zeitung	36 million	news	1991-1993
Die Zeit	52 million	news	1995-2001
Total:	269 million		

Table 1: Composition of the dataset

Europarl will play a distinguished role in our npi identification procedure when we target semantic properties of mwes. At the beginning we will not do this yet and will only use the German part for monolingual processing. In later refinements of our method in which more linguistic knowledge is brought to bear, we also add the English, French and Swedish parts for multilingual processing. These refinements will be discussed in Section 5.2 below. For the initial identification of mwe candidates, we rely entirely on monolingual processing.

3.2 Multiword Extraction with Syntactic Patterns

In German, the constituent words of multiword constructions are not always adjacent to each other for two reasons. The possibility of constituent order variations in the middle field entails that multiword expressions may be realized discontinuously, with other constituents potentially intervening. The additional alternation of the verb between a verb second and verb final position in finite sentences means that a finite verb may precede or succeed those of its dependents that occur in the middle field. The sentence in (9) contains the npi *(k)einen blassen Schimmer haben* (lit. '(not) to have a pale gleam'; '(not) to have the faintest idea'). The finite verb form *hat* in verb second position is linearly separated from its accusative object *blassen Schimmer* with which it forms a multiword npi. Note that in a corresponding verb final construction *blassen Schimmer* would precede *hat*, thus reversing their order, and they could be adjacent.

(9) Er **hat** zum jetzigen Zeitpunkt keinen **blassen Schimmer**. . .
he has at this point no pale gleam. . .

excludes too many npis that enter the candidate lists with a lower choice of minimal frequency, Lichte (2005b) decides to choose a minimal frequency of 40, despite the ensuing increase of noise in the candidate list; Lichte (2005a) lowers the threshold even further (to 20), whereas Lichte and Soehn (2007) chooses 30.

‘At this point he does not have the faintest idea.’

Deep syntactic analysis is essential in order to reliably extract potentially discontinuous multiword constructions. In the past, we successfully used the dependency parser Fspar (Schiehlen, 2003) for a variety of mwe extraction tasks. Fspar is highly efficient and relies on a large lexicon. It processes about 10 million words in 30 minutes, is very robust and includes enough morphological information for our task (see Figure 1a, 5th column), which means that we do not need a separate morphological analyzer. Fspar is compatible with the German orthographical conventions before and after the spelling reform of 1996. An example analysis of Fspar is given in Figure 1. It shows a dependency structure for the sentence *Und er hat keinen blassen Schimmer, was gerade vor sich geht* (‘And he doesn’t have the faintest idea what is going on’).

The dependency tree representation in Figure 1b is not provided by the parser in this format (but can be inferred from its analysis); we insert it here in order to enhance the readability of the example. The original Fspar output given in Figure 1a contains the following information (from left to right): position of the token in the sentence, token, part of speech tag, lemma, morpho-syntactic information, dependency relation (the numbers refer to sentence positions in the first row), and grammatical function.

The dependency parses provide all necessary information for building a collection of multiword items that we can investigate for their distributional properties. Those multiword items that occur in npi licensing environments will become our npi candidates. To obtain syntactically meaningful units we first identify in the corpus certain patterns of verbs and their dependents. The patterns we collect are verbs and dependents that are nouns, adjective-noun combinations, preposition-noun combinations, preposition-adjective-noun combinations, or noun plus preposition-noun combinations. These patterns are chosen because the class of verbal mwes and verb phrase idioms is known to be comparatively large, and we expect to find a sizable number of npis among them.

In order to extract the target patterns from the dependency analyses, we employ Perl scripts. Starting with all lexical verbs found in a sentence (such as *haben* in the example in Figure 1), these scripts collect the subject and objects (*Schimmer*), including modifying adjectives (*blassen*), and the prepositional phrases related to the initial verb. To accomplish this task, the extraction scripts refer to part of speech tags and morphological features, and to the dependency structure given in the second to last column of the Fspar output. While no other information is needed to identify the dependency patterns of interest, we still gather additional syntactic features for later use. All accessible morpho-syntactic information including the type of determiners, syntactic number features, and the presence of comparative forms (for adjectives) is collected at this point already for linguistic post-processing at a later stage (Section 5.2).

The extracted candidate items, consisting of the lemmas of the verb, objects, the subject, and prepositional phrases, are stored together with the available morpho-syntactic information in a PostgreSQL database (see Weller and Heid (2010) for details). The database entry thus obtained for the verb+object pair *Schimmer haben* in Figure 1 is shown in Table 2.

(a) Fspar output

0	Und	KON	und	|	2	ADJ
1	er	PPER	er	Nom:M:Sg	2	NP:1
2	hat	VVFIN	**haben**	3:Sg:Pres:Ind	-1	TOP
3	keinen	PIAT	kein	|	5	SPEC
4	blassen	ADJA	**blaß**	|	**5**	ADJ
5	Schimmer	NN	**Schimmer**	Akk:M:Sg	**2**	PCMP
6	,	$,	,	|	2	PUNCT
7	was	PRELS	was	Nom:N:Sg	11	NP:1
8	gerade	ADV	gerade	|	9|11	ADJ
9	vor	APPR	vor	Dat|Akk	11	ADJ
10	sich	PRF	er|sie|es	Dat|Akk	9	PCMP
11	geht	VVFIN	gehen	3:Sg:Pres:Ind*	2	ADJ
12	.	$.	.	|	-1	

(b) Tree representation

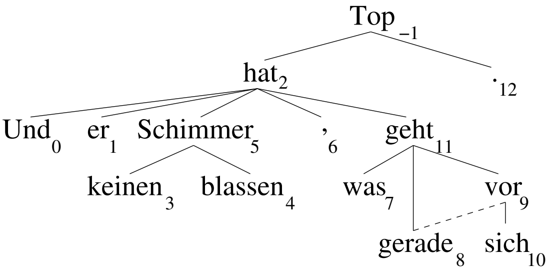

Figure 1: Dependency analysis of a sentence

The PostgreSQL database contains every dependency relation that the dependency parser makes available in its parse output and could be relevant to our NPI discovery procedure. With the database in hand it is now possible to work with patterns of varying form and length. In the present study, we choose to investigate five patterns: verb+object (NV), adjective+object+verb (ANV), preposition+noun+verb (PNV), noun+preposition+noun+verb (NPNV) and preposition+adjective+noun+verb (PANV). Examples for each of the patterns are shown in Table 3a; their occurrence frequencies can be seen in Table 3b. Since at this stage we have not yet done any checks regarding their occurrence inside or outside of NPI licensing contexts, the large majority of items are not NPIs. Most items are trivial combinations of words in the sense that they are not even MWEs but consist of free combinations of words that simply obey the general grammatical mechanisms of syntactic and semantic selection. Other items are statistical collocations with compositional semantics, and some are idiomatic expressions. Finally,

v_lem	subj	acc_obj	acc_obj_det	acc_obj_num	acc_obj_mod
haben	er	Schimmer	kein	sg	blaß

Table 2: Database entry for *Schimmer haben*

a few of our items are NPIs. In Table 3a we see one example of a compositionally constructed phrase (‘trivial combination’), one of an idiomatic expression and one of an NPI for each of the five patterns. Due to the fact that some patterns are subsets of others (ANV obviously forms a subset of NV), their respective candidates occur in the results of their super-patterns as well (e.g. *Faden verlieren* (‘lose one’s train of thoughts’), is in the NV result list, while the complete expression, *roten Faden verlieren*, is contained in the results for ANV[8]). The table also reflects the fact that we extract lemmas instead of word forms.

(a) Examples for investigated patterns

pattern	trivial combination, idiomatic expression, NPI
NV	*Frau danken*, *Rede halten*, *Hehl machen*
ANV	*sachlich Grund sehen*, *rot Faden verlieren*, *blass Schimmer haben*
PNV	*auf Agenda stehen*, *unter Druck setzen*, *über Herz bringen*
NPNV	*Herr für Rede danken*, *Wind aus Segel nehmen*, *Blatt vor Mund nehmen*
PANV	*zu neu Debatte führen*, *für bar Münze nehmen*, *mit recht Ding zugehen*

(b) Occurrence frequencies of the patterns

	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

Table 3: Overview of syntactic patterns

Having completed the extraction of all expressions from the German corpus that meet our five pre-defined syntactic dependency patterns, everything is set up for the identification of syntactically complex NPIs. As complex NPIs are (possibly idiomatic) collocations, this step will ultimately filter out trivial combinations, leaving collocations and idiomatic phrases in negative environments.

Before we describe the statistical processing and linguistic refinements for targeting idiomatic NPIs in Section 5, Section 4 discusses how we identify NPI licensing environments. This is another nontrivial task, as we assume that NPI licensing is effected by

[8] This particular example is even more intricate, since *Faden verlieren* might be considered an independent idiomatic expression with a meaning that differs slightly from *roten Faden verlieren* (‘losing the central idea’). In this case we could say that we did in fact find two idiomatic expressions.

a mixture of semantic, syntactic and pragmatic conditions which cannot be read off directly from the dependency information available in the extracted patterns.

4 Modelling Negative Contexts

Following the lead of Lichte (2005a) and Lichte and Soehn (2007), we identify the negative licensing contexts of multiword NPIs on the basis of certain syntactic configurations and a finite list of determiners, verbs, adverbs and other lexical elements. A few examples are listed in Figure 2.

sentential negation adverb	*nicht*
negative determiner	*kein*
nouns	*niemand, nichts*
adverbs	*kaum, nur, selten, wenig, ebensowenig, nie, niemals, nirgendwo, nirgends, nirgendwohin, nirgendwoher, keinesfalls, keineswegs*
inherently negative verbs	*ablehnen, anzweifeln, abstreiten, bestreiten, bezweifeln, dementieren, verhindern, verweigern, weigern*
negative conjunctions[9]	*ohne zu, ohne dass, ob, bevor*

Figure 2: A selection of lexical licensing contexts

Our extraction procedure comprises a component that recognizes negative contexts by the presence of at least one of our lexical or structural criteria for licensing contexts. Whenever an MWE occurs in such a context, that occurrence of the MWE is labelled with NEG, otherwise with NONEG. This format meets the requirements of the statistical association measures that are applied (Section 5) to distinguish multiword NPIs from other MWEs that might occasionally occur in a negative polar environment.

Let us consider four examples for licensing contexts we found for the NPI *alle Tassen im Schrank haben* (lit. 'to have all cups in the cupboard'; 'to have lost one's marbles'), which illustrate the wide variety of licensing possibilities found in corpora:

(10) Nicht **alle Tassen im Schrank** zu **haben** mag ja durchaus
not all cups in-the cupboard to have may PART indeed
produktiv sein für derlei Theater.
productive be for such theater
'Being somewhat insane may in fact be an advantage for this type of theater.'

[9] Recall that we use the word 'negative' loosely to designate environments which license NPIs. Here we mean to characterize subordinating conjunctions which license NPIs in the embedded clause.

(11) Kein Mörder, der **alle Tassen im Schrank hat**, würde mich
no murderer who all cups in-the cupboard has would me
umbringen.
kill
'No sane murderer would kill me.'

(12) ...sollte sich darüber hinaus allerdings fragen lassen, ob
...should himself moreover however ask let if
Vorstansdsmitglied P. S. noch **alle Tassen im Schrank hat**
board member P. S. still all cups in-the cupboard has
'...should seriously be wondering if board member P. S. has lost his marbles.'

(13) Jeder, der noch seine **fünf Tassen im Schrank hat**, weiß,
everybody who still his five cups in-the cupboard has knows
daß...
that...
'Any sane person knows that...'

In (10) the verb *haben* ('to have') is simply modified by the sentential negation adverb *nicht* ('not'), exemplifying the most straightforward case. Similarly, in (11), the subject noun phrase *kein Mörder* is a negative quantifier due to the determiner *kein* ('no'). In the construction in (12), the clause containing the expression *alle Tassen im Schrank haben* is an indirect question, which is a legitimate nonveridical licensing environment of the NPI. (13) is an instance of NPI licensing in the restrictor of a universal quantifier, in this case the nominal quantifier *jeder* ('everyone'). Restrictors of universal quantifiers are downward entailing, which is the most important semantic licensing condition. Replacing the universal with a proper noun or a definite noun phrase removes this semantic property and results in an ungrammatical utterance.[10]

The choice of lexical and structural indicators of negative environments for our extraction procedure is determined by two considerations: First, we use some of the lexical (and structural) licensers which CoDII lists. These elements and structural environments reflect the available linguistic knowledge in the NPI literature about licensing environments. Their practical usefulness for semi-automatic NPI extraction was confirmed by the results of Lichte and Soehn's work. Second, our choice of lexical and structural licensing environments is influenced by the type of structural information we expect to be available after running FSPAR. Here we follow our judgment about the reliability of the output of the dependency parser.

There are some obvious limitations to our selective and rather syntactic approach to modelling negative contexts. Since there are, in principle, infinitely many forms of valid licensing environments, it is impossible to define a syntactic pattern for every

[10] The substitution of *seine fünf* ('his five') for *alle* ('all') in the phrase *alle Tassen* in (13) is an instance of creative language use and does not change the (relevant aspects of the) meaning of the expression.

single one of them. The situation would become even more difficult if we decided to try to systematically detect cases in which a given pattern is not a licenser due to additional effects such as intervening quantifiers between a licenser and a potential licensee. This task would minimally presuppose some analysis of quantifier precedence conditions and would involve a closer investigation of the interplay between word order and syntactic structure. Moreover, some licensing environments are just not reliably identifiable without deep syntactic or semantic analysis. Examples in German are extraposed relative clauses (which might be in a downward entailing environment depending on the noun phrase they attach to), comparative clauses with adjectives plus *als*-clause (which require a reliable semantic analysis), subjunctive clauses, and opaque conditionals of the form *You say anything, and I kill you* (with *anything* being an NPI licensed by the conditional construction).[11] Our working assumption is that our model captures a sufficiently large portion of NPI licensing environments to produce good enough candidate lists. As long as we recognize enough actually occurring licensing environments and do not miss too many, and as long as the corpus is large enough, the statistical analysis should be able to cope with the noise caused by the lack of sophisticated semantic annotation.[12]

5 Optimization

At this point of our procedure, we have extracted a very large number of NPI candidates. The figures in Table 3 show that this is not a list that a human annotator could effectively work with to identify true NPIs. Amongst the items in the list are valid NPIs and other idiomatic multiword constructions, but the vast majority are trivial combinations of words, i.e. syntactically regular and semantically transparent constructions such as *auf Stuhl setzen* ('on chair sit'; 'to sit down on a chair'). Some of them might have an accidental high co-occurrence ratio with negative contexts in our corpus, and it is important to face the fact that there is no automatic procedure to validate NPIs. Manual annotation remains an indispensable step for our extraction method. A native speaker has to check if the use of a candidate expression without a negative context always leads to ungrammaticality. The question to decide is whether it is categorically impossible to use a candidate expression felicitously (under constant meaning) without a licensing context. Even strong statistical tendencies in large corpora cannot guarantee that this is the case for a given expression. Especially for idiomatic NPI candidates that permit a related literal meaning it can even be very hard for a native speaker to verify

[11] This list of difficult cases is taken from a slide presentation by Timm Lichte.

[12] As a reviewer succinctly remarks, our considerations here are *full of speculative assumptions*. In the presumed absence of an even remotely complete list of NPIs in German (or any other language), and confronted with a complete lack of the type of deep semantic analysis of large text collections that we would need to be able to identify all possible semantic licensing environments known from the linguistic literature, the only justification of our optimistic tone is the actual success of the method, measured by the number of new German NPIs that we find. There is much room for improvement.

that the idiomatic reading strictly requires a licensing context, because this fact might be concealed by the literal reading, which does not.

These complications aside, the key to success for semi-automatic NPI extraction is that some features that are characteristic for NPIs such as their significant co-occurrence with the licensing contexts described in Section 4, and the syntactic fixedness of idiomatic expressions can be automatically accessed. In the following sections, we describe how we used some of these features to create a list of manageable size with an enhanced number of valid NPIs at the top of the list by sorting candidates according to associative strength with their respective negative contexts and with linguistic features (morpho-syntactic fixedness or translational behavior). This preprocessing considerably reduces the necessary but time-consuming manual annotation efforts, and makes human annotation feasible.

5.1 Statistical Processing

A number of statistical association measures such as *log likelihood ratio* or *t-score* have been successfully applied to identify MWEs (see e.g. Evert (2004)). They indicate the associative strength of a word pair by taking into account the observed vs. expected frequencies of pairs and of their components in isolation. Assuming that NPIs are significantly associated with their negative context, we compute the associative strength between each MWE and its context label (which is NEG for negative contexts, and NONEG otherwise) to determine whether a negative context is obligatory for an expression. An example pair is: (*blassen::Schimmer::haben*, NEG).

We used the UCS toolkit[13] to calculate five standard association measures for each of our five candidate lists (cf. Table 3, with each candidate represented in lemma form). These lists were then sorted in decreasing order according to the resulting scores. Finally, the 500 highest scoring candidates with a strong statistical tendency to be associated with a NEG context label were manually annotated: '+' for valid NPIs and '−' for MWEs or trivial combinations. Since longer patterns are extensions of shorter patterns, there is a certain overlap between the items we find in longer and shorter patterns.

[13] UCS toolkit: www.collocations.de (Evert, 2004)

NPIs in top 500	NV	ANV	PNV	NPNV	PANV
log-likelihood	21	74	28	5	4
t-score	16	65	21	5	4
z-score	21	76	29	5	4
poisson	29	77	31	5	4
chi-squared	21	76	30	5	4

Table 4: NPIs found for each of the syntactic patterns when sorted according to standard association measures

The numbers of valid NPIs found amongst the top 500 candidates can be seen in Table 4. Even though *poisson* slightly outperforms the other measures, all results turned out to be quite similar. Furthermore, we also found that the NPIs were often the same: All 16 NPIs of the category NV found in the *t-score* sorting are a subset of those found by *log-likelihood*, *z-score* and *chi-squared*, while all 21 NPIs found by the latter ones are contained in the results for *poisson*. Similar observations were made for the other syntactic patterns.

(a)

	NPNV-pattern with negative context	f	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	868
(+)	**Hehl** aus Abneigung **machen**	28	15	17	868

(b)

	PNV-pattern with negative context	f	pos (POISSON)	pos (f)
+	**aus Staunen herauskommen**	60	48	8998
+	**über Weg trauen**	91	51	6941
+	**mit Wimper zucken**	26	289	33412

Table 5: Samples of log-likelihood orderings for two patterns: (a) NPNV: poisson and log-likelihood and (b) PNV

Table 5a shows the top 15 entries of the NPNV pattern that are labelled with NEG. The candidates are ordered according to their *poisson* scores. The first column contains the manual annotation that reflects the judgment of the human annotators whether or not the expression is an NPI (+/−). The entries marked with '(+)' would be complete with only one noun, and therefore belong to the NV class rather than NPNV. Conversely, there are patterns containing candidates that are not yet complete. The absolute frequency of the NPI candidates is indicated in column 3 (labeled *f*) while the last columns give the ranks of each expression according to different association measures (*poisson*, *log-likelihood* and *frequency* ordering, respectively). Note that the ranks obtained by the *poisson* method and *log-likelihood* do not differ substantially.

Since most NPIs are relatively infrequent, they would be hard to find in a list sorted by absolute frequency.[14] Sorting according to association measures moves NPIs towards the top of the list, as candidates that hardly ever occur in a non-negative context are considered to be highly associated with their negative context. Table 5b illustrates the huge differences between ranking positions of NPIs in the two different lists for three selected NPIs.

5.2 Linguistic Processing

In order to further improve the ordering of the lists, we add more linguistic knowledge to the statistical method. We enriched our result lists with the following linguistically motivated scores:

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

The nature and function of these scores will now be explained one by one. First of all, we use the percentage of the candidates' negative occurrences (#NEG) as a possible indicator for NPIs in our extraction process (cf. Table 6).

	NPI candidate	contexts		freq.	#NEG
+	**aus Kopf gehen**	NEG: 47	NONEG: 0	47	100%
+	**Wald vor Baum sehen**	NEG: 46	NONEG: 4	50	92%
+	**von Fleck kommen**	NEG: 111	NONEG: 14	125	88.8%
-	zu Schaden kommen	NEG: 247	NONEG: 198	445	55.3%

Table 6: Illustration of #NEG score calculation

The morpho-syntactic fixedness score (FIX) is motivated by previous work on the extraction of idiomatic MWEs (Bannard, 2007). Since many multiword NPIs have properties similar to idiomatic expressions, we expect them to be syntactically frozen to a certain degree, which means that they should not permit the usual morphological range of variation of the noun with respect to syntactic features like number, or the use

[14] This is different from the task of retrieving MWEs in general, for which ordering a list of patterns by frequency can already lead to good results. The use of association measures can then further improve initial results.

of all syntactically compatible determiners. Recall that during the extraction of the list of potential candidates, information on the nouns' number and their accompanying determiner is retrieved and stored. For each candidate, we compute the frequency distribution of the number values (sg, pl) and possible determiners (e.g. def, indef, none). The highest percentages of both categories are taken to represent the candidate's preferences. In the case of pnv triples, we also measure the distance of verb, noun and preposition, as idiomatic pnv triples tend to be most often (immediately) adjacent.

The fix score is calculated for each npi candidate based on the average of

(i) the #Neg score,

(ii) the percentage of number and article setting, and

(iii) in case of pnv triples: the averaged adjacency-scores.

In order to approximate the semantics of npi candidates, we use translational entropy (te) and the proportion of default alignments (pda). Both scores rely on the assumption that multiword npis have a non-compositional semantics, which means that they are to be translated as a whole while compositional combinations of the same syntactic form would exhibit literal translations of their components. To model the translations, we take word equivalences from the Europarl corpus (Koehn, 2005). Roughly speaking, the te score indicates the degree of diversity in a word's translation, while the pda expresses the percentage of literal (or default) translations. Descriptions of the these two scores can be found in Villada Moirón and Tiedemann (2006).

The linguistic scores are used as follows: We take the top 500 of the lists ordered by *poisson* and re-order these lists according to each of the linguistic scores. In order to measure the quality of the different orderings, we use the uninterpolated average precision (uap, see Manning and Schütze (1999) for details). Figure 3 shows the results for selected syntactic patterns. Note that for the te and pda values, we could only use the Europarl corpus (30 million words). As a consequence, some of the npi candidates cannot be assigned either score (te or pda), and are thus skipped in the calculation. The rightmost column contains the resulting uap value when sorted according to a combination of morpho-syntactic fixedness and translational behavior.[15]

sorted by	poisson	Neg	fix	te	pda	te+pda+fix
nv	0.105	0.069	0.121	0.1	0.124	0.157
anv	0.233	0.26	0.212	0.174	0.165	0.307
pnv	0.118	0.125	0.145	0.103	0.163	0.2

Figure 3: uap scores for re-orderings of top 500 (*poisson*)

For the npi candidates of all three patterns, the orderings according to the linguistic score based on both (monolingual) morpho-syntactic and multilingual features outperform the respective *poisson* orderings. The morpho-syntactic and translational features

[15] The maximal uap index for a perfectly ordered list would be 1.

are independent and thus benefit from each other when combined. While we achieved our goal to enhance the sorting quality of the candidate lists, the improvement is not great. This may be mainly due to the fact that most NPIs in the lists are relatively low-frequent. The TE and PDA score are not designed for low-frequent data, and computing morpho-syntactic preferences is known to work better for high-frequent data as well.

5.3 Remaining Challenges

There are many expressions that collocate with negation but are not grammatically dependent on it. This is partially due to the nature of newspaper text: For the NPNV-triple *Zugang zu Trinkwasser haben* ('to have access to potable water') we found 29 occurrences all of which appear in a negative context. This is straightforward to explain if we consider that we do not expect a journalist to write about existing access to potable water.

Another obstacle for the statistical approach are contexts that we still cannot model reliably, as well as creative use of language: The NPI *Wald vor Baum sehen* (lit. 'not to see the forest for the trees'; 'not to see the obvious') (contained in Table 5 and in Table 6) occurred 46 of 50 times in a straightforwardly negative context. The complete expression, as it might be listed in a dictionary, is *den Wald vor lauter Bäumen nicht sehen*. In this basic citation form the verb is negated by the sentential negation adverb, *nicht*. Interestingly, this is the form that we observe in 46 cases.

The remaining four occurrences are more difficult: The first, a question (14), is a known nonveridical licensing environment (which we modelled), while the second and third occurrences are a modal context (15) and a conditional clause (16), which are not among the contexts we modeled. In the last sentence, however, there is no clear licensing context at all (17). Regardless of the lack of an obvious negative licensing environment, the sentence is well-formed.[16]

(14) Sieht er dann den Wald vor lauter Bäumen?
see he then the forest despite all trees
'Does he see the obvious?'

(15) Doch wie immer sollte man zunächst einmal den Wald vor Bäumen sehen.
but as always should one first the forest despite trees see
'As always one should first note the obvious.'

[16] One might want to speculate that the well-formedness of (17) is contingent on the existence of a presupposition denying that M. L. manages to see the obvious. It has been observed before in the literature that at least some NPIs tolerate this type of indirect and possibly contextual licensing (Richter and Soehn, 2006, pp. 338–339).

(16) Hätte die Kommission eindeutige und anerkannte Prioritäten und könnte
had the commission clear and recognized priorities and could
sie den Wald vor Bäumen sehen, hätten wir nicht diese Aussprache
it the forest despite trees see, had we not this meeting
heute Nachmittag.
today afternoon

‘If the commission had clear and recognized priorities and if it could see the obvious, we would not have to meet this afternoon.’

(17) Manchmal sieht M. L. vor lauter Bäumen dennoch den Wald.
sometimes sees M. L. despite all trees still the forest

‘Occasionally M. L. does manage to see the obvious.’

It is necessary to keep in mind that for the recognition of each type of negative context, a syntactic pattern of this context—be it a question, some form of conditional or a preceding inherently negative verb—has to be specifically implemented. The examples above illustrate negative contexts that are not easy to detect automatically. As shown in (17), in some cases we might even find constructions with clear NPIs that are used in contexts which cannot be easily categorized as being negative.

6 Results and Discussion

CoDII, the largest collection of German NPIs, comprises 165 entries. Subtracting duplicates that occur in different extraction patterns, our method retrieved 141 NPIs. 25 of these are in CoDII, 116 are new.[17] To appreciate the effectiveness of our method, consider a ‘normal-sized’ list such as John Lawler’s collection of English NPIs[18], which is meant to be an exhaustive listing for English and comprises roughly three dozen entries. Jack Hoeksema’s collection of Dutch expressions with strong association to negative environments, which is by far the largest known collection of NPI-like items and has been developed for 15 years, reportedly contains 670 entries.[19] However, Hoeksema’s collection is not limited to grammatical NPIs in the narrow sense, i.e. it is not restricted to expressions that are perceived as ungrammatical by native speakers when presented outside of an appropriately negative context. Beyond such expressions, Hoeksema also collects expressions that are statistically strongly associated with negation, which means that they tend not to occur outside of a negative context, although they would still be perceived grammatical if they did.

Lichte and Soehn (2007) do not report how many NPIs their method found. They say that they retrieved 112 items from Kürschner’s list of 344 items. However, according to them Kürschner’s list contains about 200 pseudo-NPIs, i.e. items which exhibit a

[17] The complete list, including information about which of the retrieved items are in CoDII, is in the appendix.

[18] www-personal.umich.edu/∼jlawler/NPIs.pdf (September 2010)

[19] www.let.rug.nl/∼hoeksema/lexicon_bestanden/v3_document.htm (retrieved in September 2010)

high collocational association with negation but can still occur felicitously in contexts without negation (which makes the empirical scope of Kürschner's list comparable to Hoeksema's). Given that Lichte and Soehn's extraction algorithm primarily targeted single-word NPIs, and that all NPIs that they identified are in CoDII, the overlap between the items they extracted and ours must be small (equal to or below 25).

The main reason for the small overlap should come from the different strategies of selecting multiword NPI candidates. Our procedure is based on syntactically meaningful patterns which enter statistical processing as a whole. Lichte's extraction procedure starts with single lemmata which are extended one lemma at a time, and only those chains of length $n + 1$ that exhibit higher association to negation than the nucleus chain of length n are further considered. Apart from the computational inefficiency of considering arbitrary other lemmata (in the same clause) for extending a lemma chain, this procedure can only suggest as candidates those expressions of length m such that a sequence of of chains exists where each succeeding chain has a higher association to negation than its shorter predecessor. The existence of such a sequence of chains cannot be guaranteed for each multiword NPI, and if it does not exist, Lichte's procedure will not find the NPI.

The types of NPIs found with the three most successful search patterns, NV (29), PNV (31), and ANV (77) show interesting differences. The PNV list contains a high number of idiomatic expressions (*(mit etw.) hinter dem Berg halten, (sich) in die Karten schauen (lassen), auf den Mund gefallen (sein)*), and only a small number of semantically transparent MWEs (*mit Vorwürfen sparen, mit keinem Wort erwähnen*). The list NV is similar, containing a somewhat smaller but still sizable number of non-decomposable idioms. Finally, the third list, ANV, is markedly different, containing mostly non-idiomatic, semantically transparent MWEs such as *wesentliche Änderungen erwarten, (sich) einen anderen Rat wissen, eine andere Wahl sehen*, and a smaller number of clearly idiomatic expressions (*einen blassen Schimmer haben, schlafende Hunde wecken*). These differences between the lists could explain the varying success with reordering the top 500 by taking linguistic knowledge about the fixedness of expressions into account. The more idiomatic an expression is, the more restricted is its syntactic flexibility, and our linguistic processing favors syntactically frozen expressions. The NPIs found in the PNV pattern should thus be promoted more than the ones in the ANV pattern. A last, more general observation concerns the tendency in the ANV pattern (also visible in the PNV pattern) of expressions forming clusters which only differ in the verb, such as *geringsten Zweifel {aufkommen / geben / haben / hegen / lassen}*. From a theoretical point of view, one might be tempted to consider these items variants of one and the same underlying NPI, and it might be interesting to investigate the types of verbs that can enter into these variants, and their properties.

Our success quota of finding true NPIs in five top 500 lists (141 in 2500) clearly shows that we have not designed a fully automatic NPI extraction procedure. There is considerable human effort and expert judgment involved in finding NPIs in candidate lists. This state of affairs echoes the early apprehensions by Hoeksema (1997), who feared that difficulties such as those arising from inaccurate recognition of licensing environments or from polysemy of words and ambiguity of constructions could defeat automatizing NPI detection. Looking at the experience gathered with the two candidate extraction procedures that have been designed in the meantime, we agree that what we have achieved is probably very coarse-grained and might even suffer from inherent shortcomings that cannot be completely overcome by simply pursuing the same strategy further. Despite these imperfections, the methods that we applied are still highly successful insofar as they contribute dramatically to improving our database of German NPIs. As this is still only a beginning, we can hope for many more NPIs to be found by considering other promising syntactic patterns and by refining the candidate extraction procedure. In particular, we only applied the linguistic processing methods of Section 5.2 to candidate lists after they were annotated by human experts. Since linguistic processing improved the rankings in these pre-sorted candidate lists, it should be checked if their application at an earlier stage in the extraction pipeline improves the candidate lists given to human annotators.

7 Conclusion and Future Work

As we mentioned several times, many NPI licensing environments exhibit the logical property of being downward entailing, which means that they support inferences from supersets to subsets (see the example in Section 2). For this reason, detecting downward entailing environments is highly relevant for determining textual entailments. In a recent paper, Danescu-Niculescu-Mizil et al. (2009) exploit the licensing requirements of NPIs and use a set of English NPIs to extract downward-entailing operators from text. In a sense, this is the converse task to ours, but it presupposes a lexicon of NPIs. Knowledge of a larger set of NPIs in a given language, as provided by our method, should help improve extraction of downward-entailing operators, and may thus ultimately contribute to improving textual entailment tasks in language processing.

We showed that by sorting candidate-context pairs according to their log-likelihood scores, NPIs could be retrieved with considerable precision. In a second step, we applied linguistically motivated scores in order to enhance sorting quality for the top 500 entries of the log-likelihood sorting. We saw that our results were very promising, as we managed to increase the number of known NPIs in German by more than two thirds. However, we also believe that there is still much room for improvement by integrating linguistic knowledge and statistical processing more tightly. With a more fine-grained definition of negative contexts, as provided by the linguistic literature, we would hope to obtain better candidate lists.

Looking at our results from the perspective of theoretical linguistics, there should be much to gain from semi-automatic NPI extraction methods. Many questions about the syntactic, semantic and pragmatic nature of NPIs and their licensing environments are still open. Having a much larger empirical base for investigating these issues should contribute significantly to improving the linguistic theory. For example, one major hypothesis of pragmatic NPI theories claims that their behavior can be attributed to their property of being minimizers. This seems hard to maintain considering that many items of the joined item list consisting of CoDII and our newly extracted items are not end-of-scale elements in any obvious way. Expressions such as *jemandem (nicht) grün sein* (lit. 'so. (not) green be'; '(not) to be well-disposed toward someone'), *jemandem (nicht) über den Weg trauen* (lit. 'so. (not) over the path trust'; '(not) to trust someone') or *jemandem (nicht) von der Seite weichen* (lit. 'so. (not) from the side leave'; 'to tag along after someone') are not at an endpoint of any easily imaginable scale; many similar examples can be found by simply going through the list. Any claim to universality of theories that explain NPIs from their supposed property of being minimizers seems to be doomed considering the full range of data.

Having a large repository of NPIs also opens up new avenues for psycholinguistic research. Among the problems of researching properties of NPIs such as the distinction between *weak* and *strong* NPIs (i.e. between those items that are satisfied with weaker forms of negation in their licensing contexts and those which require stronger forms of negation) has been the diversity of syntactic categories and syntactic form of multiword NPIs. In psycholinguistic experiments we typically want to vary exactly one feature under investigation to make sure that other variation does not interfere or mask those effects that we want to study. With only a dozen or two NPIs, it is very hard or impossible to construct enough items for an experiment which only vary in one dimension. The kind of NPI database that we have now compiled makes it much easier to address the kinds of questions that psycholinguists and linguists might want to ask about the nature of NPIs, because we are now in a much better position to construct syntactically more uniform item sets for experiments. Richter and Radó (2010) have already used items from our extraction procedure in the study of demonstrating the psycholinguistic reality of the weak/strong classification, the behavior of strong and weak NPIs in Neg-Raising contexts (Horn, 1978), and so-called intervention effects of proportional quantifiers in Neg-Raising constructions with NPI licensing. These experiments would not have been possible without a large resource of syntactically similar NPIs. For this very reason, corresponding experiments can currently not be conducted for English.

Our success with finding many previously unobserved NPIs among five patterns of MWEs supports our initial intuition of a deeper relationship between the property of being an NPI and idiomatic expressions. In normal idiomatic expressions there is a strong association between the set of words that make up the idiomatic expressions, whereas multiword NPIs additionally exhibit a strong association to a more abstract grammatical feature, *viz.* (various degrees of) negation. Investigating the relationship between these apparently distinct types of grammatical association might reveal interesting, hitherto unnoticed properties of idiomatic expressions, and might lead to a re-evaluation of the function of NPIs in the grammatical system.[20]

[20] The theoretical implications of these considerations are pursued further in Richter et al. (2010).

Acknowledgments

We are very grateful to Manfred Sailer for his generous help in checking our lists of NPI candidates for true NPIs, and to Timm Lichte for giving us very useful background information on his work on NPI extraction, including the original Perl script of the NPI extraction algorithm of Lichte and Soehn (2007) with his complete definition of licensing environments. Thank you to Janina Radó for proofreading.

References

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the ACL Workshop on a broader perspective on multiword expressions*, pages 1–8, Prague, Czech Republic.

Chierchia, G. (2006). Broaden your views. Implicatures of domain widening and the 'logicality' of language. *Linguistic Inquiry*, 37(4):535–590.

Danescu-Niculescu-Mizil, C., Lee, L., and Ducott, R. (2009). Without a 'doubt'? Unsupervised discovery of downward-entailing operators. In *Proceedings of NAACL HLT*, pages 137–145.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Fauconnier, G. (1975). Pragmatic scales and logical structure. *Linguistic Inquiry*, 6(3):353–375.

Fritzinger, F. (2009). Using parallel text for the extraction of German multiword expressions. *Lexis - E-journal in English Lexicology*, 4.

Fritzinger, F. and Heid, U. (2009). Automatic grouping of morphologically related collocations. In *Online Proceedings of the Corpus Linguistics Conference 2009*, Liverpool/UK.

Fritzinger, F., Richter, F., and Weller, M. (2010). Pattern-based extraction of negative polarity items from dependency-parsed text. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*. European Language Resources Association.

Hoeksema, J. (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: www.let.rug.nl/hoeksema/docs/barcelona.html.

Horn, L. R. (1978). Remarks on Neg-Raising. In Cole, P., editor, *Pragmatics*, volume 9 of *Syntax and Semantics*, pages 129–220. Academic Press, New York, San Francisco, London.

Kadmon, N. and Landman, F. (1993). Any. *Linguistics and Philosophy*, 16:353–422.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th MT Summit 2005*, Phuket, Thailand.

Krifka, M. (1995). The semantics and pragmatics of polarity items. *Linguistic Analysis*, 25:209–257.

Kürschner, W. (1983). *Studien zur Negation im Deutschen*. Gunter Narr, Tübingen.

Ladusaw, W. A. (1980). On the notion 'affective' in the analysis of negative-polarity items. *Journal of Linguistic Research*, 1(2):1–16.

Lichte, T. (2005a). Corpus-based acquisition of complex negative polarity items. In Gervain, J., editor, *Proceedings of the Tenth ESSLLI Student Session*, Edinburgh. Heriot-Watt University.

Lichte, T. (2005b). Korpusbasierte Acquirierung negativ-polärer Elemente. Master's thesis, Seminar für Sprachwissenschaft, University of Tübingen.

Lichte, T. and Soehn, J.-P. (2007). The retrieval and classification of negative polarity items using statistical profiles. In Featherston, S. and Sternefeld, W., editors, *Roots: Linguistics in Search of its Evidential Base*, pages 249–266. Mouton de Gruyter, Berlin.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Richter, F. and Radó, J. (2010). NPI licensing in German: Some experimental results. Unpublished manuscript. Eberhard Karls Universität Tübingen. October 2010.

Richter, F., Sailer, M., and Trawiński, B. (2010). The collection of distributionally idiosyncratic items: An interface between data and theory. In Ptashnyk, S., Hallsteinsdóttir, E., and Bubenhofer, N., editors, *Korpora, Web und Datenbanken. Computergestützte Methoden in der modernen Phraseologie und Lexikographie*, volume 25 of *Phraseologie und Parömiologie*, pages 247–261. Schneider Verlag Hohengehren GmbH.

Richter, F. and Soehn, J.-P. (2006). 'Braucht niemanden zu scheren': A survey of NPI licensing in German. In Müller, S., editor, *Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, pages 421–440. CSLI Publications.

Schiehlen, M. (2003). A cascaded finite-state parser for German. In *Proceedings of the 10th EACL*, Budapest, Hungary.

Trawiński, B., Soehn, J.-P., Sailer, M., and Richter, F. (2008). A multilingual electronic database of distributionally idiosyncratic items. In Bernal, E. and DeCesaris, J., editors, *Proceedings of the XIII Euralex International Congress*, volume 20 of *Activitats*, pages 1445–1451, Barcelona, Spain. Universitat Pompeu Fabra.

van der Wouden, T. (1997). *Negative Contexts. Collocation, Polarity and Multiple Negation*. Routledge, London.

Villada Moirón, B. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL 2006 Workshop on multiword-expressions in a multilingual context*, Trento, Italy.

Weller, M. and Heid, U. (2010). Extraction of German multi-word expressions from parsed corpora using context features. In *Proceedings of the Linguistic Resources and Evaluation Conference, LREC 2010*, Valetta, Malta.

Zwarts, F. (1995). Nonveridical contexts. *Linguistic Analysis*, 25:286–312.

Zwarts, F. (1997). Three types of polarity. In Hamm, F. and Hinrichs, E. W., editors, *Plurality and Quantification*, pages 177–237. Kluwer Academic Publishers, Dordrecht.

Appendix

This appendix lists all NPIs that were extracted from the corpora in Table 1 and confirmed by human annotators of the candidate lists. They are sorted according to our five syntactic patterns. For each NPI, the tables show if it is already contained in the CoDII collection ('+'), or if it is a new NPI not noted there ('–'). The annotation '(+)' marks partial NPIs, i.e. expressions that are not complete NPIs yet but can be recognized as parts of true NPIs contained in CoDII. For example, the PNV pattern contains *über Tatsache hinwegtäuschen*, which expands to the NPI *über Tatsache hinwegtäuschen können*. With adjectives, 'C' denotes comparative morphology, and 'S' denotes superlative forms.

PNV	in CODII
vor Anfrage retten	–
vor Auftrag retten	–
hinter Berg halten	–
mit Ding zugehen	–
von Eltern sein	(+)
von Fleck kommen	–
in Haut stecken	–
in Karte schauen	(+)
aus Kopf gehen	–
mit Kritik sparen	–
in Moment sagen	–
auf Mund fallen	(+)
vor Mund nehmen	(+)
hinter Ofen hervorlocken	–
an Schlaf denken	–
von Seite weichen	+
mit Silbe erwähnen	–
aus Staunen herauskommen	+
von Stelle kommen	–
auf Stuhl halten	–
über Tatsache hinwegtäuschen	(+)
in Traum denken	+
in Traum einfallen	–
ohne Tücke sein	–
mit Vorwurf sparen	–
über Weg trauen	–
in Weise entsprechen	–
in Weise rechen	–
an Wiege singen	–
mit Wimper zucken	+
mit Wort erwähnen	–

NV	in CODII
Abbruch tun	+
Ahnung haben	+
Aufschub dulden	–
Berührungsangst kennen	–
Blumentopf gewinnen	+
Erbarmen kennen	–
Finger rühren	+
Haar krümmen	+
Haar lassen	(+)
Halten geben	–
Hauch haben	–
Hehl machen	+
Kosten scheuen	(+)
Mördergrube machen	(+)
Mühe scheuen	(+)
Pfennig erhalten	–
Pfennig haben	–
Pfennig sehen	–
Pfennig zahlen	–
Pfifferling geben	–
Rede sein	(+)
Sekunde zweifeln	–
Stein lassen	(+)
Tabu kennen	–
Träne nachweinen	–
Welt verstehen	(+)
Wort glauben	–
Wort verlieren	–
Wort verstehen	–

ANV	CODII
geringS Abstrich machen	−
geringS Ahnung haben	+
leisS Ahnung haben	+
geringS Anhaltspunkt geben	−
geringS Anlass geben	−
ganz Aufregung verstehen	−
weitC Aufschub dulden	−
ander Ausweg lassen	−
ander Ausweg sehen	−
ander Ausweg wissen	−
nennenW Auswirkung haben	−
recht Bezug finden	−
ander Chance haben	−
ander Chance sehen	−
geringS Chance haben	−
gewiß Charme absprechen	−
blass Dunst haben	−
gut Faden lassen	−
geringS Einfluss haben	−
nennenW Einfluss haben	−
recht Freude haben	−
groß Gedanke machen	−
geringS Grund sehen	−
einzig Grund geben	−
einzig Grund haben	−
erkennenB Grund geben	−
geringS Grund geben	−
zwingend Grund geben	−
gut Haar lassen	+
schlafende Hund wecken	−
geringS Hinweis finden	−
geringS Hinweis geben	−
groß Illusion machen	−
geringS Interesse haben	−
sonderlich Interesse haben	−
gut Licht werfen	−
geringS Lust haben	−
recht Lust haben	−
ander Möglichkeit sehen	−
geringS Problem haben	−
ander Rat wissen	−
recht Reim machen	−
gering Rolle spielen	−
nennenW Rolle spielen	−
gutS Ruf haben	−
gutS Ruf genießen	−
halb Sache machen	−
blass Schimmer haben	+
ganz Schritt halten	−
klein Seitenhieb verkneifen	−
recht Sinn ergeben	−
recht Spaß machen	−
groß Sprung erlauben	−
groß Sprung machen	−
groß Sprung zulassen	−
leicht Stand haben	−
gutS Tag erwischen	−
nennenW Unterschied geben	−
nennenW Veränderung erwarten	−
geringS Verständnis haben	−
ander Wahl bleiben	+
ander Wahl geben	−
ander Wahl haben	−
ander Wahl lassen	−
ander Wahl sehen	−
gut Wille absprechen	−
eigen Wort verstehen	−
einzig Wort verlieren	−
einzig Wort verstehen	−
groß Wort verlieren	−
weitC Wort verlieren	−
geringS Zweifel aufkommen	−
geringS Zweifel geben	−
geringS Zweifel haben	−
geringS Zweifel hegen	−
geringS Zweifel lassen	−
leisS Zweifel hegen	−

NPNV	in CODII
Anspruch auf Vollständigkeit erheben	−
Blatt vor Mund nehmen	+
Mördergrube aus Herz machen	+
Wald vor Baum sehen	−
Zweifel an Haltung lassen	−

PANV	in CODII
über mangelnde Arbeit beklagen	−
von schlecht Eltern sein	+
in kühnS Traum erwarten	−
auf grün Zweig kommen	+

For the lists of items above, it is important to note that many of the newly found items are partial NPIs: For example, in the PNV pattern, the verb *können* has to be appended to each of the first two items (*vor Anfrage retten, vor Auftrag retten*) and *mit Ding zugehen* has to be extended to *mit recht Ding zugehen* in order to obtain complete NPIs; among the first 4 items on this list only *hinter Berg halten* is already complete in the form in which it was extracted with the PNV pattern.

Most NPIs should be relatively easy to recognize from the form in which they occur on the lists. For *in Weise rechen* and *an Wiege singen* it might be somewhat harder to identify their citation forms, *in (k)einer Weise gerecht werden* and *an der Wiege gesungen sein.*

Examples of NPIs that occur partially in one list and completed in another are *von (schlecht) Eltern sein* (PNV, PANV), *(Blatt) vor Mund nehmen* (PNV, NPNV), *(gut) Haar lassen* (NV, ANV), and *Mördergrube (aus Herz) machen* (NV, NPNV). All four appear in CoDII. A special case are two partial items in the NV pattern, *Kosten scheuen* and *Mühen scheuen*, which combine to the complete NPI *Kosten und Mühen scheuen*, which belongs to a pattern that we were not investigating here.

The items *Ahnung haben, Aufschub dulden, Wort verlieren* and *Wort verstehen* in the NV pattern occur in extended forms in the ANV pattern, three of them with several extensions. We consider their NV forms independent complete NPIs, as they are found in corpora in these short forms. However, the longer forms also seem to be independent collocational units, which justifies listing them separately in larger patterns.