Pattern-Based Extraction of Negative Polarity Items from Dependency-Parsed Text

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

[1]Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik–
Azenbergstr. 12
D 70174 Stuttgart
fritzife@ims.uni-stuttgart.de
wellermn@ims.uni-stuttgart.de

[2]Universität Tübingen
Seminar für Sprachwissenschaft
Abteilung Computerlinguistik
Wilhelmstr. 19
D 72074 Tübingen
fr@sfs.uni-tuebingen.de

Abstract

We describe a new method for extracting Negative Polarity Item candidates (NPI candidates) from dependency-parsed German text corpora. Semi-automatic extraction of NPIs is a challenging task since NPIs do not have uniform categorical or other syntactic properties that could be used for detecting them; they occur as single words or as multi-word expressions of almost any syntactic category. Their defining property is of a semantic nature, they may only occur in the scope of negation and related semantic operators. In contrast to an earlier approach to NPI extraction from corpora, we specifically target multi-word expressions. Besides applying statistical methods to measure the co-occurrence of our candidate expressions with negative contexts, we also apply linguistic criteria in an attempt to determine to which degree they are idiomatic. Our method is evaluated by comparing the set of NPIs we found with the most comprehensive electronic list of German NPIs, which currently contains 165 entries. Our method retrieved 142 NPIs, 114 of which are new.

1. Introduction

The goal of the research presented here is to extend the set of known Negative Polarity Items (NPIs) in German, and to improve our understanding of their properties and their distribution in corpora. To this aim we develop a new method for extracting NPI candidates from corpora. The immediate goal is to expand the subcollection of 165 German NPIs in the electronic *Collection of Distributionally Idiosyncratic Items* (CoDII, (Trawiński et al., 2008)), compiled in a subproject of the former *Sonderforschungsbereich 441*.[1] Due to the striking frequency of multi-word 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 multi-word NPI candidates. We adapt an extraction pipeline that was previously successfully applied in the identification of multi-word expressions (MWEs, (Heid et al., 2008) 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. Our approach will be compared to results of the extraction algorithm in (Lichte and Soehn, 2007), which is the only other work on NPI extraction from corpora that we are aware of. However, in contrast to our method, the algorithm by Lichte and Soehn, in its basic form, targets single-word NPIs. They capture multi-word 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 methods for multi-word expressions, 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.

2. NPIs and NPI Licensing

NPIs are defined as single words or multi-word 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*. 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; van der Wouden, 1997) for details).

For the present research, we follow an idea applied in the

[1] www.sfb441.uni-tuebingen.de/a5/codii/

NPI extraction algorithm by Lichte and Soehn (2007) and exploit the fact that a finite set of particular lexemes (determiners, adverbs, 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. 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 most licensing environments sufficiently well to get good statistical results when using our heuristics in large corpora.

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 students* 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 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.

3. Extraction Methodology

3.1. Data

In order to avoid sparse data issues when extracting MWEs and in particular when focusing on the subset of NPIs (which occur rarely in everyday language), we need a very large text corpus to start from. An overview of the corpus collection we used is given in Table 1. It contains about 269 million words (tokens), including text from several German newspapers and the proceedings of the European parliament debates (Koehn, 2005) of which we used the German part for our monolingual processing, and the English, French and Swedish parts for multilingual processing (cf. Section 5.2. below).

name	size	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	–
Total:	269 million		

Table 1: Composition of the dataset used.

3.2. Multi-Word Extraction

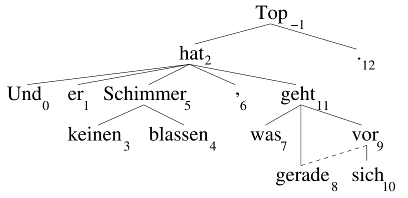
In German, the constituent words of multi-word constructions are not always adjacent to each other. The following example contains the NPI *(k)einen blassen Schimmer haben* (lit. '(not) to have a pale gleam': '(not) to have the faintest idea'):

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

Deep syntactic analysis is essential in order to reliably extract such discontinuous multi-word constructions. In the past, we successfully used the dependency parser FSPAR (Schiehlen, 2003) for different MWE extraction tasks. FSPAR is highly efficient and relies on a large lexicon. An example analysis of FSPAR is given in Figure 1, which 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 1(a) is not provided by the parser; we inserted it here in order to enhance readability of the example. The FSPAR output given in Figure 1(b) is to be read as follows (from left to right): sentence position, token, POS-tag, lemma, morphosyntactic information, dependency relation (numbers refer to sentence positions in the first row) and grammatical function.

(a) Tree representation



(b) FSPAR output presentation

```
0   Und       KON    und       |               2     ADJ
1   er        PPER   er        Nom:M:Sg        2     NP:1
2   hat       VAFIN  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
```

Figure 1: Dependency analysis example.

For the extraction task we apply PERL SCRIPTS that, starting from all lexical verbs found in a sentence (e.g. *haben* in Figure 1), collect the subject, objects (*Schimmer*) and modifying adjectives (*blassen*) and/or prepositional phrases related to the initial verb. The extraction scripts make use of part of speech tags, morphological features and the dependency structure given in the second to last column. Furthermore, all accessible morpho-syntactic information such as the type of determiner, syntactic number, and comparative forms (for adjectives) is collected.

These features will be used in linguistic post-processing as described in Section 5.2. below. All extracted items, i.e. the lemmas of objects, subject or prepositional phrases, are stored together with their 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* from Figure 1 is shown in Table 2.

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*.

Having stored every dependency relation relevant to our task listed in the parse output, we can work with patterns of varying form and length. In this study, we investigate the following 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 given in Table 3(a), their occurrence frequencies in Table 3(b). Due to the fact that some patterns are subsets of others (e.g. ANV is a subset of NV), their respective candidates occur in the results of their super-patterns as well (e.g. *Faden verlieren*, is part of the NV result list, while the complete expression, *roten Faden verlieren*, is contained in the results for ANV).

(a) Examples for investigated patterns

pattern	trivial, idiomatic,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.

4. Modelling Negative Contexts

Following the lead of (Lichte and Soehn, 2007), we identify the negative licensing contexts of multi-word NPIs on the basis of a finite list of determiners, verbs, adverbs and other elements (and syntactic structures) occurring with the NPI. A few examples are listed below:

overt negation	sent. adverb: *nicht*, determiner: *kein*
nouns	*niemand, nichts*
adverbs	*kaum, selten, nur, wenig, ebensowenig, nie, niemals, nirgendwo, nirgends, nirgendwohin, nirgendwoher, keinesfalls, keineswegs*
inherently negative verbs	*bezweifeln, anzweifeln, abstreiten, bestreiten, verhindern, weigern, verweigern, ablehnen, dementieren*

The extraction method 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, the respective 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 multi-word NPIs from regular, occasionally negative constructions.

In the following, we give 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:

1) Nicht **alle Tassen im Schrank** zu **haben** mag ja durchaus produktiv sein für derlei Theater.
'not to have all cups in the cupboard might be quite productive for this sort of fuss.'

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

3) ...sollte sich darüber hinaus allerdings fragen lassen, ob Vorstansdsmitglied P.S. noch **alle Tassen im Schrank hat**
'and besides (he) should be asked if the member of the executive board, P.S., still has all cups in the cupboard.'

4) Jeder, der noch **seine fünf Tassen im Schrank hat**, weiß, daß ...
'Everybody who still has his five cups in the cupboard, knows that ... '

In (1), the verb *haben* ('to have') is simply modified by the sentential negation adverb *nicht* ('not'), exemplifying the most straightforward case. Similarly, in (2), the subject noun phrase is a negative quantifier due to the determiner *kein* ('no'). In the construction in (3), the clause containing the expression *alle Tassen im Schrank haben* is an indirect question, which is a legitimate nonveridical licensing environment of the NPI. (4) 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.[2]

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 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 depending on word order. 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, 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).[3] Our working assumption is that our model

[2] The substitution of *seine fünf* ('his five') for *alle* ('all') in the phrase *alle Tassen* in (4) is an instance of creative language use and does not change the meaning of the expression.

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

captures a sufficiently large portion of NPI licensing environments to produce good enough candidate lists.

5. Optimization

At this point of our procedure, we have extracted a huge number of potential NPI candidates (cf. Table 3). Amongst these are valid NPIs and other idiomatic multi-word constructions, but the vast majority are trivial combinations, i.e. they are semantically transparent constructions such as *auf Stuhl setzen* ('on chair sit': 'to sit down on a chair'), possibly with an accidental high co-occurrence ratio with negative contexts in our corpora. There is no automatic procedure to validate NPIs, and manual annotation is 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, i.e. if 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 this. However, some of the NPI-characteristic features (e.g. negation, fixedness, significant co-occurrence) 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 by sorting candidates according to associative strength with their respective negative contexts and linguistic features (morpho-syntactic fixedness or translational behavior). This preprocessing considerably reduces the necessary but time-consuming manual annotation efforts.

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)). Such measures 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 for others) to determine whether a negative context is obligatory for an expression. An example pair is: (*blassen::Schimmer::haben*, NEG).

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.

We used the UCS toolkit[4] to calculate five standard association measures for each of our candidate lists (cf. Table 3; there is a candidate list for each syntactic pattern). In the next step these lists were sorted in decreasing order according to the resulting scores. Finally, the highest scoring 500 candidates with a strong statistical tendency to be associated with a NEG context label were manually annotated: '+' for valid NPIs and '−' for other MWEs or trivial combinations.

The numbers of valid NPIs found amongst the top 500 candidates are given in Table 4. Even though *poisson* slightly outperformed the other measures, all results turned out to be quite similar. Furthermore, we also found that the NPIs often were 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 5(a) 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 in terms of NPI validity (+/−). The absolute frequency of the NPI candidates is indicated in column 3 ('f') while the last columns give the ranks according to *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 frequency. Sorting according to association measures seems to move NPIs towards the top of the list, as candidates hardly ever occurring in a non-negative context are considered to be highly associated with their negative context. Table 5(b) illustrates the huge differences between ranking positions of NPIs in the two different lists.

5.1.1. Difficult Cases

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*

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

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 can still not model reliably, and the creative use of language: the NPI *Wald vor Baum sehen* (lit. 'not to see the forest for the trees': 'not to see the obvious') (cf. table 5, table 6) occurred 46 of 50 times in a trivially negative context. The complete expression, as it might be listed in a dictionary is *den Wald vor lauter Bäumen nicht sehen*, i.e. the verb is negated with *nicht*, which corresponds exactly to the 46 observed negative forms.

The remaining four occurrences are more difficult cases: the first, a question (5), is a known nonveridical licensing environment (which we can model), while the second and third occurrences are a modal context (6) and a conditional clause (7), which are not among the contexts we modeled. In the last sentence, however, there is no real negative context. Regardless of the lack of an obvious negative licensing environment, the sentence is well-formed.

5) Sieht er dann den Wald vor lauter Bäumen?

6) Doch wie immer, sollte man zunächst einmal den Wald vor Bäumen sehen.

7) Hätte die Kommission eindeutige und anerkannte Prioritäten und könnte sie den Wald vor Bäumen sehen, hätten wir nicht diese Aussprache heute Nachmittag.

8) Manchmal sieht M. L. vor lauter Bäumen dennoch den Wald.

One has to keep in mind that to allow for extraction of negative contexts, the syntactic pattern of this context – be it a question, some form of conditional or a preceding verb – has to be specifically implemented. The examples above illustrate negative contexts that are not easy to detect automatically. As shown in (8), in some cases we might even find constructions with clear NPIs that are used in contexts which cannot be easily categorized as being negative.

5.2. Linguistic Processing

In order to further improve the ordering of the lists, we add more linguistic knowledge to the statistical method. This may also help to handle the problem of overlapping results: the entries marked with '(+)' in table 5(a) 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.

We enriched our result lists with the following linguistically motivated scores:

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

We use the percentage of the candidates' negative occurrences (#NEG) as a possible indicator for NPIs in our extraction process (cf. table 6).

The morpho-syntactic fixedness score (FIX) is motivated by previous work on the extraction of idiomatic MWEs. Since many multi-word 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 of all syntactically compatible articles. While extracting candidates, information on the nouns' number and article use is collected. For each candidate, we compute the frequency distribution of the number values (SG, PL) and possible determiners (e.g. DEF, INDEF, NONE). Then, 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 are most often (immediately) adjacent. The FIX score is then calculated for each NPI candidate based on the average of:

	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.

(i) the #NEG score

(ii) the percentage of number and article setting

(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 multi-word NPIs have a non-compositional semantics, i.e. they are to be translated as a whole while trivial 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 top500 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), for details see (Manning and Schütze, 1999). Table 7 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.

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

Table 7: UAP scores for re-orderings of top500 *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 are independent and thus benefit from each other when combined. While we achieved our aim to enhance the sorting quality of the candidate lists, the improvement is not great. This is mainly due to the fact that most NPIs are relatively low-frequent: the TE and the PDA score are not designed for low-frequent data, and computing morpho-syntactic preferences works better for high-frequent data as well.

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 142 NPIs. 28 of these are in CoDII, 114 are new. Another relevant comparison is John Lawler's collection of English NPIs[5], which comprises roughly three dozen entries. Jack Hoeksema's collection of Dutch NPIs, which is by far the largest known collection of NPIs and has been developed for 15 years, reportedly contains 670 entries.[6] 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, 1983)'s list of 344 items. However, according to them, Kürschner's list contains about 200 pseudo-NPIs, i.e. items which exhibit a 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 28).

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 top500 by taking linguistic knowledge about the fixedness of expressions into account.

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.

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 top500 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 for even better results.

Looking at NPI research from the perspective of theoretical linguistics, there should also be much to gain from extraction methods such as ours: 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.

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.

8. References

Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Richard Ducott. 2009. Without a 'doubt'? Unsupervised discovery of downward-entailing operators. In *Proceedings of NAACL HLT*, pages 137–145.

Stefan Evert. 2004. The statistics of word cooccurrences: word pairs and collocations. Master's thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

[5] www-personal.umich.edu/~jlawler/NPIs.pdf

[6] www.let.rug.nl/~hoeksema/lexicon_bestanden/v3_document.htm

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

Ulrich Heid, Fabienne Fritzinger, Susanne Hauptmann, Julia Weidenkaff, and Marion Weller. 2008. Providing corpus data for a dictionary for German juridical phraseology. In Angelika Storrer, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner, editors, *Text Resources and Lexical Knowledge – Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008*, Berlin, New York. Mouton de Gruyter.

Phillip Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of the 10th MT Summit 2005*, Phuket, Thailand.

Wilfried Kürschner. 1983. *Studien zur Negation im Deutschen*. Gunter Narr, Tübingen.

Timm Lichte and Jan-Philipp Soehn. 2007. The Retrieval and Classification of Negative Polarity Items using Statistical Profiles. In Sam Featherston and Wolfgang Sternefeld, editors, *Roots: Linguistics in Search of its Evidential Base*, pages 249–266. Mouton de Gruyter, Berlin.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

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

Beata Trawiński, Jan-Philipp Soehn, Manfred Sailer, and Frank Richter. 2008. A Multilingual Electronic Database of Distributionally Idiosyncratic Items. In Elisenda Bernal and Janet DeCesaris, editors, *Proceedings of the XIII Euralex International Congress*, volume 20 of *Activitats*, pages 1445–1451, Barcelona, Spain. Universitat Pompeu Fabra.

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

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

Marion Weller and Urlich Heid. 2010. Multi-parametric extraction of German multiword expressions from parsed corpora. In *Proceedings of the Linguistic Resources and Evaluation Conference, LREC 2010*, Valetta, Malta.

Frans Zwarts. 1995. Nonveridical Contexts. *Linguistic Analysis*, 25:286–312.

Frans Zwarts. 1997. Three types of polarity. In Fritz Hamm and Erhard W. Hinrichs, editors, *Plurality and Quantification*, pages 177–237. Kluwer Academic Publishers, Dordrecht.