Corpus-based Acquisition of Complex Negative Polarity Items

Timm Lichte
Collaborative Research Centre 441 — University of Tübingen
timm.lichte@student.uni-tuebingen.de

Abstract. This paper presents the functionality and results of an extraction mechanism for German negative polarity items (NPIs) which uses a large partially parsed corpus of written German and integrates usual collocation acquisition methods. Particular attention is paid to the extraction of complex NPIs.

1 Introduction

In this paper I will address a special group of lexical elements which show a particular affinity with negative contexts. Such elements, usually referred to as *negative polarity items* (NPI), have been widely studied in the linguistic literature since Klima (1964). The classic example of an NPI is the English indefinite determiner *any*. As demonstrated in (1) a sentence containing *any* and negation is grammatical. Without the negation the sentence is ungrammatical. Following standard terminology I will refer to the negation as the *licenser* of the NPI. I will underline NPIs and print the licensers in bold face.[1]

(1) a. He has**n't** seen any students.
b. *He has seen any students.

Since I will be focusing on German, an analogous German example is presented in (2). These sentences differ only in that sentence (a) contains a

[1]There is a particular use of *any*, called *free-choice any*, which does not require a negative operator such as **not**. Nevertheless this use has a restricted distribution, i.e. it requires a context which is *nonveridical* (Zwarts (1995),see Section 2).

Proceedings of the Tenth ESSLLI Student Session
Judit Gervain (editor)

so-called *n-word* as licenser, whereas in (b) there is no exponent of negation; thus the NPI *jemals* (ever) is not licensed.

(2) a. **Niemand** von uns war jemals im Jemen.
nobody of us was ever in Yemen
(None of us has ever been to Yemen.)
b. *Einer von uns war jemals im Jemen.
One of us was ever in Yemen

The inventory of NPIs in English and Dutch has been documented fairly well. Jack Hoeksema (pc) has collected about 760 Dutch NPIs. For German the state of documentation is less ideal. There are only two relatively extensive lists: Welte (1978) and Kürschner (1983), neither of which comes even close to the data collected by Hoeksema.

The aim of this paper is to present a step towards an automatic corpus-based compilation of a list of German NPI candidates. To my knowledge van der Wouden (1992) was the first to explicitly point out that the relation between an NPI and its licenser bares similarities to the relation between a collocate and its collocator. This idea, then pursued in van der Wouden (1997), represents the basic motivating insight for this paper.

In Section 2 I will summarize the semantic literature on NPIs. These insights are applied to extract NPI candidates consisting of a single lexical element in Section 3, and in Section 4 to extract complex NPI candidates.

2 Linguistic Aspects

In this section I will present a summary of those aspects of NPIs which are directly related to our study.

Negation is understood as a logical operator which imports special entailment properties to the semantic value of an attached sentence. The literature on NPIs distinguishes several degrees of negativity based on their formal semantic properties.[2] I will concentrate on operators which are downward entailing to identify negative contexts. Since downward entailment is a rather general property of negative contexts and since all stronger degrees of negativity conform to downward entailment, operators of stronger negation are also included. As has been noted by Ladusaw (1980) most NPIs require a context which is at least downward entailing. Genuine downward entailing operators include words such as *höchstens* (at most), *kaum* (hardly) or

[2]See van der Wouden (1997) for an introduction to the necessary formal semantic properties of negative contexts and NPIs with rich data.

wenige (few). A downward entailing context allows one to reason from sets to subsets as demonstrated in (3):

(3) **Few** congressmen eat vegetables.
‖spinach‖ ⊆ ‖vegetables‖
Few congressmen eat spinach.

An even weaker notion of negativity is that of *nonveridicality* (Giannakidou (1998); Zwarts (1995)). Roughly put, if a statement is in the scope of a nonveridical operator, then the truth of the statement is not implied, but on the other hand reasoning from sets to subsets is not possible in general. Nonveridical contexts are triggered by direct or indirect questions, free relatives, and also by adverbials such as *vielleicht* (perhaps). Since this category appears to be rather large and not every nonveridical operator seems to license NPIs, I will only include interrogatives in my considerations here. Interrogatives are rather numerous and can be easily detected in the corpus.

Although I will ignore this issue for the time being, it should be noted that NPIs can have different distributional patterns along the degrees of negativity, which make it possible to distinguish different subclasses of NPIs. Zwarts (1997) mentions the modal verb *brauchen* (need) as an NPI that is compatible with downward entailing triggers, but excluded from questions.

(4) *__Wer__ braucht Brot zu kaufen**?**
who needs bread to buy

An NPI which can occur in all of the above-mentioned contexts is *jemals* (ever). Note that it is still an NPI because it is excluded from sentences without a licenser, as demonstrated in (2). Since I am only interested in identifying NPIs I will skip this subclassification issue and concentrate instead on downward entailing contexts and interrogative constructions, although subclassification naturally follows acquisition.

Another property of NPI licensing that will be simplified in the extraction mechanism is its scopal restrictiveness. As Ladusaw (1996) summarizes there seem to be various constraints at work such as the c-command relation, the precedence of a licenser and the *immediate scope* (Linebarger (1980)). Leaving these subtleties aside, I will define the scope of a licenser simply as the clause in which the licenser appears, including all of its sub-clauses. Thereby it holds that a more deeply embedded negative operator cannot license NPIs in a higher position. An example of such a configuration is given in (5-a). In this structural position *nicht* (not) cannot license an NPI in the matrix clause (b).

(5) a. [Was Frauen droht, [die dem Aufruf **nicht** folgen]], blieb unklar.
(It remained unclear [what was going to happen to women [who do **not** follow the call]].)

b. *[Was ... [... **nicht** folgen]] wurde jemals gesagt.
what not follow was ever said

3 The Basic Extraction Method

After having established the theoretical framework for our empirical study of German NPIs, we can now proceed to the actual corpus work. Section 3.1 discusses the corpus and the methods which I employ. The extracted candidates will be discussed in Section 3.2.

3.1 Methods

For the extraction mechanism the TüPP-D/Z corpus (*Tübingen Partially Parsed Corpus of Written German*; see Ule and Müller (2004)) was used. [3] The TüPP-D/Z corpus is based on the electronic version of the German newspaper *die tageszeitung* (*taz*). It contains lemmatization, part-of-speech tagging, chunking and clause boundaries. The section of TüPP-D/Z where the method is applied consists of about 5.5 million sentences (to be exact 5,531,168) which contain 641,035 different lemmatized forms.

The NPI extraction procedure is basically done in three steps: clause marking; lemmata counting; and quantitative evaluation.

Based on the lemmatization and the part-of-speech assignments in TüPP-D/Z the clauses are classified according to the presence of an NPI licenser. Basically, I require the licenser to impose downward entailment or to form an interrogative construction. Thus the set of NPI licensers comprises lexical licensers (e.g. *nicht* (not), *niemals* (never), *kaum* (hardly), question mark) and structural licensers (e.g. the restrictor of universal quantifiers)[4]. Future work will entail adding predicates with inherent negation and clausal complements(e.g. *bezweifeln* (to doubt)).

After clause marking, for each lemma in the corpus the number of total occurrences and the number of occurrences within the scope of a licenser are extracted. We restrict ourselves (i) to lemmata which are not lexical licensers and (ii) to lemmata which occur at least 40 times, because less frequent

[3] The internet homepage of TüPP-D/Z is `http://www.sfs.uni-tuebingen.de/tupp`.

[4] Although these do not trigger downward entailing contexts I also integrated the restrictors of superlatives as licensers, because they nevertheless license NPIs. However, their number and influence is marginal.

lemmata do not show a reliable occurrence pattern for polarity contexts. We have to concede that this is a purely heuristic threshold. The resulting data contain 34,957 lemmata.

In order to derive a list of NPI candidates, the ratio of contextual and total occurrence is calculated for each lemma. Based on these context ratios (CRs) a lemma ranking is set up. It can be shown that CR is equivalent to mutual information (MI, e.g. as defined in Krenn (1999)) in that it yields the same lemma ranking.[5]

3.2 Results: NPI candidates

The 20 highest CR-scored lemmata are shown in Table 14.1. Lemmata which also appear in Kürschner (1983) as NPIs or as parts of NPIs are printed in bold face. Lemmata that from my point of view show a tendency towards negative polarity, but are not included in Kürschners collection, are in bold face as well and are marked with an attached asterisk.

#	*Lemma*	*CR*	#	*Lemma*	*CR*
1	**verdenken** (to hold sth against sb)	1.00	11	**geheuer**(not odd)	0.96
2	**unversucht** (unattempted)	1.00	12	unähnlich (not alike)	0.95
3	***unterschätzender** (to underestimate / gerundiv form)	1.00	13	***wegdenken** (to imagine sth not there)	0.94
4	**umhin** (around)	0.98	14	***allzuviel** (too much)	0.92
5	**nachstehen** (to be inferior)	0.98	15	**sonderlich** (particular)	0.91
6	**lumpen** (to splash out)	0.98	16	***abneigen** (to be averse to sth.)	0.91
7	langgehen (to go along sth)	0.98	17	behagen (to please)	0.90
8	**verhehlen** (to conceal)	0.96	18	hinwegtäuschen (to obscure the fact)	0.89
9	**beirren** (to disconcert)	0.96	19	dagewesen (precedent)	0.89
10	genauer (more accurate)	0.96	20	hingehören (to belong)	0.88

Table 14.1: The 20 highest ranked lemmata according to their CRs. The mean of the CR values of all lemmata is 0.15 .

The candidate list as a whole looks promising and one can distinguish

[5]Given a lemma w with frequency N_w , the frequency of negative contexts N_{neg} and furthermore $N_{w,neg}$ as the frequency of w occurring in a negative context, the formal definitions of CR and MI will then appear as follows:

(i)
$$\begin{aligned} \text{CR} &:= \frac{N_{w,neg}}{N_w} \\ \text{MI} &:= \frac{P(w,neg)}{P(w)*P(neg)} = \frac{N_{w,neg}/N}{(N_w/N)*(N_{neg}/N)} = \frac{N_{w,neg}}{N_w} * \frac{N}{N_{neg}} \end{aligned}$$

$P(w, neg)$ is the probability of the co-occurrence of w and a negative context. It is obvious that $\frac{N}{N_{neg}}$ has a constant value and hence is not substantial for the computation of the ranking.

four types of candidates:

1. **Non-polysemous candidates:** Lemmata such as *sonderlich* and *nachstehen* are non-polysemous and complete NPIs. But there are also lemmata which clearly show negative polarity without being complete NPIs, i.e. they rarely occur as NPIs without certain lexical material surrounding them. Examples of non-polysemous, but incomplete NPIs are *umhin* from *umhinkönnen/umhinkommen* (to have a choice to do sth), *lumpen* from *lumpen lassen* (to splash out) and *verhehlen* from *verhehlen können* (to be able to conceal).

2. **Polysemous candidates:** Some lemmata require only a negative context when used in combination with certain other lemmata. Since these lemmata allow for non-negative contexts, their CR values are generally expected to be lower than those of the preceeding type. An instance of this type of candidate is perhaps *wegdenken*, which is an NPI when used as *wegzudenken sein*, but non-polar with an auxiliary as in *wegdenken müssen*. Many other instances are at lower ranks such as *Kram* (stuff) from *in den Kram passen* (to be welcome, lit. 'to fit in the stuff') ranked at 558, as well as *brauchen* ranked at 874, that only requires a negative context, if it has a non-finite clausal complement. The grade of polysemy culminates in complex NPIs, where the individual elements have a rather low CR value, e.g. *[nicht] alle Tassen im Schrank haben* (to have lost one's marbles)[6].

3. One finds several **"pseudo-polarity items"** (Hoeksema (1997)), that have a stylistically motivated affinity for negation, but can still occur outside negative contexts. And even here one can distinguish between items which can stand alone (*unähnlich*) and which are lexically dependent (*hinwegtäuschen können* (to be able to obscure the fact)). Since the text style of the corpus influenced my data, I expect better results from a more balanced corpus. If the extraction method is based on the TüPP-D/Z alone, however, there does not seem to be a way to automatically separate pseudo-NPIs from regular NPIs. Nevertheless, pseudo-polarity can be interesting as an early (or late) state of polarity sensitivity.

4. Finally, there are two lemmata in the candidate list (*langgehen* and *genauer*), that I cannot classify as any of the types above. In other words, they seem to be instances of **noise**.

[6] *Tasse* (cup) is the highest ranked lemma at position 6398.

4 The Extraction Method for Complex NPIs

I have presented a method for automatically extracting a list of NPI candidates from a corpus. However, these NPI candidates are only single lemmata. The need to be able to account for multi-word expressions can be shown for each of the candidate types above. First, it would enable us to acquire complex, non-polysemous NPIs. Second, it would offer the means to disambiguate polysemous candidates. Third, it would enable us to narrow pseudo-polar candidates down to complex pseudo-NPIs. Fourth, it would help us to check candidates that seem to be instances of noise. I would therefore like to propose an enhancement to the basic method, in order to account for complex expressions with an affinity for negation.

4.1 Methods

The starting point is the list of lemmata and their context ratios. We do a collocation test for every lemma and ask for other lemmata that significantly co-occur (i) in the same clause and (ii) in negative contexts. Doing this we obtain a list of collocates for each of the lemmata. Afterwards we ask whether or not there is a distribution pattern of lemma and collocate, which shows higher or equal affinity to negative contexts than the lemma individually. If that is the case, we then repeat the procedure again on the lemma-collocate pair, which is now handled the way we handled single lemmata. In doing so we obtain chains of lemmata as NPI candidates, which cannot be expanded because they lack either collocates or a stronger affinity for negation.[7] These new complex NPI candidates are added to the original lemma ranking in accordance with their context ratio.

The advantage of using the whole list of lemmata is that we have the chance to detect complex NPIs such as *alle Tassen im Schrank haben*, where the elements, taken individually, exhibit a rather free distribution with respect to negative contexts, therefore being ranked much farther away from the usual NPI suspects. The disadvantage is rather technical, but nevertheless meaningful to me: it is very time-consuming.[8]

As a collocation measure I integrate the G^2 score, a derivative of Log-likelihood (Rayson and Garside (2000)), since we are now confronted with bigrams consisting of varying items, whereas in the basic mechanism the distribution of a lemma is always evaluated with respect to negative contexts. Lemnitzer (1997) and Villada-Moirón (2005) report the successful integration

[7] I also demand an overall frequency larger than 20.

[8] The implementation processes 100 lemmata in about 4 hours using a 2x 1GHz Pentium III, 1GB RAM computer system.

of G^2 while working on comparable issues. The span of collocation testing is a clause as annotated in TüPP-D/Z. Here the question arises of which significance level to choose, as even a "strong" significance level at $p < 0.01$ (6.6) seems to be too weak (Lemnitzer (1997)). I chose a G^2 score of 200 for the examples below.

4.2 Results

Because of computational efforts, I have not yet carried out the enhanced method on the whole list of lemmata. Instead, I applied the enhanced method to the 200 highest ranked lemmata, which led to lemma chains, the 20 highest ranked of which are depicted in Table 14.2. In the second column one encounters the lemma chains as generated by the enhanced method, while in the third column I try to map those lemma chains to multi-word expressions. As for single lemmata in Table 14.1, multi-word expressions are printed in bold face, if they appear in Kürschner (1983), and are marked with an asterisk, if they are NPIs according to my intuition, but are not found in Kürschners collection.

The candidate list gives a promising impression, since most of the lemma chains can be mapped to complex NPIs. In addition, the advantage of my extraction method for complex NPIs can be exemplified for each of the four candidate types sketched above. For non-polysemous candidates such as *unversucht* and *lumpen* it can be observed that they are completed. Polysemous candidates such as *trauen* (*den Augen trauen*), *Veranlassung* (*eine Veranlassung sehen zu etw.*) or *verkneifen* (*verkneifen können*) are disambiguated. Pseudo-NPIs such as *hinwegtäuschen* are narrowed down to complex ones (*darüber hinwegtäuschen können, dass*). Finally there is also an example of a noisy candidate that successfully is checked: *genauer* is classified as being part of the pseudo-NPI *genaueres wissen*.

The examples above are based on high-ranked lemmata. What if we start from a lemma with a relatively low CR value? The results for the lemmata *Tasse* (cup) ranked at 6398 and *Kram* (stuff) ranked at 558 are depicted in the last two rows of Table 14.2. The compiled lemma chains lead to the complex NPIs already mentioned in section 3.2, while their ranking position has significantly improved compared to the individual lemmata due to a higher CR value (ranking position 23 and 52, respectively). That exemplifies how complex NPIs connected to low-ranked lemmata can enter the visual field of the researcher.

Taking this as an encouraging step towards complex NPI acquisition, there is, however, a small drop of bitterness: I suppressed three candidates

#	*Lemma chain*	*Multi-word expression*	*CR*
1	unversucht lassen	**etw. unversucht lassen** (to leave sth. undone)	1.00
2	geheuer ganz	**ganz geheuer** (not odd)	1.00
3	jedermanns Sache	***jedermanns Sache sein** (to be everyone's cup of tea)	1.00
4	umhin zu kommen	**umhinkommen** (not to be bound to do sth.)	1.00
5	lumpen lassen	**sich lumpen lassen** (not to splash out)	0.98
6	verkneifen können Sie	***sich verkneifen können** (to be able to deny oneself sth.)	0.97
7	beirren lassen Sie	**sich beirren lassen** (to let so. disconcert oneself)	0.97
8	genauer wissen	genaueres wissen (to know sth. more precisely)	0.97
9	hinwegtäuschen können der darüber dass	darüber hinwegtäuschen können, dass (to be able to deceive so., that)	0.96
10	trauen Auge	**den Augen trauen** (to belief one's eyes)	0.95
11	Veranlassung sehen	***eine Veranlassung sehen zu etw.** (to have reason to do sth.)	0.95
12	entmutigen lassen	sich entmutigen lassen (to lose heart)	0.95
13	auslassen Gelegenheit	***eine Gelegenheit auslassen** (to miss an opportunity)	0.95
14	fackeln lange	**lange fackeln** (to dither)	0.93
15	anhaben können	**etw. anhaben können** (to be able to harm so.)	0.92
16	nützen es	***es nützt etw.** (it is of use)	0.92
17	einwenden gegen	***etw. einzuwenden haben gegen** (to have an objection to sth.)	0.91
18	gar dabei	dabei ... gar (???)	0.88
19	Hehl aus machen	**einen Hehl aus etw. machen** (to make a secret of)	0.87
20	Ahnung haben	**eine Ahnung haben** (to have a clue)	0.85
xx	Tasse Schrank (6398/23)	**alle Tassen im Schrank haben** (not to have lost one's marbles)	0.86
yy	Kram passen (558/52)	***in den Kram passen** (to be welcome)	0.70

Table 14.2: The 20 highest ranked complex candidates according to their CR value and their mappings to multi-word expressions.

which undoubtedly have nothing to do with negative polarity. Their considerably high CR value arises due to locally limited characteristics of the newspaper corpus TüPP-D/Z. There is, for example, the lemma chain *notwendigerweise geben die auf Meinung wieder Seite erscheinend* with a CR value of 1. It originates from the weekly 'Letters to the editor' section of the corpus newspaper, where the following note uses to appear: *Die auf dieser Seite erscheinenden Leserbriefe geben nicht notwendigerweise die Meinung der taz wieder* (The reader's letters on this page don't necessarily reflect the opinion of the taz.). Since these highly recurrent notes are easily identified in the candidate list on the basis of the length and the high CR value of the corresponding lemma chains I have not taken them into consideration in Table

14.2. These odd candidates show, however, the dependence of the extraction method on the corpus data.

While the latter candidates are undesirable, certain complex NPIs which emerge from polysemous items cannot be identified. The verb *brauchen*, as mentioned earlier, only requires a negative context if it has a clausal complement. Separating the negative-polar from the non-polar variant requires considering the complement structure, however, the complement structure is not available in the TüPP-D/Z. When the extraction mechanism for complex NPIs is applied to *brauchen*, we obtain the lemma chain *brauchen zu* (lit. 'need to') which could be related to the negative-polar variant, but its CR value is far from being salient (0.40). Therefore, complex NPIs that contain syntactically but not lexically specified members seem to pose a problem for our extraction mechanism.

Nevertheless, it illustrates how complex NPIs can be obtained, even ones that were unnoticed so far. It also shows that the linguistic intuition of the researcher is a crucial factor since he has to interpret unordered lemma chains. After all, the output of my extraction method is a rather raw resource, but it could be a significant help in searching for NPIs.

5 Conclusion

My starting point was the insight from van der Wouden (1997) that the relation between an NPI and its licenser is of a collocational nature. Distributional profiles of lemmata in a partially parsed corpus of German were compiled, mainly with the aid of lemmatization, part-of-speech tags and clause structure annotation and with respect to negative contexts derived from the semantic literature on NPIs. These profiles were used to compile a list of NPI candidates. It was shown that a simple quantitative ranking leads to promising candidates. It was also shown that the method can be extended naturally to retrieve complex NPIs.

Acknowledgements

I am grateful to Lothar Lemnitzer and the anonymous reviewers for comments and to Janah Putnam for her help with the English. Thanks as well to Tylmann Ule, who assisted me in using the TüPP-D/Z corpus. Finally, I would like to thank Manfred Sailer for his invaluable inspiration and support in the early stages of this work.

Bibliography

Giannakidou, A. (1998). *Polarity Sensitivity as Nonveridical Dependency*. John Benjamins, Amsterdam.

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: `http://odur.let.rug.nl/~hoeksema/docs/barcelona.html`.

Klima, E. (1964). Negation in English. In J. A. Fodor and J. Katz (Eds.), *The Structure of Language*, pp. 246–323. Prentice Hall, Englewood Cliffs, New Jersey.

Krenn, B. (1999). *The Usual Suspects. Data-Oriented Models for Identification and Representation of Lexical Collocations*, Volume 7 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. Saarbrücken: DFKI and Universität des Saarlandes.

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

Ladusaw, W. (1980). *Polarity Sensitivity as Inherent Scope relations*. Garland Press, New York.

Ladusaw, W. (1996). Negation and polarity items. In S. Lappin (Ed.), *The Handbook of Contemporary Semantic Theory*, pp. 321–341. Blackwell Publishers.

Lemnitzer, L. (1997). *Akquisition komplexer Lexeme aus Textkorpora*. Tübingen: Niemeyer.

Linebarger, M. C. (1980). *The Grammar of Negative Polarity*. Ph. D. thesis, MIT. cited after the reproduction by the Indiana University Linguistics Club, Indiana, 1981.

Rayson, P. and R. Garside (2000). Comparing corpora using frequency profiling. In *Proceedings of the* Workshop on Comparing Corpora*, ACL, 1–8 October 2000, Hong Kong*, pp. 1–6.

Ule, T. and F. H. Müller (2004). KaRoPars: Ein System zur linguistischen Annotation großer Text-Korpora des Deutschen. In A. Mehler and H. Lobin (Eds.), *Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*. Opladen: Westdeutscher Verlag. to appear.

van der Wouden, T. (1992). Beperkingen op het optreden van lexicale elementen. *De Nieuwe Taalgids 85*(6), 513–538.

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

Villada-Moirón, B. (2005). *Data-driven identification of fixed expressions and their modifiability*. Ph. D. thesis, Alfa-Informatica, Rijksuniversiteit Groningen.

Welte, W. (1978). *Negationslinguistik. Ansätze zur Beschreibung und Erklärung von Aspekten der Negation im Englischen*. Wilhelm Fink Verlag, München.

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

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