The Acquisition and Classification of Negative Polarity Items using Statistical Profiles

Timm Lichte and Jan-Philipp Söhn
SFB 441 – University of Tübingen, Germany
timm.lichte@uni-tuebingen.de, jp.soehn@uni-tuebingen.de

1 Introduction

In this contribution we 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 classical 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 we will refer to the negation as the *licenser* of the NPI. we will underline NPIs and print the licensers in bold face.

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

The inventory of NPIs in English and Dutch has been documented fairly well. Hoeksema (2005) for instance presents about 700 Dutch NPIs. For German the state of documentation is less ideal. There is only one relatively extensive list in Kürschner (1983), which , however, does not even come close to the data collected by Hoeksema.

Zwarts (1997) assumes NPIs to have different distributional patterns along the degrees of negativity, which make it possible to distinguish different subclasses of NPIs. Following van der Wouden (1997), we differentiate between minimal (e.g. *few*), regular (e.g. *nobody*) and classical (e.g. *not*) negation and analogously between weak, strong and superstrong NPIs.

Negation	*NPI*		
	weak	strong	superstrong
minimal	+	-	-
regular	+	+	-
classical	+	+	+

Zwarts gives as an example the Dutch NPI *ook maar iets* (anything) which is compatible with regular negation, but excluded from minimal negation. Therefore it can be classified as an strong NPI.

The aim of this contribution is to show the use of statistics (1) to automatically acquire a list of NPI candidates from a partially parsed corpus of written German, and (2) to classify NPIs.

2 Acquisition

The basic motivating idea behind the corpus-based acquisition mechanism described here is to treat the relation between an NPI and its licenser as similar to the relation between a collocate and its collocator. This idea, going back to van der Wouden (1992) and then pursued in van der Wouden (1997), allows us to apply regular collocation acquisition techniques in order to yield a list of NPI candidates.

Turning to the acquisition method we use a part of the TüPP-D/Z corpus (*Tübingen Partially Parsed Corpus of Written German*)[1]. TüPP-D/Z is based on the electronic version of the German newspaper *die tageszeitung* (*taz*). It contains lemmatization, part-of-speech tagging, chunking and clause boundaries. The section of TüPP-D/Z that we use consists of about 2.8 Mio sentences. The NPI extraction proceeds in three steps: clause marking, lemmata counting and evaluation. Based on the lemmatization and the part-of-speech assignments in TüPP-D/Z we classify the clauses according to the presence of an NPI licenser. The licenser must impose at least minimal negation or form an interrogative construction. We use the clause-structure annotation given in TüPP-D/Z to derive scope relations in a very general manner, which guarantees that a deeper embedded negative operator cannot license NPIs in a higher position. On the other hand, a licenser of a clause is also valid for all its sub-clauses. After clause marking we extract for each lemma in the corpus the number of total occurrences and the number of occurrences in clauses which contain a licenser. In order to derive a list of NPI candidates, we calculate the ratio of contextual and total occurrence for each lemma. Based on these context ratios (CRs) we set up a lemma ranking and expect NPIs to have a significantly high CR-value.

To handle complex NPIs we need an enhancement of the current method. The starting

[1] See `http://www.sfs.uni-tuebingen.de/tupp`

point is the list of lemmata and their context ratios. We perform a collocation test for every lemma to identify other lemmata that significantly co-occur (i) in the same clause and (ii) in negative contexts. As a collocation measure we integrate the G^2 score, a derivative of Log-likelihood (Rayson and Garside (2000)). This yields a list of collocates for each of the lemmata. Next we ask whether the distribution pattern of lemma and collocate shows higher or equal affinity to negative contexts than the lemma individually. If that is the case we repeat the procedure on the lemma-collocate pair, which is now handled the way we handled single lemmata. In doing this we get chains of lemmata as new NPI candidates, which cannot be expanded because they lack either collocates or an enlarged affinity for negation. Starting with the lemma *Sicht* (sight), for instance, the enhanced acquisition method compiles the lemma chain *Sicht ein in Ende* (sight a in end) which corresponds to the negative-polar expression *ein Ende in Sicht sein* (to see an end). These new complex NPI candidates are added to the original lemma ranking in accordance with their context ratio.

Despite the limitations of the corpus and the method we obtain a list of NPI candidates that contains a considerable proportion of the items in Kürschner's collection. Furthermore our NPI candidates include many items not listed in Kürschner's collection, but worth a closer examination.

3 Classification

In this section we briefly show that our method can also be used for the subclassification of NPIs. In principle, classification is an elaboration of the acquisition method, since we perform a refinement on the distributional patterns that the acquisition method makes use of. For that, we simply split the set of negative contexts into subsets according to minimal, regular and classical negation. The distributional pattern we obtain for each NPI then treats the three subclasses of negative contexts separately. This way, we are able to investigate which degree of negation a given NPI candidate is most strongly associated with and to assign it to an NPI class.

How can we measure the association of an NPI with a subclass of negative contexts? We examine the increase of the CR-value of an NPI while extending contexts of classical negation by contexts of regular negation and while extending contexts of regular negation by contexts of minimal negation. The basic assumption is that extending a negative context this way leads to an increase of the CR-value of an NPI, since a larger number of negative sentences is taken into consideration. If an NPI is evenly distributed over the subclasses of negative contexts its CR-value should be increasing commensurately to the enlargement of the considered data set. However, if the CR-value increases in an unexpected manner we use this as a measure of association with a certain subclass of negative contexts. Mainly two cases of deviance are interesting: (1) going from classical to regular negation or (2) going from regular to minimal negation causes significantly less increase of the CR-value of an NPI than expected. In the

first case, we have evidence to classify the NPI as superstrong; in the second case, as strong. In any other case, we keep the null hypothesis, namely that the NPI is weak.

To give an example of our first results, the procedure classifies the NPI *sonderlich* (particular) as strong NPI in accordance with the literature. Interestingly, we did not find superstrong NPIs, which is, however, supported by Krifka (1995). A long-term goal is to test whether the subclasses predicted by Zwarts (1997) show up as patterns in the statistics of the data. We also plan to consider an alternative system of NPI subclasses proposed in Giannakidou (1997).

References

Giannakidou, A., 1997. The landscape of polarity items. Ph.D. thesis, Rijksuniversiteit Groningen.

Hoeksema, J., April 2005. De negatief-polaire uitdrukkingen van het Nederlands. inleiding en lexicon, manuskript, Rijksuniversiteit Groningen.

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

Krifka, M., 1995. The semantics and pragmatics polarity items. Linguistic Analysis 25, 209–257.

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

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

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. Routledge, London.

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