A Multilingual Electronic Database of Distributionally Idiosyncratic Items

Beata Trawiński
Jan-Philipp Soehn
University of Tübingen

Manfred Sailer
Universität Göttingen

Frank Richter
Universität Tübingen

We present a multilingual electronic database of lexical items with idiosyncratic occurrence patterns. Currently, our database consists of: (1)a collection of 444 bound words in German; (2)a collection of 77 bound words in English; (3)a collection of 58 negative polarity items in Romanian; (4)a collection of 84 negative polarity items in German; and (5)a collection of 52 positive polarity items in German. Our database is encoded in XML and is available via the Internet, offering dynamic and flexible access.

1. Introduction

Lexical items with idiosyncratic distribution such as bound words (BWs) and polarity items (PIs) represent a challenge for traditional (general and idiomatic) lexicography as well as computational applications. It is, for instance, unclear whether BWs have an independent lexical status and to what extent the expressions in which they occur are typical idiomatic expressions. The lexicographic treatment of PIs is even more problematic, because their occurrence requirements are not as local as those of bound words and because the obligatory collocators are not simple lexemes but abstract grammatical and semantic categories. The aim of our database is to document the information available about these items in dictionaries and linguistic literature, together with corpus data and sample queries for major text corpora and thus to provide a solid empirical basis for both theoretical-linguistic and lexicographic investigations into these expressions.

2. Bound words

2.1. *General characteristics*

Bound Words (BW) are words which can only occur as part of a fixed expression. These items are also called Cranberry Words in Aronoff (1976), in analogy to the term cranberry morph. A typical BW is *sandboy* which can only occur as part of the expression *happy as a sandboy*.

The repertoire of BWs in German and English is well documented in the literature on idioms. Dobrovol'skij (1988) provides the most exhaustive list of BWs in German, English and Dutch. In his work (Dobrovol'skij 1988 and Dobrovol'skij and Piirainen 1994), he emphasizes the difference between bound and free words, provides criteria for classifying BWs and estimates their number for German at 600. Out of these, 180 are classified as belonging to the common vocabulary of native speakers. At present, our collection includes 444 potential BWs. For English, Dobrovol'skij (1988) lists about 100 items, 77 of which are included. Our leading criterion for recording an item was whether it was discussed as a candidate for being a BW within the phraseological literature.

2.2. *Lexicographic representation*

Dictionaries often do not represent BWs in a uniform way. One of the subclasses of BWs identified in Dobrovol'skij (1989) are words with singular combinability. These appear as

constituents of fixed combinations which are not phraseologisms, such as German *Dafürhalten* which only occurs in the expression *nach Xs Dafürhalten* "according to X's opinion". Our German collection contains 12 nouns of this type. All 12 items are frequent enough to occur in general dictionaries, however, their treatment is not uniform.

A corpus-based, up to date, state of the art lexical resource for German is the *Digitales Wörterbuch der Deutschen Sprache* (DWDS, *http://www.dwds.de*, visited March 2008). It contains all 12 words. However, they are classified in various ways: Six are classified as "occurs only in the fixed expression"; two as "occurs mainly in the combination", four do not contain such a comment in the lexical description. One of these four, *Obacht* "attention", is marked for a regional restriction to Southern varieties, but the entry fails to mention that the noun is restricted to an exclamation, *Obacht!*, and a particular expression, *Obacht geben* "pay attention".

The Deutsches Universalwörterbuch (Duden 2001) also mentions all 12 BWs. For eleven it notes that they are restricted to a particular expression. The two problematic cases are the noun *Obacht* " attention' for which also only the regional restriction is mentioned and *Dafürhalten*, which does not receive an entry of its own but is mentioned in a fixed expression associated with the obsolete verb *dafürhalten*.

This short discussion shows that this type of BWs is fairly well captured in standard dictionaries. Still, the information collected in our database can be of help to achieve more consistency in the treatment of the respective items.

3. Polarity items

3.1. *General characteristics*

Negative polarity items (NPI) are words or multi-word units that prototypically occur in an environment characterized as "negative" or "affective". Typical instances of NPIs are the English word *any* or the German word *jemals* "ever". NPIs occur in the scope of negation as well as in a variety of other semantically or pragmatically related contexts (such as interrogatives, antecedents of conditionals, modifiers of superlative and universal NPs, or complements of adversative predicates, to name a few). Since Klima (1964) the discussion about (I) what makes an item sensitive to polarity and (II) what can license an NPI hasn't come to an end yet.

Positive polarity items (PPI, e.g. *ziemlich* "pretty") are words or multi-word units which cannot occur in the scope of negation. In the literature it has been observed that NPI-licensing contexts have an anti-triggering effect on PPIs. However, the cross-linguistic documentation of PPIs is still very poor and thus the discussion about PPI-licensing is based on a few examples only. To improve the situation for German, our database is currently being extended to German PPIs.

PIs reveal diverse occurrence patterns. Zwarts (1997) distinguishes three types of NPIs: superstrong, strong, and weak ones, according to the logical properties of their respective licensing contexts. That is to say, all NPIs are licensed by classical negation (*not*) but only a subset may occur in the scope of e.g. an downward-entailing expression (*few*).

3.2. *Lexicographic representation*

In lexicons, polarity sensitivity of lemmas is not always explicitly mentioned. For example, the lexical entry for *any* in Merriam-Webster's online dictionary[1] doesn't say anything about its NPI-hood, although the *some/any* distinction is highly relevant for learners of English. Where a lemma's distribution is constrained regarding polarity, the information remains quite vague. The NPI *sonderlich* "particularly" is described in Duden (2001) as "**I.** <Adj.> **1.** (nur in Verbindung mit einer Verneinung o.ä.) **a)** besonders,..." ("...only in combination with negation or the like..."). The language user is left without information as to what "or the like" actually refers to,

[1] *http://www.merriam-webster.com/dictionary.*

let alone what negation exactly means. Another example is the lexical entry for *Deut* "brass farthing". Duden (2001) specifies its distribution as "nur in der Fügung *keinen/nicht einen D.*" ("only in the fixed expression..."). Although the expression *keinen Deut* "not one bit" may be regarded as prototypical, our corpus data show that the distribution of *Deut* is by no means *restricted* to that: its occurrence in the scope of *only* and in the restrictor of a comparative lead us to classifying *Deut* as a weak NPI. Thus, a clear notion of polarity sensitivity in lexicons is still a desideratum. A more transparent treatment will help both language users and learners to better understand a word's usage in context.

4. The collection of distributionally idiosyncratic items (CoDII)

4.1. *Conceptual design and technical realization*

To collect and document distributionally idiosyncratic items such the BW *Dafürhalten* "positive consideration" or the NPI *sonderlich* "particularly" in a systematic way, we compiled an electronic database, the Collection of Distributionally Idiosyncratic Items (CoDII), which provides a uniform description format.[2] Each CoDII item is characterized by the following information blocks: General Information (including glosses, translations as well as paraphrases and information about a possible free distribution, if appropriate), Syntactic Information (including syntactic variations), Licensing Contexts (for PIs), Classification, and, optionally, search patterns (optimized for use in dynamic corpora such as the Internet).

This information is encoded in XML. The underlying DTD has been specified in such a way that: (1) the element *codii* is the document root, and different collections are identified by the attributes *type* (for specifying the collection type) and *xml:lang* (for the language of the data); (2) the content model of the element *codii* consists of two elements: *dii-list*, whose content is a list of distributionally idiosyncratic items, and *dii-examples*, which contains a list of examples. This implements the idea of separating data (i.e. the examples) and the linguistic documentation (i.e. the entries for distributionally idiosyncratic items). The two parts of each collection are linked by pointers (*idref*). The content model of the element *dii-list* consists of a list of *dii-entry* elements. The content model of each *dii-entry* element consists of a set of elements which encode the four information blocks mentioned above. Finally, the content model of the element *dii-examples* consists of a list of *example* elements, each providing an example for a given item and information on its source. The technical details of the CoDII-XML-encoding of BWs have been published in Sailer and Trawiński (2006) and a corresponding descriptlon of PIs has been provided in Trawiński and Soehn (To appear).

For the syntactic annotation of the German, English and Romanian items, the Stuttgart-Tübingen Tagset (STTS), the syntactic annotation scheme from the Syntactically Annotated Idiom Database (SAID), and the tagset from the Multilingual Text Tools and Corpora for Central and Eastern European Languages (MULTEXT-East) were used, respectively. For each context, appropriate examples are cited from corpora, the Internet and the linguistic literature. CoDII is freely accessible on the Internet at *www.sfb441.uni-tuebingen.de/~a5/codii*. Figure 1 shows the browser display of the German NPI *sonderlich* "particularly".

[2] CoDII is developed by (i) Project A5, *Distributional Idiosyncrasies*, of the Collaborative Research Center SFB 441 (*Linguistic Datastructures*) at the University of Tübingen, funded by the German Research Foundation (DFG), *www.sfb441.uni-tuebingen.de/a5/index-engl.html*, and (ii) members of the linguistics section of the English Department of the University of Göttingen.

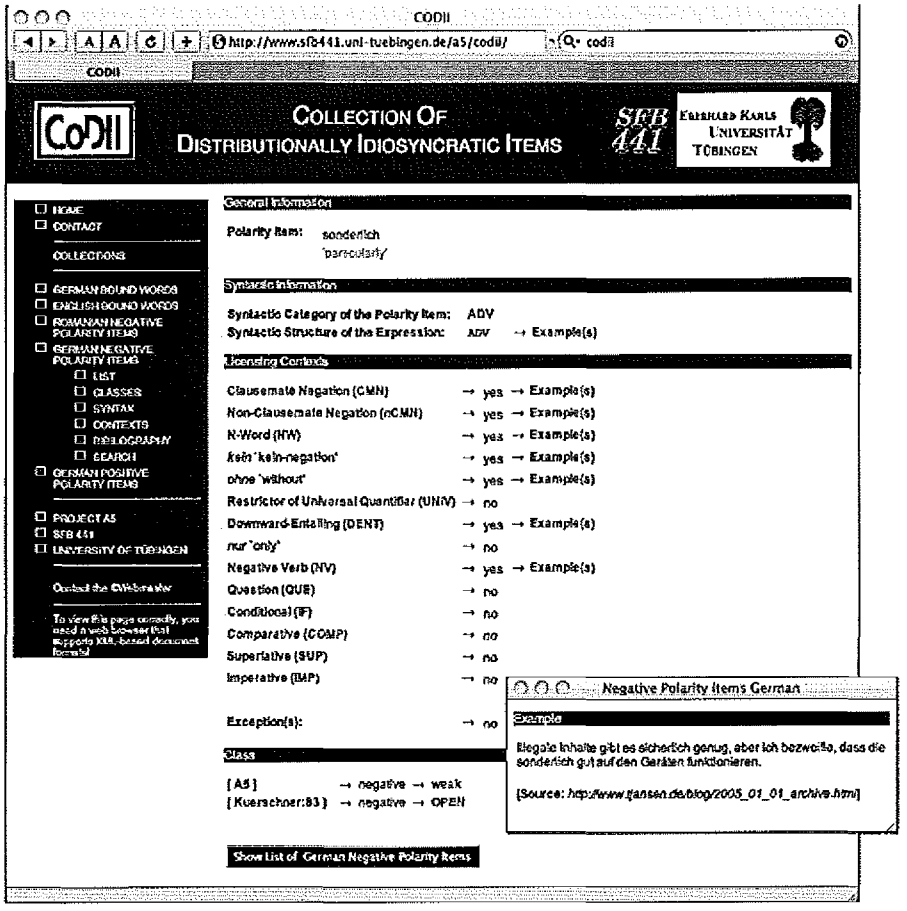

Figure 1. The browser display for the German NPI *sonderlich* "particularly"

As Figure 1 demonstrates, the user interface of CoDII displays all the linguistic information for each item, including licensing contexts together with the links to corresponding examples. Comments, information about the classification systems, licensing contexts and the relevant examples can be obtained by clicking on the links in the display.

Five collections of distributionally idiosyncratic items are currently available in CoDII: (1) a collection of 444 BWs in German, (2) a collection of 77 BWs in English, (3) a collection of 58 NPIs in Romanian, (4) a collection of 84 NPIs in German, and (5) a collection in 52 PPIs of German. The well-established international encoding standard and the linguistically motivated data structure design will make it possible to add further languages, classifications and other types of idiosyncratic items.

CoDII not only compiles, documents and (alphabetically) lists distributionally idiosyncratic items, it also offers dynamic and flexible access. Integrating CoDII into the Open Source XML database eXist (*http://exist.sourceforge.net/*), has opened the possibility of searching for particular lemmata, syntactic properties and classifications.

4.2. *Some statistics about the collected items*

The integration of CoDII's collections in a database not only allows for a flexible search but also makes it possible to quickly acquire statistical facts about the items. For example, one can see that the overwhelming majority of German BWs are nouns (79%, e.g. *Schlafittchen* "collar"), followed by predicative adjectives (7%, e.g. *sattsam* "widely"), proper names (5%, e.g. *Pandora*), and verbs (3%, e.g. *fleuchen* "fly"). VPs (66%) are the most common syntactic environment for (the typically nominal) BWs. In 87 cases (20%) a BW is the complement of a specific preposition. These "unique nominal complements" form an important subclass of BWs (e.g. *auf Anhieb* "at first attempt"). From a theoretical point of view,

these data provide excellent evidence that nonheads, including complements, can impose restrictions on the heads they combine with. In addition to syntax, one may also investigate classification issues. 26% of our BWs are specified as lexically decomposable. Furthermore, 64% are bound to their specific contexts, whereas 36% may occur freely with a different meaning (to take an English example: *lurch* in *to leave so. in the lurch*) or within domain-specific terminology. Some facts about NPIs are depicted in Table 1 (all data are as of March 2008).

Category		Classification	
verbs or verb phrases	62%	weak	71%
adverbs	18%	strong	23%
nouns or noun phrases	14%		
prepositional phrases	6%	superstrong	6%

Table 1. German NPIs in CoDII

5. Comparable collections

5.1. *The collections of bound words*

Several other projects have constructed resources for idiomatic expressions. These projects differ from CoDII by the corpora used, the kind of data and the applied methods.

Usuelle Wortverbindungen[3] (Conventionalized Word Combinations) of the Institut für Deutsche Sprache (IDS) (Steyer 2004) starts from statistically highly frequent words and subjects them to a co-occurrence analysis. This analysis serves as the basis for a linguistic and lexicographic description of the typical usage patterns of a word. In contrast to this collection, CoDII is based on linguistic intuitions and theoretical considerations. In part, this is due to the low frequency of a number of BWs. Another important difference is that the IDS project only uses the corpora of the IDS, to have full control over the frequency data. For CoDII we try to collect as much information as possible about a given item. For this reason we want to include data from different sources and retrieval strategies for different corpora.

Kollokationen im Wörterbuch[4] (Collocations in the Lexicon, completed in 2006) of the Berlin-Brandenburgische Akademie der Wissenschaft (Fellbaum et al. 2005) is based on the DWDS corpus. Similar to CoDII, the project started with idioms from the phraseological literature, but focused exclusively on German VP idioms. The database contains corpus-based linguistic descriptions of 917 idioms.

The Syntactically Annotated Idioms Database (SAID, Kuiper et al. 2003) contains a large number of English idioms, but it lists only syntactic information about them. As the encoding used there allows users to investigate structural generalizations about idioms, we used the same encoding for representing syntactic structures in the English CoDII-BW.

5.2. *The collections of polarity items*

Despite the rich literature on polarity, there are only a few collections of polarity items. Welte (1978) lists NPIs for German and English and von Bergen and von Bergen (1993) abounds with examples of English NPIs and includes some German ones as well. Yet, these listings are presumably not intended to be exhaustive. The most extensive list for German to our knowledge is provided in Kürschner (1983). However, his collection is based entirely on the author's intuitions and we have some doubts as to the NPI status of more than a half of his 344 items. Thus, a more systematic way to acquire NPIs is needed. Lichte and Soehn (2007) extracted a list of NPI candidates from the *Tübingen Partially Parsed Corpus of Written German* which not

[3] *http://www.ids-mannheim.de/ll/uwv/.*

[4] *http://kollokationen.bbaw.de/.*

only provided new German NPIs but also allowed us to validate some of the NPIs in the above-mentioned collections.

For PPIs, the empirical base is much weaker. We collected the items for CoDII-PPI.de on the basis of our own intuitions and the literature, including van Os (1989) and Ernst (2005). Our collection is currently being expanded and the items to be included are validated psycholinguistically.

6. Conlusions

Lexical items with idiosyncratic distribution patterns pose a challenge to lexicography and to formal linguistic theories alike. Partly as a result of the insufficiencies in these fields, they also remain a difficult topic in language teaching and language learning, as well as in computational applications. We believe that a considerable part of the problems is due to the lack of comprehensive, systematic and easily accessible resources which document the empirical facts. Better knowledge of the data and their relevant properties may lay the foundations for their satisfactory theoretical description, for adequate specifications of the usage and structure of distributionally restricted items for computational tools, and for useful explanations of their contextual conditions in educational materials.

With the work presented here we set out to remedy the lack of solid empirical ground by creating a modular and extensible architecture for a multilingual documentation of different kinds of distributionally idiosyncratic items. We demonstrated the potential of our framework with comprehensive databases of bound words in English and German, and negative polarity items in English, German and Romanian, and with our ongoing work on a database of positive polarity items in German. A key property of our proposal is the representation of the data according to internationally recognized standards, and the high degree of searchability of the database with user-specified queries. We hope that the architecture will enhance our knowledge of distributionally idiosyncratic items and stimulate further extensions by additional databases.

References

Aronoff, M. (1976). *Word Formation in Generative Grammar.* Cambridge: MIT Press.

von Bergen, A.; von Bergen, K. (1993). *Negative Polarität im Englischen.* Tübingen: Gunter Narr.

Dobrovol'skij, D. (1988). *Phraseologie als Objekt der Universallinguistik.* Leipzig: Enzyklopädie.

Duden (2001). *Deutsches Universalwörterbuch.* Mannheim: Dudenverlag.

Ernst, T. (2005). *On Speaker-oriented Adverbs as Positive Polarity Items.* In Polarity From Different Perspectives, poster at the Workshop held at New York University. 11.–13. March 2005.

Fellbaum, C.; Kramer, U.; Neumann, G. (2005). "Corpusbasierte lexikographische Erfassung und linguistische Analyse deutscher Idiome". In Häcki Buhofer, A.; Burger, H. (eds.). *Phraseology in Motion I. Methoden und Kritik. Akten der Internationalen Tagung zur Phraseologie (Basel, 2004).* Hohengehren: Schneider Verlag. 183-199.

Hoeksema, J. (2002). *De negatief-polaire uitdrukkingen van het Nederlands.* Manuscript, Groningen.

Klima, E. (1964). "Negation in English". In Fodor, J. A.; Katz, J. (eds.). *The Structure of Language.* Englewood Cliffs: Prentice Hall. 246-323.

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

Kuiper, K. et al. (2003). *Syntactically Annotated Idiom Database* (SAID) v.1. Documentation to a LDC resource.

Lichte, T.; Soehn, J. P. (2007). "The Retrieval and Classification of Negative Polarity Items Using Statistical Profiles". In Featherston, S.; Sternefeld, W. (eds.). *Roots: Linguistics in Search of its Evidential Base.* Berlin: Mouton de Gruyter. 249-266

Sailer, M.; Trawiński, B. (2006). "The Collection of Distributionally Idiosyncratic Items: A Multilingual Resource for Linguistic Research". In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006.* Genoa, Italy. 471-474.

Steyer, K. (2004). "Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikographische Perspektiven". In Steyer, K. (ed.). *Wortverbindungen – mehr oder weniger fest.* Berlin: de Gruyter. 87-116.

Trawiński, B.; Soehn, J. P. (to appear). "A Multilingual Database of Polarity Items". In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008,* Marrakech, Morocco.

van Os, C. (1989). *Aspekte der Intensivierung im Deutschen.* Tübingen: Gunter Narr.

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

Zwarts, F. (1997). "Three Types of Polarity". In Hamm, F.; Hinrichs, E. W. (eds.). *Plurality and Quantification.* Dordrecht: Kluwer Academic Publishers. 177-237.