A Multilingual Database of Polarity Items

Beata Trawiński and Jan-Philipp Soehn

University of Tübingen
SFB 441
Nauklerstraße 35
D-72074 Tübingen

trawinski@sfs.uni-tuebingen.de
jp.soehn@uni-tuebingen.de

Abstract

This paper presents three electronic collections of polarity items: (i) negative polarity items in Romanian, (ii) negative polarity items in German, and (iii) positive polarity items in German. The presented collections are a part of a linguistic resource on lexical units with highly idiosyncratic occurrence patterns. The motivation for collecting and documenting polarity items was to provide a solid empirical basis for linguistic investigations of these expressions. Our databe provides general information about the collected items, specifies their syntactic properties, and describes the environment that licenses a given item. For each licensing context, examples from various corpora and the Internet are introduced. Finally, the type of polarity (negative or positive) and the class (superstrong, strong, weak or open) associated with a given item is specified. Our database is encoded in XML and is available via the Internet, offering dynamic and flexible access.

1. Introduction and Motivation

This paper presents a multilingual database of polarity items (PIs), which is a part of the *Collection of Distributionally Idiosyncratic items* (CoDII, available at http://www.sfb441.uni-tuebingen.de/a5/codii). CoDII is a linguistic resource for lexical units which have highly idiosyncratic occurrence patterns. The collection has been developed by Project A5 of the Collaborative Research Center SFB 441 *"Linguistic Datastructures"* at the University of Tübingen, funded by the German Research Foundation (DFG).[1]

The motivation for the compilation of CoDII was to provide an empirical base for linguistic investigations on lexical items revealing distributional idiosyncrasies. This task includes collecting and listing the particular items, providing existing linguistic documentation, and introducing examples related to these items found in various corpora, including the Internet.

Up to this point, we were concerned with two kinds of expressions: (i) Bound Words as expressions whose distribution is restricted by lexical co-occurrence patterns, and (ii) negative and positive polarity items (NPIs/PPIs) as expressions whose distribution is restricted to certain semantico-pragmatic contexts. Five collections of distributionally idiosyncratic items are currently available in CoDII: German Bound Words, English Bound Words, Romanian NPIs, German NPIs, and German PPIs. The collections of Bound Words are described in detail in (Sailer and Trawiński, 2006). In the paper at hand, three new subcollections of CoDII are presented: Negative polarity items in Romanian (CoDII-NPI.ro), negative polarity items in German (CoDII-NPI.de), and positive polarity items in German (CoDII-PPI.de).

Despite the rich literature on polarity, there are only a few collections of PIs. 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 entirely based 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 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), (van der Wouden, 1997), and (Ernst, 2005). Our collection is currently being expanded and the items to be included are being validated psycholinguistically by a representative number of native speakers' acceptability judgements.

Section 2 provides some basic properties of negative and positive polarity items. The conceptual design and the technical realization of CoDII's PI collections are outlined in Section 3. In Section 4, the user interface and the database functionalities are presented. Some statistics about collected PIs is provided in Section 5. Final conclusions and future work are sketched in Section 6.

2. Polarity Items

In the following subsections, some basic properties of negative and positive polarity items are introduced.

2.1. Negative Polarity Items

NPIs such as *any*, *budge*, or *ever* are words or idiomatic phrases that prototypically occur in an appropriately characterized ("negative" or "affective") environment. NPIs

[1] The URL of the project's website: www.sfb441.uni-tuebingen.de/a5/index-engl.html.

have been studied intensely in several linguistic frameworks since Klima's seminal work on negation in English (Klima, 1964). They occur both in the scope of negation as well as in a variety of environments (such as interrogatives, antecedents of conditionals, the restrictor of superlative and universal NPs, complements of adversative predicates, etc.) which can more or less clearly be related to negation or affectiveness. The theory of van der Wouden (1997) conceptualizes the basic property of polarity sensitivity as collocational restrictions, regarding NPIs as collocates which have a meaning of their own and exhibit idiosyncratic restrictions on their contexts. This perspective is a consequence of the yet unresolved problem of generalizing over licensing contexts and possible inherent properties of PIs that render them sensitive to polarity. His theory predicts lexical idiosyncrasies in PIs which are related to those we observe in other elements with a varying degree of frozenness, such as idiomatic expressions. This is confirmed by our data we collected in CoDII. The fact that some lexical elements are sensitive to negativity does not follow from other properties. There are sets of (near-)synonyms (*sonderlich/besonders* 'particularly', *scheren/kümmern* 'care', *Hehl/Geheimnis* 'secret', or *von ungefähr/durch Zufall* 'by chance') in which the first item is an NPI and the second is not. Furthermore, there are cognate idioms in Dutch and German, one being an NPI and the other a PPI. In spite of all idiosyncrasies, van der Wouden classifies (analogously to Zwarts, 1997) PIs according to the logical strength of their licensing contexts: classical negation (anti-morphic, such as *not*), regular negation (anti-additive, such as *nobody* or *never*), and minimal negation, which is downward-entailing (such as *hardly*). He establishes three categories of NPIs, cf. the following table.

NPI	*negation*		
	classical	regular	minimal
superstrong	+	–	–
strong	+	+	–
weak	+	+	+

2.2. Positive polarity items

Although most attention has been paid to negative polarity items, the study of positive polarity items goes back as far as (Baker, 1970). Unlike NPIs, PPIs (such as *pretty*, *already*, or *would rather*) cannot occur in the scope of negation. Put differently, NPI-licensing contexts have an anti-triggering effect on PPIs. However, the documentation of PPIs in different languages is still very poor. Our current work serves to improve the empirical base for German. The distribution of PPIs is equally unclear as that of NPIs. Although in some theories PPIs are excluded from anti-morphic and anti-additive contexts and allowed in all other environments, van der Wouden makes the same (more finegrained) distinction as for NPIs.

PPI	*Negation*		
	classical	regular	minimal
superstrong	–	–	–
strong	–	–	+
weak	–	+	+

This concludes the introduction to polarity items. In the following sections, we describe the modeling and the representation of PIs in CoDII, and indicate how they can be accessed via the Internet.

3. Modeling and Representing PIs in CoDII

Each PI is characterized in CoDII by four information blocks: General Information, Syntactic Information, Licensing Contexts and Class.

The block **General Information** identifies PIs by providing the particular PI, its English gloss, its English translation, possible expressions in which the PI occurs, and, if appropriate, the set of possible paraphrases of these expressions. Within the block **Syntactic Information**, details on the syntactic category of the PI and, if appropriate, the syntactic structure of the expression in which the PI occurs, are specified. For the syntactic description of German NPIs and PPIs and expressions in which they occur, the Stuttgart-Tübingen Tagset (STTS)[2] has been used. For the syntactic description of Romanian NPIs, we used the (modified) tagset from the Multilingual Text Tools and Corpora for Central and Eastern European Languages (MULTEXT-East)[3].

The block **Licensing Contexts** provides information on the environment that licenses a given PI. The following licensing contexts are taken into account:

- Clausemate negation (cmn): There is a negation particle (*not*) in the same clause.
- Non-clausemate negation (ncnm): Negation particle (*not*) occurs in the matrix clause, while the PI appears in the subordinate clause.
- N-Word (nw): PI is in the scope of an N-Word such as *nobody*, *nothing*, *never* or in their equivalent terms in the other languages.
- '*kein*-negation' (kein): PI is in the scope of *kein* (for German).
- *without*: PI is in the scope of *without*.
- Universal quantifier (univ): PI is in the restrictor of a universal quantifier.
- Other downward-entailing contexts (dent): PI is in the scope of a downward-entailing expression such as *few* or *hardly*.
- *only*: PI is in the scope of *only*.
- Negative verb (nv): PI is in the scope of a non-factive predicate such as *doubt*, *fear*, or *it is impossible / improbable that* or of an adversative attitude predicate such as *be surprised* or *regret*.
- Question (que): PI occurs within a question.
- Conditional (if): PI is in the restrictor of a conditional operator such as *if*.

[2] http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html

[3] http://nl.ijs.si/ME

- Comparative (comp): PI is in the restrictor of a comparative.
- Superlative (sup): PI is in the restrictor of a superlative.
- Imperative (imp): PI occurs within an imperative clause.

In addition, exceptional cases can be specified, i. e. corpus evidence for a PI that does not occur in a licensing context. For each licensing context, corresponding examples are provided. The examples for the Romanian NPIs have been acquired from the Romanian electronic corpus developed by Rada Mihalcea from the Department of Computer Science and Engineering at the University of North Texas (USA), from the Romanian electronic corpus developed at the Romanian Academy Center for Artificial Intelligence (RACAI), as well as from the Internet via Google. A number of examples have been constructed by Gianina Iordăchioaia, a native speaker of Romanian who collaborated on CoDII-NPI.ro. To acquire the examples for the German NPIs and PPIs, we consulted corpora of the Mannheim Institute of German Language[4] and the Internet via Google.

Finally, the block **Class** specifies the type of polarity, which is *negative* in the case of NPIs and *positive* in the case of PPIs, and the class associated with a given PI. In CoDII, we use the following classes of PIs with the following definitions:

- Superstrong:
  - NPIs are superstrong if they are licensed only by antimorphic contexts (overt negation).
  - PPIs are superstrong if they are incompatible with downward-entailing, anti-additive (comprising n-words and *without*) and antimorphic contexts.
- Strong:
  - NPIs are strong if they are licensed by antimorphic and anti-additive contexts.
  - PPIs are strong if they are compatible with downward-entailing contexts but incompatible with anti-additive and antimorphic ones.
- Weak:
  - NPIs are weak if they are licensed by antimorphic, anti-additive, and downward-entailing contexts (plus the remaining ones).
  - PPIs are weak if they are compatible with downward-entailing and anti-additive contexts but incompatible with antimorphic ones.
- Open: for undefined classification.

CoDII-NPI.ro, CoDII-NPI.de and CoDII-PPI.de are internally encoded in XML. The DTD for CoDII has been specified in such a way that the element `codii` constitutes the document root and its instance is identified by the attribute `type` (for specifying the collection type) and the attribute `xml:lang` (for specifying the language the data come from). The content model of the element `codii` comprises two elements: `dii-list`, a list of PIs, and `dii-examples`, a list of examples.

The element `dii-list` consists of a list of `dii-entry` elements, whose content is composed of a set of elements which identify PIs (`dii`), describe documentation on each PI (`dii-classification`), present syntactic properties of PIs (`dii-syntax`), and specify licensing contexts of PIs (`licensers`). Figure 1 and Figure 3 present fragments of the CoDII-XML-encoding of the German PPI *ziemlich* 'pretty' and the German NPI *sonderlich* 'particularly', respectively, according to this representation model.

As the XML-representation in Figure 1 shows, the German PPI *ziemlich* can only occur within a question, in the scope of a conditional operator such as *if*, in the scope of *nur* 'only', and in the scope of a superlative. On this basis, we classified this PPI as a strong PPI. The description in Figure 3 demonstrates that the German NPI *sonderlich* can be licensed by clausemate and non-clausemate negation, can occur in the scope of *kein*, of an n-word such as *niemand* 'nobody' or *niemals* 'never', of a downward-entailing expression, of a non-affirmative predicate or of an adversative attitude predicate, and in the scope of *ohne* 'without'. On the basis of these licensing contexts, we classified this NPI as weak.

All representations of examples associated with the PIs in CoDII are contained in the same document as the PIs themselves, and, as already indicated, are encoded by means of the element `dii-examples`. The content model of the element `dii-examples` consists of a list of `example` elements. The `example` elements are identified via the attribute `id` and are linked to the appropriate PIs by means of the attribute `dii`, specified at `example`, and the attribute `hits`, specified at the corresponding `dii-entry` elements. Finally, the content model of each `example` element consists of an example for a given item (`ol`) and information on its source (`source`). Figure 2 and Figure 4 provide a corpus example for the German PPI *ziemlich* and a corpus example for the German NPI *sonderlich*, respectively.

[4] `http://www.ids-mannheim.de/cosmas2/`.

```
<dii-entry id="ziemlich">
 <dii><ol>ziemlich</ol>
      <en>pretty</en>
 </dii>
 <dii-classification>
   <dii-class category="pi"
              subcategory="ppi"
              type="PPI-Project"
              class="strong"
              original-class="no">
   <bibliography bib-item="PPI-Project"/>
   </dii-class>
   <dii-class category="pi"
              subcategory="ppi"
              type="Wouden97"
              class="OPEN"
              original-class="no">
   <bibliography bib-item="Wouden:97"/>
   </dii-class>
 </dii-classification>
 <dii-syntax hits="ziemlich-01" cat="ADV">
   <dii-expression-syntax>
       ADV
   </dii-expression-syntax>
 </dii-syntax>
 <licensers>
   <cmn given="no"/>
   <ncmn given="no"/>
   <kein given="no"/>
   <nw given="no"/>
   <dent given="no"/>
   <nv given="no"/>
   <que given="yes" hits="ziemlich-02"/>
   <imp given="no"/>
   <if given="yes" hits="ziemlich-03"/>
   <without given="no"/>
   <only given="yes" hits="ziemlich-04"/>
   <univ given="no"/>
   <comp given="no"/>
   <sup given="yes" hits="ziemlich-05"/>
   <exc given="no"/>
 </licensers>
</dii-entry>
```

Figure 1: CoDII-XML-representation of the German PPI *ziemlich*

```
<example dii="ziemlich" id="ziemlich-01">
  <source corpus="cosmasII">
Mannheimer Morgen, 18.09.2000
  </source>
  <ol>
Das bedeutet angesichts des morschen Zustandes
der Holzgebude eine Menge Arbeit.
Auch das Gelnde ist ziemlich verwahrlost.
  </ol>
</example>
```

Figure 2: CoDII-XML-representation of a corpus example for the German PPI *ziemlich*

```
<dii-entry id="sonderlich">
 <dii><ol>sonderlich</ol>
      <en>particularly</en>
 </dii>
 <dii-classification>
   <dii-class category="pi"
              subcategory="npi"
              type="A5" class="weak"
              original-class="no">
   <bibliography bib-item="A5"/>
   </dii-class>
   <dii-class category="pi"
              subcategory="npi"
              type="Kuerschner83"
              class="OPEN"
              original-class="no">
   <bibliography bib-item="Kuerschner:83"/>
   </dii-class>
 </dii-classification>
 <dii-syntax hits="sonderlich-01" cat="ADV">
   <dii-expression-syntax cat="ADV">
       ADV
   </dii-expression-syntax>
 </dii-syntax>
 <licensers>
   <cmn given="yes" hits="sonderlich-01"/>
   <ncmn given="yes" hits="sonderlich-02"/>
   <kein given="yes" hits="sonderlich-03"/>
   <nw given="yes" hits="sonderlich-04"/>
   <dent given="yes" hits="sonderlich-05"/>
   <nv given="yes" hits="sonderlich-06"/>
   <que given="no"/>
   <imp given="no"/>
   <if given="no"/>
   <without given="yes" hits="sonderlich-07"/>
   <only given="no"/>
   <univ given="no"/>
   <comp given="no"/>
   <sup given="no"/>
   <exc given="no"/>
 </licensers>
</dii-entry>
```

Figure 3: CoDII-XML-representation of the German NPI *sonderlich*

```
<example dii="sonderlich" id="sonderlich-06">
  <source corpus="google">
http://www.tjansen.de/blog/2005_01_01_archive.html
  </source>
  <ol>
     Illegale Inhalte gibt es sicherlich genug,
     aber ich bezweifle, dass die sonderlich
     gut auf den Gerten funktionieren.
  </ol>
</example>
```

Figure 4: CoDII-XML-representation of a corpus example for the German NPI *sonderlich*

4. Accessing the Database via the Internet

CoDII-NPI.ro, CoDII-NPI.de and CoDII-PPI.de are freely accessible on the Internet at `http://www.sfb441.uni-tuebingen.de/a5/codii`. Figure 5 shows the browser display of the German NPI *sonderlich*, including an example.

The browser display in Figure 5 presents the CoDII-XML-description of the NPI *sonderlich* in Figure 3 and the CoDII-XML-representation of the examples for this NPI in Figure 4. Comments, information about the classification systems, used tags, licensing contexts and the relevant examples can be obtained by clicking on the links in this display.

CoDII-NPI.ro, CoDII-NPI.de and CoDII-PPI.de not only compile, document and (alphabetically) list PIs, but they also offer 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 lemmas, syntactic properties, licensing contexts and classifications.

5. Some Statistics about the Collected PIs

The integration of CoDII collections in a database not only allows for a flexible search but also makes it possible to quickly acquire statistical facts about the items.[5] For example, one can see that the overwhelming majority of German NPIs are verbs or verb phrases (62%, e. g. *ausstehen können* 'can stand'), followed by adverbs (18%, e. g. *jemals* 'ever'), nouns or noun phrases (14%, e. g. *Deut* 'a tittle'), and prepositional phrases (6%, e. g. *von ungefähr* 'by chance'). The classification by (Zwarts, 1997) or (van der Wouden, 1997) is reflected in our collection as follows: 71% are categorized as weak NPIs (e. g. *Deut*), 23% as strong (e. g. *sich beirren lassen* 'to let oneself be misled') and 6% as superstrong NPIs (e. g. *weiter verwunderlich* 'very remarkable'). Turning to PPIs, the situation is quite different. 46% of all German PPIs are adverbs (e. g. *durchaus* 'absolutely'), followed by 38% verbs or verb phrases (e. g. *munkeln* 'rumor'). 16% of PPIs in our collection are adjectives (e. g. *ziemlich* 'pretty'). Concerning the classification, we find 3 weak PPIs (e. g. *schon* 'already'), 36 strong ones (e. g. *durchaus*), and 11 superstrong PPIs (e. g. *stockdumm* 'utterly stupid').

6. Conclusions and Outlook

CoDII-NPI.de currently includes 84 German NPIs, CoDII-PPI.de includes 52 German PPIs, and CoDII-NPI.ro includes 58 Romanian NPIs. The items of CoDII-NPI.ro correspond to the English, German and Dutch NPIs discussed in linguistic literature, since there is no specific collection of Romanian NPIs available. The collections are work in progress and as such are currently being expanded. They already reveal the great variety that exists among PIs. The obvious question is whether to treat the different parts-of-speech, idioms and non-idioms, quantifiers, verbs, etc. – all being PIs – in a uniform way. Discovering subclasses and (re)categorizing PIs is work that is on our agenda as well. Thus, lexical items with an idiosyncratic distribution are challenging data for both lexicographers and theoretical linguists. In our opinion, a considerable part of the problems is due to the lack of comprehensive, systematic and easily accessible resources which document the empirical facts. A better knowledge of their relevant properties may provide a basis for their satisfactory theoretical description and for adequate specifications of their usage which in turn will be helpful for developing educational materials and computational tools for natural language processing.

The well established international encoding standard and the linguistically motivated internal data structure of CoDII will make it possible to add further languages, classifications and to create collections of other types of items.

7. References

C. Lee Baker. 1970. Double Negatives. *Linguistic Inquiry*, 1:169–186.

Thomas Ernst. 2005. On Speaker-Oriented Adverbs as Positive Polarity Items. Electronic Poster for the Workshop: Polarity From Different Perspectives, New York University, 11.–13.03.2005.

Edward Klima. 1964. Negation in english. In J. A. Fodor and J. Katz, editors, *The Structure of Language*, pages 246–323. Prentice Hall, Englewood Cliffs, New Jersey.

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*. Mouton de Gruyter.

Manfred Sailer and Beata Trawiński. 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*, pages 471–474, Genoa, Italy.

Ton van der Wouden. 1997. *Negative Contexts. Collocation, polarity and multiple negation*. Routledge, London and New York.

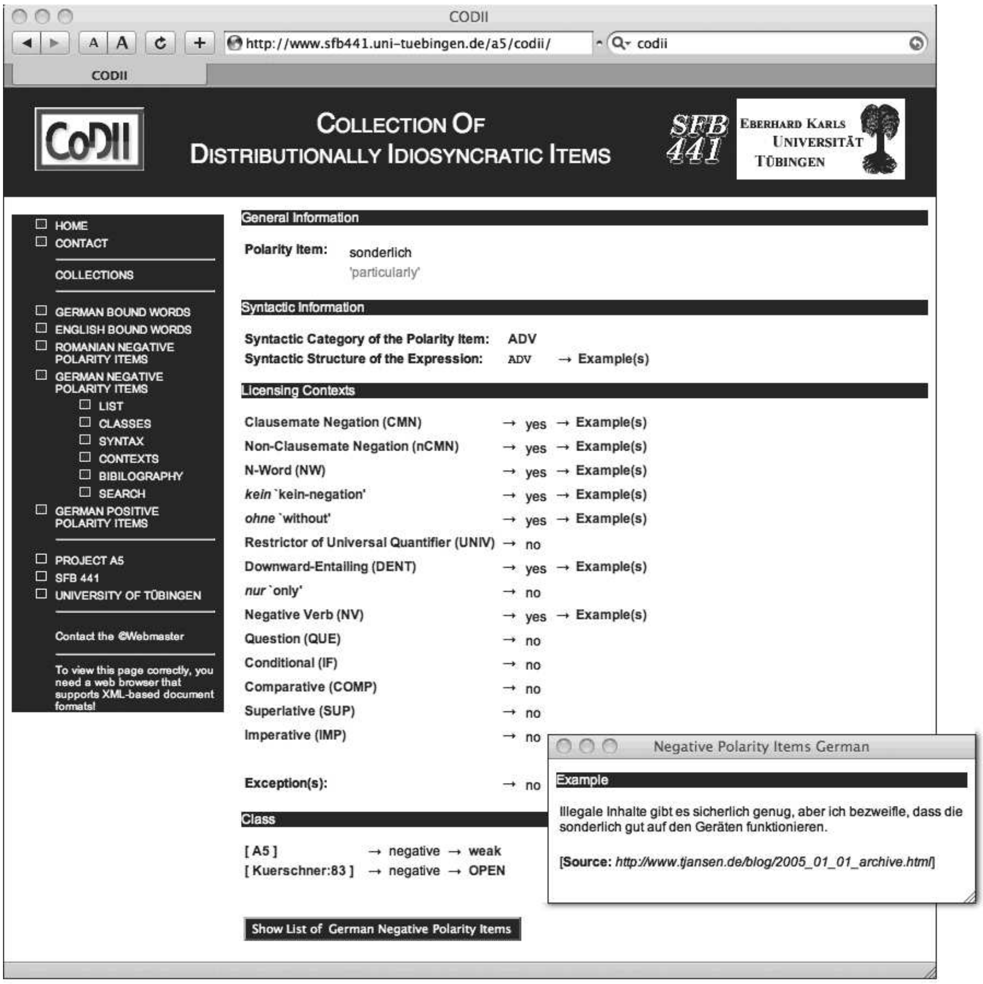
Charles van Os. 1989. *Aspekte der Intensivierung im Deutschen*. Gunter Narr, Tübingen.

Anke von Bergen and Karl von Bergen. 1993. *Negative Polarität im Englischen*. Gunter Narr, Tübingen.

Werner Welte. 1978. *Negationslinguistik. Ansätze zur Beschreibung und Erklärung von Aspekten der Negation im Englischen*. Wilhelm Fink Verlag, Munich.

Frans Zwarts. 1997. Three Types of Polarity. In F. Hamm and E. W. Hinrichs, editors, *Plurality and Quantification*, pages 177–237. Kluwer Academic Publishers, Dordrecht.

[5] All data are as of March 2008.


CODII
http://www.sfb441.uni-tuebingen.de/a5/codii/
codii
CODII
CoDII
COLLECTION OF DISTRIBUTIONALLY IDIOSYNCRATIC ITEMS
SFB 441
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
□ HOME
□ CONTACT
COLLECTIONS
□ GERMAN BOUND WORDS
□ ENGLISH BOUND WORDS
□ ROMANIAN NEGATIVE POLARITY ITEMS
□ GERMAN NEGATIVE POLARITY ITEMS
□ LIST
□ CLASSES
□ SYNTAX
□ CONTEXTS
□ BIBILOGRAPHY
□ SEARCH
□ GERMAN POSITIVE POLARITY ITEMS
□ PROJECT A5
□ SFB 441
□ UNIVERSITY OF TÜBINGEN
Contact the ©Webmaster
To view this page correctly, you need a web browser that supports XML-based document formats!
Show List of German Negative Polarity Items
Negative Polarity Items German

General Information

Polarity Item: sonderlich
'particularly'

Syntactic Information

Syntactic Category of the Polarity Item: ADV
Syntactic Structure of the Expression: ADV → Example(s)

Licensing Contexts

Clausemate Negation (CMN)	→ yes	→ Example(s)
Non-Clausemate Negation (nCMN)	→ yes	→ Example(s)
N-Word (NW)	→ yes	→ Example(s)
kein `kein-negation'	→ yes	→ Example(s)
ohne `without'	→ yes	→ Example(s)
Restrictor of Universal Quantifier (UNIV)	→ no	
Downward-Entailing (DENT)	→ yes	→ Example(s)
nur `only'	→ no	
Negative Verb (NV)	→ yes	→ Example(s)
Question (QUE)	→ no	
Conditional (IF)	→ no	
Comparative (COMP)	→ no	
Superlative (SUP)	→ no	
Imperative (IMP)	→ no	
Exception(s):	→ no	

Class

[A5]	→ negative	→ weak
[Kuerschner:83]	→ negative	→ OPEN

Example

Illegale Inhalte gibt es sicherlich genug, aber ich bezweifle, dass die sonderlich gut auf den Geräten funktionieren.

[Source: *http://www.tjansen.de/blog/2005_01_01_archive.html*]

Figure 5: Browser display for the German NPI *sonderlich* and for one of its examples