THE COLLECTION OF DISTRIBUTIONALLY IDIOSYNCRATIC ITEMS. AN INTERFACE BETWEEN DATA AND THEORY

FRANK RICHTER, MANFRED SAILER AND BEATA TRAWIŃSKI

Abstract. Dieses Papier gibt einen Überblick über CoDII, die *Collection of Distributionally Idiosyncratic Items*. CoDII ist eine elektronische Sammlung verschiedener Untergruppen lexikalischer Elemente, die sich durch idiosynkratische Distribution auszeichnen. Das bedeutet, dass sich die Verteilung dieser Lexeme im Text nicht alleine aufgrund ihrer syntaktischen Kategorie vorhersagen lässt. Die Methoden, die in der Entwicklung von CoDII angewandt werden, greifen über traditionelle Fachgrenzen hinaus und umfassen Korpuslinguistik, Computerlinguistik, Phraseologie und theoretische Sprachwissenschaft. Ein wichtiger Schwerpunkt unserer Diskussion liegt auf der Darstellung, inwiefern die in CoDII gesammelten, annotierten und unter anderem mit Suchwerkzeugen abfragbaren Daten dazu beitragen können, die linguistische Theoriebildung durch die Bereitstellung sorgfältig aufbereiteter Datensammlungen bei der Überprüfung ihrer Datengrundlage zu unterstützen.

1. Introduction

The Collection of Distributionally Idiosyncratic Items (CoDII) is an electronic resource for linguistic research. In its very design it crosses traditional boundaries of several linguistic subdisciplines. The methods and techniques that were used in its creation come from corpus linguistics, computational linguistics, phraseology and theoretical linguistics. Its goal is to provide a resource that is useful for researchers working in areas as diverse as lexicography, syntax, semantics and psycholinguistics. In this paper, we will present the main features of CoDII. An important part of this discussion will be to show that beyond being a valuable data repository that may be used for building specialized (electronic) resources or applications in specific areas of interest, CoDII can support theoretical linguistics by giving researchers structured access to a wealth of data to test and improve their theories.

Distributionally idiosyncratic items (DIIs) are special from two perspectives: First, they don't follow the distribution pattern that would be expected based on their syntactic categorial properties. Because of their irregular distributional properties, they are accessible to statistical corpus linguistic methods. Second, since they are expressions with strict context requirements, their failure to occur in their respective licensing context triggers clearcut ungrammaticality judgments by native speakers. Their location in an area which is simultaneously accessible to measurements of statistical distribution and to the investigation of the human grammatical system by grammaticality judgments makes DIIs ideally suited for gaining new insight into human language.

Section 2 gives an overview of the structure and content of the five subcollections that CoDII currently consists of, and of the sources that were used to collect the data. Section 3 outlines some of the linguistic questions that can be addressed with the data in

CoDII. We will show how these collections can be useful in approaching long-standing problems in linguistics from a new angle and on the basis of new types of empirical evidence. In Section 4 we conclude with a summary of the most important features of CoDII.

2. Data

2.1. **Five Collections.**

CoDII[1] comprises five subcollections: (1) 446 bound words in German (CoDII-BW.de); (2) 77 bound words in English (CoDII-BW.en); (3) 58 negative polarity items in Romanian (CoDII-NPI.ro); (4) 165 negative polarity items in German (CoDII-NPI.de); and (5) 88 positive polarity items in German (CoDII-PPI.de).

Bound words (BWs) are words which may only be used in combination with a fixed set of other words. Typical examples are the English word *headway*, which only occurs in the idiom *to make headway*, and the German word *Bärendienst*, which may only be used in the idiom *jemandem einen Bärendienst erweisen* (‘to do so. a disservice’). In a first informal characterization, NPIs are words (or multi-word expressions) which require the presence of some form of negation in their context. A good example is the verb *scheren*, which is only acceptable in negative contexts as provided by the negation adverb *nicht* (‘not’), the adverb *niemals* (‘never’), or a nominal phrase such as *wenige Studenten* (‘few students’). We will take a closer look at the licensing environments of NPIs below. PPIs are in a sense the positive counterpart to NPIs in that they shun negation in their immediate semantic environment.

The main source of the bound words in CoDII-BW.en and CoDII-BW.de are the studies Dobrovol’skij (1988, 1989) and Dobrovol’skij and Piirainen (1994). The items in CoDII-PPI.de were taken from van Os (1989), van der Wouden (1997) and Ernst (2005), with additional items from our own research. The sources for acquiring the NPIs for CoDII-NPI.de include the collections of NPIs in Welte (1978) and Kürschner (1983). To extend the coverage beyond previous literature, NPI candidates were extracted automatically from the *Tübingen Partially Parsed Corpus of Written German (TüPP)*[2]. The extraction algorithm is described in Lichte (2005) and Lichte and Soehn (2007). The items in our smallest collection, CoDII-NPI.ro, are mostly counterparts to the English, German and Dutch NPIs in the linguistic literature, since no specialized collection of Romanian NPIs was available as data source.

2.2. **Data Format.**

Each CoDII item has four basic information blocks: ‘General Information’, ‘Information about Licensing Contexts’, ‘Syntactic Information’, and ‘Classificatory Information’. The optional block ‘Sample Queries’ recommends search patterns that are optimized for important publicly available corpora.

The block ‘General Information’ identifies an item by providing its word form, an English gloss and a translation (for the non-English items), expressions in which the item occurs, and, if appropriate, paraphrases. This is also where we report occurrences of an NPI outside its theoretically expected licensing environments.

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

[2] URL: `www.sfs.uni-tuebingen.de/en/de_tuepp.shtml`

Within the block 'Syntactic Information', each item is assigned a syntactic category. The syntactic structure of the expression in which the item occurs is added where appropriate. Possible syntactic variations are listed, including passivization, pronominalization, modification, topicalization, occurrence in raising or control constructions, and appearance within relative or interrogative clauses. For each syntactic variation, examples from corpora, Internet or the linguistic literature are included. Three tagsets provide the theory and notation for the syntactic description of CoDII items. The *Stuttgart-Tübingen Tagset (STTS)*[3] is used for the syntactic description of German items and of expressions in which they occur. The English BWs are annotated with the syntactic annotation scheme from the *Syntactically Annotated Idiom Database* (SAID, cf. Kuiper et al. (2003)). For the syntactic description of Romanian NPIs we take the (modified) tagset from the *Multilingual Text Tools and Corpora for Central and Eastern European Languages (MULTEXT-East)*[4].

The block 'Licensing Contexts' contains information on the licensing environment of each item. In the case of polarity items, the licensing contexts are chosen from general, descriptive categories rather than from classifications in a particular theoretical framework. We distinguish the following licensing environments: clausemate (sentential) negation, non-clausemate negation, n-words (such as *nobody*, *never*), the scope of negation expressed by the determiner *kein-*, the scope of *without*, interpretation in the restrictor of universal quantifiers, other contexts of interpretation which are logically downward-entailing (and are not subsumed by one of the more specific categories), the scope of *only*, the complement clause of negative verbs (such as *doubt, fear* and *regret*), questions, antecedents of conditionals, comparative constructions, superlative constructions, and imperatives. To allow the documentation of all available data, exceptional cases that do not fit any of these predetermined categories are listed as 'Exceptions'. Some of the licensing environments will be discussed in more detail below.

The examples for the usage in their licensing contexts of the items listed in CoDII were collected from electronic and printed sources. The Romanian examples were gathered from Rada Mihalcea's Romanian electronic corpus, and from Internet search with Google. Some examples were constructed by Gianina Iordăchioaia, a native speaker of Romanian, who worked on CoDII-NPI.ro. The sources of the German BWs, NPIs and PPIs were corpora of the Institute of German Language in Mannheim[5], the corpus of the *Digitales Wörterbuch der Deutschen Sprache (DWDS)*[6], and Internet search with Google. The examples in CoDII-BW.en mainly come from dictionaries, from the Internet and from the *British National Corpus* (via the SARA software package[7]).

The last block, 'Classificatory Information', reports, for each item, classifications found in the literature. For English and German BWs, the classifications were taken from Dobrovol'skij (1988, 1989), Dobrovol'skij and Piirainen (1994), and Nunberg et al. (1994). Polarity items are classified as positive or negative, and are subdivided in three semantic classes according to the theory of Zwarts (1997) (see the discussion below).

[3] URL: `www.sfs.uni-tuebingen.de/Elwis/stts/stts.html`

[4] URL: `nl.ijs.si/ME`

[5] URL: `www.ids-mannheim.de/cosmas2/`

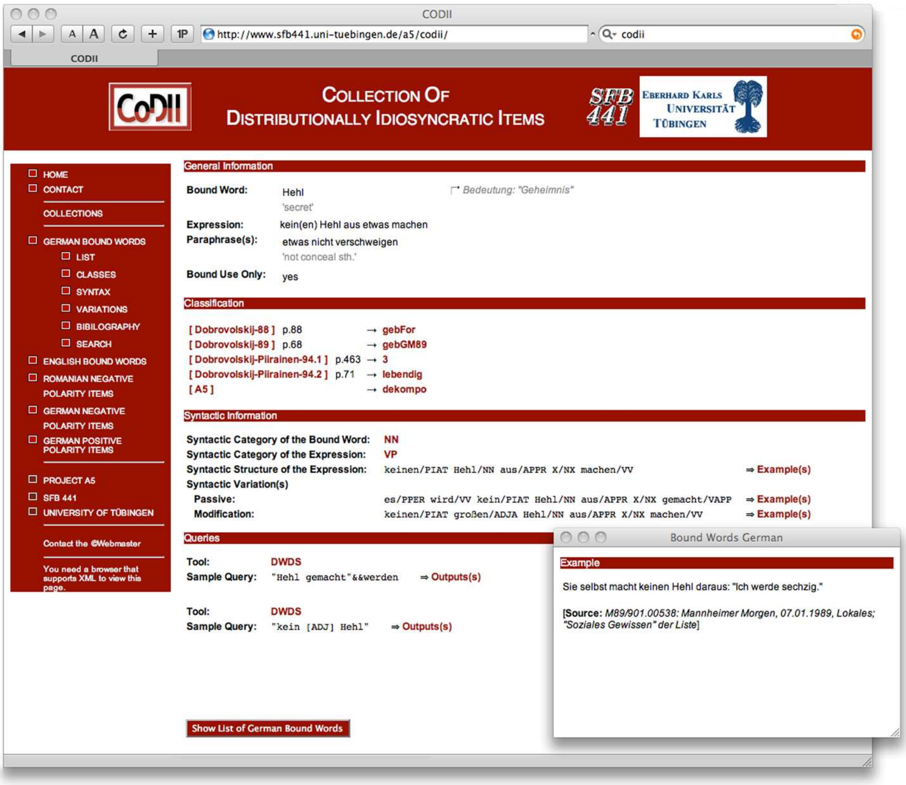
[6] URL: `www.dwds.de`

[7] URL: `www.natcorp.ox.ac.uk/sara`

For citations from literature that does not use these semantic distinctions, we use the classification tag 'open'.

The five CoDII collections are encoded in XML with a uniform schema. Technical details for BWs are described in Sailer and Trawiński (2006) and Trawiński et al. (2008b), and for PIs, in Trawiński and Soehn (2008). Design and data structure of CoDII are conceived in such a way that further types of distributionally idiosyncratic items, such as anaphora, can be modeled, and collections from various languages can easily be integrated using the existing schema.

CoDII not only compiles, documents and (alphabetically) lists distributionally idiosyncratic items. Due to the integration into the Open Source XML database *eXist*,[8] it also offers dynamic and flexible access. The design of the internal data structure and the annotation with syntactic and (partial) semantic information make it possible to query our resource with respect to particular lemmata, syntactic properties and linguistically interesting classifications. First statistical observations on the data in our collections which were obtained by using these database functionalities are reported in Trawiński et al. (2008a, 2008b) and Trawiński and Soehn (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
□ LIST
□ CLASSES
□ SYNTAX
□ VARIATIONS
□ BIBILOGRAPHY
□ SEARCH
□ ENGLISH BOUND WORDS
□ ROMANIAN NEGATIVE POLARITY ITEMS
□ GERMAN NEGATIVE POLARITY ITEMS
□ GERMAN POSITIVE POLARITY ITEMS
□ PROJECT A5
□ SFB 441
□ UNIVERSITY OF TÜBINGEN
Contact the @Webmaster
You need a browser that supports XML to view this page.
Show List of German Bound Words
Bound Words German

General Information

Bound Word: Hehl 'secret' ↱ Bedeutung: "Geheimnis"

Expression: kein(en) Hehl aus etwas machen

Paraphrase(s): etwas nicht verschweigen 'not conceal sth.'

Bound Use Only: yes

Classification

[Dobrovolskij-88] p.88 → gebFor
[Dobrovolskij-89] p.68 → gebGM89
[Dobrovolskij-Piirainen-94.1] p.463 → 3
[Dobrovolskij-Piirainen-94.2] p.71 → lebendig
[A5] → dekompo

Syntactic Information

Syntactic Category of the Bound Word: NN
Syntactic Category of the Expression: VP
Syntactic Structure of the Expression: `keinen/PIAT Hehl/NN aus/APPR X/NX machen/VV` ⇒ Example(s)
Syntactic Variation(s)
Passive: `es/PPER wird/VV kein/PIAT Hehl/NN aus/APPR X/NX gemacht/VAPP` ⇒ Example(s)
Modification: `keinen/PIAT großen/ADJA Hehl/NN aus/APPR X/NX machen/VV` ⇒ Example(s)

Queries

Tool: DWDS
Sample Query: `"Hehl gemacht"&&werden` ⇒ Outputs(s)

Tool: DWDS
Sample Query: `"kein [ADJ] Hehl"` ⇒ Outputs(s)

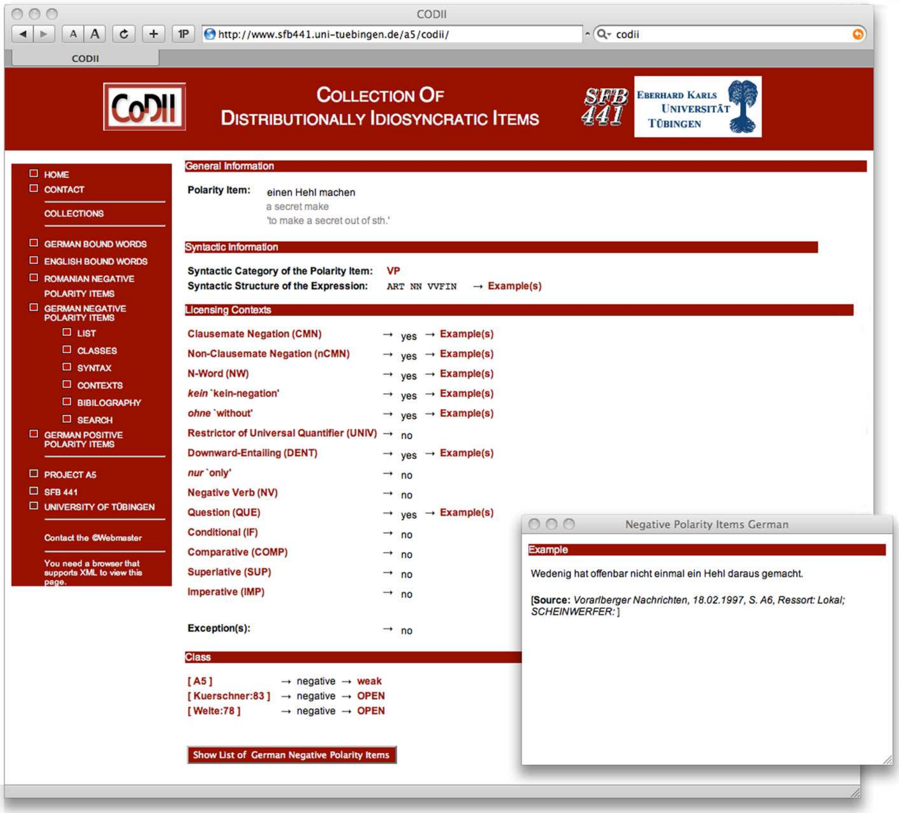
Example

Sie selbst macht keinen Hehl daraus: "Ich werde sechzig."

[**Source:** *M89/901.00538: Mannheimer Morgen, 07.01.1989, Lokales; "Soziales Gewissen" der Liste*]

FIGURE 1. CoDII web interface for the German BW *Hehl* ('secret')

[8] URL: `exist.sourceforge.net`


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
You need a browser that supports XML to view this page.
Show List of German Negative Polarity Items
Negative Polarity Items German

General Information

Polarity Item: einen Hehl machen
a secret make
'to make a secret out of sth.'

Syntactic Information

Syntactic Category of the Polarity Item: VP
Syntactic Structure of the Expression: ART NN VVFIN → 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)	→ no	
Question (QUE)	→ yes	→ Example(s)
Conditional (IF)	→ no	
Comparative (COMP)	→ no	
Superlative (SUP)	→ no	
Imperative (IMP)	→ no	
Exception(s):	→ no	

Class

[A5]	→ negative	→ weak
[Kuerschner:83]	→ negative	→ OPEN
[Welte:78]	→ negative	→ OPEN

Example

Wedenig hat offenbar nicht einmal ein Hehl daraus gemacht.

[**Source:** *Vorarlberger Nachrichten, 18.02.1997, S. A6, Ressort: Lokal; SCHEINWERFER:*]

FIGURE 2. CoDII web interface for the German NPI *ein(en) Hehl aus etw. machen* ('to make a secret out of sth.')

The user interface of CoDII displays all the linguistic information, including syntactic structure and licensing contexts together with the links to corresponding examples (see FIGURE 1 and FIGURE 2). Comments, information about the classification systems, licensing contexts, and examples of the usage of each item in context can be obtained by clicking on the links in the display. All bibliographic references in CoDII are linked to two electronic bibliographies, the 'Bound Words Bibliography'[9], and the 'Polarity Items Bibliography'[10].

2.3. **Context Classification and Variation.** FIGURE 1 and FIGURE 2 show the web interface of CoDII for two entries in different subcollections: The German BW *Hehl* ('secret') and the German multi-word NPI *ein(en) Hehl aus etw. machen* ('to make a secret out of sth.').[11] In CoDII-BW.de, *Hehl* is recorded as a word without the usual free

[9] URL: `www.sfb441.uni-tuebingen.de/a5/bwb`

[10] URL: `www.sfb441.uni-tuebingen.de/a5/pib/XML2HTML/list.html`

[11] With the idea of decomposing the meaning of the idiom *ein(en) Hehl aus etw. machen* and assigning *Hehl* a meaning contribution of its own, we follow the analysis of decomposable VP idioms in Sailer (2003) and Soehn (2006). This does not mean that we suggest to decompose every idiomatic phrase. For example, the words in the non-decomposable VP idiom *die Flinte ins Korn werfen* ('to give up') and

distribution of a noun; it may only occur as part of the multi-word expression *ein(en) Hehl aus etw. machen.* FIGURE 1 shows the information blocks that CoDII records for a bound word, including a window available through the 'Output(s)' link which illustrates one result to a given sample query. Note the links in this CoDII entry, which provide background information about the various categorizations offered on this page.

Before looking at *Hehl* as item in CoDII-NPI.de, a more precise explanation of NPIs and their semantic subclasses is in order. The three classes of NPIs we distinguish in CoDII, weak NPIs, strong NPIs, and superstrong NPIs, were introduced by Zwarts (1997). In the formulation of the theory given by van der Wouden (1997) they are algebraically defined as follows: (1) NPIs are *superstrong* if they are licensed only by antimorphic contexts (overt negation).[12] An example of an antimorphic operator is sentential negation. (2) NPIs are *strong* if they are licensed by antimorphic and anti-additive contexts.[13] Examples of anti-additive operators are the expressions *nobody* and *never.* The word *nobody* is shown to be anti-additive by checking that the sentence *Nobody complained or resisted* is true in exactly those situations in which *Nobody complained and nobody resisted* is true. (3) NPIs are *weak* if they are licensed by antimorphic, anti-additive, and downward-entailing contexts (and possibly some others).[14] An example of a plain downward-entailing operator is the phrase *few students.* This phrase is shown to be downward entailing by checking that *Few students complained or resisted* implies *Few students complained and few students resisted.* Moreover, *Few students complained or few students resisted* implies *Few students complained and resisted.* According to the definition of the three NPI classes, any NPI is licensed by sentential negation. Strong NPIs need to be in the scope of an operator that is at least strong. The German strong NPI *einen blassen Schimmer haben* ('to have the faintest idea') is thus licensed by sentential negation and *niemals* ('never') but not by *wenige Studenten* ('few students'). Weak NPIs are already satisfied in the presence of a weak licenser.

Hehl is recorded in CoDII-NPI.de because apart from being a bound word, it is also a lexeme which occurs in a multi-word expression that behaves like an NPI: FIGURE 2 shows the corresponding CoDII entry and a window with a corpus example reachable through the link 'Example(s)' of the licensing context 'Clausemate Negation (CMN)'. The reader will notice that the '[A5]' classification categorizes the item as a weak negative polarity item. This means that it is an item which only needs a logically weak form of negation as licenser. A reflex of this fact is the existence of corpus evidence in the category 'Downward-Entailing (DENT)'. These are licensing environments that are weaker than the antimorphic 'Clausemate Negation (CMN)' environment or the restrictor of the determiner *kein-*. The NPI classification in CoDII into weak, strong and superstrong is preliminary in the sense that it strictly follows the corpus evidence that we found: It can (and does) happen that an item which is generally considered a weak NPI is classified as strong in CoDII, because we only found corpus evidence for

the bound word *klipp* in the frozen expression *klipp und klar* ('point-blank') have a very different status with respect to the interpretation of the overall expression.

[12] An operator f is antimorphic iff for each set X and for each set Y, $f(X \cup Y)$ equals $f(X) \cap f(Y)$ and $f(X \cap Y)$ equals $f(X) \cup f(Y)$.

[13] An operator f is anti-additive iff for each set X and for each set Y, $f(X \cup Y)$ equals $f(X) \cap f(Y)$.

[14] An operator f is downward-entailing iff for each set X and for each set Y, $f(X \cup Y)$ implies $f(X) \cap f(Y)$ and $f(X) \cup f(Y)$ implies $f(X \cap Y)$.

its occurrence with sentential negation, *kein-*, and *ohne* ('without'). It is important to realize that CoDII deliberately stays within the limited horizon of its data base and leaves it to the user's judgment and research to revise this preliminary categorization where it is appropriate or necessary.

3. Theory

3.1. **Grammar Theories.** Idioms are treated very differently in different areas of linguistics. Two opposite extremes within the overall spectrum are the constructional (holistic) approach, and the collocational approach. The constructional perspective views idioms as syntactic and semantic units which are usually treated as fixed, stored chunks. They are basically conceived of as lexical items, differing from words primarily in that they may be syntactically complex. This perspective is common in the phraseological literature such as Fleischer (1997) and in formal linguistics, be it Generative Grammar (Chomsky, 1981), or Construction Grammar (Fillmore et al., 1988). The collocational perspective originates from corpus linguistic research. Under this perspective, the co-occurrence patterns of individual words are studied. If a word co-occurs with a second word more often than expected on the basis of their syntactic category, the two words form a collocation. This perspective is common in computational corpus linguistic research on idioms, such as in computational lexicography (Sinclair (1991), Moon (1998)), and in more general computational linguistic approaches such as Krenn (1999).

Interestingly for us, there is a natural area of overlap between these two perspectives: The constituents of what would traditionally be called an idiom may show high cooccurrence ratios in corpora. However, the two perspectives do not cover the same ground. Many idioms are very infrequent in corpora (see Moon (2007)), which makes them invisible to the collocational method. On the other hand, many high-frequency cooccurrence pairs do not show any degree of syntactic irregularity or semantic idiomaticity, which makes them irrelevant from the constructional perspective. One of the important missions of CoDII is to demonstrate that this last point is not just an innocent blind spot of the constructional approach. CoDII sets out to contribute to the development of a theory that can overcome this shortcoming and supply a picture of the missing landscape.

Formal grammars usually strive to formulate linguistic generalizations, whereas collocations (and idioms) are by definition idiosyncratic and lexeme-specific. In formal grammars, context effects are occasionally encoded when they capture generalizations about syntactic structures or systematic differences between lexical items. The concept of *selection* plays an important role in this. Selection is responsible for binary combinations of a (syntactic or semantic) functor and its argument: A syntactic head imposes restrictions on its complement(s), and an adjunct imposes restrictions on the syntactic head it combines with. Formal grammars have developed sophisticated means to express these and only these relations and restrictions. They are primarily realized in subcategorization frames or valence specifications in lexical entries. Collocations, however, do not necessarily respect the directions of selection or other grammatical relations. A good example are light-verb constructions, i.e. verb-noun collocations such as *take a shower*, *do the dishes*, *make a mistake*. In these cases the noun is syntactically realized as the complement of the verb. Nonetheless it can be argued that it is the noun that determines which verb must be used in the combination.

In previous work (Richter and Sailer, 2003; Sailer, 2004) we argued that so-called *bound words* show that at least some collocations should be included within the empirical domain of formal grammar. The underlined words in (1) are *bound* in the sense that they can only occur in this particular context.

(1) a. wend one's way (= make one's way)
b. make headway (= make progress); take/ have a dekko (= take a look)
c. without fail (= fully predictable, with no exception or cause for doubt)
d. the whole caboodle (= the whole lot)
e. flotsam and jetsam (=pieces from a wrecked ship floating in the sea or scattered on the floor)

The data in (1) show that the relation between a bound word and its required context cannot be captured with the means of selection: In (1-b) we see the same pattern as in support verb constructions, i.e. the noun determines which verb has to be chosen. In (1-d) the noun requires the presence of a certain modifier, and in (1-c) the noun must occur as the complement of a particular preposition. In (1-e) there are two bound words occurring in a conjunction. Normally no mutual selection relation is assumed among conjuncts. The data in (2) add a crucial second dimension to the behavior of bound words:

(2) a. achieve progress/ *headway
b. with exceptions/ *fail
c. (i) *jetsam and flotsam (ii) collect *(flotsam and) jetsam

Native speakers of English can give grammaticality judgments about the distribution of bound words. In particular, combinations as in (2) are judged ungrammatical. If it is a central goal of formal grammars to capture the grammaticality judgments of native speakers, the distribution of bound words can certainly not be ignored. The theoretical significance of bound words, thus, lies in their property to exhibit a firm co-occurrence with a particular other word. Crucially, the necessity of this co-occurrence is observable in the grammaticality judgments of native speakers. Despite their clear grammatical relevance, these co-occurrence patterns are not captured by the theory of syntactic selection.

When we turn to polarity items, a different, but equally puzzling picture emerges. We saw in Section 2 that NPIs require the presence of a licensing element, which is prototypically — but not necessarily — sentential negation. Many NPIs are idioms, but negation is an abstract part of the idiom rather than a lexicalized component. For this reason a holistic view on idioms lacks the means to express the negation requirement correctly.

The majority of the formal and theoretical research on polarity items focus on a small number of expressions, primarily on English *any* and *ever*. The contexts in which these NPIs may occur are carefully characterized and categorized. The limitation of this line of research to very few selected items is acknowledged in important contributions to the field, such as Kadmon and Landman (1993), von Fintel (1999), and Chierchia (2004). It is unclear whether the distribution of polarity items in general is captured, or merely the idiosyncratic distribution of certain items. In addition, it remains an open question how the negation requirement can be linked to the relevant lexical items (or multi-word expressions), let alone what the connection could be to a theory of idioms.

Within the collocational tradition, NPIs have been largely ignored. Sinclair (2004) discusses the verb *budge* (a weak NPI) and observes that it occurs in negative contexts. However, this insight is based on an inductive inspection of corpus data. No precise characterization of the term *negative context* is provided. Hoeksema (1997) presents corpus studies of individual polarity items that confirm the impression that the distribution of polarity items in their potential licensing contexts is not as homogeneous as many theoretical approaches suggest.

3.2. **Distribution Profiles.** Bearing in mind the general picture outlined in the previous section, we can now look at the contribution of CoDII to the field. The information on contexts and variation collected in CoDII is chosen in a way that makes it easy for researchers to check their theories against more data. Within phraseological research it has often been claimed that the syntactic flexibility of an idiom is related to semantic properties. While passivization as a semantically neutral operation is possible with many VP idioms, spreading an idiom over a main clause and a relative clause seems to be restricted to semantically decomposable idioms, i.e. to idioms whose parts can be assigned a meaning that contributes to the overall meaning of the idiom in a regular way (McCawley, 1981; Schenk, 1995). The study of bound words in relative clauses is, consequently, of great theoretical importance. If a bound word can occur in a relative clause constellation, it should be possible to assign this word a meaning.

The distributional profiles for polarity items can help us provide a better classification of NPIs. As we have seen earlier, the distinction between our three classes of NPIs is based on the entailment properties of their licensing contexts. Weak NPIs such as German *jemals* ('ever') may occur in all potentially NPI-licensing contexts. Strong NPIs such as German *eine Miene verziehen* ('show emotions') are restricted to sentences that contain the negation adverb *nicht* ('not') or a negative constituent such as *kein- N* ('no N') or *niemals* ('never'). They may also occur in the restrictor of a universal quantifier. Superstrong NPIs such as English *one bit* are claimed to occur only with *not*.

In CoDII we document polarity item data in exactly those contexts that have been looked at in the literature. We assign classifications to the items based on the distribution profiles found for these contexts. The resulting profiles do not confirm the predictions of Zwarts' tripartite theory. For instance, NPI modals such as German *brauchen* and English *need* may occur in many NPI contexts, but are banned from the restrictor of universal quantifiers (Hoeksema, 1997). This distributional gap is not predicted in Zwarts' theory. The distribution profiles in CoDII can be used to check whether this unexpected behavior is idiosyncratic to *brauchen* or attested with other NPIs as well.

Pragmatic theories of NPI licensing such as Krifka (1995) and Israel (2004) assume that NPIs are admitted whenever certain pragmatic conditions are met. As a consequence, they predict the availability of NPIs in contexts which are not traditionally considered NPI licensing contexts. CoDII also includes a field for unusual occurrences.

If large numbers of examples can be found in this category, this may be taken as support for pragmatic theories.

In sum, CoDII attempts to provide reliable, qualitative profiles for NPIs. These profiles can be used to confirm or to challenge the predictions of NPI theories.

3.3. **Collocations in a Formal Theory of Grammar.** Let us finally look at a line of research which tries to encode the data collected in CoDII in a formal theory of grammar, Head-driven Phrase Structure Grammar (HPSG, Pollard and Sag (1994)). HPSG is a framework that has its roots in context-free phrase structure grammars. For this reason, the original formulation of the theory in Pollard and Sag (1994) did not provide the means to encode idiosyncratic syntactic and semantic units that span more than a local tree. It was only in recent developments that a link from HPSG to Construction Grammar was established (Ginzburg and Sag, 2000; Riehemann, 2001), indicating that HPSG takes a constructional perspective on idioms.

In various publications (including Sailer (2003), Soehn (2006), and Richter and Soehn (2006)) we developed a collocational module for HPSG that can model the data documented in CoDII. Let us illustrate this collocational module with the NPI *ein Hehl aus etw. machen* that contains the bound word *Hehl*. All distributional idiosyncrasies can be located in the lexical entry, sketched in FIGURE 3.

HPSG is a constraint-based theory that employs feature structures (or similar appropriate mathematical structures) as linguistic representations (Richter, 2004). These structures encode all linguistically relevant components of a sign, including the phonological representation, a semantic representation, the syntactic category, and valence information. A prominent part of this is indicated in FIGURE 3. The value of the feature PHON(onology) specifies the phonological representation of the word. In SYNSEM HEAD the syntactic category of the noun is given. We also use a feature LISTEME which provides a unique identification label for each listeme. The relevant notion of a listeme is borrowed from Di Sciullo and Williams (1987), and is meant to subsume simple word lexemes as well as phrasal lexemes. The CONTENT value is the semantic representation of the sign. In FIGURE 3 we simplify and only mention the logical semantic constant that belongs to the word.

We enrich this conventional HPSG architecture with a new feature, COLL (context of lexical licensing), whose value specifies the co-occurrence requirements of a lexical item. The word *Hehl* has two requirements. First, it must occur as the direct object to the support verb *machen* (‘make’). Second, the semantics of *Hehl* must occur in the scope of an NPI-licensing operator. These two requirements are expressed in the two elements on the COLL list. Each element defines the syntactic domain within which the collocational restriction has to be met. The first one must hold within the minimal clause that contains the word *Hehl*. The second one only needs to be satisfied within the overall utterance.

The first COLL-element has a feature LOC-LIC. By means of its complex feature value, *Hehl* is collocationally restricted to co-occur with a particular lexeme. The collocating lexeme is identified on the basis of its LISTEME value. In FIGURE 3 this is specified as *machen*$_2$, which we assume to be the LISTEME value of the required support verb *machen*. More examples of this kind of collocational restrictions are discussed in Soehn (2006), which focuses on decomposable and non-decomposable VP idioms.

$$\begin{bmatrix}
\textsc{phon} & \langle \text{he:l} \rangle \\
\textsc{synsem} & \begin{bmatrix} \textsc{head} & \begin{bmatrix} \textit{noun} \\ \textsc{listeme} \; \textit{hehl} \end{bmatrix} \\ \textsc{content} & \textit{secret\_relation} \end{bmatrix} \\
\textsc{coll} & \left\langle \begin{bmatrix} \textit{minimal-clause} \\ \textsc{loc-lic} \; [\textsc{listeme} \; \textit{machen}_2] \end{bmatrix}, \begin{bmatrix} \textit{utterance} \\ \textsc{lf-lic} \; \textit{downward-entailing-operator} \end{bmatrix} \right\rangle
\end{bmatrix}$$

FIGURE 3. Sketch of the lexical entry of the bound word *Hehl*

The second element on the COLL list has a feature LF-LIC, which is short for LOGICAL-FORM-LICENSER. This indicates a semantic restriction. The value of this feature specifies that the semantic contribution of the word must be in the scope of an operator with a particular semantic property. In the figure we simply write *downward-entailing-operator* as a shorthand for a logical specification. A precise version for this type of collocational requirement can be found in Richter and Soehn (2006).

The lexical specifications of COLL values are complemented with a general Licensing Principle. This principle ensures that in each utterance, the collocational requirements of the lexical items that occur in the utterance are satisfied.

Our brief sketch of the collocational module shows that the information included in CoDII can be incorporated into the grammar architecture of HPSG. The better we understand the distributional patterns of bound words and polarity items the more adequately we can state the collocational constraints in a formal grammar framework. It should be clear that the analysis of bound words carries over to collocations with free words such as *take a shower.* A collocation module is a first step towards a formal theory that embodies both a constructional and a collocational perspective and thus, stands a chance to model the implicit linguistic knowledge of native speakers, including knowledge of idiosyncrasies.

4. Conclusion

We presented the Collection of Distributionally Idiosyncratic Items (CoDII), an electronic resource which collects and presents different types of lexical items that exhibit distributional idiosyncrasies. It is a characteristic feature of CoDII that it is open for the inclusion of new subcollections. It is also dynamic on the level of the items in the different collections: Not only may items be added but also new corpus evidence, which can broaden the empirical documentation and, as a consequence, change the theoretical categorization of the items. The flexibility of an electronic resource provides added value to linguists by the various search functions in the system which can be used when researching empirical evidence or counter-evidence to theoretical claims about universal properties of idiomatic expressions or polarity items. We also illustrated how the corpus evidence collected in CoDII can highlight distribution patterns that went unnoticed before and might lead to new generalizations on the behavior of the recorded classes of items. Most importantly, we believe that CoDII emphasizes the need for a comprehensive study of collocational patterns in language between mere statistical tendencies and phenomena that have acquired the status of grammatical facts that are subject to categorical grammaticality judgments and should consequently also be subject to grammatical description in formal grammar frameworks.

References

Chierchia, Gennaro (2004). Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface. In A. Belletti (Ed.), *Structure and Beyond. The Cartography of Syntactic Structures*, Volume 3, pp. 39–103. Oxford, New York: Oxford University Press.

Chomsky, Noam (1981). *Lectures on Government and Binding.* Foris, Dordrecht.

Di Sciullo, Anna-Maria and Williams, Edwin (1987). *On the Definition of Word.* Number 14 in Linquistic Inquiry Monographs. MIT Press.

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

Dobrovol'skij, Dmitrij (1989). Formal gebundene phraseologische Konstituenten: Klassifikationsgrundlagen und theoretische Analyse. In W. Fleischer, R. Große, and G. Lerchner (Eds.), *Beiträge zur Erforschung der deutschen Sprache*, Volume 9, pp. 57–78. Leipzig, Bibliographisches Institut.

Dobrovol'skij, Dmitrij and Piirainen, Elisabeth (1994). Sprachliche Unikalia im Deutschen: Zum Phänomen phraseologisch gebundener Formative. *Folia Linguistica 27*(3–4), 449–473.

Ernst, Thomas (2005). On Speaker-Oriented Adverbs as Positive Polarity Items. Elektronic Poster for the Workshop: Polarity From Different Perspectives, New York University, 11.–13.03.2005. URL: www.nyu.edu/gsas/dept/lingu/events/polarity/posters/ernst.pdf.

Fillmore, Charles, Kay, Paul, and O'Connor, M. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone. Language 64*, 501–538.

von Fintel, Kai (1999). NPI-Licensing, Strawson-Entailment, and Context-Dependency. *Journal of Semantics 16*, 97–148.

Fleischer, Wolfgang (1997). *Phraseologie der deutschen Gegenwartssprache* (2nd, revised edition ed.). Niemeyer, Tübingen.

Ginzburg, Jonathan and Sag, Ivan A. (2000). *Interrogative Investigations. The Form, Meaning, and Use of English Interrogatives.* CSLI Publications.

Hoeksema, Jack (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.

Israel, Michael (2004). The Pragmatics of Polarity. In L. Horn and G. Ward (Eds.), *The Handbook of Pragmatics*, pp. 701–723. Oxford: Blackwell.

Kadmon, Nirit and Landman, Fred (1993). 'Any'. *Linguistics and Philosophy 16*, 353–422.

Krenn, Brigitte (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.

Krifka, Manfred (1995). The Semantics and Pragmatics of Weak and Strong Polarity Items. *Linguistic Analysis 25*, 209–257.

Kuiper, K., McCann, H., Quinn, H., Aitchison, Th., and van der Veer, K. (2003). Syntactically Annotated Idiom Database (SAID) v.1. Documentation to a LDC resource.

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

Lichte, Timm (2005, October). Korpusbasierte Acquirierung negativ-polärer Elemente. Master's thesis, Seminar für Sprachwissenschaft, University of Tübingen.

Lichte, Timm and Soehn, Jan-Philipp (2007). The Retrieval and Classification of Negative Polarity Items using Statistical Profiles. In S. Featherston and W. Sternefeld (Eds.), *Roots: Linguistics in Search of its Evidential Base*, pp. 249–266. Berlin: Mouton de Gruyter.

McCawley, James D. (1981). The Syntax and Semantics of English Relative Clauses. *Lingua 53*, 99–149.

Moon, Rosamund (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach.* Oxford: Clarendon Press.

Moon, Rosamund (2007). Corpus Linguistic Approaches with German Corpora. In H. Burger, D. Dobrovol'skij, P. Kühn, and N. R. Norrick (Eds.), *Phraseologie/ Phraseology*, Volume 2 of *Ein internationales Handbuch der zeitgenössischen Forschung/An International Handbook of Contemporary Research*, Chapter 88, pp. 1045–1059. Berlin, New York: de Gruyter.

Nunberg, G., Sag, I.A., and Wasow, Th. (1994). Idioms. *Language 70*, 491–538.

Pollard, Carl and Sag, Ivan A. (1994). *Head-Driven Phrase Structure Grammar.* Chicago and London: University of Chicago Press.

Richter, Frank (2004). *A Mathematical Formalism for Linguistic Theories with an Application in Head-Driven Phrase Structure Grammar.* Phil. dissertation (2000), Universität Tübingen.

Richter, Frank and Sailer, Manfred (2003). Cranberry Words in Formal Grammar. In C. Beyssade, O. Bonami, P. Cabredo Hofherr, and F. Corblin (Eds.), *Empirical Issues in Formal Syntax and Semantics*, Volume 4, pp. 155–171. Paris: Presses Universitaires de Paris-Sorbonne.

Richter, Frank and Soehn, Jan-Philipp (2006). *Braucht niemanden zu scheren*: A Survey of NPI Licensing in German. In S. Müller (Ed.), *The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, Stanford, pp. 421–440. CSLI Publications.

Riehemann, Susanne Z. (2001). *A Constructional Approach to Idioms and Word Formation.* Ph. D. thesis, Stanford University.

Sailer, Manfred (2003). Combinatorial Semantics and Idiomatic Expressions in Head-Driven Phrase Structure Grammar. Phil. Dissertation (2000). Arbeitspapiere des SFB 340. 161, Universität Tübingen.

Sailer, Manfred (2004). Distributionsidiosynkrasien: Korpuslinguistische Erfassung und grammatiktheoretische Deutung. In K. Steyer (Ed.), *Wortverbindungen — mehr oder weniger fest*, Institut für Deutsche Sprache, Jahrbuch 2003, Berlin, New York, pp. 194–221. de Gruyter.

Sailer, Manfred and Trawiński, Beata (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, pp. 471–474.

Schenk, André (1995). The Syntactic Behavior of Idioms. In M. Everaert, E.-J. v. d. Linden, A. Schenk, and R. Schreuder (Eds.), *Idioms. Structural and Psychological Perspectives*, pp. 253–271. Lawrence Erlbaum Associates, Hillsdale.

Sinclair, John (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, John (2004). *Trust the Text. Language, Corpus and Discourse.* London and New York: Routledge.

Soehn, Jan-Philipp (2006). *Über Bärendienste und erstaunte Bauklötze. Idiome ohne freie Lesart in der HPSG.* Frankfurt am Main: Peter Lang.

Trawiński, Beata, Sailer, Manfred, Soehn, Jan-Philipp, Lemnitzer, Lothar, and Richter, Frank (2008b). Cranberry Expressions in English and in German. In E. L. R. A. (ELRA) (Ed.), *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco, pp. 35–38.

Trawiński, Beata and Soehn, Jan-Philipp (2008). A Multilingual Database of Polarity Items. In E. L. R. A. (ELRA) (Ed.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Trawiński, Beata, Soehn, Jan-Philipp, Sailer, Manfred, and Richter, Frank (2008a). A Multilingual Electronic Database of Distributionally Idiosyncratic Items. In E. Bernal and J. DeCesaris (Eds.), *Proceedings of the XIII Euralex International Congress*, Volume 20 of *Activitats*, Barcelona, Spain, pp. 1445–1451. Universitat Pompeu Fabra.

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

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

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

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

Appendix A. List of abbreviations

BW	bound word
CoDII	Collection of Distributionally Idiosyncratic Items
DII	distributionally idiosyncratic item
HPSG	Head-driven Phrase Structure Grammar
iff	if and only if
MWE	multi-word expression
N	noun
NPI	negative polarity item
PI	polarity item
PPI	positive polarity item
VP	verb phrase
XML	Extensible Markup Language