The Collection of Distributionally Idiosyncratic Items

Beata Trawiński

SFB 441
Universität Tübingen
trawinski@sfs.uni-tuebingen.de

Seminarium „Przetwarzanie języka naturalnego"
Instytut Podstaw Informatyki PAN
Warszawa, 20.12.2004
Overview

- The Project
- Purpose and Design of CoDII
- The Collection of Bound Words in German
- First Evaluation
- Outlook
The Project: General Information

CoDII has been developed in the framework of Project A5 *Distributional Idiosyncrasies* of the Collaborative Research Center (Sonderforschungsbereich) 441 *Linguistic Data Structures: On the Relation between Data and Theory in Linguistics* at the University of Tübingen.

The following people have contributed to the design and the compilation of CoDII: Monique de Jong, Manfred Sailer, Jan-Philipp Soehn and Beata Trawiński.

The formulation of basic design principles of CoDII and the compilation of the first subcollection of CoDII, the collection of German bound words, were carried out during the first project phase from 2002 to the end of 2004.
The project deals with linguistic expressions showing distributional idiosyncrasies; i.e. expressions whose distribution follows not only from their syntactic, semantic and pragmatic properties and general principles of grammar.
Based on various empirical data, one can observe two types of irregularities in language:

- internal irregularities,
- (external) distributional irregularities.

Richter and Sailer (2003) and Sailer (2003, 2004) argue for such a distinction.
Internally Irregular Expressions

An internally irregular expression (construction) can be characterized as follows:

- there is no transparent way to account for the meaning of the whole expression on the basis of the meaning of its parts,
- no parts of the expression can be modified,
- the expression cannot undergo regular syntactic processes such as passivization (in the case of VPs).

Example:

kick the bucket ‘die’
Externally Irregular Expressions

An externally (distributionally) irregular expression (collocation) can be characterized as follows:

- it behaves regularly with respect to interpretation,
- it behaves regularly with respect to modification,
- it behaves regularly with respect to syntactic processes such as passivization (in the case of VPs),
- but it provides certain requirements on the linguistic environment this expression occurs in.

Examples:

*make / do a decision, do / *make a favour*
Continuum of Irregularity

Thereby, based on various example types, a continuum can be observed of

- the internal irregularity,
- the external (distributional) irregularity.
Construction Continuum

The construction continuum relates to the question to what extent the structure of a given syntagma is formed according to general grammar principles and to what extent it can be interpreted compositionally.

**Regular**

- general phrase structure rules:
  - ID-Schemata
  - \( S \rightarrow NP \ VP \)

- special phrase structure rules:
  - relative clause constructions (Sag, 1997)
  - the "What's \( X \) doing \( Y \)" construction (Kay and Fillmore, 1999)

- marked syntagmas:
  - *let alone* construction (Fillmore et al., 1988)

- phraseological patterns:
  - *Me worry*? (Lambrecht, 1990)

- fixed word combinations:
  - *by and large*

**Irregular**
Collocation Continuum

The collocation continuum relates to the question of how far lexical elements are limited with regard to their contextual appearances in addition to their syntactic, semantic and sectional characteristics.

Regular

“free words”: read, book, …
“free phrases”: read a book
empty elements: traces
ellipses
anaphora: himself
polarity items: anybody
nothing
“ordinary” collocations: strong / *powerful coffee
heavy / *weighty smoker
bound words: make headway
by dint of

Irregular
Constructions in the Linguistics

For internally irregular expressions (constructions) there are analytic proposals in formal systems such as

- Construction Grammar (CG, (Fillmore et al., 1988)),
- Tree Adjoining Grammar (TAG, (Joshi, 1987)),
- (constructional) Head-driven Head Phrase Structure Grammar (HPSG, (Riehemann, 2001))
In contrast to constructions, distributional irregularities have not systematically been treated in the theoretical linguistics. Only a few groups of lexical items have been investigated such as

- anaphora and pronouns (Binding Theory),
- polarity elements,
- theoretical constructs such as traces (the HPSG Empty Category Principle).
The aim of the project is to elaborate a general theory of distribution within the formal paradigm of HPSG in the tradition of (Pollard and Sag, 1994).

To provide an empirical basis for the distribution theory, a large number of distributionally idiosyncratic items should systematically be considered.

CoDII (the Collection of Distributionally Idiosyncratic Items) should provide an empirical basis of this kind.
CoDII: The Essential Idea

The essential idea of CoDII is to provide a basis for linguistic investigations of lexical items showing distributional idiosyncrasies.

This includes

- listing appropriate items,
- providing existing linguistic documentation,
- specifying possibilities of data ascertainment relating to these items.
CoDII: Conceptual Design and Data Structure

The conceptual design and the data structure of CoDII have been conceived in such a way that

- subcollections of various types of distributionally idiosyncratic items can be modeled (anaphora, negative and positive polarity items, bound words, etc.),

- collections of distributionally idiosyncratic items from various languages can be compiled.
The Collection of German Bound Words (CoDII-BW.de) is the first completed subcollection of CoDII. The motivation for the compilation of CoDII-BW.de at first was following:

- In the project we first wanted to focus on the end of the collocation continuum, where the distributional restrictions are the strongest. Bound words are clearly lexical elements with extremely limited distribution.
- German bound words are relatively well documented in the German phraseological literature.
- Extraction of bound words from corpora is very trivial.
CoDII-BW.de: The Starting Basis

The initial basis for CoDII-BW.de are collections and classifications of bound words in

- (Dobrovol’skij, 1988),
- (Dobrovol’skij, 1989),
- (Dobrovol’skij and Piirainen, 1994b),
- (Dobrovol’skij and Piirainen, 1994a)

and our own observations.

Currently, CoDII-BW.de includes about 450 items.
CoDII-BW.de: Modeling CoDII-BW.de entries

Each item in CoDII-BW.de is characterized by four information blocks:

- general information,
- classification,
- syntactic information,
- queries.
CoDII-BW.de: General Information

The General Information block identifies bound words by providing

- a particular bound word,
- the English translation of the bound word,
- the expression in which the bound word occurs,
- the set of possible paraphrases of this expression.
CoDII-BW.de: Classification

The field Classification specifies classes associated with a given bound word according to the following classifications:

- classification in (Dobrovol’skij, 1988),
- classification in (Dobrovol’skij, 1989),
- classification in (Dobrovol’skij and Piirainen, 1994b),
- Nunberg et al. (1994) oriented classification,
- classifications used in the project.
The field Syntactic Information provides information on

- the syntactic category of a bound word,
- the syntactic structure in which the bound word occurs,
- possible syntactic variations such as passivization, pronominalization, modification, occurrence in raising constructions, etc.

For syntactic description of bound words and expressions in which they occur, the Stuttgart-Tübingen Tagset (STTS) has been used (http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html).

For each context, examples from various corpora, from Internet and the linguistic literature are provided.
Finally, hints on further data search are given by providing optimized queries for various publicly available corpora of German such as

- corpora of the Institut of German Language in Mannheim ([http://www.ids-mannheim.de/cosmas2/](http://www.ids-mannheim.de/cosmas2/)),
- the Tübinger Sammlung nutzbarer empirischer linguistischer Datenstrukturen – TUSNELDA ([http://www.sfb441.uni-tuebingen.de/tusnelda.html](http://www.sfb441.uni-tuebingen.de/tusnelda.html)),
- TIGERSearch, a search engine for retrieving information from a database of graph structures ([http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/](http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/)),
- Internet via Google.
CoDII-BW.de: Technical Realization

CoDII-BW.de has internally been encoded in XML. The DTD has been specified in such a way that:

- The element **codii** is the document root and its instance is identified by attributes **type** (for specifying collection type) and **xml:lang** (for specifying language the data come from).

- 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**, whose content is a list of examples.

- The content model of the element **dii-list** consists of a list of **dii-entry** elements, whose content model consists of a set of elements which
  - identify distributionally idiosyncratic items (**dii**),
  - provide expressions in which they appear (**dii-expression**),
  - describe documentation on each item (**dii-classification**),
  - present syntactic properties of items and the expressions (**dii-syntax**),
  - give query hints (**dii-queries**).
<dii-entry id="zampano">
  <dii>
    <ol>Zampano</ol>
    <en>golden boy</en>
  </dii>
  <dii-expression>
    <ol>der große Zampano</ol>
    <en>the big doer</en>
  </dii-expression>
  <dii-classification>
    <dii-class class="dekompo" type="A5">
      <bibliography bib-item="A5"/>
    </dii-class>
  </dii-classification>
  <dii-syntax hits="zampano-Bsp zampano-apposition" cat="NE">
    <dii-expression-syntax cat="NP">
      der/ART große/ADJA Zampano/NE
    </dii-expression-syntax>
    <variation kind="OPEN" hits="zampano-ecclestone">
      Spitzname von Formel-1-Manager Bernie Ecclestone
    </variation>
  </dii-syntax>
  <dii-queries>
    <query type="cosmasII">
      <query-text><![CDATA[Zampano]]></query-text>
    </query>
  </dii-queries>
</dii-entry>
CoDII-BW.de: Encoding of Examples

- The content model of elements \texttt{dii-examples} consists of a list of \texttt{example} elements.
- \texttt{example} elements are linked to appropriate distributionally idiosyncratic items by dint of attributes \texttt{dii} and \texttt{id}.

A CoDII-XML-description of a corpus example for \textit{Zampano}:

```xml
<example dii="zampano" id="zampano-Bsp">
  <source corpus="cosmasII">
    R97/APR.32703 Frankfurter Rundschau, 29.04.1997, S. 15,
    Ressort: WIRTSCHAFT; F\#252;r eine lohnende \&#220;bernahme
    sind einige H\#252;rden zu nehmen
  </source>
  <ol>
    "Ich glaube nicht, da\&#224; da Manna vom Himmel f\&#228;llt
    und der gro\&#224;e Zampano f\&#252;r diverse neue Stellen
    sorgt", meint der Betriebsratschef der Vegesacker Werft,
    Wolfgang Dettmer.
  </ol>
</example>
```
CoDII-BW.de is available via the Internet in the form of a set of XHTML files generated by an XSLT script.

The URL of CoDII-BW.de is

http://www.sfb441.uni-tuebingen.de/a5/codii
The General Characteristic of CoDII

- CoDII is a research platform for linguistic investigations.
- CoDII does not provide its own corpus but refers to existing corpora and gives sample queries.
- The target audience of CoDII are primarily linguists.
- Information gathered in CoDII allow for a systematic study of distributionally idiosyncratic items.
- On the basis of empirical and theoretical data contained in CoDII-BW.de, a first evaluation is already possible.
First Evaluation

A bound word (BW) requires the presence of a particular Lexeme (LEX), where:

- **BW** selects **LEX**:
  - *Angst einjagen* ‘to scare someone’

- **LEX** selects **BW**:
  - *Tacheles reden* ‘to talk straight’

- **LEX₁** selects **LEX₂** and **LEX₂** selects **BW**:
  - *[zu₂ Pott] kommen₁* ‘to manage a task’

- but not **LEX₁** selects **LEX₂** and **LEX₂** selects **LEX₃** and **LEX₃** selects **BW**:
  - *[zu₃ BW] kommen₂ lassen₁*
Grammar Theoretical Question

How can these generalizations be integrated in a formal grammar architecture?

Architecture of linguistic signs according to Pollard and Sag (1994)

It is the aim of the project to elaborate a general theory of distribution based on data collected in CoDII.
Outlook

- So far, only information on particular lemmata can be found. In order to make the access to the data dynamic and the search more detailed, CoDII will be converted into a data base.
- Existing CoDII entries will be completed.
- Further linguistic information such as dependency structure and logical form will be added to the structure of CoDII entries.
- A collection of bound words in English will be compiled.
- In the next project phase, a collection of German polarity items will be developed.
- The data structure of CoDII is designed in such a way that further collections of distributionally idiosyncratic items in any language can be modelled.
References


