The Collection of Distributionally Idiosyncratic Items: A Multilingual Resource for Linguistic Research

Manfred Sailer and Beata Trawiński

University of Göttingen
University of Tübingen

manfred.sailer@phil.uni-goettingen.de trawinski@sfs.uni-tuebingen.de

LREC 2006 Conference
May 24-26, 2006, Genoa, Italy
Overview

- Background
- CoDII (Idea, Design and Realization)
- The Collections of Bound Words
- Similar Collections
- Outlook
Background: The Research Groups

CoDII (the Collection of Distributionally Idiosyncratic Items) has been developing

- in Project A5 *Distributional Idiosyncrasies* of the Collaborative Research Center 441 *Linguistic Data Structures: On the Relation between Data and Theory in Linguistics* at the University of Tübingen and

- at the Department of English Studies at the University of Göttingen.
The principal concerns are 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.
Background: The Motivation

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.
Background: The Aim

- The aim is to elaborate a general theory of distribution (within the formal paradigm of HPSG).
- To provide an empirical basis for the distribution theory, a large number of distributionally idiosyncratic items should systematically be considered.
- CoDII should provide an empirical basis of this kind.
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 for extraction of data related to these items.
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.
Currently, two collections of distributionally idiosyncratic items are available in CoDII:

- Bound Words of German (CoDII-BW.de) and
- Bound Words of English (CoDII-BW.en).

Both collections can be entered at
http://www.sfb441.uni-tuebingen.de/a5/codii.
Motivation

The motivation for the compilation of the collections of BWs at first was the following:

- BWs are lexical elements with extremely restricted distribution, e.g.:
  - to make headway, by dint of, to play footsie with somebody
  - in Anbetracht von ‘in view of’, der große Zampano ‘the big doer’, noch und nöcher ‘a lot’

- BWs are relatively well documented in (in particular German) phraseological literature.

- Extraction of BWs from corpora is quite trivial.
The Initial Basis

The initial basis for CoDII-BW.de and CoDII-BW.en are

- collections and classifications of BWs in
  - (Dobrovol’skij, 1988),
  - (Dobrovol’skij, 1989),
  - (Dobrovol’skij/Piirainen, 1994b),
  - (Dobrovol’skij/Piirainen, 1994a),

- idiom dictionaries and regular dictionaries and

- our own observations.

Currently, CoDII-BW.de includes about 450 items and CoDII-BW.en includes about 100 items.
Modeling of CoDII-Entries

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

- general information,
- classification,
- syntactic information,
- queries.
General Information

The General Information block identifies BWs by providing:

- a particular BW,
- the English translation of the (German) BW,
- the expression in which the BW occurs,
- the set of possible paraphrases of this expression.
Classification

The block Classification specifies classes associated with a given BW according to the following classifications:

- classification in (Dobrovol’skij, 1988),
- classification in (Dobrovol’skij, 1989),
- classification in (Dobrovol’skij/Piirainen, 1994b),
- Nunberg et al. (1994) oriented classification,
- project-internal classification.
Syntactic Information

The block Syntactic Information provides information on

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

For syntactic description of BWs and expressions in which they occur,

- the *Stuttgart-Tübingen Tagset* (STTS) (for German BWs) and
- the syntactic annotation scheme from the *Syntactically Annotated Idiom Database* (SAID) (for English BWs)

have been used. 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 corpora of German and English such as

- corpora of the Institut of German Language in Mannheim (http://www.ids-mannheim.de/cosmas2/),
- the corpus of das Digitale Wörterbuch der Deutschen Sprache – DWDS (http://www.dwds.de/),
- the Tübinger Sammlung nutzbarer empirischer linguistischer Datenstrukturen – TUSNELDA (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/),
- British National Corpus (via the SARA software package; http://www.natcorp.ox.ac.uk/sara/),
- Internet via Google.
Technical Realization

CoDII-BW.de and CoDII-BW.en have been internally 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).
A Fragment of the CoDII-XML-Encoding of the BW Zampano

```xml
<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-syntaxhits="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-syntaxhits>
  <dii-queries>
    <query type="cosmasII">
      <query-text><![CDATA[Zampano]]></query-text>
    </query>
  </dii-queries>
</dii-entry>
```
Encoding of Examples

- The content model of elements dii-examples consists of a list of example elements.
- example elements are linked to appropriate distributionally idiosyncratic items by dint of attributes dii and id.
- A CoDII-XML-description of a corpus example for Zampano:

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

The URL:

http://www.sfb441.uni-tuebingen.de/a5/codii
Browser Display for the Entry *dint*
Browser Display for the Entry Zampano
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.
- CoDII is becoming a multilingual resource.
Similar Collections for German I

*Usuelle Wortverbindungen* (Conventionalized Word Combinations) of the IDS (*Steyer, 2004*), URL: 
http://www.ids-mannheim.de/lexik/UsuelleWortverbindungen

The project starts from statistically highly frequent words which undergo a co-occurrence analysis. This analysis serves as the basis for a linguistic and lexicographic description of the typical usage patterns of a word. (CoDII is based on linguistic intuitions and theoretical considerations.)

The project only uses the corpora of the IDS — to have full control over the frequency data. (In CoDII, data from different sources and retrieval strategies for different corpora are included.)
Similar Collections for German II

*Kollokationen im Wörterbuch* (Collocations in the Lexicon) of the Berlin-Brandenburgische Akademie der Wissenschaft (Fellbaum et al., ta), URL: http://www.bbaw.de/bbaw.Forschung/Forschungsprojekte/kollokationen/en/Startseite

- The collection is based on the DWDS corpus. *(CoDII uses different corpora.)*
- The project starts with idioms from phraseological literature. *(CoDII does it as well.)*
- The project focuses exclusively on German VP idioms. *(CoDII is a multilingual resource and collects items of any syntactic category.)*
Similar Collections for English I

Syntactically Annotated Idioms Database (SAID), (Kuiper et al., 2003):

- The project encodes the syntactic structure of a huge number of idioms. (CoDII also includes many other information about the expressions.)

- The SAID can be used to investigate structural generalizations about idioms. (For this reason its encoding was also used for representing syntactic structures in the CoDII-BW.en)
An interface for an interactive multilingual resource primarily designed to encode translation equivalents, (Villavicencio et al., 2004), URL: http://lingo.stanford.edu/cgi-bin/annotate.mli.cgi

It allows external users to contribute to the collection. (In CoDII, there is no interactive mode for inserting information from outside users directly into the database.)
The data structure design of CoDII makes it possible to add further classifications, corpora and search tools, as well as further collections of distributionally idiosyncratic items.

An extension to more languages is equally possible. A natural candidate for another language would be Dutch, for which (Dobrovol’skij, 1988) lists a large number of BWs. Furthermore, (Feyaerts, 1994) presents a detailed investigation and classification of Dutch BWs.

It is also planned to extend the collection to other types of distributionally idiosyncratic items. In particular, we intend to include a documentation of the use of polarity items (such as English *any*, or *lift a finger*), i.e., items which require a negative context (cf. (van der Wouden, 1997)).

On the technical side, CoDII will be converted into a database to allow for a dynamic and more flexible access to the data. This database will then be integrated into the TUSNELDA collection of the Collaborative Research Center 441.
References


