Cranberry Expressions in English and in German

Beata Trawiński
Manfred Sailer
Jan-Philipp Soehn
Lothar Lemnitzer
Frank Richter

University of Tübingen & University of Göttingen

Towards a Shared Task for Multiword Expressions (MWE 2008)
Workshop at the LREC 2008 Conference
Marrakech, Morocco
June 1, 2008
Overview

- Our Data Packages
- Cranberry Expressions
- The Source: CoDII
- Possible Applications for our Data Sets
- Summary
Our Data Packages

- **English_CE_Trawinski**
  - CWen.txt: 77 cranberry words in English
  - CCen.txt: 77 corresponding cranberry expressions
  - README.txt

- **German_CE_Trawinski**
  - CWde.txt: 444 cranberry words in German
  - CCde.txt: 444 corresponding cranberry expressions
  - README.txt

**Cranberry Expressions: Definition**

- **Cranberry Word (CW):** an item that only occurs within a specific expression (CE)
  - Aronoff (1976): in analogy to Cranberry Morph
  - Alternatively: *(phraseologically) bound words, unique words, unique lexemes, hapax legomena,* German: *Unikalia*

- **Cranberry Expression (CE):** an expression which contains at least one Cranberry Word
  - Moon (1998): Cranberry Collocations
Cranberry Expressions: Examples

- English CEs:
  - happy as a sandboy
  - kith and kin
  - to play footsie with somebody

- German CEs:
  - in Anbetracht von ‘in view of’
  - jemandem Kattun geben ‘to reprimand somebody’
  - noch und nöcher ‘a lot’
Cranberry Expressions: Idioms versus Collocations

- CEs versus idioms (such as *spill the beans*):
  - lexical fixedness
  - **but** no literal meaning

- CEs versus collocations (such as *take a shower*):
  - word co-occurrence of markedly high frequency
  - **but** hard distributional restrictions rather than preferences

- Idiom-like CEs (*die Spendierhosen anhaben*, wear the offering-trousers, ‘be generous’):
  - idiomatic interpretation of the non-CW components
  - possible syntactic variations and internal modifications
  - non-decomposable meaning

- Collocation-like CEs (*happy as a sandboy*):
  - literal interpretation of the non-CW components
  - structurally parallel to collocations
  - sometimes the CW is interchangeable
Cranberry Expressions: Syntactic Categories

- **Syntactic categories of CEs:**
  - VP: *make headway*
  - PP: *on tenterhooks*
  - AP: *happy as a sandboy*
  - NP: *the whole caboodle*

- **Syntactic categories of CWs:**
  - V: *wend one’s way*
  - A: *spick and span*
  - N: *run the gamut*
Cranberry Expressions: Frequency Classes

- FC 12: *Anhieb* (*auf Anhieb* ‘right away’)
- FC 21: *Kattun* (*jemandem Kattun geben* ‘to reprimand somebody’)

where an FC $n$ indicates that the most frequent German word (*der* ‘the’) is $2^n$ times more frequent than the word in question

URL: [wortschatz.uni-leipzig.de](http://wortschatz.uni-leipzig.de)
The Collection of Distributionally Idiosyncratic Items (CoDII)

Cranberry Expressions

MWE 2008
CoDII: The Database

- Integration into the Open Source XML database eXist, URL: exist.sourceforge.net
- Querying CoDII with respect to
  - particular lemmas
  - syntactic properties
  - licensing environments
  - classifications
CoDII: Some Details on the Collected CEs

- **German CWs:**
  - Common Nouns (80 %) > (predicative) Adjectives (7 %) > Proper Names (5 %) > Verbs (3 %)

- **German CEs:**
  - VPs (83 %) > PPs (20 %)

- **English CWs:**
  - Common Nouns (67 %) > (attributive) Adjectives (21 %) > (predicative) Adjectives (7 %) > Verbs (4 %)

- **English CEs:**
  - NPs (41 %) > VPs (31 %)
Possible Applications for our Data Sets

- Possible Applications:
  - The exhaustive retrieval of CEs from arbitrary text possible
  - This can be used to explore distributional differences between collocations and idioms based on the idiom-like CEs versus collocation-like CEs distinction.

- Other Applications:
  - Further linguistic investigations into MWEs
  - Development of other lexical resources (such as wordnets)
  - Training of methods for extracting semantically related lexical items
  - Acquisition and/or tuning of rules and methods for the automatic detection, annotation, and extraction of idioms
  - Automatic retrieval of otherwise similar MWEs without CWs
We presented two data sets: a list of 444 CWs in German and a list of 77 CWs in English, accompanied by corresponding lists of CEs in which the CWs occur.

We showed that CEs are interesting for the research into MWEs due to their middle position between idioms and collocations, a wide variety of syntactic categories they comprise, and different frequency classes they cover.

We introduced CoDII, the source of our data sets.

We argued that our resource may be useful for theoretical linguistic investigations into MWEs and a number of computational linguistic tasks.


