Cranberry Expressions in English and in German

Beata Trawiński[*], Manfred Sailer[#], Jan-Philipp Soehn[*], Lothar Lemnitzer[†] and Frank Richter[†]

[*]University of Tübingen
SFB 441
Nauklerstraße 35
D-72074 Tübingen
trawinski@sfs.uni-tuebingen.de
jp.soehn@uni-tuebingen.de

[#]University of Göttingen
Department of English Studies
Käte-Hamburger-Weg 3
D-37073 Göttingen
manfred.sailer@phil.uni-goettingen.de

[†]University of Tübingen
Department of Linguistics (CL)
Wilhelmstraße 19
D-72074 Tübingen
lothar@sfs.uni-tuebingen.de
fr@sfs.uni-tuebingen.de

Abstract

We describe two data sets submitted to the database of MWE evaluation resources: (1) cranberry expressions in English and (2) cranberry expressions in German. The first package contains a collection of 444 cranberry words in German (`CWde.txt`) and a collection of the corresponding cranberry expressions (`CCde.txt`). The second package consists of a collection of 77 cranberry words in English (`CWen.txt`) and a collection of the corresponding cranberry expressions (`CCen.txt`). The data included in these packages was extracted from the Collection of Distributionally Idiosyncratic Items (CoDII), an electronic linguistic resource of lexical items with idiosyncratic occurrence patterns. Each package contains a `readme` file, and can be downloaded from `multiword.wiki.sourceforge.net/Resources`.

1. Background and Motivation[0]

The original impetus for compiling the present data[1] came from research into the relationship between the regular syntactic and semantic combinatorial system of grammar and irregular, or exceptional, lexical items. Some expressions, such as reflexive pronouns and negative polarity items, appear quite freely in sentences as long as certain occurrence requirements are fulfilled – there must be an appropriate antecedent or a negation, respectively. While those items specify their occurrence requirements in terms of grammatical notions, there is another group of items, cranberry words, that require the presence of a specific lexeme. A typical cranberry word is *sandboy*, which can only occur as part of the expression *happy as a sandboy*. These items are of particular interest in our research on distributional idiosyncrasies, but they are also of interest for the study of multiword expressions in general, as we will explain below. In Section 2, important properties of cranberry expressions and their position between idioms and collocations will be discussed. In Section 3, the linguistic resource will be presented from which the sets of cranberry expressions were extracted. Section 4 will provide a few statistical details on the collected expressions. In Section 5, the potential of the described data sets for computational lexicography and information extraction will be outlined. Section 6 will summarize the discussion.

2. Cranberry Expressions

Cranberry Expressions (CE) are multiword expressions which contain an item that is not found in the language outside this expression. This item is called a *Cranberry Word* (CW) in (Aronoff, 1976), in analogy to "cranberry morph". Alternatively CWs are also called *(phraseologically) bound words* or *unique words* (German: *Unikalia*).

The repertoire of CEs in German and English is well documented in the literature on idioms. (Dobrovol'skij, 1988) contains the most exhaustive list of CEs in German, English and Dutch. Emphasizing the difference between bound and free words, (Dobrovol'skij, 1988) and (Dobrovol'skij and Piirainen, 1994) provide criteria for classifying CEs and the expressions in which they occur. In fact, it is not always clear whether an item should count as a CW or not. For example, the noun *Abstellgleis* (*holding track*) is in our list of CWs because it usually occurs in the CE *jn. aufs Abstellgleis stellen/schieben* (literally: *put so. on the holding track*). This expression receives the metaphorical interpretation *to put so. on inactive reserve* or *to deprive so. of his/her influence*. In contrast to the constituents of typical idioms, the word *Abstellgleis* stems from a technical domain (railway systems) and is not used in everyday language outside the CE.

Dobrovol'skij and Piirainen estimate the number of CEs in German at 600. They classify 180 as belonging to the common vocabulary of native speakers. At present, we have included 444 potential CEs in our collection. For English, (Dobrovol'skij, 1988) lists about 100 items, 77 of which are included. The leading criterion for recording an item was whether it was discussed as a candidate of containing a CW within the phraseological literature. In the CoDII resource, sketched in Section 3 below, we document the linguistic classifications and properties of the CEs.

CEs take a middle position between idioms (such as *spill the beans*) and collocations (such as *take a shower*). Due to their restricted occurrence CWs fulfill the criterion of lexical fixedness typically found with idioms (*spill the peas* is not a variant of *spill the beans*). However, in contrast to typical idioms, there is no (synchronically used) literal meaning. CEs share with collocations a linguistically significant co-occurrence of the CW with the other components of the

[0]We would like to thank Janina Radó for comments and suggestions concerning the content and style of this paper.

[1]The data packages originate from (i) project A5, *Distributional Idiosyncrasies* (2002–2008), of the Collaborative Research Center SFB 441 (*Linguistic Datastructures*) at the University of Tübingen, funded by the German Research Foundation (DFG), `www.sfb441.uni-tuebingen.de/a5/index-engl.html`, and (ii) the linguistics section of the English Department of the University of Göttingen.

CE. However, in the case of CWs this is not a question of preference but a hard restriction.

These differences notwithstanding, some CEs should be grouped with idioms, others with collocations. The idiom-like CEs show an idiomatic interpretation of their non-CW components. They also manifest a small range of possible modifications, and the expression as a whole can be assigned one non-decomposable meaning. This meaning can be indicated by a synonym, an antonym or a paraphrase, cf. *Schiffbruch erleiden*, synonym: *scheitern* ('to fail'); *die Spendierhosen anhaben*, synonym: *großzügig sein* ('to be generous'), antonym: *geizig sein* ('to be thrifty'). The collocation-like CEs have a literal interpretation of the non-CW components. They are also structurally parallel to collocations (CE: *make headway*, *happy as a sandboy*; collocation: *make progress*, *dark as night*). Sometimes the CW is interchangeable. A typical example is *Tacheles/Klartext/Fraktur reden* ('to state sth. clearly and with some force'). The range of interchangeable words is always rather small.

CEs comprise a wide variety of syntactic categories (VP: *make headway*; PP: *on tenterhooks*; AP: *happy as a sandboy*; NP: *the whole caboodle*). Similarly, CWs are of all major syntactic categories (V: *wend one's way*; A: *spick and span*; N: *run the gamut*). They also cover different frequency classes:[2] The German CW *Anhieb* (in *auf Anhieb* ('*right away*')) is of frequency class 12 (i.e. the most frequent German word is 2^{12} times more frequent), the CW *Kattun* (in *jm. Kattun geben* ('*to reprimand so.*')) is of frequency class 21.

The reported properties indicate that, at least in German and English, while defined on the distributional properties of one component, CEs comprise instances of a great number of types of the multiword expressions in the language.

3. The Collection of Distributionally Idiosyncratic Items (CoDII)

The data packages we present here were extracted from the Collection of Distributionally Idiosyncratic Items. CoDII is an electronic multilingual resource for lexical items with idiosyncratic occurrence patterns. It was originally designed to provide an empirical basis for linguistic investigations of these items. CoDII compiles and lists items of interest, providing linguistic documentation and corpus evidence, and specifying possibilities for extracting more context data for the items in the collection. When we created CoDII, we were concerned with two kinds of expressions: (i) negative and positive polarity items as expressions whose distribution is grammatically restricted, and (ii) cranberry expressions as expressions whose distribution is restricted by lexical co-occurrence patterns. Design and data structure of CoDII have been conceived in such a way that subcollections of various types of distributionally idiosyncratic items can be modeled (such as anaphora, negative and positive polarity items, and cranberry words), and collections of distributionally idiosyncratic items from various languages can be integrated.

[2]The frequencies are taken from the data of the project *Deutscher Wortschatz* at the University of Leipzig, `wortschatz.uni-leipzig.de`.

Five collections of distributionally idiosyncratic items are currently available in CoDII: CWs in German, CWs in English, Negative Polarity Items in Romanian, Negative Polarity Items in German, and Positive Polarity Items in German. The collections of cranberry words are based on (Dobrovol'skij, 1988; Dobrovol'skij, 1989) and (Dobrovol'skij and Piirainen, 1994), and are described in (Sailer and Trawiński, 2006). The resources for polarity items are described in (Trawiński and Soehn, To appear).

Each CoDII entry contains the following information blocks: General Information (including glosses and translations, if appropriate, as well as the expression in which the item occurs together with a set of possible paraphrases of this expression), Classification, Syntactic Information (including syntactic variations) and, optionally, search patterns. For the syntactic annotation of German and English items, the *Stuttgart-Tübingen Tagset* (STTS) and the syntactic annotation scheme from the Syntactically Annotated Idiom Database (SAID) were used, respectively. For each context, appropriate examples are provided from various corpora, the Internet and the linguistic literature.

CoDII is encoded in XML and is freely accessible on the Internet at `www.sfb441.uni-tuebingen.de/a5/codii`. A fragment of the XML encoding of the English CW *sandboy* in the CoDII format is presented in Figure 1. The elements `dii` and `dii-expression`, `dii-classification`, `dii-syntax` and `dii-queries` model the information blocks specified above.

```
<dii-entry id="sandboy">
 <dii><ol>sandboy</ol></dii>
 <dii-expression>
   <ol>happy as a sandboy</ol>
   <ol-paraphrase>very happy</ol-paraphrase>
 </dii-expression>
 <dii-classification>
   <dii-class category="bw"
              class="dekompo"
              type="A5">
   <bibliography bib-item="A5"/>
   </dii-class>
    [...]
   <dii-class category="bw"
              type="Dobro88"
              class="gebWB">
   <bibliography bib-item="Dobrovolskij88"/>
   </dii-class>
 <dii-syntax cat="NN"
             hits="sandboy01 [...] sandboy02">
   <dii-expression-syntax cat="AdjP">
      [AP[AP[Ahappy]][COMPas][NP[DETa][NP[Nsandboy]]]]
   </dii-expression-syntax>
 </dii-syntax>
 <dii-queries>
    <query type="google" hits="sandboy01">
      <query-text>
          "happy as a sandboy"
      </query-text>
    </query>
 </dii-queries>
</dii-entry>
```

Figure 1: The CoDII-XML-encoding of *sandboy*

CoDII not only compiles, documents and (alphabetically) lists distributionally idiosyncratic items, it also offers dynamic and flexible access. Taking advantage of the theoretically grounded internal data structure and an annotation scheme which involves syntactic and (partial) semantic information, a comfortable interface for querying the

database was created with the Open Source XML database eXist (`exist.sourceforge.net/`). At present, possible search criteria comprise lemmas, syntactic properties, and classifications. Searching for expressions with particular licensing contexts is also possible. With these tools, the two data sets which are presented here as pure (alphabetically ordered) lists of expressions can be modified and enriched if this is necessary for a particular task.

Several other projects have constructed resources for idiomatic expressions. These projects differ from CoDII by the corpora used, the kind of data and the applied methods. The project *Usuelle Wortverbindungen* (Conventionalized Word Combinations, URL: `www.ids-mannheim.de/ll/uwv/`) of the Institut für Deutsche Sprache (IDS) (Steyer, 2004) starts from statistically highly frequent words which undergo a co-occurrence analysis. It only uses the corpora of the IDS. In contrast to this collection, CoDII is based on linguistic intuitions and theoretical considerations and includes data from different sources. The project *Kollokationen im Wörterbuch* (Collocations in the Lexicon, URL: `kollokationen.bbaw.de`) of the Berlin-Brandenburgische Akademie der Wissenschaft (Fellbaum et al., 2005) is based on the corpus *Das digitale Wörterbuch der deutschen Sprache*. Like CoDII, the project starts with idioms from phraseological literature, but focuses exclusively on German VP idioms. For English, the *Syntactically Annotated Idioms Database* (SAID) encodes the syntactic structure of a large number of idioms (Kuiper et al., 2003), but it contains no other information about the expressions.

4. Some Details on the Collected CEs

As noted above, CWs are of all major syntactic categories. However, the overwhelming majority of German CWs are nouns (80%, e. g. *jn. beim Schlafittchen packen*, 'to take so. by the scruff of the neck'), followed by predicative adjectives (7%, e. g. *sattsam bekannt*, 'widely known'), proper names (5%, e. g. *Büchse der Pandora*, 'Pandora's box'), and verbs (3%, e. g. *alles, was da kreucht und fleucht*, 'everything that crawls and flies'). VPs (83%) are the most common syntactic environment for (the typically nominal) CWs in German CEs. In 87 cases (20%) a CW is the complement of a specific preposition. These "unique nominal complements" form an important subclass of CWs (e. g. ***auf** Anhieb*, 'right away' or ***on** tenterhooks*, cf. (Soehn and Sailer, 2003)). From a theoretical point of view, these data provide excellent evidence that non-heads, including complements, can impose restrictions on the heads they combine with.

English CWs reveal a different pattern. Although the most common category is again nouns (67%, e. g. *at first blush*), the second most common one is attributive adjectives (21%, e. g. *curule chair*). Predicative adjectives and verbs play only a minor role with 7% and 4%, respectively. The leading syntactic category of CEs is not VP (31%) but NP (41%). This is a consequence of the fact that free nouns form compounds with bound nouns. Compounding is a morpho-lexical process which works differently in English and in German: English compounds consist of several orthographic words which are categorized as multi-word expressions (NPs). In German compounds form one orthographic unit. The difference leads to many English NPs with bound nouns; additional bound adjectives in NPs further increase their frequency.

5. Our Data Sets and Other Resources

Our cranberry expression data sets for English and German are a valuable resource for the documentation of a special aspect of these languages as well as an empirical base for investigations into multi-word expressions. However, we believe that one should think beyond these applications and explore how these data can a) inform the development of other lexical resources and b) be useful for data-driven information extraction experiments.

The information provided in our data sets goes well beyond the mere listing of the CEs and includes semantic glosses which contain synonyms, antonyms and examples. Linking those CEs which behave like non-decomposable idioms to semantically related lexical items, i. e. to their synonyms, antonyms, and hypernyms, would make it possible to enrich other lexical resources.[3] Many CEs in our collection contain links to semantically related words through their paraphrases or glosses. Admittedly, there is a wide variety of glosses and not all of them consist of a single lexically related word. Nevertheless, they are a good starting point for creating more explicit lexical-semantic relations.

Moreover, systematically connecting the CEs with the English and German wordnets would benefit both resources: a) Wordnet users gain access to an interesting set of multi-word expressions. These are currently underrepresented in wordnets, and in particular in GermaNet; b) on the other hand, the CEs of our collections would be embedded into broader semantic fields, e. g. the CE *aufs Abstellgleis schieben* would be related to the verb *abschieben* and its hypernym, other hypernyms of this verb etc. (Lüngen et al., 2008) present an approach of linking general language wordnets with specialized lexical resources which can also be applied to our resources.

Regarding the use of our data for information extraction purposes, the CEs and their lexical-semantic neighbors – either in the glosses or through the links to wordnets – become an interesting resource for the training of methods for extracting semantically related lexical items from corpora, which is an active research area in the field of lexical acquisition, cf. (Hendrickx et al., 2007; Snow et al., 2006). It is comparatively easy to collect a set of contexts, in the form of concordances, for CWs because most of them are unambiguous or at least occur most often in (semi-)fixed contexts. From the concordance, context vectors can be derived and abstracted which represent the distributional characteristics or "fingerprint" of these lexical items. This can be the basis for a comparison with other, semantically related, lexical items. Currently, there are corpus citations for only a few CEs in the collection. For some of them, however, we provide search patterns which can be applied to a corpus to extract a larger set of examples.

Our collection of idiom-like CEs is suitable as training material in yet another lexical acquisition task. It is well-

[3] One might also consider including those collocation-like CEs as well whose meaning can be mapped to one single concept.

known that many idioms may undergo a range of mainly syntactic variations and internal modifications (cf. (Nunberg et al., 1994), and for German (Lemnitzer and Kunze, 2007, chapter 11) and (Soehn, 2006)). Rules and methods for the automatic detection, annotation, and extraction of idioms must take this variability into account. As said before, it is relatively easy to build a collection of examples for our CEs. As they represent several types of multiword expressions, the data can be used to capture variations and modifications in idiomatic expressions and help to acquire and / or fine tune these rules.

6. Summary and Outlook

Cranberry expressions are multiword expressions with special properties that make them interesting for the theoretically oriented linguist as well as for use as an electronic resource for lexical acquisition, and for evaluation and extraction tasks. In this paper we presented two resources, a list of 444 cranberry words in German and a list of 77 cranberry words in English, accompanied by corresponding lists of cranberry collocations in which the cranberry words occur.

What makes CEs interesting is their middle position between idiomatic expressions and collocations. Their special property is the obligatory occurrence of a unique cranberry word in each CE. Once a cranberry word has been identified, the obligatoriness of this lemma and its categorial and robust lexical occurrence restriction makes the exhaustive retrieval of its CEs from corpora and the Internet much easier and much more reliable than the retrieval of idiomatic expressions in general. It follows that well-documented CEs with particular properties may be good candidates for a gold standard in a retrieval task for otherwise similar multiword expressions without CWs. It is at this point that the middle position of CEs between idioms and collocations becomes particularly interesting, since it opens the door for using appropriate subclasses of CEs in both research contexts. The additional documentation of our CEs in CoDII, comprising linguistic classification and access by means of various search categories, further enhances the usefulness of the resource.

7. References

Mark Aronoff. 1976. *W*ord Formation in Generative Grammar. MIT Press, Cambridge, Massachusetts and London, England.

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

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

Dmitrij Dobrovol'skij. 1989. Formal gebundene phraseologische Konstituenten: Klassifikationsgrundlagen und theoretische Analyse. In Wolfgang Fleischer, Rudolf Große, and Gotthard Lerchner, editors, *Beiträge zur Erforschung der deutschen Sprache*, volume 9, pages 57–78. Leipzig, Bibliographisches Institut.

Christiane Fellbaum, Undine Kramer, and Gerald Neumann. 2005. Corpusbasierte lexikographische Erfassung und linguistische Analyse deutscher Idiome. In Annelies Häcki Buhofer and Harald Burger, editors, *Phraseology in Motion I. Methoden und Kritik. Akten der Internationalen Tagung zur Phraseologie (Basel, 2004)*, pages 183–199. Schneider Verlag, Hohengehren.

Iris Hendrickx, Roser Morante, Caroline Sporleder, and Antal van den Bosch. 2007. ILK: Machine learning of semantic relations with shallow features and almost no data. In *P*roceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 187–190, Prague, Czech Republic, June. Association for Computational Linguistics.

Koenraad Kuiper, Heather McCann, Heidi Quinn, Therese Aitchison, and Kees van der Veer. 2003. Syntactically annotated idiom database (SAID) v.1. Documentation to a LDC resource.

Lothar Lemnitzer and Claudia Kunze. 2007. *C*omputerlexikographie. Gunter Narr Verlag, Tübingen.

Harald Lüngen, Claudia Kunze, Lothar Lemnitzer, and Storrer Angelika. 2008. Towards an Integrated OWL Model for Domain-Specific and General Language WordNets. In *P*roceedings of the Fourth Global WordNet Conference, pages 281–296. University of Szeged, Hungary, Department of Informatics.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *L*anguage, 70:109–132.

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.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *A*CL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pages 801–808, Morristown, NJ, USA. Association for Computational Linguistics.

Jan-Philipp Soehn and Manfred Sailer. 2003. At First Blush on Tenterhooks. About Selectional Restrictions Imposed by Nonheads. In Gerhard Jäger, Paola Monachesi, Gerald Penn, and Shuly Wintner, editors, *P*roceedings of Formal Grammar 2003, pages 149–161.

Jan-Philipp Soehn. 2006. ¨Uber Bärendienste und erstaunte Bauklötze – Idiome ohne freie Lesart in der HPSG. Phd dissertation (2005), Friedrich-Schiller-Universität Jena.

Kathrin Steyer. 2004. Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikographische Perspektiven. In Kathrin Steyer, editor, *Wortverbindungen – mehr oder weniger fest*, pages 87–116. de Gruyter, Berlin and New York.

Beata Trawiński and Jan-Philipp Soehn. To appear. A Multilingual Database of Polarity Items. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.