Expressions with Bound Words
Linguistic Resource and Relevance for PARSEME

Manfred Sailer (Frankfurt), Frank Richter (Düsseldorf), Beata Trawiński (Mannheim) — **WG 1**

Bound words (BW) are words that only occur in a highly restricted set of environments, typically in just one multiword expression. Well-known examples from English are *headway* as in *make/ *show headway* and *umbrage* as in *take/ *feel umbrage*.
In this poster we will show the usefulness of expressions with BWs for the classification and investigation of MWEs, we will present an online collection of BWs in German and English and point to additional connections between the study of BWs and interests inside PARSEME.

1 Linguistic aspects of bound words

In the phraseological literature it has been shown that BWs exist in all languages for which phraseological research exist. Lists and classifications of BWs have been published for Dutch, English, German, Russian at least (Dobrovol'skij, 1988; Dobrovol'skij and Piirainen, 1994; Feyaerts, 1994; Fleischer, 1989). In the context of PARSEME it is relevant to show that MWEs with BWs exists for all types of MWEs, if we take the classification in Sag et al. (2002) and Baldwin and Kim (2010). This is done in the following list, using expressions with BW from German, the BW being underlined.

1. fixed expressions: *by and large*. with BWs: *Krethi und Plethi* (*people of all kind*)
2. semi-fixed expressions: complex proper names; compound nominals; non-decomposable idioms with BWs: *der gordische Knoten* (*the Gordian knot*); *Schorn-stein* (*chimney*); *wider den Stachel löcken* (*kick against the goads*)
3. syntactically flexible idioms: verb particle constructions; decomposable idioms; light verb constr. with BWs: *klein bei-geben* (*to give in*); *die Spendierhosen anhaben* (*be generous*); *jmdn. zum Hahnrei machen* (*cheat on s.o.*)
4. institutionalized phrases: idioms of encoding, not decoding ("collocations"); conversational formalæ with BWs: *Eier abschrecken* (*dip eggs into cold water*), *klar wie Kloßbrühe* (*quite plain/obvious*); *'n Abend* (*Good evening*)

The study of expressions with BW has a number of advantages: (1) As a BW is restricted in its distribution, it is easy to extract naturally occurring examples of such expressions by using the BW in virtually any corpus, be it annotated or not. (2) As the use of a BW outside its usual context leads to ungrammaticality, expressions with BW can be used as a strong argument that a collocational component needs to be present in competence-oriented language descriptions, including parsing systems. (3) As is typical for collocations, some expressions with BW are fairly frequent (such as *auf Anhieb* (*at first go*)), however there are transparent but infrequent expressions with BWs (such as *jmdn. etwas madig machen* (*spoil s.o.'s pleasure in s.th.*), which shows that collocational strength is not a mere frequency phenomenon.

2 The Collection of Distributionally Idiosyncratic Items (CoDII)

The basic idea of CoDII is to provide a linguistic documentation of items that have an unexpected distribution. BWs are a prime case of this. CoDII has been developed since 2002 with contributions from the University of Tübingen and the University of Göttingen, and is now permanently maintained at the University Frankfurt. The resource has been presented for example in Trawiński et al. (2008) and Richter et al. (2010). At present CoDII contains two collections with BWs: CoDII-BW.de with 444 German BWs and CoDII-BW.en with 77 English BWs.[1]
Each CoDII entry contains the following information blocks: (i) general information (glosses, translation to English), (ii) syntactic information (including possible syntactic variation), (iii) classification found in the literature, (iv) usage examples from corpora and the Internet (including search patterns to make it easy for users to find more examples). All information is encoded in XML, with a uniform underlying DTD for all collections. Because of this architecture, new collections on different languages can easily

[1] There are three additional collections with polarity items for German and Romanian (Trawiński and Soehn, 2008), but these will not be addressed in the present poster.

be added. All collections of CoDII can be freely accessed online at www.english-linguistics.de/codii, together with an online bibliography on BWs at www.english-linguistics.de/sfb441/a5/bwb.
We plan to integrate CoDII in a relational database to allow for a quick search and to retrieve quantitative information on the types of documented items. In addition, the collection will be dynamically updated as we become aware of missing items or errors.

3 BWs: Lexical representation and parsing

We expect that our poster contributes not only to the question of classifying MWEs but also to other concerns of PARSEME:

Lexical representation Expressions with BWs do not receive a uniform treatment in lexicography (Trawiński et al., 2008), sometimes the BW is used as a keyword, sometimes the syntactic head of the expression. While this might be defensible, current lexicographic practice is far from being systematic. One possibility is to relate the question of the lexicographic keyword to the psycholinguistic notion of an *idiomatic key* (Titone and Connine, 1999), i.e., what is/are the elements in a MWE that are most important in the recognition of the MWE. Identifying this key may be easier for expressions with BW than for MWEs in general. We hope that the results on BWs can also be used for MWEs that are composed of free words.

Parsing Expressions with BWs may also provide indications on useful parsing strategies for MWEs. As the primary aim of PARSEME is parsing, it does not do any harm if a computational grammar assigns a parse and an interpretation to structures containing a BW in an inappropriate context (such as **show headway*). In fact, over-parsing might even be desirable, as some speakers are more flexible in the use of some BWs than others. Such a treatment, however, is only possible if a "decomposed" representation is chosen, i.e., if parts of a MWE are treated as independent lexical items, at least for some MWEs.[2]

References

Baldwin, Timothy and Kim, Su Nam Kim (2010). Multiword Expressions. In N. Indurkhya and F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2 ed.)., pp. 267–292. Boca Raton: CRC Press.

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

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

Feyaerts, Kurt (1994). Zur lexikalisch-semantischen Komplexität der Phraseologismen mit phraseologisch gebundenen Formativen. In *Sprachbilder zwischen Theorie und Praxis*, pp. 133–162. Bochum.

Fleischer, Wolfgang (1989). Deutsche Phraseologismen mit unikaler Komponente — Struktur und Funktion. In G. Gréciano (Ed.), *Europhras 88*, pp. 117–126.

Richter, Frank, Sailer, Manfred, and Trawiński, Beata (2010). The Collection of Distributionally Idiosyncratic Items. An Interface between Data and Theory. In S. Ptashnyk, E. Hallsteinsdóttir, and N. Bubenhofer (Eds.), *Korpora, Web und Datenbanken. Computergestützte und korpusbasierte Methoden in der Phraseologie, Phraseografie und der Lexikografie*, pp. 245–261. Hohengehren: Schneider Verlag.

Sag, Ivan A., Baldwin, Timothy, Copestake, Ann, and Flickinger, Dan (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. F. Gelbukh (Ed.), *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, London, pp. 1–15. Springer.

Titone, Debra A. and Connine, Cynthia M. (1999). On the Compositional and Noncompositional Nature of Idiomatic Expressions. *Journal of Pragmatics 31*, 1655–1674.

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

Trawiński, Beata, Soehn, Jan-Philipp, Sailer, Manfred, and Richter, Frank (2008). A Multilingual Electronic Database of Distributionally Idiosyncratic Lexical Items. In *Proceedings of Euralex 2008*, Barcelona.

[2]The two BW collections are available to the MWE community at Souceforge as text files and can be used for experimenting. The data sets can be downloaded from multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets.