Creating a phraseme matrix based on a Tertium Comparationis[1]

Cerstin Mahlow

Keywords: *tertium comparationis*, *meta-index*, *phrasemes*.

Abstract

Diachronic exploration of linguistic resources like collections and dictionaries from different time periods allows researchers to get first impressions on language change and define specific research questions to investigate further, for example by integrating empirical data. However, manual inspection of large collections is exhausting and error prone. Automatic extraction and comparison of the keywords of dictionary entries from several dictionaries can be used to create a combined index, allowing to easily access respective dictionary entries to extract related information. As a case in point we consider information on German phrasemes in dictionaries and collections from the 18th to the 21st century. We use a concept-driven semi-automatic approach to create a matrix based on a Tertium Comparationis to allow users to easily look up related phrasemes.

1. Introduction

Phrasemes—sometimes also referred to as phraseological units, idioms, or set phrases—are the subject of investigation in phraseology. During the last centuries, several researchers collected and investigated phrasemes, and presented them in printed books with information on meaning, structure, typical usage, examples from texts, etc. For German, we have for example dedicated collections by Wander (1867–1880) (= DSL); Borchardt (1888); Friedrich (1976); Röhrich (2002) or Dudenredaktion (2008) (= Duden11). Additionally, phraseological information is included in generic dictionaries, for example by Adelung (1793–1801) (= Adelung); Campe (1807–1812); Sanders (1859–1865); Grimm and Grimm (1852–1971) or Küpper (1997). Today, some of these collections and dictionaries are available electronically, but all of them were originally intended to be used on paper. In the 21st century, we also have online collections of phrasemes; academic collections like the “Idiomdatenbank”[2] or amateur collections like the “Redensartenindex” (RA-I)[3]. Each mentioned collection presents dense, (mostly) scholarly information at a given point in time, that is, from a synchronic point of view.

For a phraseologist interested in deducing diachronic development from synchronic dictionary entries in dictionaries created at different points in time, a huge effort is necessary to gather all relevant information from all available resources (see for example Dräger (2010)). Even to identify relevant resources requires looking up the phraseme under investigation in all dictionaries.

Lexical entries have a distinct structure, often made visually explicit by using typographical features like different fonts, font styles, or font sizes, using indentation or special characters, etc. In most cases it is very easy to identify the keyword; as more and more resources are available in digitized form, it is possible to extract these keywords automatically. Given this possibility, we might aim to create a meta-index, consisting of all keywords used and pointing to the respective original resources (or more specifically to the full dictionary entry). Researchers could then use this meta-index to easily identify the dictionary entries to consider for further exploration related to the phraseme under investigation; and to get an impression about the inventory of the relevant dictionaries.

In this paper we will present our concept-driven semi-automatic approach to create such a meta-index. First we describe characteristics of the dictionaries and collections used and identify some challenges. Then we present our approach to develop a suitable Tertium Comparationis (see Connor and Moreno (2005)) as a basis for such an index.

2. Resources

For the SNSF funded project “German Proverbs and idioms in language change. Online-dictionary for diachronic phraseology” (OLdPhras), we had to decide which phrasemes to investigate in detail. It is not possible to provide a high-quality, dense diachronic dictionary including every phraseme found in a large text corpus or included in at least one of the dictionaries mentioned in section 1 (see Mahlow and Juska-Bacher (2011) and Juska-Bacher and Mahlow (in print) for details). A first idea was to get an overview of which phrasemes (or variants) are mentioned in at least two or three dictionaries. Creating this kind of index manually was impossible. However, since the keywords of a dictionary can be extracted easily, we could gather the indices of the phraseological collections automatically.

Generic dictionaries like Küpper (1997) or Grimm and Grimm (1852–1971) mention phrasemes within dictionary entries; the keywords are single words, and therefore not of interest for our purpose. However, some authors use distinct typographical features to mark phrasemes, thus allowing for automatic extraction as well. Others like Adelung (1793–1801) use specific abbreviations in paragraphs mentioning phrasemes. We used these to identify potentially interesting sentences and then marked phrasemes and variants manually. From this annotation we could then extract all phrasemes automatically.

We considered two contemporary and two historic collections: Duden11 and RA-I, and Adelung and DSL—all in the digitized version. From these we extracted:

- 3’974 manually annotated phrasemes and variants from Adelung,
- 45’729 potential phrasemes from DSL (using the explicit marker of the author),
- 13’287 phrasemes from Duden11 (using the typographical markup of phrasemes), and
- 11’542 phrasemes from RA-I (using the provided index and removing duplicates)

Given that the resources were created by different scholars at different points in time, following different phraseological schools or theories, printed by different publishers, following different lexicographic and editing guide lines, we assumed to find a phraseme to appear in different dictionaries in various forms, in line with the results of the in-depth study by Stantcheva (2003). Expected and observed variation includes: (a) spelling due to different language stages like in example 1; (b) word order (examples 2 or 3); (c) variation in words used resulting in some kind of “similarity” (example 4); (d) a somewhat artificial formulation—for example, using always the infinite verb form—or using a prototypical example with inflected word forms; (e) handling of valencies, for example, whether a valency can be filled by various nouns, pronouns, noun phrases, verbs, etc. (example 5); are valencies filled with a prototypical example or marked with an explicit place holder like *etwas* (‘something’) or *sich* (reflexive personal pronoun), place holders can be written as abbreviation, too; (f) providing morphosyntactic information on valency fillers, for example, explicit mentioning of expected case like “+ Dat.” or using a marker in the placeholder like *jemandem* (the use of the Dativ case has to be deduced by the user); (g) inclusion of possible modifications like using modal verbs, adverbs, or negation.

(1) (a) *Alle Hände voll zu thun haben* (Adelung)
(b) *alle/beide Hände voll zu tun haben* (Duden11)
‘to have both hands buisy (to have one’s hands full)’[4]

(2) (a) *Augen haben wie ein Luchs* (Duden11)
(b) *Augen wie ein Luchs haben* (RA-I)
‘to have eyes like a lynx (to have eagle-eyes)’

(3) (a) *Mein kleiner Finger hat es mir gesagt* (Adelung)
(b) das sagt mir mein kleiner Finger (Duden11)
'my little finger told me (a little bird told me)'

(4) (a) *Über Hals und Kopf* (Adelung)
(b) *Hals über Kopf* (Duden11)
'neck over head (to be in a hurry)'

(5) *auf schwachen/schwankenden/tönernen/wackligen Füßen stehen* (Duden11)
'to stand on weak/shaky/fictile/wobbly feet (to stand on shaky ground)'

(6) (a) *Augen im Kopf haben* (Duden11)
(b) *hast du/haben Sie keine Augen im Kopf?* (Duden11)
'don't you have eyes in your head? (don't you have eyes to see?)'

However, we also noted variation within single collections, probably due to the time needed to compile all data and write the dictionary entries, or due to multiple authorship (example 6). Some phrasemes, like example 7 or 8, are indeed identical through all resources.

(7) *Haare auf den Zähnen haben*
'to have hairs on the teeth (to be a tough customer)'

(8) *Lügen haben kurze Beine*
'lies have short legs (you won't get far by lying)'

For some variants, normalization would be easy, for example, by expanding abbreviated placeholders or generating infinite verb forms from inflected verb forms. Deciding whether two phrasemes belong together—based on shared vocabulary and identical or similar syntactical structure—, would involve more effort; Geyken and Boyd-Graber (2004) show some approaches. However, their approach, and other natural language processing (NLP) methods, can be applied to modern text only—they are not suitable to handle texts from the 18th or 19th century; for general remarks on NLP and historical German texts see, for example, Dipper (2010) or Scheible et al. (2011). We thus had to find another solution.

3. Approach

Taking into account the characteristics of phrasemes (i.e., polylexicality, relative stability, and idiomaticity (Burger, 2010, 36ff)), we could identify a potential Tertium Comparationis—the meaning of phrasemes. Since phrasemes are non-Fregian multi-word units, the meaning of the whole cannot be deduced from the meaning of the parts; on the other hand, phrasemes that are similar in meaning don't necessarily share vocabulary, as shown in example 9 (they all express that the time is running out).

(9) (a) *Matthäi am Letzten* 'the last of Matthew'
(b) *höchste Eisenbahn* 'highest train'
(c) *fünf vor zwölf* 'five before twelve'

In addition to the variants of phrasemes listed in section 2—which all involve lexical or structural similarity—, a phraseme could be considered a variant or synonym of another phraseme because of a shared meaning, independent from lexical or syntactic aspects. Detecting those groups automatically is impossible, since NLP methods rely on surfaces—for example, when lemmatizing word forms or extracting noun phrases—, or operate on the meaning by consulting ontologies like WordNet or GermaNet, but they do not operate on the underlying idiomatic meaning. However, grouping phrasemes according to their meaning would be a very useful strategy for providing a network of phrasemes, which could serve as the basis of a meta-index for phraseological collections from different centuries: Although syntactic structures or vocabulary might have changed over time, it is very probable that the expressed meaning was preserved, as in example 4.

As a starting point, we assumed that phrasemes sharing autosemantica (e.g., nouns), could probably share meaning to some extent. We therefore automatically lemmatized all phrasemes from our four resources resulting in pairs of phrasemes and nouns. If a phraseme had more than one noun, we manually chose the most representative one—that is, the noun you would expect the phraseme to be listed under in a general dictionary. We could then sort phrasemes according to their shared noun. Each noun-phraseme pair was then manually assigned an identifier; seeing phrasemes sharing a common noun grouped together, speeds up the assigning of identifiers.

Identifiers consist of two parts: (a) a five-character code representing the semantic category, and (b) a three-character index representing a prototypical instantiation of this category. After first experiments, we decided to not give an explicit verbal designation for semantic categories: The inventory of the semantic index is growing while annotating noun-phraseme pairs, the verbal designation would have to be adjusted all the time. The semantic category is therefore only given implicitly by the prototypical instantiations. As this resource is intended to be used primarily by human experts, there is no need for explicit categorization.

We also decided to not explicitly mark the relationship between phrasemes belonging to the same semantical category. The relation always involves semantic similarity of the underlying concept, but could involve various formal aspects concerning vocabulary, syntactic structure, transformations, morphosyntactic features of the whole multi-word unit, possible syntactic roles of the whole multi-word unit, etc.

4. Results and Conclusion

In this paper we presented our approach for creating a meta-index of phrasemes in German dictionaries and collections from various points in time. We use the semantic concept of phrasemes as a Tertium Comparationis to be able to group phrasemes expressing the same meaning by using different words and syntactical structures. This meta-index presents an overview of the inventory of special-purpose collections and general-purpose dictionaries for German with respect to phrasemes. The phrasemes included in these collections are listed in the meta-index according to their underlying idiomatic meaning. The index can be sorted by the kind of information involved, that is, concepts, collections, or nouns. Human experts can easily see at a glance (a) if a specific semantic concept is included in all collections (by browsing concepts), (b) the degree of variation concerning various aspects like vocabulary or syntactic structure (by browsing concepts and inspecting listed original entries), (c) the semantic concepts a noun is part of (by browsing or searching nouns), or (d) the semantic concepts presented in a specific collection (by browsing collections).

Our meta-index presents a first step towards an ontology of semantic concepts expressed by idiomatic multi-word units. Usual ontologies express semantic relations based on literal meaning of single words, restricting the use of these ontologies to be applied on single words or Fregian multi-word units.

Notes

[1] We thank Marcel Dräger, Britta Juska-Bacher, Sixta Quassdorf, Noemi von der Crone, and David Schreiber for collaboration on concepts as well as for the thorough manual annotation and correction of the various extracts described in this paper.

[2] http://kollokationen.bbaw.de/htm/idb_de.html

[3] http://www.redensartenindex.de

[4] We always give the literal translation first and then the equivalent English phraseme or a paraphrase for the meaning in parentheses.

References

A. Dictionaries

Adelung, J. C. 1793–1801. *Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart*. Leipzig: Breitkopf & Sohn. (Adelung)

Borchardt, W. 1888. *Die Sprichwörtlichen Redensarten im deutschen Volksmund nach Sinn und Ursprung erläutert*. Leipzig: Brockhaus.

Campe, J. H. 1807–1812. *Wörterbuch der deutschen Sprache*. Braunschweig: Schulbuchverlag.

Dudenredaktion 2008. *Redewendungen: Wörterbuch der deutschen Idiomatik*. Mannheim: Dudenverlag. (Duden11)

Friedrich, W. 1976. *Moderne deutsche Idiomatik. Systematisches Wörterbuch mit Definitionen und Beispielen*. München: Huber.

Grimm, J. and W. Grimm 1852–1971. *Das deutsche Wörterbuch*. Leipzig: Hirzel.

Küpper, H. 1997. *PONS Wörterbuch der Deutschen Umgangssprache*. Stuttgart: Klett Verlag.

Röhrich, L. 2002. *Das große Lexikon der sprichwörtlichen Redensarten*. Darmstadt: WBG.

Sanders, D. 1859–1865. *Wörterbuch der deutschen Sprache*. Leipzig: Wigand.

Wander, K. F. W. 1867–1880. *Deutsches Sprichwörter-Lexikon*. Leipzig: Brockhaus. (DSL)

B. Other Literature

Burger, H. 2010. *Phraseologie*. Berlin: Erich Schmidt.

Connor, U. and A. I. Moreno 2005. 'Tertium Comparationis: A vital Component in Contrastive Research Methodology.' In P. Bruthiaux, D. Atkinson, W. G. Eggington, W. Grabe, and V. Ramanathan (eds.), *Directions in Applied Linguistics: Essays in Honor of Robert B. Kaplan*. Bristol: Multilingual Matters.

Dipper, S. 2010. 'POS-Tagging of Historical Language Data: First Experiments.' In M. Pinkal, I. Rehbein, S. Schulte im Walde, and A. Storrer (eds.), *Semantic Approaches in Natural Language Processing: Proceedings of the Conference on Natural Language Processing 2010 (KONVENS)*, Saarbrücken: Universaar, 117–121.

Dräger, M. 2010. 'Phraseologische Nachschlagewerke im Fokus.' In J. Korhonen, W. Mieder, E. Piirainen, and R. Pinel (eds.), *Phraseologie global - areal - regional*. Akten der Konferenz EUROPHRAS 2008 vom 13.-16.8.2008 in Helsinki, Tuebingen: Narr, 411–421.

Geyken, A. and J. Boyd-Graber 2004. 'Automatic Classification of Multi-Word Expressions in Print Dictionaries.' *Lingvisticae Investigationes*, 26.2: 187–202.

Juska-Bacher, B. and Mahlow, C. (in print). ‘Phraseological Change—a Book with Seven Seals? Tracing Diachronic Development of German Proverbs and Idioms.’ In P. Bennett, M. Durrell, S. Scheible and R. J. Whitt (eds.), *New Methods in Historical Corpus Linguistics*. Tübingen: Narr (= Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP, Vol. 3).

Mahlow, C. and B. Juska-Bacher 2011. ‘Exploring New High German Texts for Evidence of Phrasemes.’ *Journal for Language Technology and Computational Linguistics (JLCL)* 26.2.: 115–126.

Scheible, S., R. J. Whitt, M. Durrell, and P. Bennett 2011. ‘Evaluating an ‘off-the-shelf’ POS-Tagger on Early Modern German Text.’ In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR: Association for Computational Linguistics, 19–23.

Stantcheva, D. 2003. *Phraseologismen in deutschen Wörterbüchern: Ein Beitrag zur Geschichte der lexikographischen Behandlung von Phraseologismen im allgemeinen einsprachigen Wörterbuch von Adelung bis zur Gegenwart*. Hamburg: Dr. Kovač.