Characterization of German phrasal compounds based on empirical findings

Katrin Hein[1] and Adrien Barbaresi[2,3]

[1] Institute for the German Language (IDS), Department of Lexical Studies, R5, 6-13,
68161 Mannheim, Germany
hein@ids-mannheim.de
[2] Berlin-Brandenburg Academy of Sciences, Jägerstraße 22/23, 10117 Berlin, Germany
barbaresi@bbaw.de
[3] Austrian Academy of Sciences (Academy Corpora), Sonnenfelsgasse 19, 1010 Vienna
adrien.barbaresi@oeaw.ac.at

German is well-known for making extensive use of compounds. In theory „the combination of two or more lexemes […] in the formation of a new, complex word" is a productive process of German word formation (Olsen 2015: 364 f.). However, in practice, certain compounds are more frequently realized while others are nonexistent or only occasional. Based on empirical findings stemming from web corpora, the present study operates at the crossroads of qualitative and quantitative research. Our goal is to better apprehend a rarely seen phenomenon – more precisely phrasal compounds without hyphens between their component parts (e.g. *Mussichnichtmehrhabengedanken*,'I-must-not-have-it-anymore-thoughts') – by sifting through large amount of linguistic data, isolating relevant language samples and finally redefining our concept based on empirical evidence. To this end, we build a typology of real-world examples, first regarding their adequacy to the pre-defined morphosyntactic criteria and second concerning their functional characteristics. Finally, our study sheds not only light on the process of phrasal compounding, but also on the process of composition in general.

We discuss the characterization of phrasal compounds in German (PCs) (e.g. *»Man-muss-doch-über-alles-reden-können«-Credo*, 'one-should-be-able-to-talk-about-everything motto'), which can be defined as „complex words with phrases in modifier position" (Meibauer 2003, 153). The study of this specific type of prototypical determinative compounds like *Baumhaus* ('tree house') is worthwhile in theoretical terms alone (cf. Hein 2017). While PCs whose immediate constituents are separated by hyphens (e.g. *Second-Hand-Liebe*) have already been investigated in large corpus-based studies (cf. Hein 2015; Trips 2016), until now no systematic study has ever put into focus orthographic realizations like *Mussichnichtmehrhabengedanken.* We focus our study on this variant, PCs without hyphens between their component parts. Our main contributions are the investigation of this PC-type in itself and our use of web corpora (in particular more spontaneous or creative speech), as we think both are a desideratum for a deeper understanding of phrasal compounding.

In a first quantitatively oriented step, we identify PCs in large corpora. Our method is based on word segmentation and morphological analysis, it takes advantage of a data-driven learning process (cf. Barbaresi/Hein 2017). The automatic detection implies to have an operational notion of phrasal compounds as well as corpora which are large and diverse enough to contain rarely seen phenomena. Billion-token web corpora built for linguistic research comprehend a significant amount and variety of texts (cf. Barbaresi 2016), which provides a fruitful supplement to the newspaper-based investigations conducted so far (cf. Hein 2015; Trips 2016). We use metadata such as source and date in order to be able to contextualize our study.

In a second step following the manual screening of results, we characterize our findings in a linguistic perspective: What is the proportion of PCs written together in comparison to PCs with hyphens? Is there a (systematic) difference between both PC-types? Our findings indicate that the orthographic variants are not explainable in a systematic way (e.g. via formal properties). Looking at the PCs which we have automatically extracted from web corpora, one could conclude that the lack of punctuation seems to increase the degree of expressivity of the complex words (e.g. *AfterworkichraffmichgradsonochaufFeierabendSportler).* In comparison to the samples from newspapers, the web corpus inventory displays a comparatively high proportion of formal idiosyncrasies with respect to criteria linked to the PC-status (cf. Hein 2015). For example, the syntactic status of the non-head in *Schlechtwetterabernichtzuwarmschuhe* is hard to define. Moreover, numerous retrieved tokens are borderline cases between phrasal compounding and other types of phrasal word formation like phrasal derivation (cf. Lawrenz 2006), e.g. *Oberflächlichallesmögliche-abernichtsrichtiglerner*, which challenges the theoretical definition of PCs and the systematic criteria used to detect them. But the comparison between PCs from newspapers and PCs from web corpora indicates that there are characteristics common to both types: certain lexemes seem to be more suitable as head words than others, most notably denominations of a person (e.g. 'woman') being modified by a non-head expressing a stereotypical property (e.g. *Fürallesoffendochdermännlichkeitverfallenfreifrau*) (cf. Steyer/Hein 2018).

References

Barbaresi, Adrien (2016). Efficient construction of metadata-enhanced web corpora. In: Paul Cook, Stefan Evert, Roland Schäfer and Egon Stemle (eds.), Proceedings of the 10thWeb as CorpusWorkshop, 7–16.Association for Computational Linguistics.

Barbaresi, Adrien & Katrin Hein (2017). Data-Driven Identification of German Phrasal Compounds. In: Kamil Ekštein and Václav Matoušek (eds.), Text, Speech, and Dialogue. 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings. Cham: Springer, 192-200.

Hein, Katrin (2015). Phrasenkomposita im Deutschen. Empirische Untersuchung und konstruktionsgrammatische Modellierung. Tübingen: Narr.

Hein, Katrin (2017). Modeling the properties of German phrasal compounds within a usage-based constructional approach. In: Carola Trips and Jaklin Kornfilt (eds.), Further investigations into the nature of phrasal compounding. Berlin: Language Science Press, 119-148.

Lawrenz, Birgit (2006). Moderne deutsche Wortbildung. Phrasale Wortbildung im Deutschen. Linguistische Untersuchung und sprachdidaktische Behandlung. Hamburg: Dr. Kovač.

Meibauer, Jörg (2003). Phrasenkomposita zwischen Wortsyntax und Lexikon. In: Zeitschrift für Sprachwissenschaft 22, 153–88.

Olsen, Susan (2015). Composition. In: Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen und Franz Rainer (eds.), Word-formation. An International Handbook of the Languages of Europe. Volume 1, II: Units and processes in word-formation I: General aspects. Berlin/Boston: De Gruyter Mouton, 364–386.

Steyer, Kathrin/Hein, Katrin (2018). Satzwertige usuelle Wortverbindungen und gebrauchsbasierte Muster. In: Stefan Engelberg, Henning Lobin, Kathrin Steyer and Sascha Wolfer (eds.), Wortschätze: Dynamik, Muster, Komplexität. Jahrbuch des Instituts für Deutsche Sprache 2017. Berlin/New York: de Gruyter.

Trips, Carola (2016). An analysis of phrasal compounds in the model of Parallel Architecture. In: Pius ten Hacken (ed.), The semantics of compounding. Cambridge: Cambridge University Press, 153–177.