Data-driven Identification of German Phrasal Compounds

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

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

Abstract. We present a method to identify and document a phenomenon on which there is very little empirical data: German phrasal compounds occurring in the form of as a single token (without punctuation between their components). Relying on linguistic criteria, our approach implies to have an operational notion of compounds which can be systematically applied as well as (web) corpora which are large and diverse enough to contain rarely seen phenomena. The method is based on word segmentation and morphological analysis, it takes advantage of a data-driven learning process. Our results show that coarse-grained identification of phrasal compounds is best performed with empirical data, whereas fine-grained detection could be improved with a combination of rule-based and frequency-based word lists. Along with the characteristics of web texts, the orthographic realizations seem to be linked to the degree of expressivity.

Keywords: corpus linguistics, word segmentation, morphological analysis, web corpora

1 Introduction

Composition, that is "the combination of two or more lexemes (roots, stems or freely occuring words) in the formation of a new, complex word", is a productive process of German word formation [27, p. 364]. German is indeed considered as a language which makes extensive use of compounds [31], for example *Biowahn*, *Freigeist*, or *zitronengelb*. Compounding does not always operate by a simple string concatenation, it can involve linking elements (e.g. the *ens* in *Schmerzensschrei*) as well as the elision of word-final characters in the non-head constituent of a compound [18]. In the non-head position of such determinative compounds, not only lexical categories, but also syntactic units can be inserted – a phenomenon called phrasal compounding.

This paper discusses the automatic detection of german phrasal compounds (PCs) like *"Man-muss-doch-über-alles-reden-können"-Credo* or *Habdichliebpolitik*[4] in web

[4] "One-should-be-able-to-talk-about-everything motto", "I-like-you-policy". All examples appear in their original graphic realization, as found in previous studies [19] and billion-token web corpora [6,5]

corpora. PCs display a specific type of determinative compounds and can be defined as "complex words with phrases in modifier position" [24, p. 153]; their study is worthwhile in theoretical terms alone and sheds light on the general process of composition [17,16,22,25,24,34]. Phrasal compounding is not restricted to German, but can be found in other languages as well [35], for example English: "cut-and-run meal" [34].

While [16] (cf. also [29]) presented the first elaborated, large corpus-based investigation of German PCs whose immediate constituents are separated by hyphens (e.g. *Second-Hand-Liebe*, cf. [15, p. 349-353] for orthographic variants), until now no systematic study has ever put into focus PCs which are written as one word, i.e. without hyphens or blanks between their component parts (e.g. *Heileweltsache*). We want to outline the methodological challenges of their automatic detection in this paper, however the investigation of this PC variant in itself can also be seen as a desideratum.

The absence of linguistical or computational approximations to PCs can notably be explained by the lack of attested data. Within the inventory of nominal compounds, PCs account for an amount of 3.2% [28]. As it is assumed that PCs written as one word are even less prominent [16], possible hits are to be found at the lower end of the frequency spectrum. Thus, the annotated samples which we put together and use can themselves be considered as a precious and unprecedented source of linguistic evidence.

The automatic detection is indeed particularly challenging, especially the distinction between complex prototypical determinative compounds and PCs which are written in one word, or the distinction between PCs and types of phrasal derivations like *Wasser-in-Wein-Verwandler* [22]. It implies to have an operational notion of compounds which can be systematically applied as well as corpora which are large and diverse enough to contain rarely seen phenomena. Web corpora built for linguistic research relying on scientific criteria [2] seem particularly suitable for this endeavor, as they may comprehend a significant amount of texts from a large gamut of sources.

2 Method

The detection method grounds on a morphological compound analysis operating on token level. Since German is considered to be a morphologically rich language, state-of-the-art approaches do not always perform well on words absent from the dictionary, which is typically the case for phrasal compounds. Thus, in order to get information on potential segmentation patterns, we use the morphological analyzer SMOR [32] in combination with a data-driven morphological segmentation based on affix and component trees. This combination follows from two different goals: to design a robust detection process and to be able to estimate the degree of lexicalization of complex compounds.

2.1 Previous work

The combination is deemed to be necessary as SMOR can be subject to coverage issues. In previous work on non-standard text in an under-resourced variety of written German, namely retro-digitized newspaper texts from former East Germany, we showed that a data-driven method could overcome data sparsity and trump SMOR's full-fledged

morphological analysis to predict whether a given token is to be considered as part of the language or as an OCR error [4].

A similar approach has also been used to build an unsupervised morphological model for a number of different languages and language varieties for the *Discriminating between Similar Languages* shared task [23,3]. Criteria resulting from the segmentation analysis are statistically relevant and can be used as a sparse feature in a model to discriminate similar languages. A reasonable hypothesis is that they add new linguistically motivated information, dealing with the morpho-lexical logic of the languages to be classified while also yielding insights on linguistic typology.

2.2 Data-driven segmentation process

The method is based on segmentation and affix analysis. The original idea behind this simple yet efficient principle appears to go back to Harris' letter successor variety which grounds on transitional probabilities to detect morpheme boundaries [14]. The principle has proven valuable to construct stem dictionaries for document classification [13], it has been used in the past by spell-checkers [30,20], as it is both linguistically relevant and computationally efficient. Relevant information is stored in a trie [11], a data structure allowing for prefix search and its reverse opposite in order to look for sublexicons, which greatly extends lexical coverage. Forward (prefix) and backward (suffix) tries are used in a similar fashion, albeit with different constraints. This approach does not necessarily perform evenly across languages; it has for example led to considerable progress in morphologically-rich languages such as Arabic [7] or Basque [1]. Similar approaches have been used successfully to segment words into morphemes in an unsupervised way and for several languages. A more recent implementation has been the *RePortS* algorithm which gained attention in the context of the *PASCAL challenge* [21,8,9] by outperforming most of the other systems. The present approach makes similar assumptions as the work cited and adapts the base algorithm to the task at hand. In this regard, this experiment also tests if the data-driven morphological analysis of surface forms can be useful in the context of phrasal compounds.

2.3 Implementation

In order to build the corresponding model, a dictionary is composed by observing unigrams in the training data, then prefix and suffix trees are constructed using this dictionary. Additionally, an affix candidate list is constituted by decomposing the tokens present in the training data. The identification algorithm aims at the decomposition into possibly known parts. It consists of two main phases: first a prefix/suffix search over respective trees in order to look for the longest possible known subwords, and secondly sanity checks including a series of known composition rules to see if the rest could itself be an affix or a word out of the dictionary. If $\alpha\beta$ is a concatenation absent from the dictionary and if α and β are both in training data, then $\alpha\beta$ is considered to be a valid token. The segmentation can be repeated if necessary, in order to identify all necessary components of long words. It is only performed backward here since the nominal basis is a discriminative criterion in this study.

Once the word has been decomposed into potentially meaningful parts, the morphological analysis tool SMOR [32] is used to determine the grammatical category of the identified root, i.e. the last matched subword on the right. If it is considered to be a valid noun, the rest of the subwords is analyzed in the same way. Adjectival and adverbial combinations on a noun base are used as a cue that the token is a phrasal compound.

For example, the token *Allerweltsfragen* is not necessarily in the dictionary, but it can be decomposed into *Allerwelt+s+fragen* and ultimately into *aller+welt+s+fragen*. *Fragen* and *Welt* are identified as in-dictionary nouns, *aller* is a valid component, and *s* is among a fixed number of composition rules. Thus, this token is classified as PC.

3 Evaluation

The evaluation is performed on a gold standard of manually annotated samples: lists of PCs (coarse-grained and fine-grained) and a list of other similar compounds (noise). There are 123 *coarse*, 103 *fine*, and 504 *noise* tokens. The samples mainly come from experiments with billion-token web corpora [6,5], completed by results from previous studies [19]. They are annotated manually following expert criteria defined in [16].[5]

Apart from morphological criteria such as binarity or constraints for the realization of linking elements, the question whether the non-head can be considered as a phrasal element is crucial for the identification as PC. We assume a gradual understanding of the category "phrase" in this context, with congruency between the elements of the non-head being an important criterion, cf. *Harte-Jungs-Gerede* or *1000-Stunden-Jahr*, but not *Dreibettzimmerzuschlag*. In addition to these classical cases, entities whose status as a phrase is less clear are also considered here, e.g. *Coca-Cola-trink-Unterhaltungs-Freundschaft* (contains only a verb stem as verbal element). Both syntactically complete structures (e.g. *"Der-Reporter-macht-sich-langsam-auf-den-Weg-in-die-Redaktion"-Stunde*) as well as sentence-like elliptical structures (e.g. *"Jetzt geht's los"-Motto*) in non-head position are considered to be within the category "sentence".

Entities which do not correspond to our criteria are gathered in the *noise* list, whose purpose resides in emulating larger datasets by entailing long tokens such as proper names, complex compounds, and compounds which share a formal similarity with phrasal compounds, notably complex nominal compounds (*Waschsalontristesse*).

In order to do justice to the numerous entities which fulfill certain, but not all of the criteria linked to the PC-status, we make use of a *coarse* inter-category. Constructs which have something to do with PCs from a coarse-grained perspective, particularly constructs with a phrasal component, are collected in this list: *Immernacktschlafende*, product of phrasal derivation [22]; *Grünkohlinderbadewannewaschens*, product of phrasal conversion [22]; *Afterwork[ichraffmichgradsonochauf]FeierabendSportler*, phrasal element in the middle of a complex construction [10]; *Einpersonenhaushalt*, lack of congruency within the non-head, potential phrase because of the A+N-non-head; *Mehr-Aufmerksamkeitsheischerei*, realization of a non paradigmatic linking element in combination with a non-lexicalized non-head [16]. Because they share certain properties

[5] One PC-type defined is not captured by our automatic detection: Word formations whose non-head consists of not explicitly coordinated NPs, e.g. *Frage-Antwort-Stunde*, cf. p. 194 f.

with fine-grained PCs, the entities from the coarse list can be useful for the automatic detection of PCs that are fully compatible with the theoretical model.

After empirical testing, the smallest possible token length for learning and searching is fixed to 4 characters, the upper bound on token and subword length during learning phase is 16 characters. We test several lists which are expected to contain valuable information on variation at morphological level. On one hand, morphologically motivated word lists which have been made available for training and/or experiments on German words by the MarMoT [26] and GermaNet [18] systems are tested in this particular context. On the other hand, an empirical dataset is used for comparison, it stems from a combination of common tokens from a german reference corpus [12] and from newspaper corpora [2] (as described in [4]).

Measure	**Coarse+Fine**			**Coarse**			**Fine**		
	MMT	GNT	CPS	MMT	GNT	CPS	MMT	GNT	CPS
Precision	.452	.405	.546	.250	.198	.368	.330	.310	.383
Recall	.527	.199	.394	.390	.138	.350	.689	.301	.447
F1	**.487**	.267	.458	.305	.163	**.358**	**.447**	.305	.413
Accuracy	.656	.662	**.711**	.651	.721	**.754**	.710	.768	**.784**

Table 1. Results of evaluation on manually annotated samples (coarse and fine-grained). MMT = MarMoT, GNT = Germanet, CPS = corpus data. Higher is better.

Table 1 summarizes the results. The training data from the MarMoT morphological toolkit lead to the best general results (*coarse+fine*) as well as the best compromise between precision and recall (F1-measure), but the empirical data from a selection of corpora reaches the best accuracies in all three settings. Therefore, it appears that coarse-grained detection of phrasal compounds is best performed with empirical learning data, whereas fine-grained detection can benefit from a combination of rule-based and frequency-based word lists.

4 Discussion

More seldom seen combinations can be problematic and lead to a decrease in recall, mostly because of segmentation and component identification issues. First, the token *Halsringreitjungfrau* is correctly segmented into *Halsring+reit+jungfrau*, but since neither the "reit" modifier (corresponding to the verb *reiten*) nor the potential noun "Reitjungfrau" are present in the training data, the case is left undecided at the current stage of implementation. The greediness of the search algorithm could also be fine-tuned: the (non-PC) token *Ichhaballesimgrifflern* is decomposed into *ichhaballesim* – an unknown string resulting from the algorithm not being greedy enough, whereas the decomposition into *lern* and *griff* is too greedy and misses the nominal formation in dative form.

Second, the dictionary coverage obtained from reference and newspaper corpora affects precision. *Grünpanzerschildkröte* (a rare species of tortoise) is wrongly considered to be a PC (lack of congruency) since the token is decomposed correctly, *grün* is an

adjective, and *Grünpanzer* does not appear to be lexicalized. Component parts coming from other languages are also problematic, such as in *Mainstream+medien+superfrau*, where the lexical class of *Mainstream* is hard to assess automatically due to sparse data. Additionally, a systematic notion of congruence as well as a refined analysis of combining elements seem to be necessary to improve the detection process: so are *Grün+gemüse+n+spendeaktion* and *Nicht+eintreten+s+entscheid* both analyzed as PCs, although this is no clear-cut case and only the first one has been annotated as a potential/coarse-grained one.

Finally, the model does not presently yield information about frequency effects. As it is restricted to concatenative morphology, the fact that a stem has to be in the dictionary is a strong limitation impeding recall in particular [9]. However, an overall conservative setting has been kept so far as it prevents the model from overgenerating.

5 Conclusions

We have presented a study to identify and document a rare phenomenon on which there is very little empirical evidence, phrasal compounds occurring in the form of as a single token without punctuation between their component parts. Our method implies a systematic approach as well as corpora which are large and diverse enough. It operates at the crossroads of qualitative and quantitative research, in such a way that both approaches benefit from each other. On one hand, we need empirical data to draw conclusions on this rarely observed phenomenon. On the other hand, trying to replicate fine-grained decisions also makes for more stringent and thorough criteria. Our method is based on word segmentation and morphological analysis, the first takes advantage of a data-driven learning process while the latter uses existing software. As documented examples are quite scarce, machine-based scans through large web corpora are one possible way to look at these compounds in all their (so far unsuspected) variety.

Since one specific communicative function of – at least one sort of – PCs can be seen in producing expressivity effects [16,25], this word formation type seems to be predestined for a productive use in computer-based communication, which web corpora entail. Our results show that coarse-grained detection of phrasal compounds is best performed with empirical data, whereas fine-grained detection could be improved with a combination of rule-based and frequency-based word lists. Additionally, the investigation of PCs in web corpora displays a fruitful supplement to the newspaper-based investigations. If we compare results from both sources, there seem to be parallels inasmuch as certain lexemes are more predestined than others to appear as a head word in PCs. For example, there are many PCs whose head word is a denomination of a person (e.g. *Frau*) in the present findings. Such heads are often combined with non heads which express a stereotypical property of a person or a type of person [33], such as *Wäschewaschaufhängbüglezusammenlegfrau* or *buchcoverfotosmitsmartphonecamerabildermachfrau*.

Looking at the PCs which we have automatically extracted from web corpora, one can conclude that their orthographic realization, i.e. the missing use of punctuation between the component parts, seems to increase their degree of expressivity. This holds especially for very complex/long PCs like *Afterworkichraffmichgradsonochaufeier-*

abendSportler. Catching attention or being creative seems to be more important in this case than facilitating the reception for the reader by the use of punctuation. Moreover, creatively formed PCs such as *Oberflächlichallesmöglicheabernichtsrichtiglerner* contribute to reject the claim [22] that predominantly lexicalized PCs (e.g. *Armeleuteessen*) are written as one word. Future work includes giving answers to linguistically relevant open questions with respect to the proportion of PCs written together in comparison to PCs with hyphens, the potential existence of a systematic difference between both PC types, and further properties of compounds which are suitable both for qualitative analysis and automatic identification.

References

1. Agirre, E., Alegria, I., Arregi, X., Artola, X., de Ilarraza, A.D., Maritxalar, M., Sarasola, K., Urkia, M.: XUXEN: A spelling checker/corrector for Basque based on two-level morphology. In: Proceedings of the 3rd conference on Applied Natural Language Processing. pp. 119–125. Association for Computational Linguistics (1992)
2. Barbaresi, A.: Ad hoc and general-purpose corpus construction from web sources. Ph.D. thesis, École Normale Supérieure de Lyon, France (2015)
3. Barbaresi, A.: An Unsupervised Morphological Criterion for Discriminating Similar Languages. In: Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J. (eds.) Proceedings of the 3rd VarDial Workshop. pp. 212–220 (2016)
4. Barbaresi, A.: Bootstrapped OCR error detection for a less-resourced language variant. In: Dipper, S., Neubarth, F., Zinsmeister, H. (eds.) Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). pp. 21–26. University of Bochum (2016)
5. Barbaresi, A.: Efficient construction of metadata-enhanced web corpora. In: Cook, P., Evert, S., Schäfer, R., Stemle, E. (eds.) Proceedings of the 10th Web as Corpus Workshop. pp. 7–16. Association for Computational Linguistics (2016)
6. Barbaresi, A., Würzner, K.M.: For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In: Beißwenger, M., Zesch, T. (eds.) KONVENS 2014, NLP4CMC workshop proceedings. pp. 2–10. Hildesheim University Press (2014)
7. Ben Hamadou, A.: A compression technique for Arabic dictionaries: the affix analysis. In: Proceedings of the 11th Conference on Computational Linguistics. pp. 286–288. Association for Computational Linguistics (1986)
8. Dasgupta, S., Ng, V.: High-performance, language-independent morphological segmentation. In: HLT-NAACL. pp. 155–163 (2007)
9. Demberg, V.: A language-independent Unsupervised Model for Morphological Segmentation. In: Annual Meeting of the Association for Computational Linguistics. vol. 45, pp. 920–927 (2007)
10. Finkbeiner, R., Meibauer, J.: Boris "Ich bin drin" Becker ("Boris I am in Becker"). Syntax, semantics and pragmatics of a special naming construction. Lingua 181, 36–57 (2016)
11. Fredkin, E.: Trie Memory. Communications of the ACM 3(9), 490–499 (1960)
12. Geyken, A.: The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, C. (ed.) Collocations and Idioms: Linguistic, lexicographic, and computational aspects, pp. 23–41. Continuum Press (2007)
13. Hafer, M.A., Weiss, S.F.: Word Segmentation by Letter Successor Varieties. Information Storage and Retrieval 10, 371–385 (1974)
14. Harris, Z.S.: From Phoneme to Morphemes. Language 31(2), 190–222 (1955)
15. Hein, K.: Phrasenkomposita – ein wortbildungsfremdes Randphänomen zwischen Morphologie und Syntax? Deutsche Sprache 39, 331–361 (2011)

16. Hein, K.: Phrasenkomposita im Deutschen. Empirische Untersuchung und konstruktionsgrammatische Modellierung. Narr (2015)
17. Hein, K.: Modeling the properties of German phrasal compounds within a usage-based constructional approach. In: Trips, C., Kornflit, J. (eds.) Further investigations into the nature of phrasal compounding. Language Science Press, Berlin (2017), to appear
18. Henrich, V., Hinrichs, E.W.: Determining Immediate Constituents of Compounds in GermaNet. In: Proceedings of Recent Advances in Natural Language Processing. pp. 420–426 (2011)
19. IDS: Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2011-I. Tech. rep., Institut für Deutsche Sprache Mannheim (2011), www.ids-mannheim.de/dereko
20. Jones, M.A., Silverman, A.: A spelling checker based on affix classes. In: Agrawal, J.C., Zunde, P. (eds.) Empirical Foundations of Information and Software Science, pp. 373–379. Springer US, Boston, MA (1985)
21. Keshava, S., Pitler, E.: A simpler, intuitive approach to morpheme induction. In: Proceedings of 2nd Pascal Challenges Workshop. pp. 31–35 (2006)
22. Lawrenz, B.: Moderne deutsche Wortbildung. Phrasale Wortbildung im Deutschen: Linguistische Untersuchung und sprachdidaktische Behandlung. Dr. Kovač (2006)
23. Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J.: Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In: Proceedings of the 3rd VarDial Workshop (2016)
24. Meibauer, J.: Phrasenkomposita zwischen Wortsyntax und Lexikon. Zeitschrift für Sprachwissenschaft 22, 153–188 (2003)
25. Meibauer, J.: How marginal are phrasal compounds? Generalized insertion, expressivity, and I/Q-interaction. Morphology 17, 233–259 (2007)
26. Müller, T.: General methods for fine-grained morphological and syntactic disambiguation. Ph.D. thesis, LMU Munich (2015)
27. Olsen, S.: Composition. In: Müller, P.O., Ohnheiser, I., Olsen, S., Rainer, F. (eds.) Word-formation. An International Handbook of the Languages of Europe. Volume 1, II: Units and processes in word-formation I: General aspects, pp. 364–386. De Gruyter Mouton, Berlin/Boston (2015)
28. Ortner, L., Müller-Bollhagen, E.: Substantivkomposita. Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache, Schwann (1991)
29. Particke, H.J.: Phrasenkomposita: eine morphosyntaktische Beschreibung und Korpusstudie am Beispiel des Deutschen. Diplomica-Verlag, Hamburg (2015)
30. Peterson, J.L.: Computer programs for detecting and correcting spelling errors. Communications of the ACM 23(12), 676–687 (1980)
31. Schlücker, B.: Die deutsche Kompositionsfreudigkeit. Übersicht und Einführung. In: Gaeta, L., Schlücker, B. (eds.) Deutsche als kompositionsfreudige Sprache. Strukturelle Eigenschaften und systembezogene Aspekte, pp. 1–25. de Gruyter (2012)
32. Schmid, H., Fitschen, A., Heid, U.: SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In: Proceedings of LREC. pp. 233–259 (2004)
33. Steyer, K., Hein, K.: Satzwertige usuelle Wortverbindungen und gebrauchsbasierte Muster. In: Engelberg, S., Lobin, H., Steyer, K., Wolfer, S. (eds.) Wortschätze: Dynamik, Muster, Komplexität, Jahrbuch des Instituts für Deutsche Sprache 2017. de Gruyter (2018), to appear
34. Trips, C.: The relevance of phrasal compounds for the architecture of grammar. In: ten Hacken, P. (ed.) The Semantics of Compounding, pp. 153–177. Oxford University Press (2016)
35. Trips, C., Kornfilt, J. (eds.): Phrasal compounds from a typological and theoretical perspective. Special issue of STUF. Language typology and universals (2015)