Schwa-deletion in German noun-noun compounds

Tom Juzek
Saarland University
Saarbrücken, Germany
tom.juzek@
posteo.net

Jana Häussler
Bielefeld University
Bielefeld, Germany
jana.haeussler@
uni-bielefeld.de

Abstract

We report ongoing research on linking elements in German compounds, with a focus on noun-noun compounds in which the first constituent is ending in schwa. We present a corpus of about 3000 nouns ending in schwa, annotated for various phonological and morpho-syntactic features, and critically, the dominant linking strategy. The corpus analysis is complemented by an unsuccessful attempt to train neural networks and by a pilot experiment asking native speakers to indicate their preferred linking strategy. In addition to existing nouns, the experimental stimuli included nonce words, also ending in schwa. While neither the corpus study nor the experiment offer a clear picture, the results nevertheless provide interesting insights into the intricacies of German compounding. Overall, we find a predominance of the paradigmatic linking element *-n* for feminine and masculine nouns. At the same time, the results for nonce words show that *-n* is not a default strategy.

1 Introduction

German compounds and especially noun-noun compounds often include a linking element (*LE*), i.e. segmental material between the two constituents of a compound, such as *-s* in *Liebesbrief* 'love letter' (*liebe-s-brief* [love-LE-letter]) or *-er* in *Kindergarten* 'nursery' (*kind-er-garten* [child-LE-garden]). Though linking elements are a common phenomenon in Germanic languages, German is special because of its rich inventory of linking elements: *-e*, *-en*, *-ens*, *-er*, *-es*, *-n*, *-ns*, *-s*, and *-i*/*-o*. Furthermore, deletion and substitution may occur. Whether the number of linking elements can be diminished by assuming variation similar to the allomorphic variation of the homophoneous inflectional affixes (e.g. *-(e)n*) is a disputed topic (cf. Neef, 2015 and Nübling & Szczepaniak, 2013). The rules governing linking element selection are anything but obvious, even native speakers of German are sometimes unsure of the "correct" choice. Linking elements also constitute a major challenge for natural language generation and machine translation (e.g. Matthews et al. 2016)..

Our study investigates a very specific linking strategy, which has received less attention so far: the deletion of a final schwa, in the literature sometimes referred to as *subtractive linking element*. An example for this strategy is *Endpunkt* ('endpoint') which combines *Ende* ('end') and *Punkt* ('point'). Though schwa-deletion in itself does not apply to too many words, it affects some high frequency nouns like *Sache* ('thing', 'matter') and *Farbe* ('colour', 'paint'). To explore the phenomenon systematically, we created a corpus of (almost) all simple nouns ending in schwa and asked two annotators to indicate the preferred linking strategy for each of those nouns. Furthermore, we conducted a forced choice experiment to gain further insights. We report both studies in turn.

2 Linking element selection in German

It is consensus that the choice of linking element is mainly determined by the left constituent in a compound. Evidence comes from tupels like *Tag-e-buch* 'diary', lit. 'day book', *Kind-er-buch* 'children's book', *Jugend-buch* 'book for adolescents', *Liebling-s-buch* 'favorite book', *Schul-buch* (*Schule+Buch*) 'school book' which share the second constituent but differ in the first constituent and the linking element. Further evidence comes from coordinated compounds, such as *Gesundheits- und Sozialwesen*

'health care and welfare system'. Expanding this conclusion, Fuhrhop (1996) proposed that the lexical representations of nouns includes specific stems for compounding. Notably, there a several cases that weaken this proposal as they exhibit variation within a single noun, e.g. *Tag-e-buch* 'diary', *Tagtraum* 'day dream', *Tag-es-satz* 'daily rate' or *Beere-n-schnaps* 'berry liquor' and *Blaubeer-schnaps* 'blue berry liquor'. Arguably, though, some of the variation can be explained with reference to diachrony, e.g. through fossilised forms. However, in the present paper, we focus on the role of the left constituent.

There is less consensus about the function of linking elements (for a critical overview see Neef 2015) and the conditions on LE selection for a given noun as the first constituent in a compound, but see Fuhrhop (1996) and Nübling and Szczepaniak (2013) for comprehensive overviews of morpho-phonological factors. For nouns ending in a schwa, the following descriptive generalisations have been hypothesised. Feminine nouns as well as masculine nouns with weak declension pattern almost obligatorily take *-n* as the linking element, cf. Libben et al. (2002), Köpcke (1993). Schwa-deletion occurs rarely, but for some nouns regularly, cf. Ortner et al. (1991). Schwa is never deleted when it constitutes a suffix, cf. Aronoff and Fuhrhop (2002).

Previous studies examined the distribution of linking elements across the board, i.e. for all kinds of nouns and all kinds of linking elements, by counting the occurrences in compounds in text corpora (e.g. Ortner et al. 1991) or lexical resources like CELEX (e.g. Krott et al. 2007). The present study in contrast focuses on a particular type of left constituent, namely nouns ending in schwa. In this sense the present study is more limited; at the same time, it is more comprehensive since the corpus we present below captures virtually all nouns of this specific type.

3 Corpus study

There is no resource one could use to look up compound strategies of German nouns. We therefore created a new corpus, focusing on items that could in principle make use of schwa-deletion. The entire corpus can be found at: `https://gitlab.com/superpumpie/schwa_deletion`.

3.1 Corpus creation

We web scraped all nouns ending in an <e> from the German Wiktionary (The Wikimedia Foundation, 2017b), using Beautiful Soup (Richardson, 2018). Using the information provided in the corresponding Wiktionary entry, we restricted the extraction to nouns in which the final <e> represents a schwa and which are not compounds themselves. We permitted derived nouns like *Tränke* ('drinking trough') because it has been claimed that schwa-deletion is permitted when schwa represents a suffix, cf. Aronoff and Fuhrhop (2002). We manually corrected the output of the extraction scripts, and we excluded proper names but kept demonyms. In a next step, we web scraped and extracted the following features: number of phonemes, CV structure, the phoneme preceding the schwa, grammatical gender, plural marker, as well as an entry's logged frequency in discussion threads of the German Wikipedia (The Wikimedia Foundation, 2017a), an entry's most common preceding word, and most common succeeding word. Further, a native speaker tagged whether an entry is or could be derived by means of schwa-suffixing.

3.2 Corpus annotation for linking strategies

Two annotators, native speakers of German and professional linguists, tagged their preferred linking strategy for each of the items as the first constituent in a noun-noun compound. Whenever the two annotators disagreed (prevalence: 26.6% of all items), a third linguist's judgements were used as a tiebreaker. If all three judgements diverged, we noted down a disagreement (prevalence: about 5%).

3.3 Corpus analysis

3.3.1 Probabilistic analysis

The corpus consists of 2994 critical items, 9 features as independent variables, and preferred linking strategy as our dependent variable. Table 1 gives the distribution of linking elements broken down by gender, excl. items for which the gender was not specified. Overall, we see a dominance of *n*-insertion as the linking strategy in compounds, which is most pronounced in masculine nouns. Since *-(e)n* is the

	schwa-deletion	null	*n*-insertion	other	disagreement
feminine (N=2437)	6.2% (152)	18.9% (460)	69.8% (1700)	0.01% (2)	5.0% (123)
masculine (N=425)	0.0% (0)	8.5% (36)	85.6% (364)	1.2% (5)	4.7% (20)
neuter (N=132)	11.4% (15)	60.6% (80)	17.4% (23)	1.5% (2)	9.1% (12)
all (N=2994)	5.6% (167)	19.2% (576)	69.7% (2087)	0.3% (9)	5.2% (155)

Table 1: Distribution of linking strategies for nouns ending in schwa as tagged by the annotators.

plural marker for feminine nouns and marks both case (incl. genitive) and plural in masculine nouns in the weak declension, which prototypically end in a schwa (Köpcke, 1995), the dominance of *n*-insertion can be interpreted as a preference for paradigmatic linking elements.

Neuter nouns in contrast rarely form the plural and never the genitive with *-n*. Notably, only five of the 23 neuter nouns for which our annotators marked *n*-insertion as the preferred linking strategy form the plural with *-n*. And although *n*-insertion is predominant in our corpus, it is by no means the only linking strategy for feminine nouns ending in schwa – nor for masculine and neuter nouns.

The second most frequent compounding strategy is concatenation without a linking element (labelled "null" in Table 1). Previous studies counting the frequency of linking elements for all types of nouns, i.e. not just ending in a schwa, report that the majority of compounds lack an overt linking element: up to 73% in Ortner et al. (1991), 65% in Krott et al. (2007). Finding only 19% in our sample underscores the assumption that linking elements are determined by the left constituent. For the few neuter nouns in our corpus, *null* is the preferred linking strategy. Finally, as expected, schwa-deletion was rare, occurring in less than 6% of all schwa-nouns. As before, there is a considerable gender effect.

A spot check of the corpus annotations seem to confirm the claim made in Aronoff and Fuhrhop (2002) that suffix-schwa is never deleted. For all of the 19 apparent counterexamples, it seems that corresponding compounds do involve the noun ending schwa but rather an alternative or older form without the schwa (e.g. *Geschrei(e)* 'yelling', *Piss(e)* 'piss') or the base form from which the noun is derived (e.g. the adjective *süß* 'sweet' versus *Süße* 'sweetness' in *Süßholz* or the verb stem *schimpf* 'rant' rather than the noun *Schimpfe* 'ranting' in *Schimpfkanonade* 'long rant').

3.3.2 Linear mixed effects models

To gain further insights, we analysed our corpus with several multi-factorial linear models, using R (R Core Team, 2018) and the *lme4* package (Bates et al., 2015), with the linking strategy as our dependent variable and the other factors listed above, i.e. logged frequency, etc., as predictors. We vary the predictors across models to be able to estimate their importance in explaining the observed variance. Crucially, there is not a single good predictor and a great deal of the variance remains unexplained: The residual SEs are around 0.22. The full output is too lengthy to be added here and a partial output would lack context, and is thus omitted. In case of interest, it can be accessed on our GitLab (see above).

3.3.3 Machine learning models

We have also tried to train various machine learning models, incl. MLPs, CNNs, and LSTMs, using various parameter settings. The difficulty is that we are facing a scarce data problem and that our attempts result in F1-scores below 0.2. Since the results are poor and not very informative, we omit them for the sake of brevity. However, in case of interest, they can be accessed at on our Gitlab. It is an open question whether the results are due to the nature of the phenomenon or due to limitations of our set of features.

4 Production experiment

The lack of effective predictors and the dominance of *-n* suggest that *n*-insertion could be a form of default strategy for nouns ending in schwa. Under this view, *-n* should also predominate in the absence of lexical information, and schwa-deletion would be an exception that is lexically encoded. If so, compounding of nonce words ending in schwa should apply *n*-insertion as the linking strategy.

To test this prediction, we conducted a forced choice experiment with nonce words. In contrast to Dressler et al. (2000), who used existing words as the first constituent and nonce words as the second

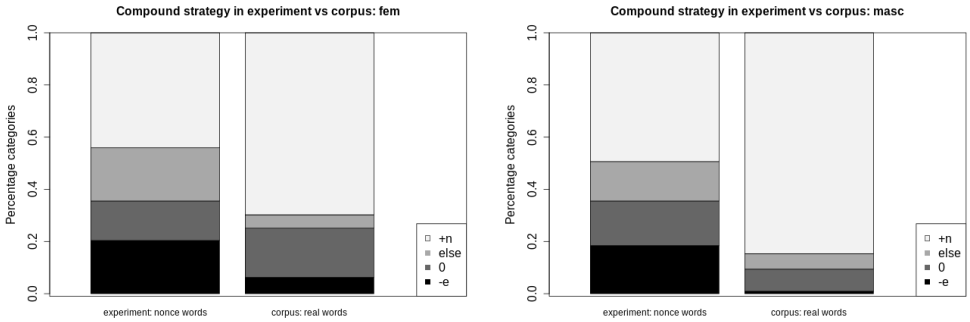

Figure 1: Results for nonce words in the experiment compared to the strategies observed in our corpus (left: feminine nouns, right: masculine nouns). For corpus data, "else" includes "disagreement".

constituent (e.g. *Suppe* 'soup' + *Fend*), we use nonce words as the first constituent and combine them with an existing word. Items were created with a script using Python (Python Software Foundation, 2018), all having the following syllable structure: one or more consonants, followed by a vowel or diphthong, followed by one or more consonants, plus a final schwa. Examples include *Trulve* and *Knüpse*. We manually checked the list for phonological and graphematic well-formedness and excluded items that were phonologically or orthographically too close to existing words. From the remaining set, we randomly sampled 16 items and varied their gender in a within-items design. We created two lists such that each list contained each item in one of its two versions (fem. or masc.) and an equal number of fem. and masc. critical items. In addition, each list contained 8 real nouns ending in schwa (4 fem., 4 masc., all the same in both lists), as well as 24 fillers, both nonce and real nouns not ending in a schwa.

Using Prolific (`https://www.prolific.co`), we recruited 24 native speakers of German. Participants were requested to choose a linking element for compounding our experimental items with *Beschreibung* ('description'). The words were presented with the corresponding article to indicate the gender (e.g. *der Knüpse* + *die Beschreibung*). Participants could choose between 7 response categories: *null* (concatenation without an LE), *+e*, *+er*, *+n*, *+s*, *schwa-deletion*, and "others". In total, we collected 304 data points for the critical nonce items. The data reveal a striking discrepancy between the distribution in the corpus of existing nouns ending in a schwa and the nonce words we tested in the experiment (Figure 1). *n*-insertion as a default strategy would have predicted that almost all nonce words select that strategy. However, this is not the case. Compared to the corpus data, the *-n* strategy is less prevalent in nonce compounding.

These surprising findings challenge the idea of *-n* as a default strategy for nouns ending in schwa. Assuming that linking strategies are encoded lexically, e.g. in form of specific compositional stems as part of the lexical representations of the nouns (cf. Fuhrhop 1996), could explain both the lack of a consistent default strategy observed with the nonce words and the failure of the LME model on the corpus data to explain a great deal of the variance.

5 Concluding remarks

While many linking elements in German are well-researched, the phenomenon of schwa-deletion is still an open question. The present paper explores the phenomenon in greater detail, by approaching it in various ways. However, the results of all our approaches paint a picture that is complex. A first analysis provides some probabilistic tendencies – pointing towards a predominance of paradigmatic linking elements. A linear mixed effects model could not identify a set of critical factors, though. The machine learning models that we trained also return poor results. And the results of the production experiment were also complex, hinting at the possibility that there is no default strategy. A plausible interpretation of all our approaches is that the choice of strategy is often encoded lexically. We hope that the results and the provided resources will be a starting point for further research and insights.[1]

[1]Both authors contributed equally. We thank the CogALex reviewers for their valuable feedback.

References

Mark Aronoff and Nanna Fuhrhop. 2002. Restricting suffix combinations in German and English: closing suffixes and the Monosuffix Constraint. *Natural Language & Linguistic Theory*, 20(3):451–490.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Wolfgang U. Dressler, Gray Libben, Jacqueline Stark, Christiane Pons, and Gonia Jarema. 2000. The processing of interfixed German compounds. *Yearbook of Morphology, 1999*, pages 185–220.

Nanna Fuhrhop. 1996. Fugenelemente. In Ewald Lang and & Gisela Zifonun, editors, *Deutsch – typologisch*, pages 525–549. De Gruyter, Berlin.

Klaus-Michael Köpcke. 1993. *Schemata bei der Pluralbildung im Deutschen. Versuch einer kognitiven Morphologie*. Niemeyer, Tübingen.

Klaus-Michael Köpcke. 1995. Die Klassifikation der schwachen Maskulina in der deutschen Gegenwartssprache. *Zeitschrift für Sprachwissenschaft*, 14(2):159–180.

Andrea Krott, Robert Schreuder, Harald R. Baayen, and Wolfgang U. Dressler. 2007. Analogical effects on linking elements in German compound words. *Language and Cognitive Processes*, 22:25–57.

Gary Libben, Gonia Jarema, Wolfgang Dressler, Jacqueline Stark, and Christiane Pons. 2002. Triangulating the effects of interfixation in the processing of German compounds. *Folia Linguistica*, 36:23–44.

Austin Matthews, Eva Schlinger, Alon Lavie, and Chris Dyer. 2016. Synthesizing compound words for machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational (Volume 1: Long Papers)*, pages 1085–1094, Berlin, Germany.

Martin Neef. 2015. The status of so-called linking elements in German: Arguments in favor of a non-functional analysis. *Word Structure*, 8:29–52, 04.

Damaris Nübling and Renata Szczepaniak. 2013. Linking elements in German: Origin, change, functionalization. *Morphology*, 23:67–89.

Lorelies Ortner, Elgin Müller-Bollhagen, Hanspeter Ortner, Hans Wellmann, Maria Pümpel-Mader, and Hildegard Gärtner. 1991. *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache, vol. 4*. de Gruyter, Berlin & New York.

Python Software Foundation, 2018. *Python: A dynamic, open source programming language*.

R Core Team, 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Leonard Richardson. 2018. Beautiful soup documentation. `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`.

The Wikimedia Foundation. 2017a. Wikipedia, the free encyclopedia. `https://www.wikipedia.de/`.

The Wikimedia Foundation. 2017b. Wiktionary, the free dictionary. `https://de.wiktionary.org/`.