Probability in the Grammar of German and Dutch:
Interfixation in Tri-Constituent Compounds

Andrea Krott[1], Gary Libben[2], Gonia Jarema[3],
Wolfgang Dressler[4], Robert Schreuder[5], and Harald Baayen[5]

1: University of Birmingham, U.K., 2: University of Edmonton, Canada,
3: University of Montreal, Canada, 4: University of Vienna, Austria
5: University of Nijmegen, The Netherlands

Address all correspondence to:
Andrea Krott PhD
School of Psychology
University of Birmingham
Edgbaston
Birmingham B15 2TT
United Kingdom
phone +44 (0)121 4144903
e-mail a.krott@bham.ac.uk

Abstract

This study addresses the possibility that interfixes in multi-constituent nominal compounds in German and Dutch are functional as markers of immediate constituent structure. We report a lexical statistical survey of interfixation in the lexicons of German and Dutch which shows that all interfixes of German, and one interfix of Dutch, are significantly more likely to appear at the major constituent boundary than expected under chance conditions. A series of experiments provides evidence that speakers of German and Dutch are sensitive to the probabilistic cues to constituent structure provided by the interfixes. Thus, our data provide evidence that probability is part and parcel of grammatical competence.

Introduction

Multi-constituent compounds such as German *Schachweltmeistertitel* ('chess world champion title', i.e., world chess championship) pose a special challenge to theories of lexical access. While a clause such as *the title of being the world's champion in chess* provides a great many syntactic cues guiding interpretation, no such cues are available for the interpretation of *Schachweltmeistertitel*. Nevertheless, speakers of German understand this compound just as well as its phrasal paraphrase, assigning it a dependency structure such as [[[chess [world champion]]title]. Note that this compound as a whole has a left-branching structure, but that its left constituent itself has a right-branching structure. The issue addressed in this study is which principles might guide readers and listeners when interpreting German and Dutch multi-constituent compounds.

Our leading hypothesis is that comprehension is guided by the statistical distributional properties of the existing compounds in the mental lexicon, even when the differences in the distributions are subtle and subject to substantial variation. One such distributional property might be the distribution of left-branching and right-branching compounds. If the majority of compounds in the mental lexicon is left-branching, this might bias comprehension towards expecting existing and novel multi-constituent compounds to be left-branching as well.

Another type of distributional information that might be crucial for comprehension is the availability of smaller compounds in the mental lexicon that might serve as larger building

blocks. Instead of trying to construct an interpretation for *Fussballweltmeistertitel* from the individual simplex nouns *Fuss, Ball, Welt, Meister*, and *Titel*, one might in fact need only construct an interpretation for the sequence of *Fussball*, *Weltmeister*, and *Titel*. The greater the frequency of an embedded compound, the more likely it might be to serve as a pre-existing unit for interpretation.

The distributional information addressed in the present study is that provided by the interfixes that are found in roughly a third of German and Dutch compounds. Interfixes, also referred to in the literature as linking elements or linking morphemes, can be traced to older case and number endings. In Dutch, they are, with a few exceptions, all that is left of the rich inflectional system that characterized medieval Dutch. In German, the interfixes are homophonous with present-day case and number endings, but their distributional properties, as will described in more detail below, differ from those of genuine inflectional formatives.

scheep-s-bouw+maatschappij	s left	left-branching
ship building company		
arbeid-s-vraagstuk	s left	right-branching
employment question-issue		
krijg-s-man-s-eer	s left and right	left-branching
war man honor		
rijk-s-arbeid-s-bureau	s left and right	right-branching
national employment office		
grondwet-s-artikel	s right	left-branching
ground law article, constitution		
hoofd+verkeer-s-weg	s right	right-branching
main traffic road		

Table 1: Examples of left-branching and right-branching tri-constituent compounds in Dutch with the interfix *-s-*.

Table 1 provides some examples of Dutch tri-constituent compounds with the interfix *-s-*. Note that this interfix may occur between the first and the second constituents, between the second and the third constituents, or even at both these positions. In addition, a compound

can be right-branching or left-branching. If the distribution of the *-s-* were as uniform as Table 1 would have us believe, then the *-s-* would not be of any use for constructing the interpretation of multi-constituent compounds. However, Grimm (1878) and in his wake, Žepić (1970), Fleischer (1976), and Fuhrhop (1998; 2000), among others, have pointed out that this is not the case. In many left-branching compounds, the interfix occurs in the second interfixation slot between the two immediate constituents of the compound. These scholars suggest that it is the function of the *-s-* to separate the compound into its immediate constituents. We will refer to this possibility as the separation hypothesis.

The present study first presents a lexical statistical survey of the distribution of the interfixes of Dutch and German as a function of their position in left and right branching compounds. The goal of this survey is to establish to what extent the *-s-* interfix and the other interfixes of German and Dutch indeed reveal a preference for the main constituent boundary. We will refer to this aspect of the separation hypothesis as the distributional separation hypothesis. In addition, this study presents a series of experiments that seek to clarify whether speakers of German and Dutch indeed make use of interfixes to determine the main constituent boundary in compounds. We will refer to this aspect of the separation hypothesis as the functional separation hypothesis. Before addressing the distributional and functional separation hypotheses, however, we first provide additional details of the distributional properties of the interfixes in German and Dutch.

Interfixes in German and Dutch

The German interfixes *-s-, -(e)n-, -es-, -e-, -er-, -ens-* and the Dutch interfixes *-s-*, *-e(n)-*, *-er-* are found in roughly a third of nominal compounds. Historically, interfixes are case and number endings of left constituents. In Dutch, the *-e-* may also have formed part of the stem before schwa-apocope applied (Booij, 2002). In some compounds, one can still interpret the interfix as a plural or genitive marker (e.g., German *Kapitän+s+kajüte* 'captain+GENITIVE+cabin'; Dutch *abrikoz+en+taart* 'apricot+PLURAL+tart'). However, although homophonous with modern case and number endings (in German) and with plural markers (in Dutch), the interfixes do not have the same distributional properties. For instance, the German noun *Liebe* 'love' cannot be combined with the case marker *-s*,

nor with the plural marker *-s-*, but it does occur with the interfix *-s-* in *Liebe+s+brief* 'love+INTERFIX+letter'. The same is true for the Dutch noun *schaap* 'sheep' that occurs with *-s-* in *schaap+s+leder* 'sheep+INTERFIX+leather' but that occurs with *-en-* in the plural *schaap-en*. This is not to say that interfixes are by definition meaningless. To the contrary, Schreuder, Neijt, Van der Weide, & Baayen (1998) have shown that in Dutch the interfix *-en-* induces a plural interpretation for the left constituent.

An indication that the interfixes are not mere relics of the past, as in the case of the English interfix *-s-* in *helmsman*, is that interfixes are used productively in novel compounds. In fact, the distribution of interfixes in novel and existing compounds has been one of the enigmas of German and Dutch morphology, as there are no hard and fast syntagmatic rules governing their use, especially when the left constituent is a monomorphemic noun. Dressler, Libben, Stark, Pons, and Jarema (2001) have argued for German that the choice of the interfix is sensitive to inflectional microclasses. Krott, Baayen, and Schreuder (2001) have argued for Dutch, and Krott, Schreuder, Baayen, and Dressler (in Krott, 2001, see also Libben, Jarema, Dressler, Stark, and Pons, 2002) have similarly argued for German, that an even more tightly constrained paradigmatic systematicity is at issue. They show that the distribution of interfixes in the sets of compounds sharing the left constituent (and also to some extent those sharing the right constituent) is predictive of the choice of the interfix in novel compounds and of the time required to make this choice. One of the questions to be considered in the present study is to what extent the paradigmatic systematicity governing the distribution of interfixes leaves room for the separation hypothesis as an additional principle shaping this distribution.

Another important factor governing the choice of the interfix is whether the left constituent has a derivational suffix as head. If so, this suffix may govern the selection of the interfix. This is the only circumstance in which some clear syntagmatic rules are operative. In Dutch, for instance, all compounds that have a diminutive as left constituent contain the *-s-* interfix. Aronoff and Fuhrhop (2002) have pointed out for German that especially suffixes that do not allow further word formation, i.e., that have a morphological valency of zero, are most likely to require an interfix when embedded as a left constituent in a compound. They suggest that for such compounds the interfix has the function of opening the

left constituent for further word formation. Krott, Schreuder, and Baayen (2002) provide distributional evidence that Dutch interfixes may serve a similar opening function. Note that wherever the choice of the interfix is grammaticalized deterministically, as in the case of Dutch diminutive left constituents, the question of the additional explanatory power of the separation hypothesis again arises.

In what follows, we first present the results of our survey of the lexical statistics of interfixation in German and Dutch. This survey is restricted to compounds consisting of simple nouns. By excluding all compounds from the analysis in which the presence of the interfix is determined by other factors, such as derivational suffixes on non-final constituents, the relation between the hierarchical organization of the compound and the distribution of interfixes can be established most clearly.

Following the lexical statistical survey, we present a series of experiments investigating whether and how the interfixes affect the way in which speakers of German and Dutch interpret the structure of multi-constituent compounds. All experimental studies of compound processing that we are aware of have addressed the production or comprehension of compounds consisting of two constituents. For such compounds, the question of their hierarchical structure does not arise. The experiments presented here address, for the first time, the assignment of structure to compounds with three constituents. The goal of these experiments is to establish the extent to which the distributional properties of existing compounds in the lexicon, and specifically the distribution of interfixes within these compounds, guide peoples' interpretations.

Lexical statistics

Table 2 lists, for German and Dutch, the number of compounds with two, three, and four simple nominal constituents (henceforth simple nominal compounds), as well as the proportions of compounds without interfixes. The counts for German are based on newspaper corpora comprising some 76 million words, as well as on the German data in the CELEX lexical database, which are based on a corpus of some 6 million words. The counts for Dutch are based on the Dutch section of the CELEX lexical database, which is based on a corpus of approximately 42 million words.

First note that the number of compounds with two constituents is much larger than the number of compounds with three constituents, and that the number of compounds with four constituents is extremely small, both in German and in Dutch.

Surprisingly, the number of two-constituent compounds in German is less than half the corresponding number in Dutch (6643 versus 14644). This suggests that compounding might be more productive in Dutch than in German. This possibility receives further support from the total counts of noun-noun compounds (including derived constituents) in our data sets: 44,000 for Dutch versus 32,000 for German, in spite of the fact that the German data are based on corpora that together are twice as large as the Dutch corpus. It should be kept in mind, however, that the Dutch corpus samples a much broader range of registers. Moreover, comparative studies of morphological acquisition studies suggest that compounds emerge more rapidly with German speaking children than with children learning Dutch (see, e.g., Dressler, Kilani-Schoch, & Klampfer, 2003). We leave this issue to further research.

language	number of constituents	number of compounds	proportion without interfixes
German	2	6643	0.65
	3	442	0.57
	4	5	0.80
Dutch	2	14644	0.73
	3	546	0.67
	4	6	0.67

Table 2: Compounds in German and Dutch consisting of simplex nouns and optionally interfixes.

Focusing on the compounds with two and three constituents, we find that the proportion of two-constituent compounds with no interfix is larger than the corresponding proportion for three-constituent compounds, both in German ($X^2(1) = 11.19, p = 0.0008$, here and elsewhere, all chi-squared tests are run with continuity correction), and in Dutch ($X^2(1) = 7.36, p = 0.0067$). In other words, three-constituent compounds are more likely to contain interfixes (one, possibly even two) than two-constituent compounds. This observation is in line with what Žepić (1970) and Fuhrhop (1998) report for German.

This increase in the use of interfixes in tri-constituent compounds constitutes a first potential piece of evidence in favor of the separation hypothesis, namely that in tri-constituent compounds interfixes have the function of marking the major constituent boundary. However, in order to evaluate this potential evidence, we first have to consider other possible factors that might lead to such an increase. For instance, this increase might be due simply to there being two positions available for interfixation instead of just one. In fact, the 442 German tri-constituent compounds provide 884 slots for interfixation, of which 232 are used, which amounts to 26.2%. (There are 190 compounds with one interfix and 21 compounds with two interfixes.) This is a significantly reduced realization rate compared to that of the bi-constituent compounds (35.0%, $X^2(1) = 26.3, p < 0.0001$). Similarly, the 546 tri-constituent Dutch compounds provide 1092 slots for interfixation, of which only 214 are actually used: 18.0%. (There are 180 compounds with one interfix and 17 compounds with two interfixes). Compared to the bi-constituent compounds, for which 27% of the positions are realized, we again have a significant reduction ($X^2(1) = 28.2, p < 0.0001$). Thus it is not the case that the increase in the number of interfixes in tri-constituent compounds is simply the result of the doubling of the available slots.

Interestingly, there are surprisingly few compounds with two interfixes: 21 for German, and 17 for Dutch. Given the realization rates of interfixes in bi-constituent compounds, 0.35 for German and 0.27 for Dutch, binomial tests suggest that the probabilities of having 21 or fewer compounds with two interfixes in German and of having at most 17 such compounds in Dutch are vanishingly small ($p < 0.00001$). This shows that the increased proportions of interfixation for the tri-constituent compounds compared to the bi-constituent compounds are also not due to the existence of large numbers of compounds with two interfixes. In fact, the opposite is true: There are far fewer such compounds than one would expect under chance conditions. We conclude that the overrepresentation of interfixes in tri-constituent compounds compared to bi-constituent compounds is a non-trivial fact that requires further explanation. The separation hypothesis provides one such explanation. To evaluate this hypothesis, we must take into account whether a tri-constituent compound is left branching or right-branching, as the direction of branching determines the position of the major constituent boundary. We therefore begin by surveying the distributional properties of left and

right-branching compounds in German and Dutch in our databases, which are summarized in Table 3. (Note that due to the different sizes of the corpora for German and Dutch, the numbers of types and their token frequencies cannot be directly compared across languages.)

language	branching	types	C12	median C12	C23	median C23
German	left	287	287	62	67	44
German	right	155	22	21	155	126
Dutch	left	347	341	190	25	20
Dutch	right	199	20	25	193	365

Table 3: Type frequency for left and right branching compounds with three simplex nominal stems in Dutch and German. Types: number of such compounds; C12: number of compounds for which the first two nouns form an existing compound; median C12: their median frequency; C23: number of compounds for which the second and third noun form an existing ompounds; median C23: their median frequency.

The first thing to note is that there are more left-branching compounds than right-branching compounds, both in German and in Dutch. Ortner and Bollgahen-Müller (1991) also reports a preference for left-branching compounds in German. This suggests that left branchingness is the unmarked structure for a tri-constituent compound.

The fourth column of Table 3, labeled 'C12', lists the number of tri-constituent compounds for which the first two nouns form a compound (henceforth a C12 compound) that is also attested in our databases. In German, all 287 left-branching compounds have a C12 compound matching the left branch. In Dutch, 341 of the 347 compounds similarly consist of a C12 compound followed by another noun.

A similar pattern emerges for right-branching compounds. In German, all 155 right-branching compounds have a right branch that is an attested C23 compound. For Dutch, 193 of the 199 right-branching compounds have an attested right-branch C23 compound.

Occasionally, existing compounds straddle the major constituent boundary. In our German database, there are 67 left-branching compounds for which the C23 compound exists. The median frequency (44) of these spurious compounds, however, is lower than the median frequency (62) of the correct C12 compounds. This difference, however, is not significant

($W = 9849.5, p = 0.408$, Wilcoxon test). The reverse pattern holds for the right-branching compounds, for which the correct C23 compounds have a higher median frequency (126) than the spurious C12 compounds (median frequency 22), a difference which is significant ($W = 883, p = 0.0008$). Table 3 shows exactly the same pattern for Dutch, with more pronounced differences among the median frequencies, both of which are significant (left branching: $W = 6827$, right-branching: $W = 534.5$, $p < 0.0001$ in both cases).

What these distributional data show is that there is a strong bi-directional implicational relation between frequency and left versus right branching. If the C12 but not the C23 compound exists, the tri-constituent compound is almost always left-branching. If only C23 exists, the tri-constituent compound is almost certainly a right branching compound. Conversely, if a compound is left-branching, we can infer that the C12 exists. Similarly, if a tri-constituent compound is right-branching, the C23 compound is extremely likely to exist.

The relation between the frequency of the embedded compound and its position in the tri-constituent compound raises the question of what the added functionality of interfixes as separators of the immediate constituents might be. After all, nearly all our tri-constituent compounds are composed of an existing compound and a monomorphemic noun.

With this correlation of frequency and immediate constituent structure in mind, we now consider whether interfixes indeed show a strong preference for the major constituent boundary, as predicted by the separation hypothesis.

language	branching	types	Slot 1 only		Slot 2 only		Slot 1 and Slot 2	
			all interfixes	s	all interfixes	s	all interfixes	s
German	left	287	24	3	89	60	10	4
German	right	155	41	10	14	2	11	0
Dutch	left	347	50	13	39	25	10	2
Dutch	right	199	60	38	11	3	7	1

Table 4: Number of interfixes realized after the first noun only (Slot 1 only), after the second noun only (Slot 2 only), or after both the first and the second noun (Slot 1 and Slot 2) for German and Dutch left and right branching compounds with three simplex noun constituents. The numbers of compounds with the interfix *-s-* are listed separately.

The distribution of interfixes in left and right-branching German and Dutch compounds is summarized in Table 4. The third column of this table again lists the number of left and right-branching compounds. Columns four and five show the counts of interfixes that fill the first interfixation slot (the position following the first noun), with the second interfixation slot (the position following the second noun) being empty. Column four lists the total number of compounds in which any interfix (*-s-, -en-, -n-, -e-, -er-, ...*) occurs after the first noun. Column 5 lists the number of such compounds in which *-s-* is realized, as the separation hypothesis has been advanced primarily for the *-s-* interfix (Grimm, 1878; Fuhrhop, 1998; 2000). Columns six and seven document the counts of compounds for which only the second interfixation slot is filled. The last two columns list the counts of compounds in which both interfixation slots are filled.

The literature on German interfixation is concerned primarily with interfixation in left-branching compounds. The basic observation has been that especially the *-s-* interfix shows a preference for the second interfixation slot in left-branching compounds. Table 4 shows that in German there are indeed more interfixes in slot 2 (89/287) than in slot 1 (24/287) for left-branching compounds (0.08 versus 0.31, $X^2(1) = 45.1, p < 0.0001$), and the same observation holds for the *-s-* interfix by itself (3/287=0.01 vs. 60/287=0.21, $X^2(1) = 55.9, p < 0.0001$). This distributional asymmetry supports the separation hypothesis, especially in the case of the *-s-*, which hardly occurs at all at the minor boundary position.

Interestingly, our data reveal not only a distributional asymmetry for left-branching compounds, but also a similar asymmetry for right-branching compounds. Interfixes in right-branching compounds show a preference for the first slot (41/155) rather than the second slot (14/155, $X^2(1) = 14.9, p = 0.0001$), a preference that is also manifest for the *-s-* interfix by itself (10/155 versus 2/155, $X^2(1) = 4.2, p < 0.0393$). This suggests that in German there is an overall asymmetry in the distribution of interfixes, with preferential occurrence in the slot of the major constituent boundary.

In Dutch, no distributional assymmetry appears to be present for left-branching compounds when we count all interfixes (50/347 is not significantly different from 39/347, $X^2(1) = 1.29, p = 0.2563$). In fact, more interfixes are realized in slot 1 than in slot 2, contrary to what the seperation hypothesis predicts. However, when we count the com-

pounds with the -*s*- interfix, we find a tendency in the predicted direction, although it does not reach significance (13/347 versus 24/347, $X^2(1) = 3.37, p = 0.06645$). For Dutch right-branching compounds, we do see a clear distributional asymmetry, both for all interfixes counted jointly (60/199 versus 11/199, $X^2(1) = 39.5, p < 0.0001$) and for the -*s*- by itself (38/199 versus 3/199, $X^2(1) = 31.4, p < 0.0001$). The overall pattern suggests that in Dutch the distributional asymmetry expected under the separation hypothesis is carried primarily by the -*s*- interfix, and that it is most clearly present in right-branching compounds.

Thus we see that the distribution of the -*s*- interfix in German and Dutch across the first and second interfixation slots in left and right-branching compounds reveals a clear preference for the slot coinciding with the major constituent boundary. It should nevertheless be kept in mind that the number of cases where the -*s*- reliably marks a major constituent boundary is quite restricted. Consider the -*s*- in German. Taking both right and left-branching compounds together, we see that -*s*- is used in only 79 out of 442 tri-constituent compounds (18%). Nine of these 79 compounds contain an -*s*- at the minor constituent boundary (a noise rate of 11%). For Dutch, we have a realization rate for the -*s*- of 79/546=15% and a noise rate of 19/79=24%.

There is a further observation to be made when we consider Tables 2 and 4 together. Table 2 shows that 2325/6643=35% of the bi-constituent compounds in German and 3954/14644 = 27% of such compounds in Dutch contain an interfix. However, the degree of realization of interfixes in bi-constituent compounds that appear as constituents in tri-constituent compounds (the C12 compounds in left-branching compounds and the C23 compounds in right-branching compounds) is much lower. Collapsing over left and right-branching compounds in German which have an interfix at the minor boundary, we count a maximum of 24+14+10+11 = 59 compounds. (This number represents a maximum because we have not checked whether each of the pairs of constituents surrounding the -*s*- is in fact an existing compound of the language.) For Dutch, the corresponding count is 50+11+10+7 = 78. Hence, the degree of realization of interfixes in embedded compounds is at most 59/442=13.3% for German and 78/546=14.3% for Dutch. The within-language reductions in degree of realization are significant ($X^2(1) = 86.0, p < 0.0001$ for German, $X^2(1) = 43.0, p < 0.0001$ for Dutch). In other words, bi-constituent compounds with an interfix are

relatively unproductive as base words in tri-constituent compounds. If a compound happens to already have an interfix, it apparently is less likely to be used as a constituent of a tri-constituent compound. The low degree of productivity of interfixed compounds as immediate constituents of tri-constituent compounds provides further support for the separation hypothesis. Unconstrained use of interfixed bi-constituent compounds would lead to, other thing being equal, roughly one third of tri-constituent compounds having an interfix at the minor constituent boundary, which would be detrimental to the cue-validity of the interfix as the marker of the major constituent boundary.

As mentioned above, both German and Dutch show remarkably few compounds with two interfixes. This holds both for arbitrary combinations of interfixes, as discussed above, as well as for compounds with two *-s-* interfixes. Under chance conditions, one would expect a larger number of such compounds than are actually observed. Consider the *-s-* interfix, the interfix for which the distributional asymmetry is most clearly present. There are four *-s-s-* German tri-constituent compounds and three in the case of Dutch (see the last column of Table 4). To gauge whether these numbers are indeed surprisingly small, we make three simple assumptions. First, we assume that in all cases of double interfixation, the interfix within the embedded compound is inherited from the lexicon, leaving one slot for further interfixation. Second, we assume that we can estimate the probability of *-s-* interfixation by calculating the proportion of *-s-* interfixation in bi-constituent compounds. In German, this probability is 0.062. For Dutch, it is 0.145. Third, we assume that the probability of having X compounds with doubly filled slots is (546, 0.062)-binomially distributed in the case of German, and (442, 0.145)-binomially distributed in the case of Dutch. Under these assumptions, the probabilities of having not more than four *-s-s-* compounds in German or three *-s-s-* compounds in Dutch is extremely small ($p < 0.000001$). This suggests that there is a probabilistic constraint disfavoring having two interfixes in a tri-constituent compound. Interestingly, in compounds with two interfixes, the interfixes can no longer uniquely identify the main constituent boundary. Therefore, the scarcity of compounds with two *-s-* interfixes is again exactly in line with the separation hypothesis.

Summing up, the scarcity of tri-constituent compounds with two interfixes, the low degree of productivity of bi-constituent interfixed compounds as constituents in tri-constituent

compounds, and the particular preference for the *-s-* to occupy the slot coinciding with the major constituent boundary, for both left and right-branching compounds, are all consistent with the separation hypothesis.

The question now arises to what extent language users employ these distributional properties in their parsing of tri-constituent compounds. What might obscure the functionality of *-s-* interfixation is that for tri-constituent compounds, the immediate constituent structure falls out naturally given that almost all tri-constituent compounds consist of an existing compound and a monomorphemic noun. From this perspective, it is quite unlikely that any parsing ambiguities would arise and that an interfix would be necessary to disambiguate the immediate constituent structure. In fact, the scarcity of tri-constituent compounds with an *-s-* interfix that might carry the separation function might be due to precisely the fact that structural ambiguities seldom arise.

In what follows, we therefore investigate experimentally whether language users nevertheless show sensitivity to the specific distributional properties of the *-s-* interfix.

Experiments

The lexical statistics show that the interfixes, and notably so the *-s-*, reflect the hierarchical structure of tri-constituent compounds. The question to be addressed now is whether the reverse might also be the case. Does the interfix co-determine how speakers of German and Dutch assign structure to tri-constituent compounds?

In order to test this possibility, we make use of tri-constituent compounds consisting of three non-existing but orthographically and phonologically legal monomorphemic and monosyllabic simple words. We have opted for pseudo-compounds for three reasons. The use of pseudo-compounds has the advantage that we can study the role of the interfix in the absence of the great many lexical factors that co-determine the structure of existing compounds, factors such as the frequencies of the simple constituents, the frequencies of the embedded compounds, and the analogical preferences of nouns for a given interfix (or its absence). The interfix studied in our experiments is the *-s-*, as this interfix most clearly emerges from the lexical statistics as a boundary marker in both German and Dutch. The use of the *-s-* has the additional advantage that, unlike the *-en-* interfix, it does not alter the

syllable structure of the stimuli. We were thus able to construct, for each pseudo-compound, four interfixation conditions, namely a condition with no interfix (No-S), a condition with an interfix in the left interfixation slot (Left-S), a condition with the interfix in the right slot (Right-S), and a condition with the interfix in both slots (Both-S). An example of a German pseudo-compound under these four conditions is:

Belkfliemguhr	No-S
Belksfliemguhr	Left-S
Belkfliemsguhr	Right-S
Belksfliemsguhr	Both-S

The last condition, with both interfixation slots filled, is one that is hardly ever realized in the language (see Table 4). We nevertheless included this condition because it provides us a means for checking that participants will not merely base their response on the first (or the last) interfix they encounter in a given compound.

Experiments 1,2 and 4 make use of an implicit task to tap into the intuitions of speakers of German and Dutch regarding the hierarchical structure of tri-constituent compounds. We asked participants to hyphenate the stimuli, the idea being that hyphenation is a natural task that does not explicitly invoke meta-linguistic grammatical skills. The prediction is that participants will tend to hyphenate the compounds at the perceived major constituent boundary. In Experiment 3, we further verified these intuitions by asking participants explicitly to assign constituent structure to the stimuli.

Experiment 1

Materials. We constructed 60 pseudo-words, each consisting of three non-existing syllables that follow German phonotactic rules. The syllables were all constructed so that the initial and final consonant clusters were clear indicators of syllable boundaries. In this way, all pseudowords could unambiguously be interpreted as tri-constituent compounds. The initial letters were capitalized as required by German spelling conventions. The coda and onset clusters of the syllables were chosen such that an *-s-* could only be part of the preceding syllable. Since the interfix *-s-* is homographic with the inflectional suffix *-s*, we made sure

Table 5: Number of left hyphenations and right hyphenations for each of the four interfixation conditions No-S, Left-S, Right-S, and Both-S for German (Experiment 1).

position of *-s-*	left hyphenation	right hyphenation	errors
Left-S	176	161	8
No-S	140	198	7
Right-S	93	239	13
Both-S	116	217	12

that the phonology of the constituents ensured that the *s* could be interpreted as an interfix only.

Each participant saw all compounds under one of the four interfixation conditions described above. We counterbalanced the materials to guard against practice effects by grouping the items into four lists and assigning each of the four variants of a compound to a different list. The resulting four lists each contained 60 compounds, 15 compounds for each of the four conditions No-S, Left-S, Right-S, and Both-S. Each list was presented in one of three random orders, resulting in a total of 12 random orders of presentation for the experiment. Each of these lists was preceded by the same set of 10 practice items.

Procedure. Participants were told that they were going to read non-existing three-constituent compounds and were informed that their task in the experiment was to identify optimal hyphenation points. They were instructed to indicate the optimal hyphenation point by drawing a vertical bar between the relevant constituents. Letters were in 12 point font and were separated by 3 points of space to allow the vertical lines to be placed unambiguously. The experiment lasted approximately 5 minutes.

Participants. Twenty-four participants, all students of the University of Vienna, volunteered to take part in the experiment. All were native speakers of German.

Results. One participant marked all compounds as left-branching. His responses were excluded from the analyses. Other participants occasionally either did not mark any boundary or marked an impossible boundary. These responses were counted as errors. The counts of left hyphenation and righ-hyphenation responses, together with these errors, are presented in Table 5.

An analysis of the log odds ratio of left and right hyphenations revealed significant differences between the four hyphenation conditions ($F1(3, 66) = 7.15, p = 0.0003$; $F2(3, 177) = 14.68, p < 0.0001$). Participants placed more hyphenation marks after the second constituent than after the first (815 right responses, 60.8%, versus 515 left response, 39.2%, $\chi^2(1) = 124.7, p < .0001$). This is in line with the lexical statistics of German, which also show that the major constituent boundary occurs more often after the second than after the first constituent.

Pairwise comparisons of the counts of left and right responses in the four experimental conditions (with alpha=0.05 after Bonferroni correction, using Chi-squared tests on the summed left and right counts for each condition) revealed significant differences between all conditions except two comparisons involving the Both-S condition. This condition differed significantly from the Left-S condition, but not from the No-S condition nor from the Right-S condition. The upper left panel of Figure 1 shows the pattern of results. The vertical axis in this figure represents the proportion of right hyphenation responses on the total of non-erroneous responses.

First note that the greatest percentage of right hyphenation responses is found for the compounds in which a unique interfix follows the second constituent. Conversely, a unique interfix following the first constituent attracts left hyphenation responses. The compounds in which no interfix is present pattern in between the compounds with a unique interfix. This pattern in the experimental data matches the distributional pattern of interfixes in German.

The experiment also included a condition that is not widely used in the language, namely, compounds with two interfixes. Our lexical statistical survey revealed that there are surprisingly few of such compounds. The present experiment suggests that they pattern along with both the compounds with no interfix, as well as with the compounds with an interfix in the right interfixation slot. Because the proportion of right responses for the Both-S condition does not differ significantly from the corresponding propertion for the No-S condition, one might argue that the interfixes in the Both-S condition have no cue validity for the hierarchical structure, just as when there is no interfix at all. On the other hand, the Both-S condition also does not differ from the Right-S condition, suggesting that having the interfix in the right slot is sufficient to attract as many right responses as when there is a unique

interfix in the right interfixation slot. The present data do not allow us to decide between these two interpretations.

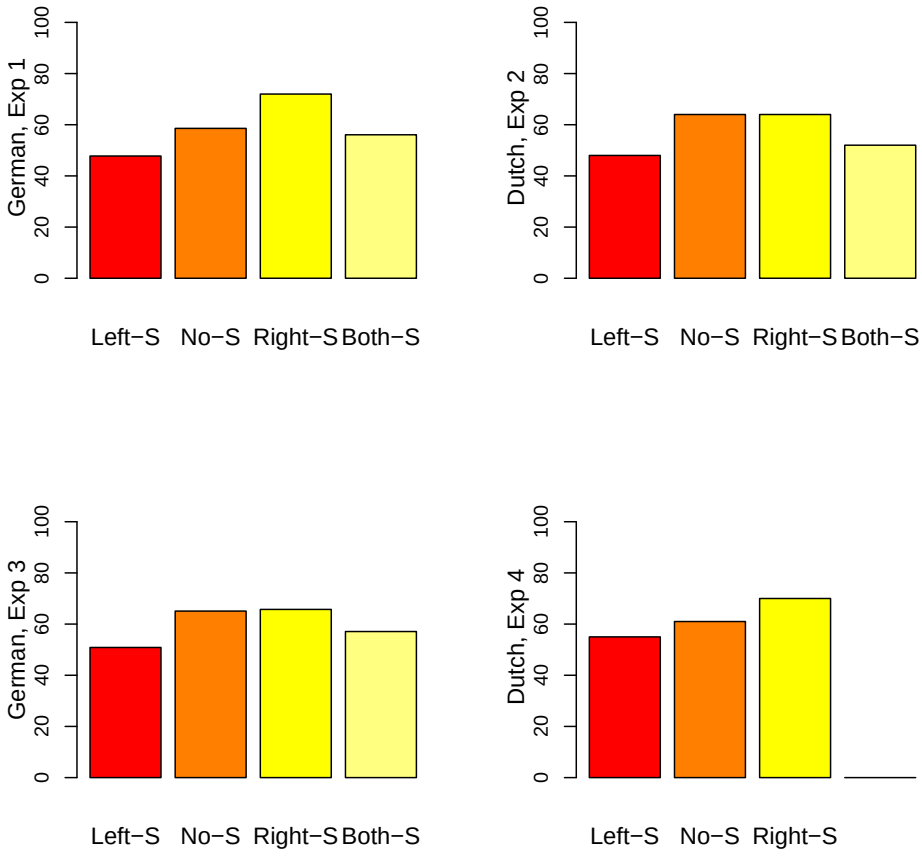

Figure 1: Percentages of right boundary responses for Experiments 1 to 4 as a function of interfixation.

Table 6: Number of left hyphenations and right hyphenations for each of the four interfixation conditions No-S, Left-S, Right-S, and Both-S for Dutch (Experiment 2).

position of *-s-*	left-hyphenation	right-hyphenation	errors
Left-S	171	156	33
No-S	112	214	32
Right-S	113	214	33
Both-S	150	171	39

Experiment 2

Materials. Similarly to the German stimuli in Experiment 1, we constructed 60 pseudo-words, each consisting of three non-existing syllables that follow Dutch phonotactics. The syllables were again of a form such that the initial and final consonant clusters were clear indicators of syllable boundaries so that each word could be unambiguously interpreted as a tri-constituent compound. The coda and onset clusters of the syllables were chosen such that an *-s-* could only be part of the preceding syllable and not a plural suffix. As in Experiment 1, we presented each pseudo-compound in each of the four interfixation conditions (No-S, Both-S, Left-S, Right-S), using the same counterbalancing procedure as in Experiment 1.

Procedure. The procedure was identical to that of Experiment 1.

Participants. Twenty-four participants, all students of the University of Nijmegen, were paid to take part in the experiment. All were native speakers of Dutch.

Results. Two participants marked all compounds as left-branching. Their responses were excluded from the analyses. As in Experiment 1, participants sometimes either not marked any boundary at all or marked an impossible boundary. These responses were counted as errors and were discarded in the analyses below. There were also two missing data points, for which no response was provided. Table 6 lists the number of left and right hyphenation responses for the four conditions (No-S, Left-S, Right-S, Both-S), the upper left panel of Figure 1 shows the proportion of right hyphenation responses for each condition.

An analysis of the log odds ratio of left and right hyphenation responses revealed significant differences between the four hyphenation conditions ($F1(3, 63) = 8.44, p < 0.0001$;

$F2(3, 177) = 11.01, p < 0.0001$). Participants placed more hyphenation marks after the second constituent than after the first (755 right responses, 58.0%, versus 546 left response, 42.0%, $\chi^2(1) = 66.5, p < .0001$). The higher proportion of right hyphenation responses suggest a preference for a left-branching interpretation of the compounds, a preference reflecting the predominance of left-branching compounds in Dutch.

Pairwise comparisons of the number of left and right responses in the four conditions (with alpha=0.05 after Bonferroni correction, using Chi-squared tests on the summed left and right counts for each condition) revealed significant differences between all conditions except two comparisons. First, the proportions of right responses do not differ significantly for the No-S and Right-S conditions. Second, the Left-S and Both-S conditions elicited very similar proportions of right hyphenation responses.

As can be seen in the upper right panel of Figure 1, the Left-S condition elicits many left hyphenation responses, while the Right-S condition elicits many right-hyphenation responses. This pattern of results is very similar to that observed in Experiment 1 for German, and as in German, this pattern is what the distribution of the *-s-* interfix in the existing compounds of Dutch leads us to expect.

Unlike in the German experiment, the No-S condition patterns along with the Right-S condition. This would suggest that an interfix in the left interfixation slot is marked and functional, and that an interfix in the right interfixation slot is unmarked and not functional compared to compounds with no interfixes.

The Both-S condition also patterns differently than in the German experiment. In that experiment, the proportions of left and right hyphenation responses in the Both-S condition did not differ significantly from the corresponding proportion in the Right-S condition. In the present experiment, the proportions in the Both-S and Left-S conditions do not differ significantly. This suggests that perhaps the presence of an *-s-* in the left interfixation slot is the crucial variable for Dutch.

These experiments assess the intuitions of speakers of German and Dutch concerning the hierarchical structure of novel compounds by means of a task that implicitly asks participants to decide on the main constituent boundary. The hyphenation task calls upon an orthographic skill and harnesses this skill to indirectly perform a meta-linguistic immediate

Table 7: Number of left and right main constituent boundary responses for each of the four interfixation conditions No-S, Left-S, Right-S, and Both-S for German (Experiment 3).

position of *-s-*	left-boundary	right-boundary	errors
Left-S	142	147	26
No-S	103	192	20
Right-S	98	188	29
Both-S	115	153	47

constituent analysis. In order to ensure that the results obtained do not depend on specifically this implicit task, we repeated Experiment 1, but now explicitly asking participants to parse the compounds into their immediate constituents.

Experiment 3

Materials. The materials were identical to those of Experiment 1.

Procedure. Instead of asking participants to indicate the position where they would hyphenate the word, we provided explicit parsing instructions. After introducing the notions of left-branching and right-branching compounds, we asked them to mark the main constituent boundary.

Participants. Twenty-four participants, all students of the University of Vienna, were paid to take part in the experiment. All were native speakers of German.

Results. We excluded the responses from three participants from the analyses since they marked all compounds as left-branching. Responses with missing boundary marks or with boundary marks at improbable positions in the compound were counted as errors. Table 7 lists the number of left and right constituent boundary responses for the four conditions (No-S, Left-S, Right-S, Both-S), the lower left panel of Figure 1 shows the proportion of right hyphenation responses for each condition.

An analysis of the log odds ratio of left and right responses revealed a marginally significant difference between the four interfixation conditions for the by-subject analysis and a significant main effect for the by-item analysis ($F1(3, 60) = 2.66, p = 0.0562$; $F2(3, 177) = 5.23, p < 0.0017$). Participants assigned more main constituent boundaries

marks to the second interfixation slot than to the first slot (669 right responses, 59.7%, versus 452 left response, 40.3%, $\chi^2(1) = 83.2, p < .0001$), as expected.

Pairwise comparisons of the number of left and right responses in the four conditions (with alpha=0.05 after Bonferroni correction, using Chi-squared tests on the summed left and right counts for each condition) revealed significant differences only for two pairs of conditions: Left-S and Right-S, and No-S and Left-S. In other words, in this experiment, it is a unique *-s-* in the first interfixation slot that elicited more left constituent boundary responses compared to the other conditions that are represented in the language, No-S and Right-S.

This brings us to the responses in the Both-S condition, the condition with doubly interfixed compounds that are hardly ever used in German or Dutch. Unlike in Experiment 1, in which this condition elicited fewer left hyphenation responses than the Left-S condition, no such difference is present in Experiment 3. In fact, the proportion of right responses in the Both-S condition is not statistically different from the corresponding propertion in any of the other conditions. When we consider Experiments 1–3 jointly, we find that the Both-S condition has the most variable response pattern of all conditions.

While in all experiments, one quarter of all trials contained two interfixes, the total number of compounds with two *-s-* interfixes in our database is 4 for German and 3 for Dutch (see Table 4). The variable patterning of the responses in the Both-S condition across experiments may reflect the fact that there is no consistent and well-established distributional pattern in the language for participants to fall back on.

As explained above, we included the condition with two interfixes to provide us with some control for tracing task effects, allowing us to ascertain whether participants make use of some *-s-* spotting strategy leading them to base their response on the first or last *-s-* encountered. However, a possible consequence of the scarcity of doubly interfixed compounds in the language and their overrepresentation in the experiments is that it may have reduced the cue validity of the interfix for the other experimental conditions. In order to ascertain to what extent the presence of doubly interfixed compounds might have biased the participants' responses, we conducted a final experiment using the same materials as in Experiment 2 but excluding the doubly marked compounds.

Experiment 4

Materials. The materials were identical of those of Experiment 2, but excluding the words in the Both-S condition.

Procedure. The procedure was identical to that of Experiments 1 and 2.

Participants. Seventeen participants, all students of the University of Nijmegen, were paid to take part in the experiment. All were native speakers of Dutch and none of them had taken part in Experiment 2.

Results. The responses of three participants were excluded as they produced right hyphenation responses exclusively. No hyphenation responses and hyphenations at improbable positions in the compound were again counted as errors. The counts of left hyphenation and righ-hyphenation responses, together with the error counts, are presented in Table 8.

An analysis of the log odds ratio of left and right hyphenations revealed significant differences between the three hyphenation conditions ($F1(2,26) = 6.61, p = 0.0048$; $F2(2,118) = 13.1, p < 0.0001$). Participants placed more hyphenation marks after the second constituent than after the first (766 right responses, 62.2%, versus 465 left response, 37.8%, $\chi^2(1) = 146.2, p < .0001$), as expected.

Pairwise comparisons of the counts of left and right responses in the three experimental conditions (with alpha=0.05 after Bonferroni correction, using Chi-squared tests on the summed left and right counts for each condition) revealed significant differences between the No-S and Right-S conditions as well as between the Left-S and Right-S conditions. The upper left panel of Figure 1 summarizes the pattern of results. The vertical axis in this figure represents the proportion of right hyphenation responses on the total of non-erroneous responses.

As in the preceding experiments, the Left-S condition elicits more left hyphenation responses while the Right-S condition elicits more right hyphenation responses. Recall that in Experiment 2, the No-S condition patterned along with the Right-S condition. We suggested that this might indicate that an interfix in the right interfixation slot might be non-functional. In the present experiment, omission of the Both-S condition has resulted in a different positioning of the No-S condition, which is now positioned in between the Left-S and Right-S

Table 8: Number of left and right hyphenation responses for each of the three interfixation conditions No-S, Left-S, and Right-S (Dutch, Experiment 4).

position of *-s-*	left-hyphenation	right-hyphenation	errors
Left-S	189	224	7
No-S	160	253	7
Right-S	116	289	15

conditions, differing significantly from the Right-S condition. In fact, the Left-S condition is now no longer significantly different from the Left-S condition. Given the present experiment considered by itself, one might argue that it is precisely an interfix in the right interfixation slot which is functional. What is constant across both experiments, importantly, is the consistent difference between the Left-S and Right-S conditions.

General Discussion

This study addressed the question of whether interfixes in German and Dutch compounds might be functional as probabilistic markers of the immediate constituent structure of multi-constituent compounds, a possibility first suggested by Grimm (1878) for German and known as the separation hypothesis.

A survey of tri-constituent compounds in the lexicons of German and Dutch revealed that the immediate constituent structure of compounds falls out naturally from the existing compounds that the lexicon provides as constituents. This would suggest that there would be no added functionality for interfixes. Nevertheless, it turns out that interfixes occur surprisingly often at the main constituent boundary. In German this holds for all interfixes, in Dutch, it holds only for the *-s-* interfix. We also observed surprisingly few tri-constituent compounds with two interfixes, and that bi-constituent compounds that themselves contain an interfix are significantly underrepresented as constituents of tri-constituent compounds. This suggests that German and Dutch avoid interfixation at both interfixation slots in tri-constituent compounds. Since an interfix can be a marker of immediate constituent structure only by virtue of there being an empty interfixation slot, we believe this provides further support for the separation hypothesis.

Having used lexical statistics to ascertain whether and how robustly main constituent boundaries attract interfixes in the existing compounds of German and Dutch, we further probed the functionality of interfixation by means of a series of experiments. We investigated whether the presence of an interfix, in circumstances in which no other information could potentially guide the assignment of hierarchical structure, might prompt speakers of German and Dutch to posit a main constituent boundary at the interfixed position.

Using monosyllabic pseudo-words brought together in tri-constituent pseudo-compounds, we observed across all experiments and tasks a significant difference in the number of posited right and left boundaries as a function of the presence of a single *-s-* interfix. When there is a single interfix following the first constituent, this interfix gives rise to significantly more left boundary responses than when there is a single interfix following the second constituent.

The experiments investigated two further conditions, one in which no interfix was present, and one in which both interfixation slots were occupied by the *-s-*. Across experiments, we found that when the interfix is at the right interfixation slot, a number of right responses is generated that is at least as great or greater than that for the No-S condition. Conversely, when the interfix is at the left interfixation position, the number of left responses is as large as or larger than the corresponding responses for the condition with no interfix. This in-between behavior of the No-S condition is in line with the marked nature of the interfix, which according to the functional separation hypothesis provides an explicit indication of the compound's structure.

The last condition in our experiments was one in which both interfixation slots were filled. This condition does not allow us to test the correspondence between the distributional properties of interfixes in the lexicon and the behavior of the participants, as it is a configuration that hardly ever is made use of in the language. Nevertheless, the condition does provide insight into the nature of the participants' performance in the experiment. If, for example, their performance was guided by a left-to-right processing algorithm positing a major boundary at the first interfix encountered, then the Both-S condition would show a predominance of left breaks, contrary to fact. Likewise, it does not reveal a systematic preference for right breaks. In fact, this condition is the most variable of all conditions in the experiment. We think this reflects the artificiality, perhaps the ungrammaticality of this

condition. The contrast of this condition with the consistent pattern of responses in the two conditions with a unique interfix reinforces our confidence that the latter conditions have some real ecological validity, in that the experiments tap into the way in which the interfix in novel 'grammatical' compounds guides structural interpretation.

Finally, we observed, across experiments and conditions, that there is a preference for right boundary responses. We interpret this preference as a preference for left-branching structures, the default hierarchical structure that emerged from the lexical statistics.

It is important to note that, across experiments, participants do not respond in an all-or-nothing way to the presence or absence of an interfix. The grammar of interfixation is not deterministic, in which case the position of the interfix would have resulted in nearly categorical response pattern with nearly all boundary markers following the interfix. To the contrary, the presence of the interfix serves as a probabilistic cue to immediate constituent structure, just as in German and Dutch the distribution of interfixes in multi-constituent compounds is probabilistic rather than deterministic.

Our study therefore bears further witness to the amazing sensitivity of the human brain to subtle differences in the probability distributions of linguistic variables (see, e.g., Albright & Hayes, 2003; Ernestus & Baayen, 2003). More than half of the tri-constituent compounds in German and Dutch do not contain any interfix. The distribution of the interfixes among the remaining tri-constituent compounds that do have an interfix is noisy, as was documented by Table 4. Moreover, which interfix (*-s-, -en-, -er-, ...*) is appropriate for a given compound is itself governed by a probabilistic paradigmatic system (Krott et al., 2001). Nevertheless, speakers of German and Dutch turn out to be sensitive to the probability distribution of the interfixes in compounds, even though their functionality as markers of immediate constituent structure is almost fully redundant. This suggests to us that probability estimation is an integral part of grammatical competence, essential for the optimization of verbal communication.

Author Note

The authors are indebted to Arne Fitschen and Ulrich Heid (University of Stuttgart)

for providing a list of 34,000 German compounds. This study was made possible by an an SSHRC MCRI grant to G. Libben, G. Jarema, E. Kehayia, B. Derwing, and L. Buchanan, and was also supported by a PIONIER grant of the Dutch Research Council NWO to Baayen.

References

Albright, A. and Hayes, B.: 2001, Rules vs. analogy in English past tenses: A computational/experimental study, *Manuscript UCLA*.

Aronoff, M. and Fuhrhop, N.: 2002, Restricting suffix combinations in English: Closing suffixes and the monosuffix constraint, *Natural Language and Linguistic Theory* **20**(3), 451–490.

Baayen, R. H., Piepenbrock, R. and Gulikers, L.: 1995, *The CELEX lexical database (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Booij, G. E.: 2002, *The morphology of Dutch*, Oxford University Press, Oxford.

Dressler, W., Kilani-Schoch, M. and Klampfer, S., 2003, How does a small child detect morphology, *in* R.H. Baayen and R. Schreuder (eds), *Morphological Structure in Language Processing*, Mouton de Gruyter, Berlin (to appear).

Dressler, W. U., Libben, G., Stark, J., Pons, C. and Jarema, G.: 2001, The processing of interfixed German compounds, *in* G. E. Booij and J. Van Marle (eds), *Yearbook of Morphology 1999*, Kluwer, Dordrecht, pp. 185–220.

Ernestus, M. and Baayen, R.: 2003, Predicting the unpredictable: Interpreting neutralized segments in Dutch, *Language (to appear)*.

Fleischer, W.: 1976, *Wortbildung der deutschen Gegenwartssprache*, Bibliographisches Institut, Leipzig.

Fuhrhop, N.: 1998, *Grenzfälle Morphologischer Einheiten (Border Cases of Morphological Units)*, Stauffenburg, Tuebingen.

Fuhrhop, N.: 2000, Zeigen Fugenelemente die Morphologisierung von Komposita an?, *in* R. Thieroff, M. Tamrat, N. Fuhrhop and O. Teuber (eds), *Deutsche Grammatik in Theorie und Praxis*, Max Niemeyer Verlag, Tuebingen, pp. 201–213.

Grimm, J.: 1877 (1967), *Deutsche Grammatik*, Olms, Hildesheim.

Krott, A.: 2001, *Analogy in Morphology: The selection of linking elements in Dutch compounds*, University of Nijmegen, Nijmegen.

Krott, A., Baayen, R. H. and Schreuder, R.: 2001, Analogy in morphology: modeling the choice of linking morphemes in Dutch, *Linguistics* **39**(1), 51–93.

Krott, A., Schreuder, R. and Baayen, R. H.: 2002, A note on the function of Dutch linking elements, *in* G. E. Booij and J. Van Marle (eds), *Yearbook of Morphology 2001*, Kluwer, Dordrecht, pp. 237–252.

Libben, G., Jarema, G., Dressler, W., Stark, J. and Pons, C.: 2002, Triangulating the effects of interfixation in the processing of german compounds, *Folia Linguistica* **26**, 23–43.

Ortner, L. and Mueller-Bollhagen, E.: 1991, *Deutsche Wortbildung. Vierter Hauptteil: Deutsche Substantivkomposita*, de Gruyter, Berlin.

Schreuder, R., Neijt, A., Van der Weide, F. and Baayen, R. H.: 1998, Regular plurals in Dutch compounds: linking graphemes or morphemes?, *Language and cognitive processes* **13**, 551–573.

Žepić, S.: 1970, *Morphologie und Semantik der deutschen Nominalkomposita*, Philosophische Fakultät der Universität Zagreb, Zagreb.