Semantic Influence on Linkers in Dutch Noun-Noun Compounds

Andrea Krott, Loes Krebbers, Robert Schreuder & R. Harald Baayen*

Abstract

As in many other languages, the constituents of nominal compounds in Dutch are often separated by a linking element. This study investigates to what extent form and semantic properties of the right constituents in Dutch compounds affect the choice of the linker. Using both lexical statistics and experimentation, we show that the left and right constituent families affect the choice of the linker independently of the semantic categories of the left and right constituents themselves. We also show that the choice of the linker is co-determined by the animacy and concreteness of the left constituent. No role for the semantic class of the head constituent was observed in the experiment. Apparently, linkers are non-canonical suffixes in the sense that their occurrence is codetermined by the form properties of the constituent to their right.

1. Introduction

In many languages, elements known as connectives, interfixes, linkingmorphemes, and linkers, may occur between the two constituents of compounds. Sometimes, the occurrence of such linkers can be predicted on phonological grounds as in Zoque, a Mixe-Zoquean language spoken in Mexico. Zoque has a nominal compound formation in which the linking element is a vowel that is identical to the vowel in the preceding syllable (Herrera, 1995). In Germanic languages such as German and Dutch, the principles governing their distribution are less clear. The distribution of linkers in German appears to be governed by a complex set of factors (see, e.g., Dressler, Libben, Stark, Pons, & Jarema, in press; Fuhrhop, 1998). Although Dutch is closely related to German, the linkers in Dutch have different properties, possibly because, in contrast to German, modern Dutch no longer has productive case morphology. Historically, the Dutch linkers can be traced back to the case endings that existed in medieval Dutch. However, the original functionality of the linking elements as case suffixes is absent in modern Dutch.

Folia Linguistica XXXVI/1-2 *0165-4004/02/36-7 $ 2.–*
(C) Mouton de Gruyter, Berlin — *Societas Linguistica Europaea*

More than a third of Dutch noun-noun compounds contain a linker connecting the two main constituents.[1] Usually *-s-* or one of the orthographic variants *-en-* and *-e-* appear as a linker (e.g., *bevolking+s+getal* ‘number of population’, *boek+en+kast* ‘bookcase’, *zon+e+schijn* ‘sunshine’).[2] The usage of these linkers reveals considerable variation and unpredictability. Existing rule-based descriptions report various morphological, phonological, and semantic factors which seem to govern their choice (e.g., van den Toorn, 1981a; 1981b; 1982a; 1982b; Mattens, 1984; Haeseryn, Romijn, Geerts, De Rooij, & van den Toorn, 1997; Booij & Van Santen, 1995). However, almost every rule comes with a large number of exceptions. Taking all phonological and morphological rules together that are described in the literature[3], one can apply them to only 51% of all CELEX compounds and correctly predict only 63% of this subset. We therefore may conclude that a rule-based account for Dutch linkers is observationally inadequate.

Krott, Baayen, & Schreuder (2001) and Krott, Schreuder, & Baayen (in press) argue that the choice of linkers in Dutch is governed by analogy. Using an off-line cloze task in which participants had to form novel compounds from two Dutch nouns, they show that the usage of linkers in novel compounds can be predicted with a high degree of accuracy on the basis of analogy to the forms of existing compounds sharing the left or the right constituent of a given target compound, for instance, *schaap-?-oog*, ‘sheep-eye’. We refer to the set of compounds sharing the left constituent *(schaap* ‘sheep’ in this example) as the Left Constituent Family (*schaap+en+bout*, ‘leg of mutton’, *schaap+s+kooi*, ‘sheepfold’, *schaap+herder*, ‘shepherd’, etc.), and we refer to the set of compounds sharing the right constituent (*oog* ‘eye’ in the present example) as the Right Constituent Family (*uil+e+oog* ‘owl‘s eye’, *spleet+oog* ‘slant eye‘, *glas+oog* ‘glass eye’, etc.). One can predict the choice of the linker for a novel compound on the basis of the distribution of linkers in its Left and Right Constituent Families. For instance, if *schaap* occurs as a left constituent mostly in compounds containing the linking *-en-* (70% in CELEX), there is a high chance that a novel compound with *schaap* as the left constituent would also be built with *-en-*.

The strongest analogical factor predicting linkers appears to be the bias of the Left Constituent Family. In addition to the Constituent Family, experiments with pseudo-stems followed by existing suffixes as left constituents showed effects of the bias of the suffix and the rime of the left constituent. The bias of the suffix appears to be the second strongest factor overruling the bias of the rime. Apart from the effects of the left constituent, there is also evidence for a smaller, but statistically reliable effect of the bias of the Right Constituent Family.[4] Explicit computational models for analogy (AML, Skousen, 1989; TiMBL, Daelemas, Zavrel, Van der Sloot, & Van den Bosch, 2000) provide excellent fits to the empirical data as well as to the distributional patterns in CELEX.

The analogical form effect of the Right Constituent Family on the choice of the linker is surprising as the left constituent is usually taken to be the prime determiner (see, e.g., Booij, 1996; Mattens, 1970). First, etymologically, both *-en-* and *-s-* developed out of inflectional suffixes, i.e. markers for genitive singular or nominative plural. The linker *-en-* is still restricted to first constituents that select the suffix *-en-* for the formation of noun plurals. Second, there is experimental evidence that adding a linker to a first constituent may activate plural semantics (Schreuder, Neijt, van der Weijde, & Baayen, 1998). Third, phonologically, linkers belong to the first constituents of compounds. The linker always groups with the final syllable of the first constituent (e.g., *koning+s+kind* 'king's child'), even when the second constituent is separated from the first in contractions such as *varken+s- en schap+e+vlees* 'pork and mutton'. Finally, left constituents sometimes undergo vowel alternation in combination with a linker (compare *schip+breuk,* 'shipwreck', *scheep+s+werf,* 'shipyard'), suggesting that the left constituents and their linkers might also be interpreted as allomorphs. Considered jointly, these observations strongly suggest that the linker groups with the left constituent and that compounds with linkers are left-branching structures.

The strong analogical force of the Left Constituent Family reported by Krott, Baayen, & Schreuder (2001) is in line with the above considerations, while the weaker but statistically reliable analogical force of the Right Constituent Family that they report is surprising and requires further investigation. The aim of the present paper is to explore whether the observed analogical effect of the Right Constituent Family might not be an analogical effect based on the pure forms of the Right Constituent Families, but rather an analogical effect based on the semantic properties of the Right Constituents. Thus, we focus on the question whether it is the set of compounds sharing the Right Constituent with the target compound that forms the analogical basis for the choice of the linker or whether it is the set of compounds sharing the semantic class of the Right Constituent with the target compound that forms the analogical basis. Returning to the example of *schaap-?-oog,* the question is whether we should consider the set of compounds having *oog* as right constituent, or whether we should consider the set of compounds that have, for instance, a concrete noun as right constituent.

Van den Toorn (1982a) mentions several semantic factors that might be relevant. These factors fall into two types. First, the semantic class of a constituent might play a role. First constituents that are mass nouns, for instance, seem to occur predominantly without a linker (e.g., *papier+handel* 'paper trade'), though this is not always the case (*tabak+s+rook* 'tabacco smoke').

Second, the semantic relation between the constituents seems to have an influence on the choice of the linker. For example, if the first constituent is the logical object of the second constituent, the constituents tend to be connected without a linker (e.g. *boek+verkoper* 'bookseller', but again there are many exceptions, e.g. *gezin+s+planning* 'family planning'). We will restrict our focus to the first kind of semantic factors, the semantic class of the constituents.

Some preliminary evidence for an effect of the semantic class of the constituents has already been found in post-hoc simulation studies in which responses of participants have been modeled with TiMBL. The responses were produced in two cloze tasks which orthogonally varied the bias of the Left and Right Constituent Family (Krott, Baayen, & Schreuder, 2001). Simulation studies with TiMBL revealed optimal prediction accuracies when the analogy was based not only on the left constituent, i.e. the Left Constituent Family, but also on information concerning the semantic class of the right constituent. Prediction accuracy did not improve any further by additionally taking the Right Constituent Family into account. These results suggest that the form effect of the Right Constituent Family might indeed be a semantic effect. However, these post-hoc analyses are inconclusive by themselves and require supplementation by an independent factorial experiment explicitly addressing the potential role of semantic categories.

In what follows, we first present some lexical statistics concerning the relation between the use of linkers in Dutch compounds and the semantic class of the left and right constituents. Next, we discuss a factorial experiment designed to clarify the potential effect of semantic features on the choice of linkers in novel compounds, following which we reanalyze the experiments reported in Krott, Baayen, & Schreuder (2001) with respect to the role of the semantics of the left and right constituents.

2. Lexical statistics

In order to ascertain the effect of the semantics of both left and right constituents, we investigated the 6949 compounds in the families of the first two experiments reported in Krott, Baayen, & Schreuder (2001). For these compounds, we have annotated the left and right constituents with the following semantic categories: abstract versus concrete, and animate versus inanimate. Within the category of animate nouns, we distinguished between human versus animal, and within the category of inanimate we distinguished between plant versus other. Table 1 gives an overview over the distribution of linkers across these 6949 compounds when we partition these compounds according to the semantics of the first and second constituents. A partition into abstract and concrete first constituents reveals, for instance, that *-en-* prefers concrete first constituents. An independent partitioning according to the animacy of the first constituent shows that animate first constituents prefer *-en-* and that no linker is preferred for inanimate nouns. Partitionings according to the second constituents show different distributions.

Constituent	Semantic Class	-s-	en-	-Ø-
first				
	abstract	1801	172	1471
	concrete	559	1007	1939
	animate	274	510	195
	inanimate	2086	669	3215
second				
	abstract	1818	415	1497
	concrete	542	764	1913
	animate	157	79	345
	inanimate	2203	1100	3165

Table 1: Numbers of linking possibilities for different semantic classes of the left and right constituents of 6949 Dutch compounds.

A more informative way of summarizing the distribution of the linkers as a function of semantic categories is to construct a classification tree using a non-parametric technique, CART (Breiman, Friedman, Olshen, & Stone, 1984). CART is useful for classification problems with one or more predictor variables (here: the semantic class) and one response variable (here: the linker). The statistical model is fitted by binary recursive partitioning of the data, which means that the dataset is successively split up into increasingly homogeneous subsets with different values of the predictor variable (different semantic classes). Each split partitions the data into two subsets while maximizing the difference in the relative proportions of linkers. This process results in a classification tree.

Model selection in CART analyses is accomplished by means of cost-complexity pruning, a technique for finding the smallest (most parsimoneous) tree with low heterogeneity of the leaves. The left panel of Figure 1 plots the cross-validation score function. The horizontal axis plots the size of the classification tree, the vertical axis plots the corresponding deviance (calculated using 10-fold cross-validation). The deviance is a measure of average node heterogeneity. The upper axis shows the mapping between tree size and the cost-complexity parameter α (by increasing α, the size of the tree is penalized more heavily). We chose a quite conservative α of .0145, following the advice of Breiman et al. (1984). The resulting pruned tree is shown in the right panel of Figure 1. Table 2 lists the percentages of linkers for the leaves of the pruned tree as it is presented in Figure 1. The length of the vertical lines represents the amount of deviance accounted for by a particular split. The largest deviance and therefore the largest predictive power is given by the partition into abstract and concrete first constituents. The next highest deviance is reached by the split into first animate and inanimate constituents. The latter are further divided into plants and non-plants. The semantic class of the second constituent seems to be less relevant. The only predictive split appears to be the division into abstract nouns and human beings on the one side and concrete objects that are not human beings on the other side. Concreteness and animacy of the first (left) constituent emerge as strong predictors of the linkers in our data. For right constituents, it seems to matter to some extent whether they are abstract or concrete and whether they are human beings.

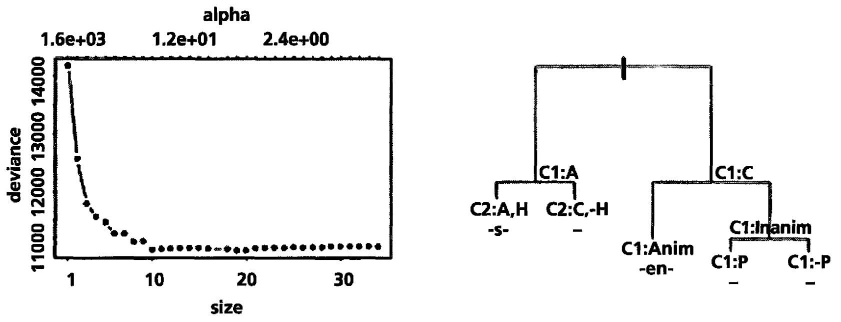

Figure 1: analysis of the semantic classes of the constituents of 6949 Dutch compounds as predictor variable and linker (-en-, -s-, and — in the case of a zero realization) as the response variable; left panel: plot of deviance versus tree size for sequences of subtrees; right panel: pruned classification tree; C1 = first constituent; C2 = second constituent; A = abstract; C = concrete; H = human being; Anim = animate; Inanim = inanimate; P = plant.

Node	*-en-* (%)	-∅-(%)	*-s-* (%)
C1:A;C2:A,H	4	36	60
C1:A;C2:C,-H	6	65	29
C1:C,Anim	52	20	28
C1:C,Inanim,P	44	56	1
C1:C,Inanim,-P	16	74	10

Table 2: Percentages of linkers for the leaves of the pruned tree of Figure 1 (see the legend of Figure 1 for further details of notation).

Summing up, the concreteness and animacy of the first constituent emerge from this analysis as reliable predictors of the linkers. The predictive force of the concreteness of the second constituent is weak. The next section addresses the question whether it is still strong enough to guide the decisions of participants in a cloze task.

3. A production experiment

3.1. Method

3.1.1. Materials

We constructed three sets of left constituents (L1, L2, L3) and four sets of right constituents (R1, R2, R3, R4). Each set contained 10 Dutch nouns. Given the results of the CART analysis, we considered animacy and concreteness as the main important semantic features and the feature 'human-being' as an additional feature potentially important for right constituents. Therefore, we chose the following experimental sets. The groups of left constituents contained abstract (L1), concrete-inanimate (L2), and concrete-animate nouns (L3). The sets of right constituents contained abstract (R1), concrete-inanimate (R2), concrete-human (R3) and concrete-animal nouns (R4). We made sure that all left constituents can be combined with the linker *-en-*. In addition, all constituents have a bias against being combined with a linker, i.e. at least 60% of all compounds in the Constituent Families occur without a linker (L1: mean: 82.7%; range: 64%—97%; L2: mean: 81.4%; range: 71.4%—100%; L3: mean: 81.0%; range: 63.6%—100%; R1: mean: 75.6%; range: 61.5%—100%; R2: mean: 83.8%; range: 60.0%—100%; R3: mean: 90.7%; range: 66.7%—100%; R4: mean: 92.6%; range:60.0%—100%).

Each of the three sets of left constituents (L1, L2, L3) was combined with the four sets of right constituents (R1, R2, R3, R4) to form pairs of constituents for new compounds in a factorial design with two factors: Semantic Class of the Left Constituent (abstract, concrete-inanimate, concrete-animate) and Semantic Class of the Right Constituent (abstract, concrete-inanimate, concrete-human, concrete-animal). None of these compounds is attested in the CELEX lexical database with a token frequency higher than zero. All have a high degree of semantic interpretability. Appendix A lists all experimental items. The 3 x 4 x 10 = 120 experimental items were divided over three lists. List 1 contained the compounds of the factorial combinations L1-R1, L2-R4, and L3-R3. List 2 contained the compounds of the combinations L1-R3, L2-R2 and L3-R4. List 3 contained the compounds of the combinations L1-R4, L2-R1, and L3-R2, and List 4 contained the compounds of the combinations L1-R2, L2-R3, and L3-R1. In this way, each participant saw a given constituent only once. We constructed a separate randomized list of the 3 x 10 = 30 pairs of compound constituents for each participant.

3.1.2. Procedure

The participants performed a cloze-task. An experimental list of items was presented to the participants in written form. Each line presented two compound constituents separated by two underscores. We asked the participants to combine these constituents into new compounds and to specify the most appropriate linker, if any, at the position of the underscores, using their first intuitions. Occasionally, the first constituent may change its form when it is combined with a linker (e.g., *schip* ('ship') appears as *scheep* in the compound *scheepswerf* ('shipyard')). The instructions clarified that these changes were not of interest and could be ignored. We told the participants that they were free to use *-en-* or *-e-* as spelling variants of the linker *-en-*. The experiment lasted approximately 10 minutes.

3.1.3. Participants

Sixty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment. All were native speakers of Dutch. The participants were divided into three groups, one for each experimental list.

3.2. Results and discussion

One participant produced unexpected, non-standard letter sequences for three stimuli. These responses were classified as errors and excluded from the statistical analyses. Table 3 summarizes the percentages of the responses for the twelve experimental conditions. Appendix A lists the individual words together with the counts of the responses.

The counts of *-s-*, *-en-*, and *-Ø-* responses for a given word are not independent—they always sum up to 20, the total number of participants. In order to bring the data in line with the requirements of standard multivariate methods, we divided the number of *-en-* and *-s-* responses by the number of *-Ø-* responses. A multivariate analysis of variance of the logarithms of the resulting ratios[5] revealed a main effect of the Semantic Class of the Left Constituent, but no effect of the Semantic Class of the Right Constituent, and no interaction of both factors (left semantic class: $F_2(2,108) = 16.8, p < .001$; right semantic class: F_2 $(3,108) = 1.1, p = .374$).

The way in which the Semantic Class of the left Constituent affects the responses of the participants is summarized in Figure 2. Responses with the linking *-s-* (solid line) occur predominantly with abstract left constituents. By contrast, *-en-* responses (dotted line) are least frequent with abstract constituents, but common for concrete, and even more common for animate concrete left constituents. Responses with *-Ø-* (dashed line) are slightly less common for concrete animate left constituents. This pattern of results is quite similar to the general pattern in the Dutch lexicon as summarized in Table 2 above.

Left Constituent		Right Constituent							
		abstract		concrete-inanimate		concrete-human		concrete-animal	
		%	#	%	#	%	#	%	#
abstract									
	en	28.5	57	34.0	68	31.0	62	30.0	60
	s	29.5	59	23.0	46	19.5	39	25.5	51
	Ø	42.0	84	43.0	86	48.5	97	44.5	89
inanimate									
	en	56.0	112	43.5	87	55.5	111	44.5	89
	s	4.5	9	4.5	9	1.5	3	9.0	18
	Ø	39.5	79	52.0	104	43.0	86	46.6	93
animate									
	en	75.0	150	77.5	155	54.0	108	52.0	104
	s	2.0	4	2.5	5	5.5	11	5.5	11
	Ø	23.0	46	20.0	40	40.5	81	42.0	84

Table 3: Percentages and numbers of selected linkers when varying the Semantic Class of the Left and Right Constituent.

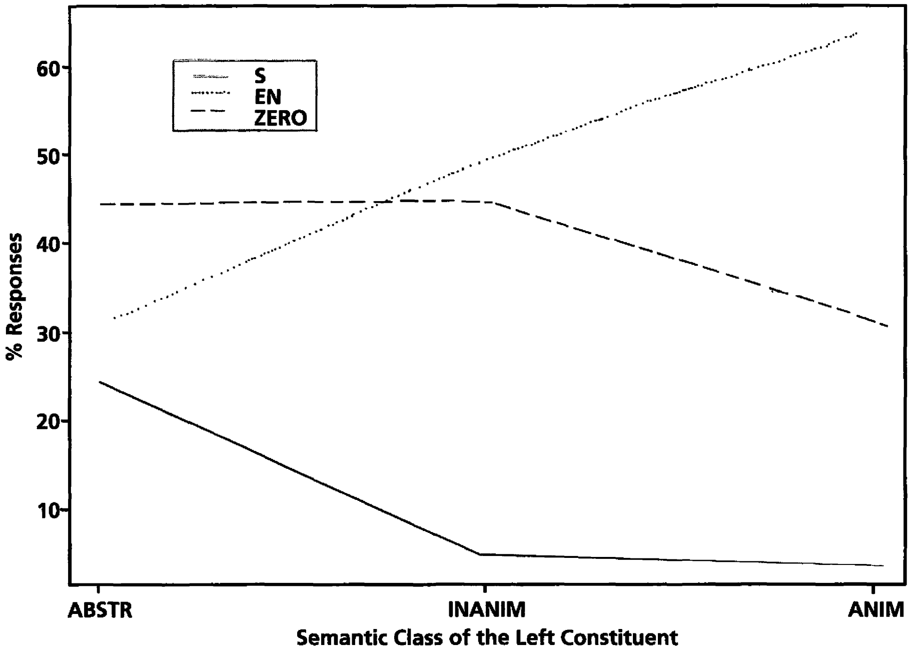

Figure 2: Percentages of -en-, -s- and -Ø- responses for different Semantic Classes of the left constituent (ABSTR: abstract; INANIM: inanimate; ANIM: animate).

Do the Left and Right Constituent biases co-determine the choice of the linker in addition to the semantic class of the Left Constituent? More specifically, does the absence of a semantic effect for the Right Constituent imply that no effect of the bias of the Right Constituent should be observable? A post-hoc multivariate analysis of covariance revealed reliable effects of both the Left and Right Constituent Bias in addition to the factorially established effect of the Left Semantic Class (Left Constituent Family: $F_2(2,109) = 6.5, p < .001$; Right Constituent Family: $F_2(2,109) = 2.5, p = .047$; Left Semantic Class: $F_2(2,109) = 18.6, p <.001$). We also observed an interaction of the Semantic Class of the Left Constituent and the Bias of the Left Constituent Family F_2 $(4,109) = 2.6$, $p = .009$). We conclude that, apparently, the Right Constituent Family is a factor in its own right that cannot be reduced to a semantic effect of the right constituent.

Recall that the CART analysis of the relation between the semantic categories and the linkers revealed a weak but reliable effect for the concreteness of the right constituent (Figure 1, Table 2). The present experimental results suggest that the variability in the lexicon is too large to allow individual language users to make use of the observed distributional pattern. It is possible that the present experimental paradigm is not sensitive enough to register potential semantic effects of the right constituent. However, given that it is sensitive enough to reveal a reliable effect for the Right Constituent bias and a clear effect of the semantics of the Left Constituent,we have to conclude that, at the very least, the effect of the Right Constituent bias is much stronger than the potential effect of the semantics of the Right Constituent.

These results raise the question whether the semantic effect reported by Krott, Baayen, & Schreuder (2001) on the basis of two cloze tasks mentioned in the introduction is reliable. A post-hoc logit analysis of the EN-experiment with semantic class as covariate revealed main effects for the Left and Right Constituent families (Left Constituent Family: $F_2(2,141) = 92.18, p < .001$; Right Constituent Family: ($F_2(2,141) = 11.68, p < .001$) as well as a main effect of the Semantic Class of the Left Constituent ($F_2(5,164) = 5.70, p < .001$). No such effect could be observed for the right constituent ($F_2(5,164) < 1$). As in the present experiment, a similar interaction between the Semantic Class of the left constituent and the Bias of the Left Constituent Family was visible ($F_2(8,141) = 2.22, p = .029$). Analyses of the S-experiment revealed the same pattern of results.[6] These post-hoc analyses parallel the results obtained in the present experiment and confirm that the Right Constituent Family bias cannot be reduced to a semantic effect. Apparently, the slight increase in prediction accuracy reported by Krott, Baayen, & Schreuder (2001) that they obtained using TiMBL is not robust, and, in fact, inclusion of the semantic information for the second constituent does not lead to a statistically significant improvement in performance in their experiments (EN-experiment: 86.6% versus 79.9%, $\chi^2_{(1)} = 2.74, p = .0977$; S-experiment: 88.4% versus 87.3%, $\chi^2_{(1)} = 02, p =.875$).

4. General discussion

This study addressed the question whether the Right Constituent Family affects the choice of linkers in Dutch noun-noun compounds, an analogical effect across complex words sharing constituents, or whether the semantic category of the right constituent is the crucial factor at issue. A statistical survey of 6949 Dutch compounds and the semantic categories of their constituents revealed that the concreteness or abstractness of the right constituent is a minor predictor of the linker compared to the semantic class of the left constituent. However, a factorial experiment using a cloze task revealed a reliable effect of the Left Semantic Class, but no effect whatsoever of the Right Semantic Class. A post-hoc analysis revealed clear effects of both the Left and Right Constituent Families and an Interaction of the Left Semantic Class and the Left Constituent Family. Re-analyses of the experiments reported by Krott, Baayen, & Schreuder (2001) yielded the same pattern of results.

The failure to find any influence of the Right Semantic Class in combination with the clearly observable robust effect of the Right Constituent Family falsifies our initial hypothesis that the effect of the Bias of the Right Constituent Family might in fact be an effect of the semantic class of the right constituent. We have to conclude that the choice of the linker in Dutch is analogically co-determined by the distribution of linkers in the set of compounds sharing the right constituent.

What then, is the morphological status of the linkers in Dutch? Clearly, Dutch linkers are not normal suffixes. Whether or not a suffix can be attached to a base word may depend on the phonological, morphological, and semantic properties of the base. But, to our knowledge, normal suffixes never depend on the properties of what follows to their right.

Although, as mentioned in the introduction, linkers resemble normal suffixes in their strong etymological, semantic, and phonological dependence on the left constituent, there is also evidence that they may not form very strong units with their left constituents. For instance, Kehayia, Jarema, Tsapkini, Perlak, Ralli, & Kadzielawa (1998) report that left constituents followed by linkers in Polish and Greek compounds are effective primes only when their combination occurs as a separate (inflected) word in the language. Without such support, left constituents followed by linkers do not prime, which is not what one would expect if the linker and the left constituent would form a unit at some level of representation in the mental lexicon.

The unexpected role for the right constituent on the choice of the linker in Dutch may be due to the absence of a clear functional role for linkers in this language. From a historical perspective, the following sequence of events may have occurred. Initially, various nominal case endings occurred in compounds. Many such compounds, especially those enjoying a frequent use, were probably stored in the mental lexicon (Van Jaarsveld & Rattink, 1988). Following the loss of the nominal case system, the only place where nominal case endings were retained in great numbers was nominal compounding, where they persisted thanks to their being stored in the mental lexicon. In the absence of a clear functional role, each new generation of language learners is faced with the problem of having to use the standard forms as in current use in the community without having recourse to a clear-cut systematicity for predicting the correct form for existing words and for the formation of new compounds. In such a situation, all possible sources of information might be useful. One such source of information might be the semantic classes of the constituents. In modern Dutch, the abstractness versus the concreteness of the modifying constituent might be a growing source of systematicity for a functional re-interpretation of the linkers from a case-marker to a marker of semantic class. But we suspect that as long as such a process of re-interpretation has not been fully completed, all available information, including the distributional information contained in the Right Constituent Family, is used to optimize the chances of the learner to conform to the current norms in the society.

Note, finally, that there are two ways in which our data on the analogical nature of the choice of linkers in Dutch can be interpreted. On the one hand, it may be argued that this kind of analogical word formation is typical for language domains that have become more or less chaotic due to historical change. On the other hand, it may be that analogy is much more pervasive and underlies phenomena traditionally analyzed as rule-governed. From this second perspective, the Dutch linkers provide an excellent window on the general properties of analogy. Future research will have to clarify the merits of these contrasting views.

Notes

* This study was financially supported by the Dutch National Research Council NWO (PIONIER grant to the fourth author), the University of Nijmegen (The Netherlands), and the Max Planck Institute for Psycholinguistics (Nijmegen, The Netherlands). Requests for reprints should be addressed to Andrea Krott, Interfaculty Research Unit for Language and Speech & Max Planck Institute for Psycholinguistics, P.O.Box 310, 6500 AH Nijmegen, The Netherlands. E-mail: akrott@mpi.nl.

[1] Of all noun-noun compounds listed in the *CELEX* lexical database (Baayen, Piepenbrock, & Gullikers, 1995) 35% are formed with a linker.

[2] A description of spelling variants *-en-* and *-e-* can be found, e.g., in the Woordenlijst (1995).

[3] For a complete list of phonological and morphological rules, see Krott, Schreuder, & Baayen (in press). Semantic rules were not taken into account because semantic information in CELEX is not available.

[4] The rules in the literature focus on the properties of the left constituent and never consider the right constituent as a possible factor. Note that the effect of the Right Constituent Family cannot be accounted for by means of rules that would be sensitive to the phonological or morphological properties of the right constituent. A statistical survey of 22966 Dutch compounds shows that the onset and, if present, the prefix of the second constituent can be used to predict only 64.5% of the linkers, which is identical to the percentage of compounds with no linker (the default) and thus could be attained by always chosing the linker with the a-priori maximum likelihood.

[5] Counts equal to zero were set to 0.1 before taking the logarithm.

[6] Semantic Class of the Left Constituent: $F_2(5,173) = 3.78, p = .003$); Semantic Class of the Right Constituent: $F_2(4,173) < 1$; Left Constituent Family: $F_2(2,153) = 124.65, p < .001$; Right Constituent Family: $F_2(2,153) = 9.34, p < .001$; Interaction between the Semantic Class of the Left Constituent and the Bias of the Left Constituent Family ($F_2(4,153) = 9.64, p < .001$.

Address of the Authors:

Andrea Krott, Loes Krebbers, Robert Schreuder & R. Harald Baayen
Max Planck Institute for Psycholinguistics &
Interfaculty Research Unit for Language and Speech
P.O.Box 310
NL - 6500 AH Nijmegen
The Netherlands

References

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

Booij, G. & Van Santen, A.: 1995, Morfologie. De woordstructuur van het Nederlands (Morphology: the word structure of Dutch), Amsterdam University Press, Amsterdam.

Booij, G. E.: 1996, Verbindingsklanken in samenstellingen en de nieuwe spellingregeling (Linking phonemes in compounds and the new spelling system), Nederlandse Taalkunde 2, 126—134.

Breiman, L., Friedman, J., Olshen, R. & Stone, C.: 1984, Classification and Regression Trees, Wadsworth International Group, Belmont, California.

Daelemans, W., Zavrel, J., van der Sloot, K. & van den Bosch, A.: 2000, TiMBL: Tilburg memory based learner reference guide. Version 3.0, Technical Report ILK 00-01, Computational Linguistics Tilburg University.

Dressler, W. U., Libben, G., Stark, J., Pons, C. & Jarema, G.: in press, The processing of interfixed German compounds, to appear in G. E. Booij and J. Van Marle (eds), Yearbook of Morphology 1999, Kluwer, Dordrecht.

Fuhrhop, N.: 1998, Grenzfälle morphologischer Einheiten (Border cases of morphological units), Stauffenburg, Tuebingen.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. & van den Toorn, M.: 1997, Algemene Nederlandse Spraakkunst (General Grammar of Dutch), Martinus Nijhoff, Groningen.

Herrera, Z. E.: 1995, Palabras Estratos y Representaciones: Temas de Fonologia Lexica en Zoque, El Colegio de Mexico.

Kehayia, E., Jarema, G., Tsapkini, K., Perlak, D., Ralli, A. & Kadzielawa, D.: 1999, The role of morphological structure in the processing of compounds: The interface between linguistics and psycholinguistics, Brain and Language 68, 370—377.

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

Krott, A., Schreuder, R. & Baayen, R. H.: in press, Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds, to appear in R. Skousen (Ed.), Advances in the Analogical modeling of language.

Mattens, W. H. M.: 1970, De indifferentialis. Een onderzoek naar het anumerieke gebruik van het substantief in het algemeen bruikbaar Nederlands (The indifferential. An inquiry into the anumeric use of the noun in general Dutch), Van Gorgum, Assen.

Mattens, W. H. M.: 1984, De voorspelbaarheid van tussenklanken in nominale samenstellingen (The predictability of linkers in nominal compounds), De nieuwe taalgids 7, 333—343.

Schreuder, R., Neijt, A., van der Weide, F. & Baayen, R. H.: 1998, Regular plurals in Dutch compounds: linking graphemes or morphemes?, Language and Cognitive Processes 13, 551—573.

Skousen, R.: 1989, Analogical Modeling of Language, Kluwer, Dordrecht.

Toorn, M. C. v. d.: 1981a, De tussenklank in samenstellingen waarvan het eerste lid een afleiding is (The linker in compounds of which the first constituent is a derived form), De nieuwe taalgids 74, 197—205.

Toorn, M. C. v. d.: 1981b, De tussenklank in samenstellingen waarvan het eerste lid systematisch uitheems is (The linker in compounds of which the first constituent systematically is non-native), De nieuwe taalgids 74, 547—552.

Toorn, M. v. d.: 1982a, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I (Tendencies for rule-based analyses of the linker in nominal compounds I), De nieuwe taalgids 75(1), 24—33.

Toorn, M. v. d.: 1982b, Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II (Tendencies for rule-based analyses of the linker in nominal compounds II), De nieuwe taalgids 75(2), 153—160.

Van Jaarsveld, H. & Rattink, G.: 1988, Frequency effects in the processing of lexicalized and novel nominal compounds, Journal of Psycholinguistic Research 17, 447—473.

Woordenlijst: 1995, Woordenlijst van de Nederlandse taal 1995 (Word list of the Dutch language), Sdu Uitgevers and Standaard Uitgeverij, 's-Gravenhage.

Appendix A

Materials for the Experiment: left constituent and right constituent (number of -s- responses, number of -en- responses, number of -Ø- responses).

L1-R1: Left Constituent: abstract; Right Constituent: abstract
taal staat (1, 7, 12); seizoen zang (11, 8, 1); loon gebrek (12, 2, 6); brand geluid (2, 3, 15); dienst toeval (2, 8, 10); vorm feest (2, 9, 9); symbool energie (4, 8, 8); naam toeslag (8, 4, 8); kracht maaltijd (8, 5, 7); contract opslag (9, 3, 8)

L1-R2: Left Constituent: abstract; Right Constituent: concrete-inanimate
dienst vliegtuig (0, 8, 12); taal fles (1, 8, 11); seizoen jurk (17, 3, 0); symbool vork (3, 12, 5); loon altaar (3, 4, 13); vorm tapijt (3, 8, 9); brand nagel (4, 1, 15); kracht muts (5, 10, 5); naam standbeeld (5, 10, 5); contract telefoon (5, 4, 11)

L1-R3: Left Constituent: abstract; Right Constituent: concrete-human
dienst consulent (0, 12, 8); taal heilige (0, 4, 15); brand leidster (1, 7, 11); seizoen zuster (15, 5, 0); contract producent (2, 4, 14); kracht idioot (2, 6, 12); symbool machinist (2, 6, 12); vorm redacteur (4, 4, 12); naam handelaar (5, 13, 2); loon violist (8, 1, 11)

L1-R4: Left Constituent: abstract; Right Constituent: concrete-animal
dienst vogel (0, 3, 17); taal aap (0, 9, 11); symbool baars (1, 15, 4); loon uil (10, 0, 10); seizoen mees (15, 4, 1); vorm aal (3, 9, 8); contract gans (4, 4, 12); brand kat (4, 5, 11); naam slak (6, 8, 6); kracht os (8, 3, 9)

L2-R1: Left Constituent: concrete-inanimate; Right Constituent: abstract
spier maaltijd (0, 10, 10); fiets staat (0, 11, 9); huis toeval (0, 11, 9); kaars energie (0, 14, 6); schoen geluid (0, 14, 6); arm gebrek (1, 14, 5); duim opslag (1, 7, 12); tand zang (2, 17, 1); trein feest (2, 6, 12); boot toeslag (3, 8, 9)

L2-R2: Left Constituent: concrete-inanimate; Right Constituent: concrete-inanimate
kaars vork (0, 12, 8); tand fles (0, 15, 5); huis jurk (0, 3, 17); fiets vliegtuig (0, 8, 12); spier standbeeld (1, 10, 9); schoen nagel (1, 7, 12); arm tapijt (1, 9, 10); duim altaar (2, 11, 7); boot telefoon (2, 4, 14); trein muts (2, 8, 10)

L2-R3: Left Constituent: concrete-inanimate; Right Constituent: concrete-human
duim handelaar (0, 15, 5); tand producent (0, 15, 5); arm zuster (0, 18, 2); kaars idioot (0, 18, 2); trein redacteur (0, 3, 17); fiets machinist (0, 6, 14); huis consulent (0, 8, 12); spier violist (0, 8, 12); schoen heilige (1, 12, 7); boot leidster (2, 8, 10)

L2-R4: Left Constituent: concrete-inanimate; Right Constituent: concrete-animal

fiets vogel (0, 6, 14); huis baars (0, 6, 14); arm slak (0, 9, 11); kaars uil (1, 11, 8); spier os (1, 11, 8); trein gans (1, 7, 12); duim mees (2, 11, 7); schoen aal (2, 12, 6); tand aap (3, 12, 5); boot kat (8, 4, 8)

L3-R1: Left Constituent: concrete-animate; Right Constituent: abstract

weduwe toeslag (0, 13, 7); vis feest (0, 14, 6); marxist geluid (0, 19, 1); prins zang (0, 19, 1); koningin staat (0, 20, 0); christen energie (0, 8, 12); wees toeval (0, 9, 11); gast opslag (1, 14, 5); leerling maaltijd (1, 17, 2); vorst gebrek (2, 17, 1)

L3-R2: Left Constituent: concrete-animate; Right Constituent: concrete-inanimate

vis altaar (0, 13, 7); wees telefoon (0, 13, 7); gast vliegtuig (0, 16, 4); marxist tapijt (0, 19, 1); prins muts (0, 20, 0); christen jurk (0, 7, 13); weduwe standbeeld (1, 14, 5); vorst nagel (1, 18, 1); koningin fles (1, 19, 0); leerling vork (2, 16, 2)

L3-R3: Left Constituent: concrete-animate; Right Constituent: concrete-human

prins machinist (0, 14, 6); vis consulent (0, 6, 14); wees zuster (0, 6, 14); vorst heilige (1, 10, 9); gast idioot (1, 12, 7); leerling leidster (1, 16, 3); koningin producent (1, 17, 2); weduwe handelaar (1, 8, 11); marxist redacteur (2, 13, 5); christen violist (4, 6, 10)

L3-R4: Left Constituent: concrete-animate; Right Constituent: concrete-animal

gast baars (0, 13, 7); vorst slak (0, 13, 7); prins vogel (0, 18, 2); koningin gans (0, 20, 0); wees uil (0, 4, 16); vis aal (0, 5, 15); marxist mees (1, 16, 3); weduwe kat (1, 7, 11); christen os (3, 5, 12); leerling aap (6, 3, 11)