Analogy in morphology: modeling the choice of linking morphemes in Dutch

Andrea Krott, R. Harald Baayen & Robert Schreuder

Interfaculty Research Unit for Language and Speech

University of Nijmegen

The Netherlands

SUBMITTED, PLEASE DO NOT QUOTE

Address all correspondence to:

Andrea Krott

Interfaculty Research Unit for Language and Speech &

Max Planck Institute for Psycholinguistics

P.O.Box 310, 6500 AH Nijmegen

The Netherlands

ABSTRACT

This study argues that a productive, but not fully-regular morphological phenomenon, the choice of linking morphemes in Dutch nominal compounds, is based on analogy. In Dutch, a linking -s- or -en- can appear between the constituents of a nominal compound. We present production experiments which reveal strong evidence that the choice of linking morphemes in novel compounds is analogically determined by the distribution of linking morphemes in what we call the 'constituent families'. A 'constituent family' is the set of existing compounds that share the first (or second) constituent with the novel compound. A further experiment shows that in the case of derived pseudo-words as first constituents, it is the family of the suffix which influences the choice of the following linking morpheme. In addition to these experiments, we present computational simulation studies in which the choices made by participants in our experiments are predicted with a high degree of accuracy using a machine-learning algorithm for analogy. These studies support the status of the constituent family as the primary basis for analogical prediction. Finally, we outline a psycholinguistic model for analogy in the mental lexicon that does not give up symbolic representations and, at the same time, captures non-deterministic variation.

INTRODUCTION

Morphological variation can often be captured by simple rules. Consider, for example, the realization of the regular plural of English nouns, which appears in three different forms, /ɪz/, /z/, and /s/. These three variants can be predicted on the basis of the phonological form of the base word. The plural is pronounced /ɪz/ after bases ending in sibilants (e.g., horses), /z/ after bases ending in vowels and voiced segments other than /z/, /ʒ/, and /dʒ/ (e.g., beds), and it is pronounced /s/ after bases ending in voiceless segments other than /s/, /ʃ/, and /tʃ/ (e.g., months).

In addition to this kind of regular variation, there are morphological domains where the choice between alternative realizations is less predictable. One such domain is the analysis of linking elements in compounds, which are also referred to as connectives, interfixes, linkers, or linking morphemes. Linking elements occur in various languages across different language families. In English, linking elements are extremely rare. We know of only a few examples, all built with the head word man: marksman, sportsman, craftsman, kinsman, tradesman, and spokesman. The last example, in which the -s- appears without any possible semantic function, best illustrates the phenomenon of linking elements. In some languages, linking elements can be fully predicted on the basis of the phonological characteristics of the preceding (and/or the following) constituent. For instance, Zoque, a Mixe-Zoquean language spoken in Mexico, has a nominal compound formation in which the linking element is a vowel that is identical to the vowel in the preceding syllable. However, in many other languages such clear rules cannot be formulated. For example, Kabardian (North Caucasian) has the linking elements -ah-, -m-, -n-, and -r-, which tend to be obligatory in some morphological

contexts and optional in others (Kuipers, 1960:78–80). In Indo-European, the Germanic languages are especially rich in non-predictable or only partly predictable variation in the use of linking elements (e.g., Danish, Norwegian, Swedish, and German). The distribution of the two main linking elements in Dutch, -en- and -s-, is likewise only partially predictable by rule.

The systematicities governing the selection of linking morphemes is a longstanding unsolved problem in the morphology of Dutch and many other Germanic and Non-Germanic languages. It is an issue that has hardly received attention in the generative tradition,[1] with the exception of Botha (1968), even though it is a problem that receives discussion in any good reference grammar (e.g., Haeseryn, Romijn, Geerts, de Rooij, & Van den Toorn, 1997; de Haas & Trommelen, 1993).

A first goal of the present study is to show that the distribution of linking morphemes in Dutch noun-noun compounds can be accounted for by means of a formal computational model of analogy with a higher degree of observational adequacy than can be achieved by means of the rules proposed in the literature. Our conclusions are based on both surveys of existing compounds in the Dutch lexicon as well as on the choices for linking morphemes in novel compounds as produced by participants under strict experimental conditions.

A second goal is to contribute to the discussion in the current literature about the nature of morphological rules, whether such rules are symbolic in nature (Clahsen, 1999; Marcus, Brinkman, Clahsen, Wiese, & Pinker, 1995; Pinker, 1991, 1997) or whether rules are an epiphenomenon of distributed storage in connectionist networks (Seidenberg, 1987; Seidenberg & Hoeffner, 1998; Plunkett & Juola, 1999; Rueckl, Mikolinski, Raveh, Miner & Mars,

1997). The phenomenon that we are dealing with is interesting in the sense that it is fully productive and yet not completely regular. As such, it poses a serious challenge to proponents of symbolic rule systems. At the same time, we will show that it is possible to predict non-deterministic aspects of human cognition without necessarily making use of distributed connectionist networks. In this sense, our present analogy-based approach provides an alternative to both symbolist and connectionist approaches to cognition.

The notion of analogy as we use it in this paper is different from its two traditional interpretations in linguistics. First, analogy is often contrasted with rules, with regular novel forms being formed by rules, and exceptional novel forms being built by analogy to individual examples (e.g., brunch by analogy to smog, see, e.g., Anshen & Aronoff, 1988). Second, analogy can also be understood as the initial basis for the acquisition of rules. In this view, analogical learning might be involved in determining the conditions under which a rule applies. But once a rule is established, the instances which led to the rule would then be irrelevant, and would not be kept in memory.

Our use of the term analogy differs from these two interpretations in the following ways. First, the kind of analogy with which we are concerned is not the kind of analogy that occasionally leads to exceptional new creatively coined words such as brunch. Instead, we are concerned with the regular phenomena that are traditionally described by means of linguistic rules. Following Skousen (1989) and Daelemans, Zavrel, Van der Sloot, & Van den Bosch (1999), we adopt a formal and computationally tractable definition of analogy that offers a new way of understanding the way in which linguistic rules actually work. Second, we hypothesize that, at least in the domain of morphological processing, there are

no rules that are formed on the basis initially stored examples of complex words, with the initial exemplars fading from memory. Instead, we assume that many fully regular complex words, both inflected and derived, remain available in the mental lexicon (e.g., Bertram, Laine, Baayen, Schreuder, 1999; Bertram, Schreuder, & Baayen, 2000; Baayen, Dijkstra, & Schreuder, 1997; Sereno & Jongman, 1995; Sandra, Frisson, & Daems, 1999; Taft, 1979; Baayen, Schreuder, De Jong, Krott, in press), and serve as exemplars for the analogical formation of novel forms. In other words, we hypothesize that rules are essentially analogical in nature (De Saussure, 1966).

In what follows, we first describe the problem of the systematicities underlying the distribution of linking morphemes in Dutch, and we show that the notion of default rules that has figured prominently in recent discussions (Marcus et al., 1995; Clahsen, 1999) is not applicable to this phenomenon. In the next section, we present the results of three production experiments, which show that, the substantial variation in the choice of linking morphemes notwithstanding, Dutch native speakers tend to converge on the same linking elements for novel compounds. These experiments show, furthermore, that the choice of a linking element for a novel compound is strongly influenced by the distribution of linking elements in the set of existing compounds sharing the first or second constituent with the novel compound (e.g., fiets 'bike' in fiets-pad 'cycle path' and fiets+bel 'bicycle bell', and winkel 'shop' in schoen+winkel 'shoe shop' and hoed+en+winkel, 'hat+plur+shop', 'hat shop'). We will refer to these sets of compounds as constituent families.

In the subsequent section, we will show that the notion of analogy based on constituent families can be formalized computationally, and that this allows us to predict the distri-

bution of linking morphemes in the Dutch lexicon and also to predict the performance of our experimental participants. In the general discussion, we outline how the computational model can be mapped onto a psycholinguistically more realistic spreading activation model along the lines of Schreuder & Baayen (1995).

LINKING MORPHEMES IN DUTCH: NO RULES BUT TENDENCIES

In this section, we describe the distributional properties of the linking elements in Dutch and their linguistic status. The two main linking elements in Dutch noun-noun compounds are -s- and -en-. The latter is occasionally realized in the orthography as -e-. Both -en- and -e- are pronounced as schwa in standard Dutch. As the present study focuses on the production of linking elements, we do not distinguish between the two orthographic realizations.

There is a long-standing discussion about the status of these linking elements. Are they just meaningless letters or do they carry semantic information? Both -s- and -en- are homographic with the two productive plural suffixes of Dutch nouns.[2] The linking element -en- may only appear after left constituents that themselves pluralize with -en. The linking element -s- is not constrained in the same way. It may appear following constituents with which it does not form a plural. There is evidence that -en- marks plurality in compounds, as shown by Schreuder, Neijt, Van der Weide, & Baayen (1998). Neijt, Baayen, & Schreuder (in preparation) show that, depending on the first constituent, the -s- may also convey plural semantics. In the light of this evidence, we will henceforth refer to -en- and -s- as linking morphemes rather than linking elements. Note, however, that the question whether the -s- and -en- forms in Dutch compounds are indeed completely identical to the Dutch plural

suffixes is not what is at issue in the present study. Our aim here is to come to grips with the distribution of these forms irrespective of their morphological status.

The literature on linking morphemes in Dutch compounds has attempted to capture the distribution of linking morphemes by means of rules operating at the levels of phonology, morphology, and semantics (see, e.g., Van den Toorn, 1981a; 1981b; 1982a; 1982b; Mattens, 1984). An example of a phonological rule is the constraint that after first constituents ending in a vowel, or ending in a schwa followed by a sonorant, or ending in a liquid followed by /k/ or /m/ (thee+bus 'tea box'; meubel+zaak 'furniture shop'), linking morphemes are not allowed. This rule is not without exceptions, however, as shown by a compound such as pygmee+en+volk, 'pygmy+plur+people', 'pygmy people'.

At the morphological level, particular affixes show preferences for specific linking morphemes. For instance, the diminutive suffix -je is always followed by the linking -s- in compounds (plaat+je+s+boek, picture+diminutive+plur+book, 'small pictures book'). Other morphemes show strong preferences, such as the suffix -heid, '-ness', which appears predominantly with -s-, but occasionally without a linking morpheme and rarely with -en-.

At the level of semantics two different kinds of constraints have been observed. First, the semantics of the first constituent may render the use of a linking morpheme unlikely. For instance, mass nouns are not followed by linking morphemes (e.g., papier+handel 'paper trade'; exception: tabak+s+rook, 'tabacco+genitive+smoke', 'tabacco smoke'). Conversely, the linking morpheme -en- often occurs when the first constituent of a compound has a plural interpretation (Haeseryn et al., 1997: 685; Schreuder et al., 1998): boek+en+kast, 'book+plur+case', 'book case', krent+en+brood, , 'currant+plur+bread', 'currant bread',

exception boek+handel, 'book shop'. Semantic factors may interact with the morphological structure of the first constituent. For instance, first constituents ending in -er denote human agents or objects. For human agents one tends to find the linking -s-, as in duik+er+s+ziekte, 'dive+er+plur+sickness', 'decompression sickness', while for inanimate objects one tends to find no linking morpheme, as in straal+jager+piloot, 'stream+hunt+er+pilot', 'fighter jet pilot'. These rules are also not without exceptions (e.g., leraar+en+opleiding ('teacher+plur +education', 'education of teachers')) (see Mattens 1984). Second, the semantic relation between the two constituents has also been argued to codetermine the choice of the linking morpheme. For instance, copulative compounds such as man+wijf, 'man+bitch', 'mannish woman' never take a linking morpheme. Similarly, compounds in which the first constituent is the object of a de-verbal agent or action noun to its right also tend to resist insertion of linking morphemes (boek+verkoper 'book seller'; exception: weer+s+verwachting, 'weather+genitive+expectation', 'weather forecast').

A final property of linking morphemes in Dutch is that they evidence a certain amount of variability. Fo instance, the word 'spelling change' has two translation equivalents in Dutch, spelling+verandering and spelling+s+verandering. Even for a single speaker, forms such as these appear to be in free variation.

Summing up, first constituents seem to have the strongest influence on the choice of linking morphemes, phonologically, morphologically, and semantically. The second constituent plays a minor role, being a codeterminant of the semantic relation between the two constituents. The numbers of exceptions to the rules describing the distribution of linking morphemes are so large that Van den Toorn (1982) has argued that we are dealing with

tendencies rather than with real rules.

It is important to note that the distribution of the linking morphemes in Dutch does not lend itself to an analysis in terms of a set of rules including a default rule. In such a system of rules, a series of positively specified cases is supplemented by a general case, the default, for which a simple and straightforward definition of its input domain (in the sense of Van Marle, 1985) cannot be given.[3] Focussing on the phonological rules for the distribution of Dutch linking morphemes, we observe only negative specifications: Linking morphemes do not appear following left constituents that end in a vowel, in a schwa followed by a sonorant, or in a liquid followed by /k/ or /m/. Crucially, the notion of a default, covering those words that do not fall under the negatively specified input domains, does not make sense for Dutch linking morphemes, as it does not have any predictive power with respect to the appropriate linking morpheme. Thus, words falling under the default, i.e. words that do not end in a vowel, in a schwa followed by a sonorant, or in a liquid followed by /k/ or /m/, can still appear in a compound with no linking morpheme, with -s- or with -en-. Clearly, none of these three possibilities can be the default choice. Turning to the level of morphology, we again find that the notion of a default is not applicable, as each suffix has its own stronger or weaker preferences. Similarly, at the level of semantics, we only observe random subgeneralizations without a well specified overall default. In spite of the absence of a rule system with a default, speakers of Dutch nevertheless have strong intuitions about which linking morpheme is appropriate for novel compounds.

PRODUCTION EXPERIMENTS

In this section, we address two related questions. First, to what extent do native speakers of Dutch agree about which linking morphemes are most appropriate to use in novel compounds? How much variability can be observed given the strong intuitions of native speakers as to what might be the appropriate choice? Second, what factors underlie these strong intuitions? We shall see that there is indeed strong agreement about which linking morpheme is most appropriate. As to the factors underlying the choice of linking morphemes, we shall see that the existing compounds sharing the left (or right) constituent with the target compound form perhaps the most important factor of all. In what follows, we will refer to these compounds as the left and right constituent families of such a target compound. An individual compound in such a family will be referred to as a constituent family member.

The next section presents experimental evidence for the important role of the constituent families for the linking morphemes -en- and -s-. The following section investigates the relevance of the morphological structure of the first constituent. We have not explicitly included semantic and phonological factors in our experimental design. However, we will show that analogical modeling of the experimental data yields slightly better results when semantic properties of the constituents are also taken into account. Including phonological information results in slightly worse performance.

The constituent family effect

The next two subsections present experiments studying the effect of the constituent family on the choice of the linking morphemes -en- and -s-.

Experiment 1: The linking morpheme -en-

If the choice of linking morphemes in novel compounds were based simply on the distribution of the linking morphemes in the lexicon as a whole, one would expect speakers to choose not to use a linking morpeme in roughly 7 out of 10 cases: 69% of all compounds listed in the celex lexical database (Baayen, Piepenbrock, & Gullikers, 1995) appear without any linking morpheme. Their second best guess would then be -s-, which occurs in 20% of the compounds in this database, and their least probable bet would be -e(n)- (11%). In the light of the linguistic description of the distribution of -en- and -s- presented in the previous section, this simple guessing behavior is unlikely. On the other hand, the linguistic rules that have been formulated tend to have so many exceptions that their explanatory value is called into question as well. In what follows, we explore the hypothesis that native speakers of Dutch base their choice on the relative frequencies of the linking morphemes as realized not in the lexicon as a whole, but in the restricted sets comprising the constituent families of individual compounds.

Method

Materials. We constructed three sets of left constituents (L1, L2, L3) and three sets of right constituents (R1, R2, R3). Each set contained 21 nouns. The constituents of L1 and R1 had constituent families with as strong a bias as possible towards the linking morpheme -en-. Conversely, L3 and R3 showed a bias as strong as possible against -en-, though we

made sure that these constituents form their plural with the suffix -en so that a linking -en- is possible. The sets L2 and R2, the neutral sets, contained nouns with families without a clear preference for or against -en-. We used the celex lexical database (Baayen et al. 1995) to determine the constituent families of the constituents in these six sets. Compounds with a token frequency of zero in a corpus of 42 million words were not included.

The constituents in the L1 set had constituent family members of which at least 70% contained the linking morpheme -en-. The mean number of compounds in these families was 12.5 (range 5–43). Their mean token frequency was 149.2 per 42 million wordforms (range 58–439). The range of choices for R1 constituents was more restricted. The constituents in the R1 set therefore had constituent family members of which at least 60% contained the linking morpheme -en-. The mean number of compounds in these families was 3.6 (range 2–7). Their mean token frequency was 49.1 per 42 million wordforms (range 20–119). Neutral left constituents are rare. The neutral set L2 included left constituents whose families contained between 35% and 65% compounds with the linking morpheme -en-. These families had a mean number of compounds of 8.3 (range 3–24) and a mean token frequency of 136.3 per 42 million wordforms (range 15–439). The constituents in the R2 set had constituent family members of which 40% to 60% contained the linking morpheme -en-. These families had a mean number of compounds of 5.3 (range 3–15) and a mean token frequency of 66.7 per 42 million wordforms (range 8–192). The remaining sets L3 and R3, the groups with a bias against -en-, contained constituents whose family members never have a linking -en-. There were in the mean 25 (range 11–66; L3) and 17.9 (range 10–47; R3) family members respectively. Their mean token frequency was 573.7 (range 98–2650; L3) and 349.8 (range

47–2290; R3). These are the maximal contrasts that allowed us to select 21 constituents for each experimental set.

Each of the three sets of left constituents (L1, L2, L3) was combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors: Bias in the Left Position (Positive, Neutral, and Negative) and Bias in the Right Position (Positive, Neutral, and Negative). 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 9 x 21 = 189 experimental items were divided over three lists. List 1 contained the compounds of the factorial combinations L1-R1, L2-R3, and L3-R2. List 2 contained the compounds of the combinations L1-R2, L2-R1 and L3-R3, and List 3 contained the compounds of the combinations L1-R3, L2-R2, and L3-R1. In this way, each participant saw a given constituent only once. We constructed a separate randomized list of the 3 x 21 = 63 compound constituent pairs for each participant.

Procedure. The participants performed a cloze-task. The 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 linking morpheme, 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 linking morpheme (e.g., schip ('ship') appears as scheep in the compound scheepswerf ('shipyard')). The instructions made clear 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 linking morpheme -en-. The experiment lasted approximately 15 minutes.

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. Each group was asked to complete one of the three experimental lists.

Results and discussion

Occasionally, participants filled in a question mark or a letter sequence other than a linking morpheme. Such responses were counted as errors. The overall error rate was extremely low (0.05%), which allowed us to include all participants and all items in the data analysis. Table 1 summarizes the percentages of en responses versus other responses for the nine experimental conditions. Appendix A lists the individual words together with the absolute numbers of en and not en responses.

A by-item logit analysis (see, e.g., Rietveld & Van Hout, 1993; Fienberg, 1980) of the en and not en responses revealed a main effect of Bias in the Left Position ($F(2,180) = 119.3$, $p < 0.0001$), a main effect of Bias in the Right Position ($F(2,180) = 12.8$, $p < 0.0001$), and no interaction of the Bias in both positions ($F(4,180)<1$). Although the Neutral Bias condition for the right constituents led to slightly higher numbers of en responses than the Positive

Table 1:

Percentages of selected linking morphemes when varying bias for -en- (Positive, Neutral, and Negative) in the left and right compound position.

		Right Position					
Left Position		Positive		Neutral		Negative	
Positive							
	en	94.8	(11.2)	96.4	(6.7)	87.4	(15.3)
	not en	5.2	(11.2)	3.6	(6.7)	12.6	(15.3)
	other	0		0		0	
Neutral							
	en	75.0	(23.7)	81.9	(15.5)	58.3	(26.9)
	not en	25.0	(23.7)	18.1	(15.5)	41.2	(26.9)
	other	0		0		0.5	
Negative							
	en	18.1	(19.1)	18.8	(19.9)	6.0	(7.7)
	not en	81.9	(19.1)	81.2	(19.9)	94.0	(7.7)
	other	0		0		0	

Note. Standard deviations between parentheses.

Bias condition, the difference between these two conditions is not reliable ($F(1,120) = 1.1$, $p = 0.2974$).

The upper panel of Figure 1 shows the effects of both biases on the percentage of *en* responses. Bias has a larger effect on the Left Position (a difference of roughly 80% between the Positive and Negative conditions) than on the Right Position (a difference of roughly 15%). This result reflects an asymmetry in the distribution of the linking elements in Dutch that is also mirrored in our experimental design. Figure 2 illustrates this asymmetry for the families of left and right constituents of compounds with the linking morpheme *-en-*. The left panel is a scattergram for the left constituents. It represents each of the 4320 constituents by a dot in the plane spanned by the number of compounds with *-en-* in which it appears (horizontal axis) and the number of compounds without *-en-* in which it appears (vertical axis). Note that the points are scattered along the two axes, indicating that there are many left constituents that occur predominantly either with *-en-* or without *-en-*. Turning to the right panel, we find a more random pattern for the 3935 right constituents: Here, the presence of a larger number of compounds with *-en-* does not imply a small number of compounds without *-en-*, and vice versa. Thus, a strong bias for *-en-* exists only for left constituents. Interestingly, this asymmetry is clearly reflected in the responses of the participants of the present experiment. If participants had chosen the linking morpheme at random on the basis of all the existing compounds (*celex*: 43413) in the language, one would have expected *-en-* (*celex*: 4744) to be selected in roughly 11% of our experimental material. The left constituents provide larger families with clearer preferences for or against *-en-*, leading to a much higher percentage of *en* responses in the Positive and Neutral conditions (58%–96%

versus 6%–19% in the Negative condition).

Place Figure 1 about here

Place Figure 2 about here

In a post-hoc analysis we also tested the overall effect of family homogeneity on the response homogeneity across the three conditions (Positive, Neutral, Negative) both for the Left and Right Bias. We calculated the family homogeneity in terms of the difference between the number of family members with -en- and the number of family members without -en-. We calculated the response homogeneity in terms of the difference between the number of en responses and other responses. The upper panels of Figure 3 reveal a non-linear correlation between response homogeneity and family homogeneity represented by a dotted line.[4] The upper left panel shows a sigmoid curve for the left constituents. The upper right panel shows a more diffuse pattern for the right constituents. Despite this difference, a Spearman correlation test revealed a significant correlation between the family homogeneity and the response homogeneity both for the Left ($r_s = 0.87$, z=6.88, $p < 0.0001$) and the Right Position ($r_s = 0.34$, z=2.70, p=0.007). The magnitude of these correlation coefficients ($r_s = 0.87$ versus $r_s = 0.34$) correspond to the difference in strength of the Left and Right Bias: In terms of rank correlations, the Left Bias explains 76% of the variance, while the Right Bias explains only 12% of the variance.

Place Figure 3 about here

Having observed clear effects of analogy on the choice of the linking morpheme -en-, we now turn to the linking morpheme -s-.

Experiment 2: The linking morpheme -s-

Method

Materials. As in Experiment 1 we constructed three sets of left constituents (L1, L2, L3) and three sets of right constituents (R1, R2, R3). Each set contained 21 nouns. The constituents of L1 and R1 sets had constituent families with as strong a bias as possible towards the linking morpheme -s-. Conversely, L3 and R3 showed a bias as strong as possible against -s-. The sets L2 and R2, the neutral sets, contained nouns with families without a clear preference for or against -s-. We used the celex lexical database to determine the constituent families of the constituents in these six sets. Compounds with a token frequency of zero in a corpus of 42 million words were not included.

The constituents in the L1 set had constituent family members of which at least 80% contained the linking morpheme -s-. The mean number of compounds in these families was 45.7 (range 15–174). Their mean token frequency was 1196.8 per 42 million wordforms (range 102–6663). The constituents in the R1 set had constituent family members of which at least 70% contained the linking morpheme -s-. The mean number of compounds in these families was 6.5 (range 4–19). Their mean token frequency was 103.5 per 42 million wordforms (range 12–409). Neutral left constituents are rare. The neutral set L2 included left constituents whose families contained between 35% and 65% compounds with the linking morpheme -s-.

These families had a mean number of compounds of 6.4 (range 2–34) and a mean token frequency of 116.9 per 42 million wordforms (range 5–915). The constituents in the R2 set had constituent family members of which 45% to 55% contained the linking morpheme -s-. These families had a mean number of compounds of 16.4 (range 4–52) and a mean token frequency of 216.4 per 42 million wordforms (range 18–527). The remaining sets L3 and R3, the groups with a bias against -s-, contained constituents whose family members never have a linking -s-. There were in the mean 31.2 (range 15–77; L3) and 2.45 (range 10–37; R3) family members respectively. Their mean token frequency was 903.1 (range 98–2874; L3) and 532.9 (range 39–2677; R3). These are the maximal contrasts that allowed us to select 21 constituents for each experimental set.

As in Experiment 1, each of the three sets of left constituents (L1, L2, L3) was combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors: Bias in the Left Position (Positive, Neutral, and Negative) and Bias in the Right Position (Positive, Neutral, and Negative). 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 B lists all experimental items. The 9 x 21 = 189 experimental items were divided over three lists. List 1 contained the compounds of the factorial combinations L1-R1, L2-R3, and L3-R2. List 2 contained the compounds of the combinations L1-R2, L2-R1 and L3-R3, and List 3 contained the compounds of the combinations L1-R3, L2-R2, and L3-R1. In this way, each participant saw each constituent only once. We constructed a separate randomized list of the 3 x 21 = 63 compound constituent pairs for each participant.

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

Participants. Sixty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment. All were native speakers of Dutch, none had participated in the previous experiment. The participants were divided into three groups. Each group was asked to complete one of the three experimental lists.

Results and discussion

The participants followed the instructions very closely so that no responses had to be counted as errors. That allowed us to include all participants and all items in the data analysis. Table 2 summarizes the percentages of s responses versus other responses for the nine experimental conditions. Appendix B lists the individual words together with the absolute numbers of s and not s responses.

A by-item logit analysis of the s and not s responses revealed a main effect of Bias in the Left Position ($F(2,180) = 150.6$, $p < 0.0001$), a main effect of Bias in the Right Position ($F(2,180) = 10.5$, $p < 0.0001$), and no interaction of the Bias in both positions ($F(4,180) = 1.6$, $p = 0.1883$). Again, the difference between the Neutral and Positive Bias conditions on the Right Position is not reliable ($F(1,120) = 1.9$, $p = 0.1687$).

The lower panel of Figure 1 shows the effects of both Biases on the percentage of s responses. As in Experiment 1, Bias has a larger effect on the Left Position (a difference of

Table 2:

Percentages of selected linking morphemes when varying bias for -s-

(Positive, Neutral, and Negative) in the left and right compound position.

		Right Position					
Left Position		Positive		Neutral		Negative	
Positive							
	s	96.7	(20.3)	97.4	(22.8)	91.7	(24.2)
	not s	3.3	(20.3)	2.6	(22.8)	8.3	(24.2)
Neutral							
	s	70.5	(10.2)	67.6	(3.7)	53.6	(9.5)
	not s	29.5	(10.2)	32.4	(3.7)	46.4	(9.5)
Negative							
	s	13.6	(5.2)	5.2	(11.5)	1.9	(5.1)
	not s	86.4	(5.2)	94.8	(11.5)	98.1	(5.1)

Note. Standard deviations between parentheses.

minimal 70% between the Positive and Negative conditions) than on the Right Position (a difference of maximal 17%). This result again reflects an asymmetry in the distribution of the linking elements in Dutch that is also mirrored in our experimental design. The left constituents provide larger families with clearer preferences for or against -s-, leading to a much higher percentage of s responses in the Positive and Neutral conditions (from 53% up to 97% versus 2% up to 14% for the Negative condition).

In a post-hoc analysis we tested the overall effect of the family homogeneity on the response homogeneity across the three conditions (Positive, Neutral, Negative) both for the Left and Right Bias. As before, we calculated the family homogeneity in terms of the difference between the number of family members with -s- and the number of family members without -s-. We calculated the response homogeneity in terms of the difference between the number of s responses and other responses. The lower panels of Figure 3 reveal a non-linear correlation between response homogeneity and family homogeneity represented by a dotted line. The lower left panel shows the data of the left constituents, the lower right panel shows the data of the right constituents. As for the -en- homogeneity, the left constituents reveal a sigmoid curve, while the right constituents show a more diffuse pattern. As in Experiment 1, a Spearman correlation test revealed a significant correlation between the family homogeneity and the response homogeneity both for the Left ($r_s = 0.89$, z=7.00, p<0.0001) and the Right Position (Spearman: $r_s = 0.42$, z=3.33, p<0.0001). The magnitude of these correlation coefficients ($r_s = 0.89$ versus $r_s = 0.42$) correspond to the difference in strength of the Left and Right Bias: In terms of rank correlations, the Left Bias explains 79% of the variance, while the Right Bias explains only 18% of the variance.

Experiment 2 addressed the question whether the families of the right and left constituent affect the choice for or against the linking morpheme -s- when building a new nominal compound. We were able to replicate the results of Experiment 1 which tested the family effect on the linking morpheme -en-. The family of the left constituent has a strong effect on the choice of the linking morpheme, while the family of the right constituent has a smaller, but also significant effect.

The suffix family effect

Experiment 3: The effect of the preceding suffix on the linking morpheme -s-

We have seen that the families of the immediate constituents of a new nominal compound have a great influence on the choice of the linking morpheme. The linguistic literature tells us that, in the case of derived words as left constituents, it is the suffix that has influence on the following linking morpheme (Van den Toorn, 1981a; 1981b). For instance, suffixes -ist (similar to English person-noun forming '-ist') or -in (similar to English '-ess') appear mainly with -en-, while suffixes -aard (similar to English '-ee') or -heid (similar to English '-ness') appear mainly with -s-. However, like the constituents, the suffix does not completely determine the linking morpheme. We therefore tested whether the suffix family, i.e. all compounds which contain a left constituent built with a particular suffix, has an effect on the choice of the linking morpheme. For this experiment we chose the linking morpheme -s- because the -s- appears much more often with a preceding suffix ($586/1004 * 100 = 58.4\%$ of all preceding derived words) than the -en- ($54/594 * 100 = 9.1\%$ of all preceding derived

words). To make sure that we test the effect of the suffix and not the effect of the left constituent, we used pseudo-derivations.

Method

Materials. We constructed two sets of left pseudo-constituents (L1, L2) and three sets of right existing constituents (R1, R2, R3). Each set contained 21 nouns. The pseudo-constituents of the sets L1 and L2 contained Dutch suffixes with pseudo-stems, none of which violated the phonotactic rules of Dutch. The suffixes of L1 were -ing (similar to English '-ing'), -heid (similar to English '-ness'), and -iteit (similar to English '-ity'). They appear in celex compounds mainly with the linking morpheme -s- (-ing: $379/406 * 100 = 93.3\%$; -heid: $65/66 * 100 = 98.5\%$; -iteit: $21/25 * 100 = 84.0\%$). The suffixes of L2 were -in (similar to English '-ess'), -sel (similar to English '-ee'), and -ster (similar to English '-ess'). They appear in celex in at least 50% without the linking morpheme -s- (-in: $0/1 = 0.0\%$; -sel: $0/6 = 0.0\%$; -ster: $1/2 = 50.0\%$). R1, R2, and R3 were the same as in Experiment 2. Thus, R1 had constituent families with as strong a bias as possible towards the linking morpheme -s-. R3 showed a bias as strong as possible against -s-. The set R2, the neutral set, contained nouns with families without a clear preference for or against -s-.

Similar to the previous experiments each of the two sets of left pseudo-constituents (L1, L2) was combined with the three sets of right constituents (R1, R2, R3) to form pairs of constituents for new compounds in a factorial design with two factors: Bias in the Left Position (Positive and Negative) and Bias in the Right Position (Positive, Neutral, and Negative).

Appendix C lists all experimental items. The 6 x 21 = 126 experimental items were divided over three lists. List 1 contained the compounds of the factorial combinations L1-R1 and L2-R2. List 2 contained the compounds of the combinations L1-R2 and L2-R3, and List 3 contained the compounds of the combinations L1-R3 and L2-R1. In this way, each participant saw each constituent only once. We constructed a separate randomized list of the 2 x 21 = 42 compound constituent pairs for each participant.

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

Participants. Sixty participants, mostly undergraduates at Nijmegen University, were paid to participate in the experiment. All were native speakers of Dutch, none had participated in the previous experiments. Each group was asked to complete one of the three experimental lists.

Results and discussion

Occasionally, participants filled in a question mark or a letter sequence other than a linking morpheme. Such responses were counted as errors. The overall error rate was extremely low (0.2%), which allowed us to include all participants and all items in the data analysis. Table 3 summarizes the percentages of s responses versus other responses for the six experimental conditions. Appendix C lists the individual words together with the absolute numbers of s and not s responses.

Table 3:

Percentages of selected linking morphemes when varying Bias for -s- in the Left Position (Positive and Negative) and Right Position (Positive, Neutral, and Negative).

		Right Position					
Left Position		Positive		Neutral		Negative	
Positive							
	s	84.0	(14.9)	86.4	(9.0)	79.5	(14.7)
	not s	16.0	(14.9)	13.6	(9.0)	20.5	(14.7)
	other	0		0		0	
Negative							
	s	24.8	(17.4)	20.0	(15.4)	16.4	(16.9)
	not s	75.2	(17.4)	80.0	(15.4)	82.6	(16.9)
	other	0		0		1.0	

Note. Standard deviations between parentheses.

A by-item logit analysis of the s and not s responses revealed a main effect of Bias in the Left Position ($F(1,120) = 276.0$, $p < 0.0001$), no effect of Bias in the Right Position ($F(2,120) = 2.2$, $p = 0.1201$), and no interaction of Bias in both positions ($F(2,120) = 0.6$, $p = 0.5726$).

Experiment 3 addressed the question whether the family of the preceding suffix affects the choice for or against the linking morpheme -s- when building a new nominal compound. We found a strong effect of the suffix family on the choice of the linking morpheme. We were not able to replicate the smaller, but significant effect of the family of the right constituent which we have seen in Experiments 1 and 2. The use of pseudo-words in the Left Position led to compounds which are difficult to interpret. Maybe the lack of a possible interpretation decreased the effect of the bias in the Right Position which was already small in the previous two experiments.

Summary: Experimental results

Experiments 1 and 2 have revealed that linking morphemes in novel compounds can be predicted on the basis of the families of both left and right constituents, and that the effect of the left family is much stronger. We have seen that the difference in strength mirrors a distributional asymmetry in the lexicon, i.e. left constituents tend to have a stronger bias for or against a linking morpheme than right constituents. Experiment 3 has shown that suffixes attached to pseudo-words to form left constituents also affect the choice of linking morphemes.

The experimental results are in line with the descriptions in the literature in so far as the

properties of the left constituent are traditionally described as the main factors influencing the choice of linking morphemes. The presence of a weaker, but significant effect of the right constituent is in line with the observation that right constituents may be important because they codetermine the semantic relation between the constituents in a compound. We have also shown that the final suffix in derived left pseudo-words plays a role, which is in line with the observations reported by Van den Toorn (1981a; 1981b) for real words. Most importantly, the results of our experiments have revealed unambiguous evidence for a strong analogical effect of the constituent family, a novel factor that is not discussed in the linguistic literature.

In the next section, we proceed to test whether it is possible to simulate the effect of the constituent families with the help of an explicit computational algorithm for analogy. The aim of this section is to ascertain whether analogy based on constituent families is computationally tractable. In the general discussion, we will outline how the computational technique that we have opted for can be mapped onto a psycholinguistically plausible architecture of the mental lexicon.

ANALOGICAL MODELING

Several techniques are available for the modeling of data which display statistical tendencies rather than discrete regularities. Connectionist models are widely used to obtain predictions for graded data where standard rule-based methods fail. Although connectionist networks are powerful nonlinear classifiers, they have the disadvantage that additional follow-up analyses of the network are required in order to understand how the network ar-

rives at its classifications. A second disadvantage of connectionist models is that it is at present unclear whether they can accommodate the family size effect reported in Schreuder & Baayen (1997) and De Jong, Schreuder, & Baayen (2000). The family size effect concerns the finding that type counts of morphologically related words for target words correlate with lexical decision times and subjective frequency ratings to these target words, while the corresponding token counts have emerged as irrelevant. Given the sensitivity of connectionist networks to frequencies of occurrence, i.e., token frequencies, it is as yet unclear how this type frequency effect might emerge in combination with the absence of the token frequency effect. As the role of the constituent family that has emerged from our experiments appears to be a similar type count effect, but now in production rather than in comprehension, we have opted for an exemplar-based approach in which type counts effects are more easily accommodated.

Exemplar-based approaches have been developed by, e.g., Skousen (1989) and Daelemans, Zavrel, Van der Sloot, & Van den Bosch (1999). Skousen has proposed an analogical model specifically for the domain of language. In his model, stored exemplars are compared with a given target word using a similarity metric defined over a series of user-specified features. Exemplars that are most similar to the target are most likely to serve as the analogical basis for its classification.

Various machine-learning techniques proceed along similar lines. We have opted for a program implementing a series of machine-learning techniques, TiMBL, developed by Daelemans et al. (1999).[5] This implementation offers powerful heuristics for finding directly the features with a strong analogical weight. In what follows, we first describe this machine-

learning technique, which we have found very useful from a computational linguistics point of view. We then discuss the results that we have obtained with this technique. In the general discussion, we outline the way in which the technical computational model can be mapped onto a psycholinguistically more plausible model of analogical processing in the mental lexicon.

Exemplar-based learning

Exemplar-based learning techniques implement the idea that the performance of cognitive processes is based on explicit storage of representations of earlier experiences. Reasoning is conducted by comparing a new instance with stored instances. Crucially, the information carried by earlier experiences is not extracted from these experiences and stored in the form of abstract rules. Instead, a general strategy for similarity-based reasoning is combined with the extensive storage of exemplars in an instance database. For example, the problem of assigning the position of the main stress to a novel Dutch word is solved by storing large numbers of multi-syllabic words in the instance database, and by using a distance measure defined over the phonological make-up of the final two syllables of these words. A search in the instance base leads to the exemplar which is most similar to that of the target noun. The stress position stored with this exemplar is suggested to be that of the target noun (see Daelemans, Gillis, & Durieux, 1994, for a detailed study). The main advantage of exemplar-based learning is that no abstract rules need to be formulated. The price to be paid is that computational load increases substantially with the size of the database, because the distance between any new instance and all exemplars in the instance database must be

computed. We will return to the issue of this computational load below.

In our experience, the k-NN algorithm with the Hamming distance measure known as ib1 in machine learning literature (Aha, Kibler, & Albert, 1991) yields the best results for the modeling of Dutch linking morphemes. Its similarity metric is very simple. Given two patterns X and Y, each represented by n features, the distance between X and Y is the number of shared features. TiMBL makes three additions to the original k-NN algorithm. First, the value of k refers to the k-nearest distances and not the k-nearest cases. In our simulation studies we have set k to unity, which means that all instances at Hamming-distance 1 are included in the set of nearest neighbors. Second, if the nearest neighbor set contains more than one instance, the linking morpheme is selected that is most often instantiated in this nearest neighbor set. Third, in case of a tie, the linking morpheme is selected that has the highest frequency in the instance base.

TiMBL has the useful possibility to add to the Hamming-distance measure a relevance weight for every feature (the ib1-ig algorithm). TiMBL accomplishes this by means of the information gain (IG) which looks at a feature and measures how much information it contributes to our knowledge of the correct linking morpheme. The information gain of a feature i is obtained by calculating the difference in uncertainty or entropy between the situations without and with knowledge of the value of that feature:

$$\text{IG} = w_i = H(C) - \sum_{v \in V_i} P(v) \cdot H(C|v). \quad (1)$$

In (1), C denotes the set of linking possibilities (-en-, -s-, $\emptyset$), and V_i the set of values for feature i (e.g., 'stressed' and 'unstressed' for the feature Stress). The entropy of the linking

possibilities is

$$H(C) = -\sum_{c \in C} P(c) \log_2 P(c), \tag{2}$$

with c ranging over {-en-, -s-, ∅}. Using information gain weights, we get the following distance metric:

$$\Delta(X, Y) = \sum_{i=1}^{n} w_i \mathrm{I}_{[x_i \neq y_i]} \tag{3}$$

(Daelemans et al., 1999: 9). By computing the information gain for the many features that one might potentially use in a particular simulation study, it becomes possible to make an informed preselection of features.

In what follows, we will apply this methodology to the materials of the first two experiments in order to ascertain to what extent machine learning techniques are able to predict the choice of linking morphemes.

Predicting linking morphemes

In order to gauge the predictive power of exemplar-based learning of Dutch linking morphemes, we first studied the preferred choices for existing compounds using 10-fold cross-validation. In 10-fold cross-validation the dataset is divided into 10 'held-out' subsets. For each held-out subset, linking morphemes are predicted on the basis of the remaining 90% of the data, which serve as the training set. The overall performance of the model is evaluated in terms of the average percentage of correctly predicted linking morphemes calculated over the 10 cross-validation runs.

A crucial determinant of the model's performance is the set of features defining its input space. In our simulation studies, we have made use of 9 features. The first and second features

code the left and right immediate constituents, which represent the left and right constituent families. The third feature represents the plural suffix selected by the left constituent. This feature can be used to extract the knowledge that the linking morpheme -en- is found only after left constituents that select -en as their plural suffix. Features 4–7 code the abstractness and animacy of the first and the second constituent. They allow us to trace whether the semantics of the constituents codetermine the choice of linking morphemes (Van den Toorn, 1982a). Feature 8 marks the presence of stress on the final syllable of the first constituent, as it might be possible that the linking morpheme -en- is inserted to avoid a stress clash between the two constituents. Finally, feature 9 codes the morphological complexity of the first constituent in terms of its number of morphemes as a greater complexity of the left constituent has been argued to give rise to a preference for -s- (see Mattens, 1984). In various simulation runs not reported here, we used the three final phonemes of the first constituent, the three initial phonemes of the second constituent, as well as the last morpheme of the first constituent instead of features 1 and 2. As the results obtained with this alternative feature set invariably turned out to yield inferior results, we do not discuss these alternative features.

We used the 22,994 Dutch nominal compounds in the celex lexical database that occur with a frequency of at least 2 per 42 million word forms as our instance base. Each of these compounds was assigned a vector of values for our 9 features. The second column of Table 4 lists the information gain for each individual feature on the basis of the training sets in the cross-validation runs. When we use all features, we predict the correct linking morpheme for 93.2% of the compounds in the held-out datasets. When we use only the first feature,

the first constituent, which has the highest information gain, we obtain an accuracy which is only slightly less, 92.5%. The linguistic literature describes the choice of linking morpheme as governed by a conspiracy of tendencies. Our cross-validation results suggest that, indeed, these tendencies allow the linking morpheme to be predicted with a high degree of accuracy. Surprisingly, most of the predictive power resides in a single feature only: the first constituent, i.e., the key for the morphological family of the first constituent.

How well does the model predict the choice of the linking morpheme for the neologisms used in Experiments 1 and 2? First consider Experiment 1 summarized in columns 4–6. The column labelled Fam1 lists information gain and accuracy when the model is trained on pooled constituent families of all experimental words. We trained the model on this subset of the compounds listed in celex for the following reason. The semantic specification for a constituent of a given compound, as we have used it for the first study, is not restricted to the meaning of the constituent in this particular compound, but provides the full range of possible meanings the constituent can have when used in isolation. For a specific compound, this range of possible feature values is too broad. For the subset of constituent families it was feasible to manually narrow down the semantics to the correct meaning for each specific compound separately. Consequently, there are two differences between this analysis and the previous analysis based on the celex data. First, the semantic features are more precise, second, the number of types on which TiMBL is trained is much smaller (celex 22,994 vs. Fam1 1864).

When we train on the pooled families using all features, we obtain an accuracy of 83.6%.

Table 4:

Features used in the simulation studies, their information gain (upper part of the table), and the corresponding prediction accuracy (lower part of the table).

No.	Feature	Celex	EN			S		
			Fam1	CELEX	Fam2	Fam1	CELEX	Fam2
1	1st C	1.11*	1.29*	1.11*	*	1.14*	1.11*	*
2	2nd C	0.41	0.96	0.41		0.70	0.41	
3	1st C: plur	0.10	0.12	0.10		0.07	0.10	
4	1st C: abst	0.07	0.13	0.07		0.13	0.07	
5	1st C: anim	0.04	0.13*	0.04	*	0.07	0.04	
6	2nd C: abst	0.02	0.06*	0.02	*	0.06*	0.02*	*
7	2nd C: anim	0.00	0.01	0.00*		0.01	0.00	
8	1st C: stress	0.07	0.13	0.07		0.07	0.07	
9	1st C: compl	0.11	0.05	0.11		0.08	0.11	
	accuracy 1–9	93.2%	83.6%	78.3%	84.7%	91.5%	82.5%	82.5%
	accuracy 1	92.5%	78.8%	75.1%	79.9%	87.8%	82.5%	87.3%
	accuracy *	92.5%	85.2%	82.0%	86.8%	91.5%	83.1%	88.4%

Note. Celex: results using 10-fold cross-validation. EN, S: results for Experiments 1 and 2, with accuracy being evaluated against the majority choice of the participants. Predictions are made on the basis of various training sets: Fam1: pooled family members of all experimental items; CELEX: all compounds in celex; Fam2: predictions based on left and right constituent families of each individual item. *: features determined as relevant by forward step-wise selection.

As we are dealing with neologisms, accuracy is evaluated in terms of the percentage of experimental words for which TiMBL predicts a linking morpheme that is identical to the majority choice of our participants. Again, we observe that the first constituent has the highest information gain, and that using this feature exclusively already leads to an accuracy of 78.8%. By adding features 5 and 6, we can increase the accuracy to 85.2%. Feature 5 concerns the animacy of the left constituent: Animate left constituents elicit higher numbers of en responses. Feature 6 represents the abstractness of the right constituent: Abstract right nouns lead to fewer en responses. The selection of these features is based on forward step-wise selection. At the first step, the feature with the highest information gain is selected. For each successive step, the feature with the next highest information gain is considered. If addition of this feature improves accuracy, it is added to the list of features. Otherwise, the feature with the next highest information gain is tested. The information gains of the features selected by this algorithm are marked with an asterisk in Table 4.

When we compare these results with those obtained with cross-validation for all compounds in celex (column 3), we observe a decrease in accuracy of roughly 10%. This loss of accuracy has three possible sources. First, the experiment made use of neologisms, non-existing compounds presented without a natural context, that may have been somewhat more artificial than existing compounds. However, whatever the nature of our materials may be, the performance of the model is similar to that of human subjects. When we calculate the average accuracy of the subjects in the same way as we evaluate the accuracy of the model, i.e., by treating the majority choice as norm, we obtain an average accuracy of 85.1%, which comes close to the maximum of the range of model accuracies (78.8–85.2).

Apparently, participants and the model find the task equally difficult.

Second, the set of words with a Neutral Bias in the experiment is atypical for the population as a whole. As we have already seen in Figure 1, most of the left constituents in celex reveal a strong bias for or against -en- (98% of all left constituents appear with the linking morpheme -en- either in less than 35% or in more than 65% of all members of the constituent family). The over-representation of left constituents without a strong bias in the experiment (30% versus 2% off all celex compounds) renders the experiment more difficult to model than the celex population of compounds using cross-validation. In fact, the accuracy scores for the subsets of words with a strong bias for or against -en- are substantially higher than those for the words with a Neutral Bias (Left Positive Bias: 92.1%; Left Neutral Bias: 71.4%; Left Negative Bias: 90.5%). Clearly, the atypical Neutral set renders the experiment more difficult.

Third, the reduced size of the training set may have led to reduced accuracy. To investigate this possibility we ran additional simulation experiments. When we train the model on all compounds in celex rather than on the subsets of words for which we checked the coding of concreteness and animacy of the constituents by hand, we observe a slight reduction in accuracy of roughly 3%. Possibly, this reduction arises because the semantic coding is less precise for the database as a whole. Interestingly, we obtain slightly improved accuracies when we train the model not on a larger but on an even smaller training set. By training on the unique family members of each experimental compound separately, we improve the average accuracy to 86.8% (column 6, Fam2), using the same features that led to the highest accuracy when training on the pooled family members.[6] It is remarkable that training on

the basis of small by-item families (with a range of 8-84 family members) results in slightly, although not significantly ($p > 0.2$, proportions test), improved performance compared to training on the 1864 pooled family members or the 22,994 compounds in celex. This suggests that the constituent families provide the analogical basis for selecting the linking morphemes in novel compounds. From a psycholinguistic perspective, this is an important result as it obviates the need to scan the complete lexicon for analogical exemplars. In the general discussion, we shall use this result to formulate a psycholinguistic spreading activation model for the analogical selection of linking morphemes.

The last three columns of Table 4 summarize the results obtained using the same procedures for the data of Experiment 2. The best results are obtained when we train TiMBL on the pooled constituent family members of all experimental compounds. On the basis of the first constituent and the abstractness of the second constituent (abstract right constituents lead to more s responses), TiMBL achieves an accuracy of 91.5%. When we train the model on the compounds in celex, accuracy decreases significantly to 83.1% (p= 0.02, proportions test). Training on the individual families of the experimental compounds leads to a slight reduction in accuracy that, however, does not differ significantly from the accuracy when trained on the pooled constituent family members. Compared to the participants of Experiment 2, who on average opt for the majority choice for 83.5% of the experimental compounds, TiMBL performs surprisingly well.

The results summarized in Table 4 are the best results that we have been able to obtain. Replacing the features for the first and second constituents by features for the last three segments of the first constituent and the first three segments of the second constituent in-

variably leads to decreasing performance. The same holds for training on the last morpheme of the first constituent.

Table 5 compares the success rate that can be achieved on the basis of the phonological and morphological rules that have been formulated for Dutch with the corresponding success rate as achieved by TiMBL (trained on the constituent families of the the individual items), for experiments 1 and 2. Note that the rules are applicable only to small subsets of the materials. The phonological rules state that no linking morpheme is allowed following a rime ending with a vowel, with a liquid preceding /k/ or /m/, or with a schwa followed by a sonorant. For words with other rime characteristics, the rules provide no predictions at all. Not surprisingly, the morphological rules apply only to the compounds in our materials which have a derived left constituent. Similarly, the semantic rules apply only to words with a mass noun and human agents ending in -er as left constituent, as well as to synthetic compounds in which the left constituent is the non-subject argument of the embedded verb to its right. From Table 5, it is clear that TiMBL outperforms the rules for all applicable words. In addition, TiMBL provides good predictions where the rules provide none. Interestingly, TiMBL reveals the animacy and abstractness of the left and right constituents to be relevant factors co-determining to some extent the choice of the linking morpheme. Further rigorous quantitative research will have to clarify which semantic factors contribute to the choice of the linking morpheme over and above the constituent families themselves.

Finally, Table 6 presents a comparison of the performance of the participants with the performance of TiMBL when trained on the constituent families of the the individual items. The first two columns specify the Bias (Positive, Neutral, or Negative) for the left and right

Table 5:

Comparison of rule-based and analogy-based predictions for experiments 1 and 2.

	EN (experiment 1)			
	applicable		not applicable	
	rules	TiMBL	rules	TiMBL
phonology	9/15	13/15	-/174	142/174
morphology	15/36	36/36	-/153	119/153
semantics	8/14	10/14	-/175	245/175
	S (experiment 2)			
	applicable		not applicable	
	rules	TiMBL	rules	TiMBL
phonology	12/24	24/24	-/165	133/165
morphology	27/51	41/51	-/138	116/138
semantics	11/34	28/34	-/155	129/155

Note. x/y: number of successful prediction/number of applicable cases; phonology: predictions based on the final rime; morphology: predictions based on the final suffix; semantics: predictions based on semantic rules for mass nouns, human agents ending in -er, and synthetic compounds in which the left constituent is the non-subject argument of the embedded verb to its right.

constituents. The third and fifth columns list the number of participants (averaged over items) that selected -en- (column 3) and -s- (column 5) in Experiments 1 and 2 respectively. TiMBL provides for each item the probabilities for the various linking options. Given that there were 20 participants in each of the two experiments, the expected number of participants selecting, e.g., -en- in Experiment 1 for a given item equals 20 times the probability of -en- for that item. The average number of participants selecting -en- for the nine experimental conditions of Experiment 1 and 2 are listed in columns 4 and 6 respectively. Note that the expected values as predicted by TiMBL are similar to the experimental values, and this impression is confirmed by goodness of fit tests.[7] Thus, the predictions of TiMBL as a computational model of analogy remain accurate even when we consider the individual conditions of our experimental design.

Note that this is not a trivial result. The model could have failed in several ways. First, it could have predicted linking morphemes at chance level. This would have indicated that constituent bias would not be the true factor underlying the choice of linking morphemes. In that case, our conclusion would have been that we failed to include the appropriate features in the input data. Second, the model could have predicted the correct choice for the wrong reasons. Suppose that the model had based its predictions not on the constituent family but on the nature of the third phoneme of the right constituent. Suppose, furthermore, that the left constituent family bias is uncorrelated with the nature of this third phoneme. In these circumstances, the model would be interesting from a technical point of view but seriously flawed from a cognitive point of view, as our experiments show that constituent bias is an important factor if not the most important factor. Third, we ran our simulation studies not

Table 6:

Comparison of the participants and TiMBL across experimental conditions

Left	Right	EN (Experiment 1)		S (Experiment 2)	
		participants	TiMBL	participants	TiMBL
pos	pos	19.0	17.8	19.3	19.7
pos	neutr	19.3	18.3	19.5	19.7
pos	neg	17.5	17.9	18.3	19.7
neutr	pos	15.0	11.3	13.3	10.4
neutr	neutr	16.4	12.4	12.7	10.4
neutr	neg	11.7	11.8	10.5	10.4
neg	pos	3.6	0.0	2.7	0.0
neg	neutr	3.8	0.0	1.0	0.0
neg	neg	1.2	0.0	0.4	0.0

Note. Number of participants (averaged over items) selecting -en- in Experiment 1 and -s- in Experiment 2 and the corresponding expectations based on TiMBL (see text).

only on the bases of the constituent families but on a great many other features as well. The simple fact that the model assigns the greatest information gain to the constituent families is not an artifact of the selection of our experimental materials, as can be seen from the cross-validation data obtained for all noun-noun compounds in the celex lexical database.

Summing up, the present simulation studies show that predictions mirroring the actual choices of human participants can be made on the basis of the families of the left constituent in combination with the semantics of both constituents. These results suggest that analogy may well underlie the strong intuitions that language users have concerning the choice of the appropriate linking morpheme.

GENERAL DISCUSSION

This study has addressed the question in what way analogy influences the choice of linking morphemes in Dutch noun-noun compounds. Even though the usage of linking morphemes in noun-noun compounds is not well predictable by rule, it can be quite well predicted analogically on the basis of the constituent families of both the left and the right constituents. It is the family of the left constituent which constitutes the primary domain of analogical prediction for existing words (Experiments 1 and 2). In the case of suffixed pseudo-words as left constituents, the suffix provides the analogical domain for the choice of the linking morpheme (Experiment 3). A series of computational simulation studies using an exemplar-based machine-learning algorithm for the modeling of analogy, TiMBL, revealed that the actual linking morphemes selected by the participants in our experiments can be predicted with a high degree of accuracy on the basis of the morphological family of the first

constituent with some additional influence of the semantics of the second constituent. These results lead us to conclude that the left constituent families provide the crucial analogical basis for selecting the most appropriate linking morpheme in Dutch. When comparing the choices made by the participants in our experiments with those made by the machine-learning algorithm, we found that the selection is equally difficult for human subjects and TiMBL.

Our results show that the choice of the linking morpheme hinges on existing exemplars with the same left constituent. At the same time, our experimental evidence suggests that the right constituent has a minor role to play. We know of three other studies that mention a possible role for the left constituent. For compounds in Afrikaans, Botha (1968) argued that nouns are lexically marked for linking morpheme when they appear as left constituents in compounds. This works fine for those left constituents that consistently occur with only one linking morpheme. However, for the many left constituents with variable realizations, Botha is forced to assume lexical listing of the full compounds. Unfortunately, Botha's theory has no predictive power with respect to neologisms which have a left constituent with variable realizations.

The idea that analogy might be involved has been suggested for German linking morphemes by Becker (1992), who, however, makes use of such a general notion of analogy that it is difficult to see how any falsifiable predictions might be obtained. Dressler, Libben, Stark, Pons, & Jarema (submitted) present experimental data that hint at a role for left constituent bias in German, but these authors mention this possibility only in passing for a small subset of their data. Since our present results show that it is possible to explicitly model analogy quantitatively and to predict its influence experimentally, we believe that we

now have a realistic methodology for studying the influence of analogy on the realization of linking morphemes across a wider range of languages.

Recall that there is considerable variation in the realization of the linking morphemes. We have seen this variation in the responses of the participants in our experiments, and it is also visible in comprehensive dictionaries, which list variants such as spelling+wijziging and spelling+s+wijziging ('spelling change') side by side. This variation is captured by our analogical model, which allows for some uncertainty with respect to the appropriate linking morpheme exactly as observed for the responses of our participants. This kind of variation is not restricted to linking morphemes, it is also found in the domain of derivational morphology. For instance, Malicka-Kleparska (1985) discusses the formation of diminutives in Polish and calls attention to the free variation between the rival forms -ik and -ek that occurs for words with a particular phonological form. Such free variation is at odds with strict rule-based systems, while it may arise in systems based on analogy in the absence of a clear bias for a particular form. We believe that such variational data provide evidence in favor of the view that morphological rules are grounded in analogy.

Thus far, we have used the machine-learning algorithm implemented in TiMBL to model the analogical selection of linking morphemes in novel compounds. From a computational linguistics point of view, TiMBL captures the analogy underlying the linking morphemes quite satisfactorily. From a psycholinguistics point of view, the question arises whether it is realistic to assume that in general analogy is really based on an exhaustive calculation of a distance metric for all forms in the lexicon. In fact, TiMBL itself does not carry out such an exhaustive calculation for a novel form. While this might be feasible on a massively

parallel machine, present-day sequential machines require alternative algorithms. TiMBL solves this algorithmic problem by constructing a decision tree during training (Daelemans, Van den Bosch, & Weijters, 1997). By dropping a novel form through the decision tree, the appropriate linking morpheme is identified.

Such a decision tree can in fact be understood as a set of rules. Given that the analogy underlying the choice of linking morphemes is based on constituent families, a separate rule for each constituent is embodied in the decision tree. Those researchers who view morphological processing as fundamentally rule-based therefore have the option of reformulating the decision tree of TiMBL as a set of morphological rules. The cost of this option is a proliferation of rules, one for each possible left constituent. As we find this cost too high, we have explored an alternative approach based on the idea of parallel co-activation of constituents in a spreading activation framework along the lines of Schreuder & Baayen (1995). Parallel co-activation is a realistic option precisely because our experimental results have revealed that it is only the constituent families that have to be inspected, and not each and every compound in the mental lexicon (or in TiMBL's instance base). Consider Figure 4.

Place Figure 4 about here

The units in the bottom layer in Figure 4 represent sets of semantic and syntactic features. For instance, the unit labelled problem is a short-hand representation for a series of syntactic and semantic representations such as noun, abstract, inanimate etc. Even though not represented graphically in Figure 4, representations such as those for noun and abstract are shared by the units life and form. The central layer contains lemma nodes, nodes that

link sets of semantic and syntactic representations to form representations. For instance, the left-hand lemma, which represents leven+s+probleem ('life problem'), is activated during production by the semantic and syntactic representations of problem and life and in turn activates the form representations $< leven >$, $< probleem >$, and $< s >$. The numbers accompanying the outgoing arrows specify the order in which the form representations have to be linearized for articulation.

In this architecture, the choice for the linking morpheme -s- for the novel compound leven+?+therapie made by 19 out of 20 participants in Experiment 2 might proceed as follows. Once the syntactic and semantic representations of life and therapy have been activated, activation spreads to their lemma nodes. In turn, activation spreads from the lemma nodes to their form representations, activating $< leven >$ and $< therapie >$. Because leven+?+therapie does not have its own lemma representation, and because the linking morphemes are not themselves addressed, the form representations of linking morphemes have not yet been activated.

It has recently been shown that in subjective frequency ratings and in visual lexical decision, morphological families of target words are coactivated (Schreuder & Baayen, 1997; De Jong et al., 2000). Our hypothesis is that in production an analogous coactivation of the constituent families takes place. Thus, we assume that the semantic and syntactic representations for the left constituent life in Figure 4 coactivates the lemmas of leven+s+vorm ('life form'), leven+s+probleem ('life problem') and other such compounds when the target word is leven+?+therapie.[8] [9] The lemmas of these constituent family members in turn coactivate their form representations, including their linking morphemes.[10]

In addition to the strong influence of the first constituent, we have also seen a somewhat weaker effect of the right constituent in our experiments, both factorially and in the correlation analyses of bias and response. We can model the prominence of the left constituent families by having the semantic and syntactic representations of the left constituent, life in our example, send extra activation to the lemma nodes with which it is connected. Possibly, the special burst of activation flowing from the first constituent to the lemma layer is a consequence of it being the first constituent that has to be articulated (Roelofs, 1996).[11] Recall that the TiMBL results revealed an effect of the semantics of the right constituent. For instance, abstract right constituents show a slight preference for the linking morpheme -s-. We assume that right abstract constituents coactivate lemma nodes for abstract nouns, and therefore also abstract noun compounds in the constituent families. The activation of these compound lemma nodes leads to some extra support for the linking morpheme -s-.

Finally, the results of Experiment 3, in which the left constituents were suffixed pseudowords, can be understood along similar lines. Under the assumption that the suffix in the pseudo-word activates its semantics, and that these semantics in turn coactivate the lemmas of the compounds with this suffix, the bias in the suffix family will lead to a preference for a given linking morpheme.

The present results challenge the idea that in order to model non-deterministic linguistic phenomena symbolic representations have to be given up and replaced by subsymbolic representations as argued by, for instance, Rumelhart & McClelland (1986a) and Seidenberg (1987); see also Zhou & Marslen-Wilson (unpublished manuscript). We have shown that it is possible to model analogy without giving up symbolic representations such as lemmas for

complex words. At the same time, we do not think it is necessary to be committed to the view that morphological rules are in essence symbolic rewrite-rules. This formal view of word formation rules is challenged by the experimental and simulation results for the compounds with neutral bias that we have studied. Here, both our participants and our model showed great uncertainty with respect to what might be the most appropriate linking morpheme. This uncertainty is difficult to reconcile with formal deterministic rules. For strongly converging, consistent domains, formal analogical models will show behavior similar to that of deterministic rules. For diverging, inconsistent domains, deterministic rules impose regularity that is not present in the data nor, if we may trust our experimental results, in the minds of speakers of Dutch. Formal models of analogy, on the other hand, reflect the inconsistency present in their input domains both in the variation in their output and in the confidence they assign to their output. This shows that formal models of analogy are not unconstrained all-powerful theories that can always predict any outcome and hence have no explanatory value. Instead, the behavior of formal models of analogy is tightly constrained by its input domain. For Dutch compounds, local family-based analogical generalization instead of global lexicon-based rule generalization has allowed us to approximate human behavior with greater precision and insight.

References

Aha, D. W., Kibler, D., & Alber, M. (1991). Instance-based learning algorithms. Machine Learning 6, 37–66.

Anshen, F., & Aronoff, M. (1988). Producing morphologically complex words. Linguistics 26, 641–655.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. Journal of Memory and Language 36, 94–117.

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

Baayen, R. H., Schreuder, R., De Jong, N. H., & Krott, A. (in press). Dutch inflection: the rules that prove the exception. In S. Nooteboom, F. Weerman, & F. Wijnen (Eds.), Storage and computation in the language faculty. Dordrecht: Kluwer Academic Publishers.

Becker, T. (1992). Compounding in German. Rivista di Linguistica 4 (1), 5–36.

Bertram, R., Laine, M., Baayen, R. H., Schreuder, R., & Hyönä, J. (1999). Affixal homonymy triggers full-form storage even with inflected words, even in a morphologically rich language. Cognition 74, B13–B25.

Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and

productivity. Journal of Experimental Psychology: Memory, Learning, and Cognition 26, 419–511.

Botha, R. P. (1968). The function of the lexicon in transformational grammar. The Hague: Mouton.

Clahsen, H. (1999). Lexical entries and rules of language: a multi-disciplinary study of German inflection. Behavioral and Brain Sciences 22, 991–1060.

Clahsen, H., Eisenbeiss, S., & Sonnenstuhl-Henning, I. (1997). Morphological structure and the processing of inflected words. Theoretical Linguistics 23, 201–249.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74, 829–836.

Daelemans, W., Bosch, A. Van den, & Weijters, A. (1997). IGTree: Using trees for compression and classification in lazy learning algorithms. Artificial Intelligence Review 11, 407–423.

Daelemans, W., Gillis, S., & Durieux, G. (1994). The acquisition of stress, a data-oriented approach. Computational Linguistics 20 (3), 421–451.

Daelemans, W., Zavrel, J., Sloot, K. van der, & Bosch, A. van den. (1999). TiMBL: Tilburg memory based learner reference guide 2.0 (Report 99-01). Computational Linguistics Tilburg University.

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2000). The morphological family size effect and morphology. Language and Cognitive Processes 15, 329–365.

De Saussure, F. (1966). Course in general linguistics. New York: McGraw.

Dressler, W. U., Libben, G., Stark, J., Pons, C., & Jarema, G. (submitted). The processing of interfixed German compounds.

Fienberg, S. (1980). The analysis of cross-classified categorical data. Cambridge, Mass.: The MIT Press.

Haas, W. d., & Trommelen, M. (1993). Morfologisch handboek van het Nederlands. Den Haag: SDU.

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

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. Keyser (Eds.), The view from building 20: essays in linguistics in honor of Sylvain Bromberger (Vol. 24, pp. 111–176). Cambridge, Mass: MIT Press.

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., Schreuder, R., & Baayen, R. H. (submitted). Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds.

Kuipers, A. H. (1960). Phoneme and morpheme in Kabardian. The Hague: Mouton and Co.

Malicka-Kleparska, A. (1985). Parallel deriviation and lexicalist morphology: the case of Polish diminutivization. In E. Gussmann (Ed.), Phono-morphology. studies in the interaction of phonology and morphology (pp. 95–112). Lublin: Catholic University of Lublin.

Marcus, G., Brinkman, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. Cognitive Psychology 29, 189–256.

Mattens, W. H. M. (1984). De voorspelbaarheid van tussenklanken in nominale samenstellingen. De nieuwe taalgids 7, 333–343.

Neijt, A., Baayen, R. H., & Schreuder, R. (in preparation). Reading relicts of the past: The semantics of linking elements in present-day Dutch orthography.

Pinker, S. (1991). Rules of language. Science 153, 530–535.

Pinker, S. (1997). Words and rules in the human brain. Nature 387, 547–548.

Plunkett, K., & Juola, P. (2000). A connectionist model of English past tense and plural morphology. Cognitive Science.

Rietveld, T., & Hout, R. Van. (1993). Statistical techniques for the study of language and language behaviour. Berlin: Mouton de Gruyter.

Roelofs, A. (1996). Serial order in planning the production of successive morphemes of a word. Journal of Memory and Language 35, 854-876.

Rueckl, J. G., Mikolinski, M., Raveh, M., Miner, C. S., & Mars, F. (1997).

Morphological priming, fragment completion, and connectionist networks. Journal of Memory and Language 36 (3), 382–405.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). Parallel distributed processing. Explorations in the microstructure of cognition. Vol. 1: Foundations. Cambridge, Mass.: MIT Press.

Sandra, D., Frisson, S., & Daems, F. (1999). Why simple verb forms can be so difficult to spell: the influence of homophone frequency and distance in Dutch. Brain and Language 68 (1/2), 277–283.

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), Morphological Aspects of Language Processing (pp. 131–154). Hillsdale, New Jersey: Lawrence Erlbaum.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. Journal of Memory and Language 37, 118–139.

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

Seidenberg, M. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy. In M. Coltheart (Ed.), Attention and Performance XII (pp. 245–263). Hove: Lawrence Erlbaum Associates.

Seidenberg, M., & Hoeffner, J. (1998). Evaluating behavioral and neuroimaging data on past tense processing. Language 74, 104–122.

Sereno, J., & Jongman, A. (1997). Processing of English inflectional morphology. Memory and Cognition 25, 425–437.

Skousen, R. (1989). Analogical modeling of language. Dordrecht: Kluwer.

Spencer, A., & Zwicky, A. (Eds.). (1998). The handbook of morphology. Oxford: Blackwell.

Stark, J., & Stark, H.-K. (1991). On the processing of compound nouns by a Wernicke's aphasic. In J. Tesak (Eds.), Neuro- und Patholinguistik. Grazer Linguistische Studien 35, 95–112. Graz.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. Memory and Cognition 7, 263–272.

Toorn, M. C. v. d. (1981a). De tussenklank in samenstellingen waarvan het eerste lid een afleiding is. De nieuwe taalgids 74, 197–205.

Toorn, M. C. v. d. (1981b). De tussenklank in samenstellingen waarvan het eerste lid systematisch uitheems is. De nieuwe taalgids 74, 547–552.

Toorn, M. v. d. (1982a). Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II. De nieuwe taalgids 75 (2), 153–160.

Toorn, M. v. d. (1982b). Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I. De nieuwe taalgids 75 (1), 24–33.

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

Zhou, X., & Marslen-Wilson, W. (unpublished manuscript). Lexical representation of compound words: cross-linguistic evidence.

Appendix A

Materials for Experiment 1: left constituent and right constituent (number of en responses, number of other responses).

L1-R1: Left Position: Positive -en- Bias; Right Position: Positive -en- Bias

student kolder (20, 0); pen prik (20, 0); advocaat geslacht (18, 2); soldaat deken (19, 1); vreemdeling buurt (20, 0); kleur tegenstelling (10, 10); sigaret knipsel (18, 2); sigaar kiosk (17, 3); pan rook (19, 1); toerist klooster (20, 0); roos gaas (20, 0); beer lever (20, 0); noot laan (18, 2); aap klauw (20, 0); tomaat moes (20, 0); kat haat (19, 1); reus hol (20, 0); gans lijf (20, 0); stier beet (20, 0); vrucht massa (20, 0); wesp ras (20, 0)

L1-R2: Left Position: Positive -en- Bias; Right Position: Neutral -en- Bias

noot dief (19, 1); sigaret bundel (20, 0); sigaar republiek (17, 3); stier kooi (20, 0); kat paar (20, 0); wesp jacht (20, 0); aap vel (19, 1); vrucht rek (20, 0); tomaat stam (20, 0); roos zee (19, 1); soldaat bond (20, 0); pen hout (20, 0); gans boter (20, 0); kleur rad (19, 1); student kas (20, 0); reus rijk (20, 0); beer galerij (17, 3); pan kaas (15, 5); vreemdeling steun (20, 0); toerist kuil (20, 0); advocaat corps (20, 0)

L1-R3: Left Position: Positive -en- Bias; right constituent: Negative -en- Bias

sigaar juffrouw (20, 0); sigaret tarief (20, 0); tomaat project (18, 2); pan lengte (11, 9); toerist gedeelte (20, 0); soldaat bevoegdheid (17, 3); beer maaltijd (19, 1); aap terrein (20, 0); vreemdeling crisis (20, 0); student voorschrift (20, 0); gans schade (18, 2); advocaat weg

(17, 3); kleur techniek (13, 7); noot gewas (11, 9); pen patroon (12, 8); vrucht kanaal (18, 2); roos kunst (20, 0); kat therapie (17, 3); wesp deskundige (19, 1); reus vrijheid (19, 1); stier psycholoog (18, 2)

L2-R1: Left Position: Neutral -en- Bias; Right Position: Positive -en- Bias

begrip tegenstelling (7, 13); bloem laan (20, 0); bom massa (14, 6); bron gaas (11, 9); buur geslacht (15, 5); god hol (13, 7); heer buurt (20, 0); kaart kiosk (20, 0); koe ras (18, 2); klas kolder (19, 1); kool moes (8, 12); leerling klauw (13, 7); lid lijf (10, 10); persoon beet (7, 13); pijp rook (11, 9); plaat knipsel (19, 1); pop klooster (19, 1); prul deken (19, 1); wolf lever (12, 8); woord haat (20, 0); ziel prik (20, 0)

L2-R2: Left Position: Neutral -en- Bias; Right Position: Neutral -en- Bias

begrip stam (11, 9); bloem boter (14, 6); bom kuil (16, 4); bron rijk (15, 5); buur steun (19, 1); god vel (11, 9); heer kaas (18, 2); kaart bundel (20, 0); klas republiek (20, 0); koe kooi (20, 0); kool rek (16, 4); leerling corps (19, 1); lid kas (15, 5); persoon bond (13, 7); pijp galerij (18, 2); plaat hout (11, 9); pop rad (19, 1); prul zee (19, 1); wolf paar (14, 6); woord jacht (19, 1); ziel dief (17, 3)

L2-R3: Left Position: Neutral -en- Bias; Right Position: Negative-en- Bias

begrip patroon (10, 10); bloem weg (20, 0); bom lengte (11, 8); bron terrein (13, 7); buur project (12, 7); god maaltijd (16, 4); heer tarief (18, 2); kaart juffrouw (18, 2); koe psycholoog (17, 3); kool gewas (3, 17); leerling bevoegdheid (12, 8); lid voorschrift (11, 9); persoon ther-

apie (3, 17); pijp schade (4, 16); plaat techniek (11, 9); pop kunst (14, 6); prul kanaal (12, 8); ziel vrijheid (6, 14). woord gedeelte (3, 17); klas crisis (19, 1); wolf deskundige (12, 8)

L3-R1: Left Position: Negative -en- Bias; Right Position: Positive -en- Bias

stad haat (2, 18); gevangenis deken (0, 20); neus knipsel (6, 14); angst prik (2, 18); industrie rook (4, 16); wijn kiosk (4, 16); kalf beet (2, 18); bevolking ras (0, 20); bier lever (8, 12); overheid geslacht (0, 20); christen klooster (6, 14); dokter klauw (0, 20); fabriek buurt (4, 16); dak gaas (5, 15); aardappel moes (3, 17); rivier massa (15, 5); citroen laan (10, 10); groep hol (0, 20); wetenschap kolder (1, 19); kwaliteit tegenstelling (3, 17); koning lijf (1, 19)

L3-R2: Left Position: -en- bias; Right Position: Neutral -en- Bias

stad republiek (0, 20); industrie corps (7, 13); bevolking stam (0, 20); dokter bond (2, 18); rivier hout (8, 12); dak kuil (6, 14); groep jacht (3, 17); kwaliteit kaas (0, 20); angst steun (6, 14); aardappel bundel (5, 15); wijn dief (2, 18); kalf kooi (4, 16); koning vel (1, 19); bier zee (5, 15); neus paar (17, 3); wetenschap rijk (0, 20); overheid kas (0, 20); gevangenis rek (3, 17); citroen boter (3, 17); christen galerij (6, 14); fabriek rad (1, 19)

L3-R3: Left Position: Negative -en- Bias; Right Position: Negative -en- Bias

aardappel juffrouw (2, 18); angst crisis (1, 19); bevolking gedeelte (0, 20); bier deskundige (1, 19); christen vrijheid (2, 18); citroen gewas (1, 19); dak lengte (4, 16); fabriek psycholoog (0, 20); gevangenis terrein (0, 20); groep bevoegdheid (0, 20); industrie weg (1, 19); kalf maaltijd

(4, 16); koning therapie (2, 18); kwaliteit kunst (0, 20); neus kanaal (2, 18); overheid project (0, 20); rivier techniek (5, 15); stad patroon (0, 20); wetenschap voorschrift (0, 20); wijn schade (0, 20)

Appendix B

Materials for Experiment 2: Left constituent and right constituent (number of s responses, number of other responses).

L1-R1: Left Position: Positive -s- Bias; Right Position: Positive -s- Bias

arbeider standpunt (20, 0); bedrijf bevoegdheid (19, 1); beslissing angst (19, 1); bestuur aangelegenheid (20, 0); fabriek norm (20, 0); gezicht dimensie (16, 4); groep afstand (19, 1); handel fractie (20, 0); investering orientatie (20, 0); leven tactiek (19, 1); macht woede (18, 2); onderzoek reden (20, 0); ontwikkeling duur (20, 0); persoonlijkheid bevordering (20, 0); regering verhouding (20, 0); staat besluit (19, 1); training toename (19, 1); veiligheid drang (20, 0); verkeer delegatie (20, 0); verzorging bijdrage (20, 0); vrede uitoefening (18, 2)

L1-R2: Left Position: Positive -s- Bias; Right Position: Neutral -s- Bias

arbeider functie (20, 0); bedrijf organisatie (20, 0); beslissing conflict (18, 2); bestuur regel (20, 0); fabriek geschiedenis (19, 1); gezicht verandering (19, 1); groep plicht (18, 2); handel project (20, 0); investering kunst (20, 0); leven therapie (19, 1); macht dienaar (20, 0); onderzoek niveau (20, 0); ontwikkeling patroon (20, 0); persoonlijkheid controle (20, 0); regering kwaliteit (20, 0); staat conferentie (16, 4); training probleem (20, 0); veiligheid mechanisme (20, 0); verkeer rust (20, 0); verzorging commissie (20, 0); vrede karakter (20, 0)

L1-R3: Left Position: Positive -s- Bias; Right Position: Negative -s- Bias

arbeider tent (20, 0); bedrijf bos (15, 5); beslissing schrift (13, 7); bestuur club (19, 1);

fabriek kaas (20, 0); gezicht tekening (17, 3); groep kast (15, 5); handel voorraad (19, 1); investering meester (20, 0); leven bel (20, 0); macht laag (19, 1); onderzoek schaal (19, 1); ontwikkeling sprong (19, 1); persoonlijkheid spiegel (19, 1); regering les (20, 0); staat eiland (13, 7); training olie (20, 0); veiligheid venster (20, 0); verkeer soort (19, 1); verzorging transport (20, 0); vrede stok (19, 1)

L2-R1: Left Position: Neutral -s- Bias; Right Position: Positive -s- Bias

begrip dimensie (14, 6); bisschop fractie (17, 3); directeur besluit (19, 1); dood reden (18, 2); generaal delegatie (13, 7); geschut afstand (19, 1); geweld bijdrage (20, 0); god woede (10, 10); heil bevordering (15, 5); klimaat verhouding (11, 9); lucifer norm (9, 11); minister bevoegdheid (12, 8); monnik aangelegenheid (0, 20); persoon angst (14, 6); plicht uitoefening (17, 3); president standpunt (12, 8); temperatuur toename (14, 6); tijd orientatie (16, 4); voordracht duur (18, 2); voorkeur drang (20, 0); wolf tactiek (8, 12)

L2-R2: Left Position: Neutral -s- Bias; Right Position: Neutral -s- Bias

begrip probleem (12, 8); bisschop karakter (18, 2); directeur commissie (12, 8); dood rust (16, 4); generaal functie (19, 1); geschut mechanisme (15, 5); geweld organisatie (14, 6); god dienaar (11, 9); heil therapie (11, 9); klimaat geschiedenis (10, 10); lucifer kwaliteit (6, 14); minister plicht (11, 9); monnik regel (6, 14); persoon kunst (17, 3); plicht verandering (18, 2); president conferentie (12, 8); temperatuur controle (15, 5); tijd conflict (19, 1); voordracht niveau (17, 3); voorkeur patroon (17, 3); wolf project (8, 12)

L2-R3: Left Position: Neutral -s- Bias; Right Position: Negative -s- Bias

begrip laag (10, 10); bisschop spiegel (18, 2); directeur stok (14, 6); dood eiland (9, 11); generaal kast (18, 2); geschut tent (12, 8); geweld soort (10, 10); god bos (5, 15); heil olie (9, 11); klimaat schaal (7, 13); lucifer voorraad (3, 17); minister club (14, 6); monnik kaas (6, 14); persoon transport (6, 14); plicht schrift (4, 16); president bel (9, 11); temperatuur venster (14, 6); tijd sprong (14, 6); voordracht les (13, 7); voorkeur tekening (18, 2); wolf meester (12, 8)

L3-R1: Left Position: Negative -s- Bias; Right Position: Positive -s- Bias

avond duur (5, 15); boek bijdrage (0, 20); christen aangelegenheid (0, 20); dak afstand (5, 15); dwang reden (0, 20); kleur verhouding (5, 15); licht dimensie (5, 15); morgen delegatie (3, 17); nacht tactiek (2, 18); natuur bevordering (6, 14); nood besluit (3, 17); slag uitoefening (1, 19); soldaat woede (3, 17); straat orientatie (1, 19); student standpunt (0, 20); vuur angst (2, 18); wapen bevoegdheid (3, 17); wijn norm (3, 17); woning fractie (4, 16); zand toename (2, 18); zang drang (4, 16)

L3-R1: Left Position: Negative -s- Bias; Right Position: Neutral -s- Bias

avond functie (1, 19); boek organisatie (0, 20); christen commissie (0, 20); dak controle (1, 19); dwang regel (1, 19); kleur kwaliteit (0, 20); licht kunst (1, 19); morgen rust (1, 19); nacht project (0, 20); natuur therapie (0, 20); nood mechanisme (1, 19); slag niveau (0, 20); soldaat dienaar (3, 17); straat karakter (0, 20); student conflict (0, 20); vuur patroon (0, 20); wapen geschiedenis (3, 17); wijn conferentie (0, 20); woning verandering (9, 11); zand

probleem (1, 19); zang plicht (0, 20)

L3-R1: Left Position: Negative -s- Bias; Right Position: Negative -s- Bias

avond sprong (0, 20); boek transport (0, 20); christen schrift (1, 19); dak kast (0, 20); dwang soort (0, 20); kleur schaal (0, 20); licht spiegel (0, 20); morgen bos (2, 18); nacht tent (0, 20); natuur eiland (0, 20); nood olie (0, 20); slag les (0, 20); soldaat stok (2, 18); straat bel (1, 19); student kaas (0, 20); vuur venster (0, 20); wapen club (0, 20); wijn laag (0, 20); woning tekening (2, 18); zand voorraad (0, 20); zang meester (0, 20)

Appendix C

Materials for Experiment 3: Left constituent and right constituent (number of s responses, number of other responses).

L1-R1: Left Position: Positive -s- Bias; Right Position: Positive -s- Bias

ontbolfing aangelegenheid (18, 2); verbrimming afstand (18, 2); bebuiping angst (18, 2); wouking besluit (12, 8); hernabbeling bevoegdheid (18, 2); struffing bevordering (18, 2); snoking bijdrage (15, 5); bronkheid delegatie (20, 0); golheid dimensie (19, 1); pritsheid drang (20, 0); dulligheid duur (20, 0); sloefheid fractie (19, 1); spreunheid norm (19, 1); vlitheid orientatie (18, 2); conviriteit reden (15, 5); descaliteit standpunt (10, 10); dipromeniteit tactiek (14, 6); illuniteit toename (15, 5); recarveniteit uitoefening (18, 2); solutaniteit verhouding (18, 2); virubaniteit woede (11, 9)

L1-R2: Left Position: Positive -s- Bias; Right Position: Neutral -s- Bias

ontbolfing commissie (18, 2); verbrimming conferentie (18, 2); bebuiping conflict (18, 2); wouking controle (15, 5); hernabbeling dienaar (18, 2); struffing functie (16, 4); snoking geschiedenis (12, 8); bronkheid karakter (19, 1); golheid kunst (18, 2); pritsheid kwaliteit (16, 4); dulligheid mechanisme (19, 1); sloefheid niveau (20, 0); spreunheid organisatie (19, 1); vlitheid patroon (18, 2); conviriteit plicht (16, 4); descaliteit probleem (16, 4); dipromeniteit project (19, 1); illuniteit regel (17, 3); recarveniteit rust (17, 3); solutaniteit therapie (18, 2); virubaniteit verandering (16, 4)

L1-R3: Left Position: Positive -s- Bias; Right Position: Negative -s- Bias

ontbolfing bel (20, 0); verbrimming bos (18, 2); bebuiping club (13, 7); wouking eiland (10, 10); hernabbeling kaas (12, 8); struffing kast (11, 9); dipromeniteit laag (17, 3); vlitheid les (19, 1); golheid meester (19, 1); pritsheid olie (15, 5); dulligheid schaal (19, 1); sloefheid schrift (19, 1); spreunheid soort (17, 3); bronkheid spiegel (18, 2); conviriteit sprong (16, 4); descaliteit stok (14, 6); snoking tekening (13, 7); illuniteit tent (15, 5); recarveniteit transport (14, 6); solutaniteit venster (17, 3); virubaniteit voorraad (18, 2)

L2-R1: Left Position: Negative -s- Bias; Right Position: Positive -s- Bias

moepsel aangelegenheid (4, 16); lirksel afstand (3, 17); steukster angst (9, 11); raalster besluit (7, 13); vilkster bevoegdheid (7, 13); girdin bevordering (4, 16); kloerdin bijdrage (3, 17); dreekster delegatie (14, 6); preuksel dimensie (3, 17); pleefster drang (11, 9); veepsel duur (3, 17); taapster fractie (8, 12); brumsel norm (4, 16); zwaagster orientatie (7, 13); borberin reden (1, 19); doerin standpunt (0, 20); darsin tactiek (5, 15); stimsel toename (2, 18); vlatsel uitoefening (5, 15); ploebin verhouding (2, 18); zwaperin woede (2, 18)

L2-R2: Left Position: Negative -s- Bias; Right Position: Neutral -s- Bias

taapster commissie (6, 14); girdin conferentie (1, 19); raalster conflict (8, 12); preuksel controle (0, 20); ploebin dienaar (3, 17); steukster functie (10, 10); stimsel geschiedenis (5, 15); dreekster karakter (5, 15); pleefster kunst (7, 13); veepsel kwaliteit (2, 18); vlatsel mechanisme (2, 18); moepsel niveau (2, 18); vilkster organisatie (8, 12); lirksel patroon (1, 19); borberin plicht (3, 17); doerin probleem (0, 20); darsin project (6, 14); zwaagster regel (9,

11); kloerdin rust (3, 17); zwaperin therapie (2, 18); brumsel verandering (1, 19)

L2-R3: Left Position: Negative -s- Bias; Right Position: Negative -s- Bias

steukster bel (11, 9); lirksel bos (2, 18); pleefster club (12, 8); zwaperin eiland (1, 19); kloerdin kaas (1, 18); zwaagster kast (6, 14); vlatsel laag (4, 16); dreekster les (5, 15); veepsel meester (3, 17); raalster olie (4, 15); moepsel schaal (1, 18); taapster schrift (7, 13); vilkster soort (2, 18); brumsel spiegel (3, 17); borberin sprong (0, 20); doerin stok (0, 19); darsin tekening (3, 17); girdin tent (0, 20); stimsel transport (3, 17); ploebin venster (1, 19); preuksel voorraad (0, 20)

Author Note

We would like to thank Wolfgang Dressler and Nivja de Jong for their helpful comments on an earlier version of this paper. This study was financially supported by the Dutch National Research Council NWO (PIONIER grant to the second 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]For instance, the 800 page handbook of morphology edited by Spencer & Zwicky (1998) devotes five lines of text to the problem of linking elements (p.81).

[2]Marcus et al. (1995) and Clahsen, Eisenbeiss & Sonnenstuhl-Henning (1997) have argued that it is impossible for a language to have more than one productive rule for a particular inflectional function. This claim is based on the distribution of noun plurals in German. The Dutch plural system provides a counterexample to this claim, as shown by Baayen, Schreuder, De Jong, & Krott (in press), a study that presents detailed linguistic and psycholinguistic evidence for the regularity and productivity of both Dutch plural suffixes.

[3]For an analysis of German noun pluralization in such a framework, see Marcus et al., 1995.

[4]We used a non-parametric regression smoother (see Cleveland, 1979), as parametric techniques based on linear models are clearly inappropriate for our data.

[5]For a detailed comparison between TiMBL and Skousen's AML model see Krott, Schreuder & Baayen, submitted.

[6]One might expect to achieve the same accuracy for Fam1, Fam2, and CELEX when training only on the first constituent (accuracy 1). However, the different numbers of training items and the resulting different structures of the three TiMBL-internal decision trees as well as

the random choice of linking morphemes in the case of ties lead to somewhat different results.

[7]For Experiment 1, $\chi^2_{(8)} = 6.44, p = 0.60$ and for Experiment 2, $\chi^2_{(8)} = 9.05, p = 0.34$. In order to avoid technical problems with zero counts for the negative left bias conditions, the chi-squared tests were actually run on the complement counts for all conditions, i.e., the number of participants not selecting -en- (Experiment 1) or -s- (Experiment 2).

[8]For evidence of storage of regular complex words in Dutch see Baayen, Dijkstra & Schreuder (1997), Bertram, Schreuder & Baayen (2000); for compounds see Van Jaarsveld & Rattink (1988).

[9]It is in principle possible that compounds are activated which contain leven as a right constituent as in student+en+leven 'student life'. However, a post-hoc analysis showed that the family homogeneity of these compounds in Experiment 2 is not correlated with the response homogeneity. This is true for compounds containing left constituents at the right position ($\underline{r}_s = 0.18$; $\underline{z} = 1.44$; $\underline{p} = 0.15$) as well as for compounds containing right constituents at the left position ($\underline{r}_s = 0.01$; $\underline{z} = 0.04$; $\underline{p} = 0.97$). These results suggest that only those family members of the left constituent are activated which share the left constituent with the novel compound, and only those family members of the right constituent which share the right constituent with the novel compound.

[10]Figure 4 illustrates the composition route of our parallel dual route model. We assume

that there is also a full-form representation $< levens >$, the plural of $< leven >$, for which support can accumulate in the same way as for $< s >$.

[11]The prominence of the first constituent is in line with the observed greater priming effects of first constituents reported by Kehayia, Jarema, Tsapkini, Perlak, Ralli, & Kadzielawa (1999). In addition, Stark & Stark (1991) report impaired production of second constituents of compounds by a Wernicke's aphasic.