Brain and Language **81,** 708–722 (2002)
doi:10.1006/brln.2001.2558, available online at http://www.idealibrary.com on IDEAL®

Linking Elements in Dutch Noun–Noun Compounds: Constituent Families as Analogical Predictors for Response Latencies

Andrea Krott,* Robert Schreuder,† and R. Harald Baayen†

**Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; and †University of Nijmegen, Nijmegen, The Netherlands*

Published online January 2, 2002

This study addresses the choice of linking elements in novel Dutch noun–noun compounds. Previous off-line experiments (Krott, Baayen, & Schreuder, 2001) revealed that this choice can be predicted analogically on the basis of the distribution of linking elements in the left and right constituent families, i.e., the set of existing compounds that share the left (or right) constituent with the target compound. The present study replicates the observed graded analogical effects under time pressure, using an on-line decision task. Furthermore, the analogical support of the left constituent family predicts response latencies. We present an implemented interactive activation network model that accounts for the experimental data.

Key Words: analogy; analogical modeling; constituent families; compounds; linking elements; interfixes; interactive activation model; graded effects.

INTRODUCTION

Dutch noun–noun compounds often contain so-called linking elements or interfixes. The two main ones are *-en-* and *-s-* as in *schaap+en+bout,* ''leg of mutton,'' or *schaap+s+kooi,* ''sheep fold.'' The linking *-en-* also occurs as the orthographic variant *-e-*. Linguistic descriptions indicate that the occurrence of linking elements seems to be characterized by tendencies instead of clear-cut morphological rules (e.g., Van den Toorn, 1982; Mattens, 1984; Haeseryn, Romijn, Geerts, Rooij, & Van den Toorn, 1997; see also Plank, 1976). A survey of the CELEX Lexical Database (Baayen, Piepenbrock, & Gulikers, 1995) reveals that all phonological and morphological rules that are reported in the linguistic literature apply to only 51% of all Dutch compounds. Of this subset, they correctly predict only 63%, which amounts to 32% of all compounds (Krott, Schreuder, & Baayen, in press). Thus, rules do not provide an adequate account of linking elements. Nevertheless, linking elements are used productively in novel compounds and, as it has been shown in Krott, Baayen, and Schreuder (2001), with substantial agreement among native speakers.

Whereas rule-based approaches have resulted in observationally inadequate analy-

This study was financially supported by the Dutch National Research Council NWO (PIONIER grant to the third author), the University of Nijmegen (The Netherlands), and the Max Planck Institute for Psycholinguistics (Nijmegen, The Netherlands).

Address correspondence and reprint requests to Andrea Krott, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands. E-mail: Andrea.Krott@mpi.nl.

0093-934X/02 $35.00

ses, an analogical approach has proved to be fruitful (Krott et al., 2001; in press). These studies, which used off-line production experiments in which participants had to choose the linking elements for novel Dutch compounds, report the crucial role of a graded, probabilistic factor: the distribution of linking elements in what we have called the left and right constituent families. The left (or right) constituent family is the set of existing compounds that share a left (or right) constituent with the novel compound. We confirmed the predictive power of the constituent families by simulating the choice of linking elements by means of the analogical models AML (Skousen, 1989) and TiMBL (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2000). In the case of the novel compounds used in our experiments, these models' choices were comparable to those of an average participant. In the case of existing compounds, these models correctly predict 92% of the linking elements in all Dutch compounds in CELEX, which is remarkable considering the mere 32% that can be accounted for by rules.

In this article we focus on three main questions. First, do the left and right constituent families affect the choice of the linking element in Dutch novel noun–noun compounds when the choice has to be made under time pressure? Second, do constituent families also affect the speed of the selection process? Third, can we formalize the processes that underlie the choice and the response latencies in terms of an implemented computational model?

In what follows, we first present an on-line production experiment in which responses have to be given under time pressure. The results show that the constituent families indeed also affect the choice of linking elements under time pressure. There is also an effect of the left constituent family on the reaction latencies. We will give an interpretation of these findings in terms of a two-stage cognitive process.

In the second part of the article, we present an interactive activation model that implements the morphological analogical processes. A simulation study of the experimental results shows that our model can account for the effect of the constituent families on the choices as well as the response latencies.

ON-LINE PRODUCTION EXPERIMENT

In order to come to grips with the influence of the constituent families on the choices of linking elements under time pressure, we focus on the linking element *-en-*.

Method

Materials. The materials were identical to those used in Experiment 1 reported in Krott et al. (2001), i.e., three sets of left constituents (L1, L2, and L3) and three sets of right constituents (R1, R2, and R3). The constituents of L1 and R1 had constituent families with as strong a bias as possible toward the linking element *-en-*. Conversely, L3 and R3 showed a bias as strong as possible against *-en-*. The sets L2 and R2, the neutral sets, contained nouns with families without a clear preference for or against *-en-*.

As in the previous experiment, each of the three sets of left constituents (L1, L2, and L3) was combined with the three sets of right constituents (R1, R2, and 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). The items were presented to each participant in a separate random order.

Procedure. The participants performed an on-line cloze task. The experimental items were presented on a computer screen as pairs of two compound constituents separated by two underscores. We asked the participants to combine these constituents into new compounds and press as quickly as possible and according to the chosen linking element a button labeled either ''E/EN'' or labeled ''S/-.'' Participants were asked to give a sign when the pressed button was not intended. We kept a protocol of these errors. All participants pressed the E/EN button with their dominant hand. Each stimulus was preceded by a fixation mark in the middle of the screen presented for 500 ms. After another 50 ms, the stimulus

appeared in the same position and remained on the screen for 2000 ms. The maximum time span allowed for the response was 2500 ms from stimulus onset. Stimuli were presented on Nec Multicolor monitors in white lowercase 21-point Helvetica letters on a dark background. The experiment lasted approximately 15 min.

Occasionally, the first constituent may change its form when it is combined with a linking element [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.

Participants. Twenty participants, undergraduates at Nijmegen University, were paid to participate in the experiment. All were native speakers of Dutch.

Results and Discussion

We distinguished two different types of errors, time-out errors and self-corrections. Taking both types of errors together, all participants performed the experiment with an error rate of maximal 10% and no item showed an error rate of more than 20%. Therefore, all participants and items were included in further analyses. Table 1 summarizes the percentages of *-en-* responses versus *not -en-* responses, the time-out errors, and the self-corrections for the nine experimental conditions. A by-item logit analysis (see, e.g., Rietveld & Van Hout, 1993) of the valid responses showed a main effect of both Bias in the Left Position [$F(2, 180) = 156.6$, $p < .001$] and Bias in the Right Position [$F(2, 180) = 8.2$, $p < .001$] and no interaction between these factors [$F(4, 180) = .4$, $p = .829$]. Thus, the linking elements chosen by the participants follow both the Right and the Left Bias. This is illustrated in the two upper left panels of Fig. 1 for both the *-en-* and the *not -en-* responses. With this result we have replicated the findings obtained with the off-line cloze task used in Krott et al. (2001). We conclude that the choice of the linking element for a novel compound is based on analogy even under time pressure. Apparently, the members of the constituent families become available quite fast.

Note that participants responded slightly more often with *-en-* than expected on

TABLE 1
Mean Percentages of Selected Linking Elements and Errors with Varying Left and Right Bias for *-en-*

	Right position		
Left position	Positive	Neutral	Negative
Positive			
en	90.0 (2.2)	92.4 (1.6)	82.1 (2.8)
not en	8.8 (2.1)	6.7 (1.5)	16.4 (2.6)
self-corr	1.0	1.4	1.4
time-out	1.2	1.0	1.4
Neutral			
en	68.1 (4.3)	75.5 (3.0)	57.9 (4.8)
not en	30.0 (4.1)	22.1 (2.9)	39.8 (5.1)
self-corr	1.2	0.7	1.4
time-out	1.9	2.4	2.4
Negative			
en	17.1 (3.5)	18.1 (3.3)	12.4 (2.9)
not en	81.7 (3.5)	79.5 (3.5)	86.4 (3.1)
self-corr	1.9	1.9	3.1
time-out	1.2	2.4	1.2

Note. Left and Right Bias split up into the experimental conditions (Positive, Neutral, and Negative). *en:* *-en-* responses; *not en:* *not -en-* responses; self-corr: self-corrections; time-out: time-out errors. Standard deviations by items in parentheses.

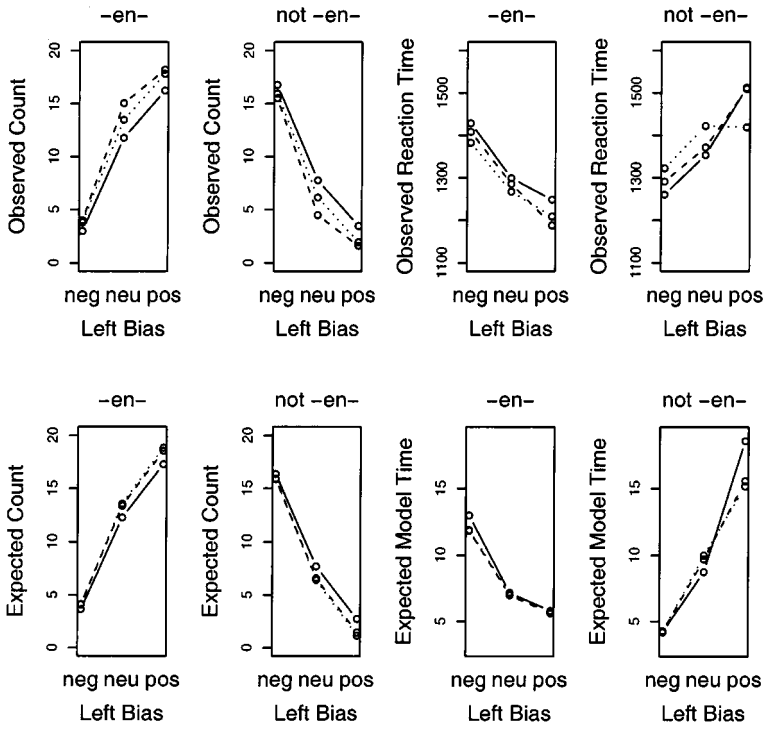

FIG. 1. Interaction plots for the observed and expected counts and response latencies of the linking element *-en-* and the other two linking elements *-s-* and *-0-* (= *not -en-*) with the left constituent bias on the horizontal axis and the right constituent bias indicated by line type (solid line: negative bias; dashed line: neutral bias; dotted line: positive bias).

the basis of the Bias. Overall, more than half of the choices were *-en-* responses (1981 of 3671, or 54%), leaving 46% for the other two linking elements. Even though the experiment was designed to elicit an equal number of responses for both push buttons, the E/EN button response represents two linking elements instead of one. Over the course of the experiment, participants may have become sensitive to *-en-* as being the most likely response. A similar response bias for *-en-* was present in the off-line cloze task reported by Krott et al. (2001). An additional factor in the present on-line experiment may be that participants pressed the *-en-* push button always with their dominant hand.

A by-item logit analysis of the time-out errors revealed no effect, not of the Bias in the Left Position [$F(2, 4) = 3.7$, $p = .124$] nor of the Bias in the Right Position [$F(2, 4) = .8$, $p = .515$].

A by-item logit analysis of the self-corrections, on the other hand, revealed a reliable effect of the Bias in the Left Position [$F(2, 4) = 11.2$, $p = .023$], but no effect of the Bias in the Right Position [$F(2, 4) = 3.7$, $p = .123$]. Participants correct their choices more often if the left constituent has a bias against *-en-* than if it has a bias for *-en-*. This result becomes even more interesting when we take the direction of the self-correction into account, i.e., corrections from *-en-* to *not -en-* or vice versa. Self-corrections occur almost exclusively when a participant has responded against the bias. A by-item logit analysis of the self-corrections from *-en-* to *not -en-* revealed a reliable effect of the Bias in the Left Position [$F(2, 4) = 14.2$, $p = .015$], but again no effect of the Bias in the Right Position [$F(2, 4) = 3.2$, $p = .150$]. A stepwise by-item logit analysis of the self-corrections from *not -en-* to *-en-* also revealed a reliable effect of the Bias in the Left Position only [$F(2, 6) = 5.8$, $p = .039$].

TABLE 2
Mean Response Latencies for Varying Left and Right Bias for *-en-*

Left position	Right position: Positive	Neutral	Negative
Positive			
RT *en*	1209 (130)	1188 (122)	1248 (129)
RT *not en*	1419 (611)	1509 (799)	1512 (310)
Neutral			
RT *en*	1267 (124)	1286 (145)	1299 (149)
RT *not en*	1422 (458)	1371 (179)	1354 (152)
Negative			
RT *en*	1382 (592)	1408 (491)	1429 (672)
RT *not en*	1322 (176)	1291 (177)	1260 (159)

Note. Left and Right Bias split up into the experimental conditions (Positive, Neutral, and Negative). Standard deviations by items are in parentheses.

Table 2 shows the mean response latencies (calculated for the valid responses) for the nine experimental conditions. An analysis of variance of the *-en-* and *not -en-* responses revealed a main effect of the Bias in the Left Position [*-en-* responses: $F1(2, 180) = 15.2$, $p < .001$; $F2(2, 180) = 16.3$, $p < .001$; *not -en-* responses: $F1(2, 180) = 10.7$, $p < .001$; $F2(2, 180) = 10.8$, $p < .001$], but no effect of the Bias in the Right Position [*-en-* responses: $F1(2, 180) = .7$, $p = .519$; $F2(2, 180) = .8$, $p = .462$; *not -en-* responses: $F1(2, 180) = 1.5$, $p = .237$; $F2(2, 180) = .9$, $p = .915$]. Apparently, the Right Bias does have influence on the choice of the linking element, but not on the response latency. The upper two right panels of Fig. 1 show the effect of the Left Bias on the reaction latencies for both *-en-* and *not -en-* responses. Participants react faster when the response follows the bias than when the response conflicts with the bias.

We also tested whether the Left and Right Bias of the preceding experimental trial and the choice made for that trial had an influence on the choice, in addition to the effects of the Left and Right Bias. A logit analysis that included the preceding Left and Right Bias and the preceding choice along with the Left and Right Bias themselves revealed a significant effect only for the Left and Right Bias, both with respect to the choices and with respect to the response latencies.

Summing up, we replicated the finding that linking elements in novel Dutch compounds are chosen on the basis of analogy. As in Krott et al. (2001), both the bias of the left constituent family and the right constituent family show a main effect on the choice. The left constituent family also plays a crucial role for the response latencies: Responses that follow the bias require less processing time. The right constituent family, however, which has been shown to have a weaker effect on the choices, does not predict the response latencies.

What kind of cognitive processes might account for these findings? In order to explain the absence of an effect on the reaction times of the right constituent family, we propose to distinguish between an early selection process and a series of processing stages during which activation accumulates up to response initiation. In the early selection process, a linking element is chosen based on maximum likelihood, i.e., on the distribution of linking elements in both the left and right constituent families. Along the lines of the interactive activation model that has been outlined in Krott et al. (2001), we hypothesize that the lemma representations of the constituents of the novel compound activate the corresponding left and right constituent families.

The compounds in these families then activate the linking elements they contain. Since the left constituent family has a stronger effect than the right constituent family, we assume that the members of the left constituent family are initially higher activated than the members of the right constituent family. The higher activation of the left constituent family implies that the linking elements receive more activation from members of the left constituent family. After the initial activation of linking elements, the activation flows back and forth between the linking elements and the constituent families. The activation accumulates until the selected linking element has become sufficiently activated to reach an awareness threshold, which initiates the response. We hypothesize that the alternating activation flow between the constituent families and the linking elements leads to an exponential increase of the activation of the already more highly activated members of the left constituent family and a comparably slow increase of activation of the member of the right constituent family. This results in response latencies that appear to be based solely on the bias in the left constituent family, with the relatively weak contribution of the right constituent family being masked.

AN INTERACTIVE ACTIVATION MODEL

Introduction

In previous studies, we used AML (Skousen, 1989) and TiMBL (Daelemans et al., 2000) as analogical tools to model the choice of linking elements in novel compounds (Krott et al., 2001, in press). The selection of a linking element can be understood as a classification problem, and both these models are very much suited to this task. However, they are restricted in that they are not designed to model response latencies. We therefore decided to develop a symbolic activation model that incorporates, in part, aspects of TiMBL.

Figure 2 illustrates the connectivity structure for a simple lexicon with 10 compounds for the situation in which the novel compound *schaap-?-oog* (''sheep's eye'') has been conceptualized, with *schaap* in the modifier position (LEFT) and *oog* in the head position (RIGHT). As outlined in the previous section, initially, activation flows from the lemma representation of *schaap* to the word forms (the lexemes according to Levelt, 1989) with which it is connected and modified by the (identical) weights w_1 (model parameter: IG-weight left constituent, γ_1). Similarly, activation flows from the lemma representation of *oog* to the word forms of the compounds in which *oog* is the head and modified by the (identical) weights w_2 (model parameter: IG-weight right constituent, γ_1). The weight w_1 is larger than the weight w_2, in accordance with the empirical finding that the left constituent family has greater analogical force than the right constituent family. Only members of the two constituent families are activated. Therefore, compounds such as the members of the left constituent family of *lam* (see Fig. 2) are not activated. From the activated word forms, activation flows further to the linking elements. The word forms with the linking element *-en-* support the linking element *-en-;* similarly, the word forms with the linking elements *-s-* and -0- support the linking elements *-s-* and *-0-*, respectively. The linking element that receives the highest activation from the word forms is the linking element that is most likely to be selected. Following selection, activation flows back from the linking elements to the word forms and from the word forms to the lemma representations. The forward activation flow from the lemmas to the linking elements and the backward activation flow from the linking elements to the lemmas jointly constitute one resonance cycle. Generally, a series of resonance cycles, the time steps of the model,

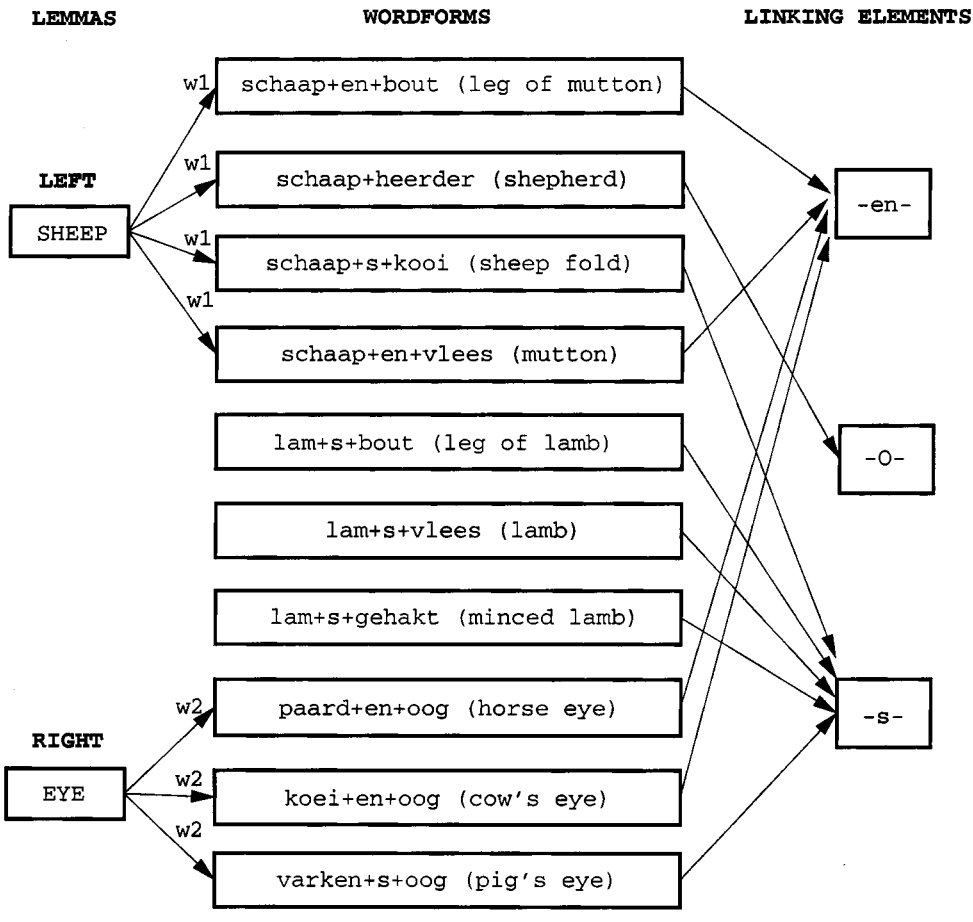

FIG. 2. Connectivity of a simple lexicon: lemmas (left), word form representations [lexemes according to Levelt (1989), middle], and linking elements (right).

are required for a selected linking element to become sufficiently activated to reach the level of awareness required for response execution.

Apart from the weights for the left and right constituents, the model contains some other parameters: The general decay δ determines the activation decay of nodes in the network. The resonance weight ρ specifies the strength of the activation resonance, while the activation is only passed on from compounds whose activation exceeds a similarity threshold ϑ. The overall bias for *-en-* that has been observed in the experiment can be adjusted by changing the parameter β. The strength of the bias increases if the parameter ξ has a value above zero. Furthermore, one can specify whether the frequency of the compounds should affect the activation increase. A linking element reaches awareness once its activation reaches a threshold θ. In order to guarantee that the model terminates, the number of maximal time steps has to be set. In the following subsection, we explain the model's details. The reader may skip that part without losing the main thread of the argument.

Technical Details

The connectivity structure of the model is defined formally by means of two matrices, **C** and **E.** Let **C** denote the connectivity matrix of n_w word forms and n_F feature–value pairs as follows:

$$\mathbf{C} = \begin{bmatrix} i_{1,1} & i_{1,2} & \cdots & i_{1,n_F} \\ i_{2,1} & i_{2,2} & \cdots & i_{2,n_F} \\ \vdots & \vdots & & \vdots \\ i_{n_w,1} & i_{n_w,2} & \cdots & i_{n_w,n_F} \end{bmatrix}, \tag{1}$$

with

$$i_{w,F} = \begin{cases} 1 & \text{if word form } w \text{ is connected to feature } F, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In the present working example (Fig. 2), $n_w = 10$ and $n_F = 2$. The relevant features are the left and right constituent positions (modifier and head); the values of these features are the lemma representations of *schaap* and *oog*, respectively. Similarly, let **E** denote the connectivity matrix of the n_w words with the n_e exponents (the three linking elements studied here) as follows:

$$\mathbf{E} = \begin{bmatrix} i_{1,1} & i_{1,2} & \cdots & i_{1,n_e} \\ i_{2,1} & i_{2,2} & \cdots & i_{2,n_e} \\ \vdots & \vdots & & \vdots \\ i_{n_w,1} & i_{n_w,2} & \cdots & i_{n_w,n_e} \end{bmatrix}, \tag{3}$$

with

$$i_{w,e} = \begin{cases} 1 & \text{if word form } w \text{ is connected to exponent } e, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

These two matrices completely define the connectivity in the model. For the present example, these two matrices have the following form:

$$\mathbf{C} = \begin{Bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{Bmatrix} \quad \text{and} \quad \mathbf{E} = \begin{Bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{Bmatrix}. \tag{5}$$

Note that the connectivity matrix **C** differs for each pair of left and right target constituents. For every such pair, we consider only that section of the lexical connectivity that is relevant for precisely this pair of constituents.

The forward activation flow from the lemmas to the linking elements is codetermined by the weights on the connections between the lemmas and the word forms as well as by the frequencies of these word forms. Let $\boldsymbol{\gamma}$ denote the vector of feature information gain weights according to Daelemans et al. (2000) as follows:

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{n_F} \end{bmatrix}, \tag{6}$$

with

$$\gamma_i = w_i = H(\boldsymbol{\varepsilon}) - H_i'(\boldsymbol{\varepsilon}). \tag{7}$$

To understand this equation, let $F_i \in \mathcal{F}$ denote the ith feature, and let this feature assume values F_{ij}, $j = 1, 2, \ldots, c(F_i)$, with $c(F_i)$ the cardinality of the set of values that F_i can assume. In the present working example, $c(F_1) = 5$ as there are five different left constituents in the lexicon and $c(f_2) = 6$. Furthermore, let $e_i \in \boldsymbol{\varepsilon}$, $i = 1, 2, \ldots, c(\boldsymbol{\varepsilon})$, with $c(\boldsymbol{\varepsilon})$ the cardinality of $\boldsymbol{\varepsilon}$, denote the ith exponent. In our working example, we have three exponents, hence $c(\boldsymbol{\varepsilon}) = 3$. The entropy of ε equals the following:

$$H(\boldsymbol{\varepsilon}) = -\sum_{i=1}^{c(\varepsilon)} P(e_i)\log_2 P(e_i), \tag{8}$$

with $P(e_i)$ the relative frequency of the ith linking element among the word forms. The entropy of $\boldsymbol{\varepsilon}$ is reduced by introducing knowledge of the value of feature F_i. The weighted entropy of $\boldsymbol{\varepsilon}$ given knowledge of the value of F_i is as follows:

$$H_i'(\boldsymbol{\varepsilon}) = \sum_{j=1}^{c(F_{ij})} P(F_{ij})H(\boldsymbol{\varepsilon}\,|\,F_{ij}), \tag{9}$$

with $P(F_{ij})$ the relative frequency of the jth value of F_{ij} among all the values that feature F_i assumes, and with $H(\boldsymbol{\varepsilon}|F_{ij})$ the entropy calculated over those exponents that are linked with word forms sharing the jth value of feature F_i. Thus, the information gain weight of feature i can be understood as the reduction in entropy achieved by introducing knowledge of the value of feature F_i. Note that all connections from the modifier position share the same weight, the information gain weight of the left constituent, and that likewise the connections from the head position share the same information gain weight. All information gain weights are easily estimated on the basis of the word forms in the lexicon. The model requires no further training.

In addition to the connection weights, the model makes use of a vector $\boldsymbol{\varphi}$ of word frequency weights as follows:

$$\boldsymbol{\varphi} = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{n_w} \end{bmatrix}. \tag{10}$$

The frequency weight φ_i is a function φ of the CELEX frequency f_i of word form w_i as follows:

$$\phi(f_i) = \frac{1.0}{1.0 + \log(f_i)}. \tag{11}$$

Inverse frequency weighting favors the analogical contribution of the lower frequency words, the words that most clearly express the regularities in the lexicon (cf. Baayen & Sproat, 1996). It is in symmetric contrast with the noninverse frequency effect that arises when word forms directly feed articulation (Jescheniak & Levelt, 1994).

The pattern of activation values of the word forms after the first forward pass of activation is as follows:

$$\mathbf{s} = (\mathbf{C} \cdot \boldsymbol{\gamma}) * \boldsymbol{\varphi} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{n_w} \end{bmatrix} \tag{12}$$

and is a vector of by-word-form similarity scores. Each similarity score specifies how much activation a given word form will pass on to the exponent with which it is connected. By applying a thresholding function Θ, we obtain the equivalent of the standard k-NN distance sets, but now defined in terms of similarities instead of distances as follows:

$$\Theta(s_i, \vartheta) = \begin{cases} s_i & \text{if } s_i \geq \vartheta \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

with ϑ representing a similarity threshold. In the present simulation, the value of ϑ is set to zero. In other words, we have allowed even distant neighbors to codetermine the selection of the linking elements. But by choosing an appropriate value for ϑ, only those words that are sufficiently similar to the target input affect the activation of the linking elements.

The activation of the word forms is passed on to the exponents. The vector of activations of the exponents **e** after the first forward pass of activation has run its course equals the following:

$$\mathbf{e} = \mathbf{E}^T \cdot \mathbf{s} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n_e} \end{bmatrix} \tag{14}$$

The probability of selecting the *i*th linking element is as follows:

$$P(i) = \frac{e_i}{\sum_{j=1}^{n_e} e_j}. \tag{15}$$

When no frequency weighting is used, the resulting probabilities of the linking elements are identical to those obtained by applying the k-NN nearest neighbor algorithm with information gain weighting as developed in TiMBL.

Maximum likelihood selection according to Eq. (15) allows us to model the choices of the linking elements, but not the time required for executing an actual response. As the constituent family of the right constituent affected the choice of the linking elements but not the response latencies, we need a mechanism that introduces noise in such a way that the strongest factor, the left constituent family, masks the effect of the weaker factor, the right constituent family. In the present model, this is accomplished by means of resonance in the network. We assume that this resonance leaves activation traces, either in the connections, or in the activation levels of the word forms. As the word forms themselves are not the forms to be produced, we prefer to view the activation traces as accumulating in the connections. However, the following formal definition is neutral with respect to these interpretations.

We assume that the activation received by the wordforms from the lemmas during the initial forward pass of activation leaves an activation trace in the network of connections between the lemma layer and the wordforms, proportional to what we call the forward activation matrix **F:**

$$\mathbf{F} = (1 + \mathbf{s}) * \mathbf{C}. \tag{16}$$

Following maximum likelihood selection of a linking element, activation flows back from the exponents to the lemma layer, again increasing the activation in this network of connections, this time proportional to what we call the backward activation matrix **B,** indexed here for the initial time step $t = 1$ as follows:

$$\mathbf{B}_1 = (\mathbf{E} \cdot \mathbf{e}) * \mathbf{C}. \tag{17}$$

Let $\mathbf{A}_t$ denote the activation pattern at time step *t*, $t = 0, 1, 2, \ldots$, with $\mathbf{A}_0 = \mathbf{C}$. For $t = 1$, the first resonance cycle, we define the following:

$$\mathbf{A}_1 = \delta(\mathbf{A}_0 + \rho(\mathbf{F} + \mathbf{B}_1)), \tag{18}$$

with δ a general activation decay and ρ a resonance weight, a parameter allowing us to specify the granularity of the resonance. The state of the model at an arbitrary time step *t* is, in summary form:

$$\begin{aligned}
\mathbf{s}_t &= (\mathbf{A}_{t-1} \cdot \boldsymbol{\gamma}) * \boldsymbol{\varphi} \\
\mathbf{e}'_t &= \mathbf{e}_{t-1} + \mathbf{E}^T \cdot \mathbf{s}_t \\
\mathbf{B}_t &= (\mathbf{E} \cdot \mathbf{e}'_t) * \mathbf{C} \\
\mathbf{A}_t &= \delta(\mathbf{A}_{t-1} + \rho(\mathbf{F} + \mathbf{B}_t)) \\
&= \delta(\mathbf{A}_{t-1} + \rho(([1 + \mathbf{s}_1] + [\mathbf{E} \cdot \mathbf{e}'_t]) * C)) \\
\mathbf{e}_t &= \mathbf{e}'_t + \mathbf{b}.
\end{aligned} \tag{19}$$

The last line specifies that the activation of the linking elements is modified by the vector **b.** This vector allows us to implement the observed response bias for the *-en-* linking element as follows:

$$\mathbf{e}_t = \begin{bmatrix} e_{\text{-en-}} \\ e_{\text{-Ø-}} \\ e_{\text{-s-}} \end{bmatrix} + \begin{bmatrix} \beta\xi^{t-1} \\ 0 \\ 0 \end{bmatrix}, \tag{20}$$

with $\beta > 0$ and $\xi \geq 1$. Note that this bias for *-en-* increases during the resonance cycles when $\xi > 1$. In other words, we assume that the response bias is a task factor that is itself external to the connectivity in the lexical network.

A selected linking element reaches awareness once its activation has reached a present threshold value θ. The time step at which this threshold is reached is taken to represent the model's response latency. Model times exceeding a preset time limit are not taken into account, just as response latencies exceeding the time-out limit are not taken into account.

Simulation Results

A reasonable fit of this model to the present experimental data was obtained with the following parameter values: IG-weight left constituent: $\gamma_1 = 1.12$; IG-weight right constituent: $\gamma_2 = 0.10$; general decay: $\delta = 0.97$; resonance weight: $\rho = 0.05$; activation threshold: $\theta = 100.0$; and *-en-* bias parameters: $\beta = 2.5$ and $\xi = 1.2$, with time-out after 25 time steps, with frequency weighting and no similarity threshold ($\vartheta = 0$). Figure 1 presents a visual summary of the goodness of fit, and Table 3 shows

TABLE 3

Goodness-of-Fit Statistics: A Logit Analysis of the Observed and Expected Counts and Analyses of Variance for the Reaction Times Corresponding to the *-en-* and the *not -en-* Responses

	Observed	Expected
	Logit analysis of counts	
Left Bias	$F(2, 180) = 156.6$ $p < .001$	$F(2, 180) = 902.99$ $p < .001$
Right Bias	$F(2, 180) = 8.2$ $p < .001$	$F(2, 180) = 12.11$ $p < .001$
Interaction	$F(4, 180) = 0.37$ $p = .829$	$F(2, 180) = 2.76$ $p = .029$
	Analysis of variance of log RT for *-en-*	
Left Bias	$F(2, 169) = 16.3$ $p < .001$	$F(2, 180) = 177.89$ $p < .001$
Right Bias	$F(2, 169) = 0.8$ $p = .462$	$F(2, 180) = 0.68$ $p = .510$
Interaction	$F(4, 169) = 0.1$ $p = .969$	$F(2, 180) = 0.19$ $p = .943$
	Analysis of variance of log RT for *not -en-*	
Left Bias	$F(2, 166) = 10.8$ $p < .001$	$F(2, 147) = 165.14$ $p < .001$
Right Bias	$F(2, 166) = 0.9$ $p = .915$	$F(2, 147) = 0.02$ $p = .978$
Interaction	$F(4, 166) = 0.4$ $p = .437$	$F(2, 147) = 0.90$ $p = .468$

that the same main effects that can be observed for the experimental data also emerge in the simulation. The same holds for the interaction term for left and right constituent bias, except for the logit analysis of the observed and expected counts. The model suggests a minor interaction that does not receive clear support from the empirical data. However, given that the model has no sources of variation other than those provided by the constituent families, this small interaction, that qualitatively is of the same kind as the nonreliable interaction visible in the empirical results, is not a source of serious concern. We conclude that our morphological resonance model provides a reasonable first approximation of the role of analogical cognition in the production of Dutch noun–noun compounds.

GENERAL DISCUSSION

In this study we addressed three related questions. First, does the distribution of linking elements in the right and left constituent families predict the choice of the linking elements in novel compounds not only in an off-line cloze task but also in a speeded decision task? Second, does this distribution also predict the speed with which these decisions are made? Third, is it possible to model the processes involved in the on-line experiment in a psycholinguistically plausible way?

The on-line experiment that we presented in this study showed that indeed the effect of the left and right constituent families on the choice of linking elements in Dutch noun–noun compounds also occurs under time pressure. This effect is not restricted to the choices made by the participants—it also emerges in their response latencies. We observed an asymmetry between the choice pattern and the reaction time pattern, however. Both the left and the right constituent families play a role in the choices, while for the response latencies it is only the left constituent family that is a predictor.

We interpreted these results in terms of a two-stage cognitive process. In the first stage, a linking element is selected on the basis of a maximum likelihood selection following initial activation spreading from the left and right constituent families to the linking elements. In the second stage, the activation of the selected linking element increases until it reaches an awareness threshold, after which the selected response can be initiated. We assume that in this process the relatively weak effect of the right constituent is masked by the additional variability of this second processing stage.

We have made this explanation more explicit by means of a computational simulation model. In this model, the first processing stage is captured by a spreading activation mechanism that is mathematically equivalent to a k-NN nearest neighbor classifier as used in computerized approaches to natural language processing (e.g., Daelemans et al., 2000). The second processing stage is captured by allowing activation to resonate in the lexical network.

A simulation study of the results of our experiment showed that our model can account for the analogical effects on both the choices and the response latencies. An advantage of the present psycholinguistic model compared to linguistic models of analogy such as AML and TiMBL is that it captures, within a spreading activation framework, the patterns in the data not only with respect to the choices, but also with respect to the reaction times.

The results that we have obtained are difficult to account for within a traditional approach based on symbolic rules. As mentioned in the introduction, the rules that have been formulated for the linking elements in Dutch have insufficient predictive power (Krott et al., in press). Given the syntagmatic nature of rules, this lack of

predictive power is not so surprising. By definition, symbolic rules do not have access to constituent families. They may be sensitive to particular properties of left and right constituents, for instance, to whether the first constituent ends in a vowel. In order to capture generalizations, rules can only be sensitive to properties of words and not to specific words.

Interestingly, the phenomenon that we have studied here is not syntactic in nature, but paradigmatic. The left and right constituent families both constitute positional paradigms. In fact, each such paradigm constitutes its own domain of markedness. A positive bias for *-en-* as linking element indicates that this linking element is the locally unmarked form.

The notion of local markedness as introduced by Tiersma (1982) concerns the fact that some marked forms may behave as unmarked forms. For instance, noun plurals denoting objects that naturally occur in pairs or groups (e.g., ''eyes'' and ''sheep'') may serve as attractors in language change, a role that is normally reserved for the unmarked singular forms of words such as ''nose'' and ''nightingale.'' Not surprisingly, locally unmarked plurals are much more frequent than their corresponding singulars than marked plurals, which tend to be less frequent than their singulars. They are also conceptually more central than their singulars. Although linking elements lack this conceptual aspect, they share the property of being locally unmarked with plural forms such as ''eyes.'' Just as ''eyes'' occurs, for the domain of the lemma EYE, more often than the singular ''eye,'' a locally unmarked linking element with a large positive bias in the relevant constituent family occurs more often than the other linking possibilities. For the local domains of constituent families, the formally unmarked linking element *-0-*, which also occurs in the majority (69%) of Dutch compounds, may be rare and, if so, locally marked. Furthermore, markedness and the constituent family bias have in common that they are both graded in nature.

Finally, markedness theory claims that unmarked forms are easier to process than marked forms (Dressler, Mayerthaler, Panagl, & Wurzel, 1987). Given that the left constituent families constitute independent markedness domains, the shorter response latencies of the locally unmarked linking elements, the dominant linking elements in their own local markedness domains, are exactly as expected. From a methodological point of view, it is interesting to find that classic structuralist notions such as markedness and paradigmatics can help to understand a graded analogical phenomenon such as the realization of linking elements in Dutch noun–noun compounds.

REFERENCES

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

Baayen, R. H., & Sproat, R. (1996). Estimating lexical priors for low-frequency morphologically ambiguous forms. *Computational Linguistics,* **22,** 155–166.

Daelemans, W., Zavrel, J., Van der Sloot, K., & Van den Bosch, A. (2000). *TiMBL: Tilburg memory based learner reference guide,* version 3.0, *Technical Report ILK 00-01*. Tilburg, The Netherlands: Computational Linguistics Tilburg University.

Dressler, W., Mayerthaler, W., Panagl, O., & Wurzel, W. (1987). *Leitmotifs in natural morphology*. Amsterdam: Benjamins.

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

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **20**(4), 824–843.

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

Krott, A., Schreuder, R., & Baayen, R. H. (in press). Analogical hierarchy: Exemplar-based modeling of linkers in Dutch noun-noun compounds. In R. Skousen (Ed.), *Analogical modeling: An exemplar-based approach to language.* London: Benjamins.

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge, MA: The MIT Press.

Mattens, W. H. M. (1984). De voorspelbaarheid van tussenklanken in nominale samenstellingen (The predictability of linking phonemes in nominal compounds). *De nieuwe taalgids*, **7,** 333–343.

Plank, F. (1976). Morphological aspects of nominal compounding in German and certain other languages: What to acquire in language acquisition in case the rules fail? In G. Drachman (Ed.), *Salzburger Beiträge zur Linguistik: Akten des 1. Salzburger Kolloquiums über Kindersprache* (pp. 201–219). Tübingen: Gunter Narr.

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

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

Tiersma, P. M. (1982). Local and general markedness. *Language*, **58,** 832–849.

Van den Toorn, M. (1982). Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I (Tendencies for the regulation of linking phonemes in nominal compounds I). *De nieuwe taalgids*, **75**(1), 24–33.