Complex words in complex words

Andrea Krott, Robert Schreuder & R. Harald Baayen
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SUBMITTED, PLEASE DO NOT QUOTE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

Constituents of complex words can themselves be complex words. Some kinds of complex constituents appear more often than others. This study presents a quantitative investigation of this phenomenon. We show that many kinds of base words are significantly overrepresented or underrepresented. This holds not only for constituents of derived words, but also for constituents of compounds. We furthermore show that the degree of overrepresentation or underrepresentation correlates with word frequency, word length, and degree of productivity. We offer a functional explanation of this correlation in terms of processing and storage in the mental lexicon.

1 Introduction

It is well known that word formation rules accept several kinds of base words as input. As pointed out by Aronoff (1976), some kinds of base words of a given word formation rule give rise to more complex forms than others. He judged these differences in overall productivity important enough to warrant explicit mention in his formal definition of word formation rules. For the English prefix un-, e.g., he proposed the following rule in which the list of base words is "given roughly in order of productivity" (Aronoff 1976:63):

(20) Rule of negative un#

a. $[X]_{Adj} \rightarrow [un\#X]_{Adj}]_{Adj}$

semantics (roughly) $un\#X = notX$

b. Forms of the base

1. $X_V en$ (where en is the marker for past participle)
2. $X_V \# ing$
3. $X_V \# able$
4. $X + y$ (worthy)
5. $X + ly$ (seemly
6. $X \# ful$ (mindful)
7. $X - al$ (conditional)
8. $X \# like$ (warlike)

Corpus based data presented in Baayen and Renouf (1996) show that there are indeed substantial and significant differences in the numbers of base word types for un-. For instance, base words ending in -ed are very common, while base words ending in -less are virtually non-existent.

Given the fact that some kinds of base words occur more frequently than others, the following questions arise. First, are such unequal distributions simply reflections of the general proportions of complex words in the language? That many words in -ed and few words in -less give rise to un- formations would not be surprising at all if there would be

many more independent words in -ed than in -less available in the language. We would only be dealing with a non-trivial phenomenon if there were relatively few formations in -ed and many formations in -less. In other words, further research is called for only if the distribution of base words for a particular kind of complex word deviates significantly from the distribution of these words as independent words in the language.

Second, if it is indeed the case that non-trivial unequal distributions exist in the domain of derivational morphology, the question arises whether similar unequal distributions can be observed in the domain of compounding as well.

Third, if unequal distributions arise both in derivation and in compounding, then we are apparently dealing with a general phenomenon. But why would this phenomenon exist? What kind of factors might give rise to such unequal distributions?

In what follows, we first examine the distribution of base words for the Dutch suffix -heid, a suffix similar to the English suffix -ness. We introduce a statistical method for testing whether the distribution of base words differs from their distribution as independent words in the language. We will show that indeed the two distributions differ significantly.

We then extend our analysis to nominal compounds. We again observe that the extent to which words from morphological categories are used as constituents in compounds differs remarkably from the extent to which these words are used on their own. This suggests that we are indeed dealing with a general phenomenon.

Finally, we will show that frequency of use, linguistic complexity, and degree of productivity are important factors underlying the observed patterns.

2 Derived words in -heid

Table 1 in the Appendix summarizes various statistics for the different kinds of base words for the suffix -heid. These statistics have been calculated using the CELEX lexical database (Baayen, Piepenbrock, and Gullikers, 1995). This database contains frequency counts for some 120,000 morphologically analyzed lemmas based on a corpus of written Dutch of 42 million words. The first part of the table lists the main derivational affixes that give rise to words in -heid. The monomorphemic base words are labelled MONO, compounds are listed

as COMP, and adjectivized participles are listed as PART. The category listed as SEMI groups together those words in CELEX of doubtful morphological complexity (marked as *I* or *U* in CELEX; in what follows we will call this set semi-derived words). Finally, the remaining category SY contains almost exclusively synthetic compounds.

The second column of Table 1 (labelled f) lists the number of types in -heid for these sets of base words. The total number of formations in -heid is 2226, including 11 affixes not listed in Table 1 because they jointly account for 16 formations only. Note that we have substantial variation. Base words in -ig (groenig, 'greenish') give rise to 255 -heid formations, while base words in -s (schools, 'schoolish') give rise to only 18 formations. The question that we now have to ask ourselves is whether these differences in the number of types are in any sense remarkable from a statistical point of view.

In the past, this question has been addressed by investigating either rival affixes (e.g., -ness and -ity, see Aronoff, 1976; Anshen and Aronoff, 1988) or a set of affixes sharing the same word category for the base word (Baayen and Renouf 1996). The idea is that if a particular kind of base word gives rise to many formations in one affix and few formations in another affix, then, provided the difference is statistically reliable, we have genuine evidence that we are observing a non-trivial phenomenon worth further investigation.

In the present study we have opted for a different approach in which we compare for one kind of word formation the numbers of observed types for its various kinds of base words with the numbers that one would expect under chance conditions. To do so we make use of the binomial model. In the case of *-heid*, we regard the 2226 -heid formations as 2226 random trials. For a given kind of base word, we consider a trial to be successful if it yields a -heid formation with that particular kind of structure, i.e., if there is at least one token in our database for that particular type. (Note that the present statistical analysis has nothing to say about the token frequencies with which the individual types appear.) In other words, the f column in Table 1 can be viewed as listing the observed number of successes out of 2226 trials for each base word type.

How can we determine the expected number of successes? In the binomial scheme the expected number of successes equals np, where n denotes the number of trials and p the probability of success. In the case at hand, n is 2226. We can estimate p for a base type

X by the relative type frequency of X in the list of all adjectives in CELEX which form the attested set of words to which -heid can be attached in principle.[1] There are 9925 such potential input words of which 528 belong to the morphological category of -ig. The column labelled *fcel* lists this number of types in CELEX for all base word types. We can now estimate the probability of success for -ig to be $528/9925 = 0.0532$ and for -s to be $111/9925 = 0.0112$. The corresponding expected values are $0.0532 * 2226 = 118.42$ and $0.0112 * 2226 = 24.90$ respectively. Column E lists the expected numbers of types for all kinds of base words.

Comparing the observed and expected values, we observe far more -ig base words (255) than expected (118), while for -s the observed count (18) is smaller than expected (25). Are these differences between the observed and expected counts significant? Because the number of trials is large we can approximate the binomial model by a normal model and calculate Z-scores. To do so we need the standard deviation in addition to the expected counts. The standard deviation in the binomial model equals $\sqrt{np(1-p)}$, listed in Table 1 in column s. The Z-scores $((f-np)/\sqrt{np(1-p)})$ are listed in column Z and the corresponding Bonferroni-adjusted significance levels in column *sign* ($*$: 0.05; $**$: 0.01). Positive Z-scores imply overrepresentation, negative Z-scores imply underrepresentation. Table 1 shows that we have significant underrepresentation or overrepresentation for almost all base word types. The only exceptions are the adjectives in -s and the set of synthetic compounds. As a group, derived words are overrepresented as base words. The only affix that is significantly underrepresented is -achtig. The only other base word type exhibiting overrepresentation is the set of monomorphemic words. Significant underrepresentation is characteristic of compounds, participles, and semi-derived words.

We conclude that the phenomenon of overrepresentation and underrepresentation observed by Aronoff (1976), Anshen and Aronoff (1988), and Baayen and Renouf (1996) for English can also be observed for Dutch.

This phenomenon receives some qualitative support from the subset of *-heid* formations coined from adjectives in -ig (groenigheid, 'greenishness'). It has been observed that in some of these formations the suffix -ig does no longer contribute its own semantics: stommig means somewhat stupid, while stommigheid means 'stupidity'. This suggests that the sequence -igheid might be analyzed as a separate affix in its own right (Schultink, 1962; but see also

De Haas & Trommelen, 1993 who do not make this distinction). If the combination of -ig and -heid is indeed developing into a single unit, then this provides qualitative evidence paralleling our quantitative evidence that the morphological structure of the base word in a complex word should be taken into account. Differences in over- and underrepresentation might then go hand in hand with subtle differences in semantics.

3 Compounds

Can we observe similar patterns of overrepresentation and underrepresentation for compounds? If we are dealing with a general phenomenon, one would expect that the left and right constituents of compounds behave in a similar way as the base words underlying formations in -heid. We have explored this possibility for Dutch and German nominal compounds using the CELEX lexical databases for Dutch and German. The German database lists some 52,000 entries based on a corpus of 6 million wordforms. Table 2 lists the same statistics as presented in Table 1 for a partition of left and right constituents into six kinds of base words: Monomorphemic base words (MONO), semi-derived words (SEMI), derived words (DER), compounds (COMP), synthetic compounds (SY), and a small heterogeneous set of other kinds of complex words (O). In both languages none of these kinds of base words occur with frequencies that one would expect under chance conditions, as shown by the Z-scores and the associated probabilities. Just as for *-heid*, monomorphemic words are strongly overrepresented, while the compounds and to a lesser degree the synthetic compounds are underrepresented. Dutch and German diverge with respect to the set of derived words. In Dutch, derived words are overrepresented, while in German they are underrepresented. Interestingly, left and right constituents reveal exactly the same pattern, even though the right headedness of most compounds might have led to an asymmetry.

4 The role of word frequency

Is there any systematicity in the patterns of overrepresentation and underrepresentation observed in the previous section? Altmann (1988) suggests that higher frequency words are

more likely to appear as constituents in compounds than lower frequency words. If this hypothesis generalizes to complex words in general, the following relation might hold:

> The higher the average word frequency for a given base word type, the higher the chance of it being overrepresented in complex words.

To test this hypothesis, we calculated the mean log frequency using the CELEX lexical database for each base word type, the column labelled *meanf* in Tables 1–2.[2] Figures 1–2 show that we indeed have a positive correlation between mean log frequency and Z-score. Figure 1 presents a scatterplot for the -heid data. Monomorphemic words have the highest mean log frequency and the highest positive Z-score, while compounds have a low mean log frequency and a large negative Z-score. The other kinds of base words are scattered between these extremes. Both a Pearson correlation analysis and a Spearman rank correlation analysis show that the correlation between mean log frequency and Z-score is reliable ($r = 0.58, t(13) = 2.56, p = 0.024; r_S = 0.53, p = 0.049$). The solid line in Figure 1 represents the corresponding mean squares regression line.

Place Figure 1 about here

Figure 2 presents similar scatterplots for the left and right constituents of Dutch and German compounds. As before, the monomorphemic words appear in the upper right corners of the scatterplots and the compounds in the lower left corners. Despite the small number of base word categories, the correlations between mean log frequency and Z-score are all reliable (left constituents Dutch: $r = 0.91, t(4) = 4.41, p = 0.012; r_S = 1, p = 0.030$; right constituents Dutch: $r = 0.91, t(4) = 4.41, p = 0.012; r_S = 1, p = 0.030$; left constituents German: $r = 0.92, t(4) = 4.55, p = 0.011; r_S = 0.94, p = 0.041$; right constituents German: $r = 0.89, t(4) = 3.96, p = 0.017; r_S = 0.94, p = 0.041$).

Place Figure 2 about here

Thus far, the data support our hypothesis that word frequency is an important factor co-determining the extent to which base words appear in complex words. As a final test, we calculated the mean log frequency and the Z-scores for the various kinds of derived words that

appear as left and right constituents in Dutch compounds. Tables 3–4 and Figure 3 summerize the results. The scatterplots reveal some outliers, notably the nominalizing suffixes -ing ('-ing') and -atie ('-ation') in the upper panel, and the nominalizing suffixe -ing ('-ing'), -er ('-er'), and -heid ('-ness') in the lower panel. Given this outlier structure, we have only calculated the Spearman rank correlations, which again show that we are dealing with reliable correlations (left derivations: $r_S = 0.71, p < 0.000$; right derivations: $r_S = 0.47, p = 0.007$). The solid lines in Figure 3 represent the least median squares regression lines.

Place Figure 3 about here

5 The role of word length

We have shown that the average word frequency of a particular kind of base word is an important factor co-determining its use in complex words. It is well known that word frequency is strongly correlated with word length. To show that this relation also holds for constituents in complex words, we divided the Dutch data in classes of different lengths. Table 5a lists the classes for left compound constituents when measuring length in terms of number of morphemes. Table 5b lists the classes when measuring length in terms of number of phonemes. In both tables, the column labelled f contains the number of words in each class, and the column $meanf$ lists their mean log frequency. Comparable data for base words used in -heid formations and for right constituents of compounds are presented in Table 6 and Table 7. To illustrate the strong negative correlation between length and frequency, we consider the left constituents of compounds in some more detail. The top left panel of Figure 4 plots mean log frequency as a function of number of phonemes ($r_s = -0.99; p < 0.000$).

Place Figure 4 about here

Given this negative correlation between word frequency and word length, we also expect the following relation to hold:

> The longer a base word, the higher the chance of it being underrepresented in complex words.

To test this hypothesis, we calculated for each length class a Z-score, as we did for the constituent types in the previous section. The results of the Z-score statistics are listed in Tables 5–7. As expected, the Z-scores reveal that both short base words and short compound constituents are indeed overrepresented, while long base words and long compound constituents are underrepresented. The top right panel of Figure 4 shows for the left constituents of compounds how the number of types in each phonemic length class (represented by dots) diverge from the expected number of types (represented by a solid line). For word length 1 the observed and expected number of types are nearly identical. For word lengths 2–7 the observed number of types exceeds the expected number of types, especially for word lengths 3–6. From lengths 8–19 the observed number of types is smaller than the expected number of types, especially for lengths 9–15. Note that there are relatively few types with very small or very large word length. We see the same pattern in the lower left panel of Figure 4 which plots the corresponding Z-scores as a function of word length.

The lower right panel of Figure 4 plots the length classes in the plane spanned by mean log frequency and Z-score. The underrepresented sets of constituents consist of words which are infrequent and long, while the overrepresented sets of constituents consist of words which are frequent and short. In sum, constituents in complex words reveal a correlational system in which word length, mean log frequency, and number of types are all interrelated.[3]

6 A productivity paradox

We have seen that word frequency and word length co-determine how often complex words appear as constituents in other complex words. Especially short and frequent words give rise to overrepresentation. Paradoxically, this suggests that those categories of base words that have a low category-conditioned degree of productivity are relatively more productive as constituents in other complex words than base words that have a high category-conditioned degree of productivity. The category-conditioned degree of productivity is defined as follows (Baayen 1992; see Baayen, 1994 for experimental evidence):

$$\mathcal{P} = \frac{V(1,N)}{N}, \tag{1}$$

with $V(1, N)$ the number of hapax legomena (types occurring once only) in a sample of N tokens of a given category. This statistic estimates the probability of sampling a word that has not yet been observed in the previous N tokens of the morphological category. Thus, a base word category with 1000 tokens and 50 hapax legomena has a category-conditioned degree of productivity equal to $\mathcal{P} = 0.05$. Another category with 10000 tokens and 50 hapax legomena has a category-conditioned degree of productivity equal to $\mathcal{P} = 0.005$. Note that the probability of sampling new unobserved types decreases as N increases. A category with many short and high-frequency words will have a large value of N and hence a lower $\mathcal{P}$ compared to a category with only a few high-frequency forms. This leads to the following paradox:

> The more productive an affix, the greater the degree to which it is underrepresented in other complex words. The less productive an affix, the greater the degree to which it is overrepresented in other complex words.

In other words, the relative productivity of an affix, i.e., the degree to which it is overrepresented, is negatively correlated with its category-conditioned degree of productivity.

To test this prediction, we first investigated the relation between underrepresentation and overrepresentation expressed in Z-scores with estimates of the category-conditioned degree of productivity. [4] Figure 5 plots categories in the plane of $\mathcal{P}$ and Z for base words of -heid formations (upper panel) and for left constituents of Dutch compounds (lower panel). The particular values of $\mathcal{P}$ are listed in Table 1 and Table 3 in the column labelled *prod*.

For the base categories of words in -heid we observe a trend in the expected direction. The category with the highest $\mathcal{P}$-value (-achtig, '-like') has the lowest Z-score. Conversely, the category with the lowest $\mathcal{P}$-value (-(e)lijk, '-able') has the highest Z-score. However, due to the small number of observations, the Spearman rank correlation is not fully reliable ($r_s = -0.52; p = 0.06$, one-tailed test). Interestingly, the object-modifying rival affixes -(e)lijk (verwerpelijk, 'objectionable') and -baar (*toepasbaar*, 'applicable') behave exactly as expected. Van Marle (1988) and Hüning and van Santen (1994) point out that -baar is productive and semantically transparent, while -(e)lijk is unproductive and appears in many semantically opaque words. This difference is reflected in the $\mathcal{P}$-values of these suffixes, and

indeed we observe that -(e)lijk has the higher Z-score.

Place Figure 5 about here

For the base categories appearing as left constituents in compounds (shown in the lower panel of Figure 5) we observe a very clear negative correlation between category-conditioned degree of productivity and Z-score ($r_s = -0.69, p = 0.0001$): the more productive categories have the lower Z-scores. These data show that word frequency and word length have to be considered in combination with degree of productivity when studying the contribution of morphological categories to the productivity of other complex words.

7 General Discussion

The aim of this paper has been to study the extent to which the productivity of derivation and compounding is influenced by the morphological structure of base words. We have first shown that the unequal contributions of different kinds of base words are extremely unlikely to be a chance phenomenon. We have further shown that the phenomenon of unequal contributions is not limited to derivation, but that it likewise occurs in the domain of compounding, both for left and right constituents. Finally, we have shown that the extent to which particular kinds of base words are overrepresented or underrepresented correlates with their mean frequency of use and their length (measured in number of phonemes or morphemes). Shorter and more frequent words are overrepresented, longer and less frequent words are underrepresented. Paradoxically, categories with a low degree of productivity are relatively more productive as constituents in other complex words.

The correlation of word frequency, word length, and category-conditioned degree of productivity on the one hand with the degree of overrepresentation (Z-scores) on the other hand explains 1/5 up to 1/3 of the variance in the data. This observation raises the following question. How can we understand this non-trivial role of word frequency, word length, and productivity as explanatory variables?

In all our calculations of expected numbers of types, we have assumed the null-hypothesis that all word types are equiprobable. The observed underrepresentation and overrepresen-

tation show that this null-hypothesis is incorrect. This raises the question in what way some words are more likely to be selected as a constituent than other words. From a psycholinguistic point of view, we can understand the finding that base word categories which comprise frequent words are overrepresented compared to categories comprising less frequent words in terms of the word frequency effect (e.g., Scarborough, Cortese, and Scarborough, 1977; Hasher and Zacks, 1984). The word frequency effect is the finding that higher frequency words are recognized and produced more quickly and accurately than lower frequency words. Assuming that a wide range of complex words is stored in the mental lexicon, the same word frequency effect applies to complex words as well (Baayen, Dijkstra, and Schreuder, 1997; Sereno and Jongman, 1997). This means that higher frequency complex words are more accessible as potential constituents than lower frequency words. A category of base words that contains many frequent formations will then be overrepresented.

Similarly, shorter words are easier to produce and recognize than longer words (e.g., Henderson, 1985, p. 470-471). Since higher frequency words tend to have more meanings and shades of meanings (Köhler, 1986; Altmann, Beöthy, and Best, 1982; Paivio, Yuille, and Madigan, 1968; Reder, Anderson, Bjork, 1974), they are also more likely to be selected during the process of conceptualization and lexical selection in speech production. Note, furthermore, that less productive and unproductive categories typically comprise higher frequency formations that tend to have more, and more opaque meanings. Such formations have to be stored in the mental lexicon in any case where they are readily available for further word formation. This explains the paradox that less productive categories are relatively more productive as constituents, a paradox that is entirely unexpected on the basis of the combinatorial properties of word formation rules only. From this perspective, any summary description of a word formation rule is incomplete without a quantitative description of the pattern of overrepresentation and underrepresentation of its base words.

In traditional analyses of morphological productivity, the role of phonological, semantic and syntactic constraints has figured prominently (Van Marle, 1985; Booij, 1977). The morphological restrictions formalized by Aronoff (1976) as part of generative word formation rules have received little attention. The present results, however, show that these morphological restrictions are statistically non-trivial: constituent length, constituent frequency,

and the productivity of the morphological category to which the constituent belongs form a correlational complex that codetermines the overall productivity of a word formation rule. We have offered a quantitative, partial explanation in terms of the mental lexicon, but further qualitative research is neccessary in order to fully understand how such morphological restrictions arise.

Author Note

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). 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.

Footnotes

[1]Our counts of the number of types in CELEX to which -heid can attach in principle are raw counts. Our counts do not differentiate between base words for which a -heid formation is plausible versus implausible (Matthews, 1974:221–222), nor do they take possible semantic restrictions on the affixation of -heid (Bertram, Baayen, & Schreuder, 1999) into account. Here we simply assume that the effects of such constraints are uniformly distributed over the input domains. Further quantitative research is required here.

[2]The logarithmic transformation largely eliminates the Zipfian skewness from the word frequency distributions and allows us to gauge more precisely the central tendency in the data. In addition, the human processing system is also sensitive to log frequency rather than absolute frequency.

[3]In the presented data, word frequency and length are so strongly correlated that it proved to be impossible to ascertain the extent to which these factors might play an independent role.

[4]The CELEX lexical database does not provide counts of hapax legomena. We have therefore approximated the category-conditioned degree of productivity by the ratio of dis legomena (words occurring twice) to the total number of tokens of a category in CELEX.

References

Altmann, Gabriel (1988). Hypotheses about compounds. In Glottometrika 10, Rolf Hammerl (ed.). Brockmeyer, Bochum, 100–107.

Altmann, Gabriel, Beöthy, Erzsebet and Best, Karl-Heinz (1982). Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35 (5), 537–543.

Anshen, Frank and Aronoff, Mark (1988). Producing morphologically complex words. Linguistics 26, 641–655.

Aronoff, Mark (1976). Word Formation in Generative Grammar, MIT Press, Cambridge, Mass.

Baayen, R. Harald (1992). Quantitative aspects of morphological productivity. In Yearbook of Morphology 1991, Geert E. Booij and Jaap van Marle (eds.). Kluwer Academic Publishers, Dordrecht, 109–149.

Baayen, R. Harald (1994). Productivity in language production. Language and Cognitive Processes 9, 447–469.

Baayen, R. Harald and Renouf, Antoinette (1996). Chronicling The Times: Productive Lexical Innovations in an English Newspaper. Language 72, 69–96.

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

Baayen, R. Harald, Piepenbrock, Richard and Gulikers, Leon (1995). The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bertram, Raymond, Baayen, R. Harald and Schreuder, Robert (1999). Effects of family size for derived and inflected words. To appear in Journal of Memory and Language.

Booij, Geert E. (1977). Dutch Morphology. A Study of Word Formation in Generative Grammar. Foris, Dordrecht.

De Haas, Wim and Trommelen, Mieke (1993). Morfologisch handboek van het Nederlands. SDU, Den Haag.

Hasher, Lynn and Zacks, Rose T. (1984). Automatic processing of fundamental information. The case of frequency of occurrence. American Psychologist 39, 1372–1388.

Henderson, Leslie (1985). Issues in the modelling of pronunciation assembly in normal reading. In Surface dyslexia: Neuropsychological and cognitive studies on phonological reading, K. Patterson, J. Marshall and M. Coltheart (eds.). Lawrence Erlbaum, London, 459–508.

Hüning, Matthias and Van Santen, Arianne (1994). Productiviteitsveranderingen: de adjectieven op -lijk en -baar. Leuvense Bijdragen 83 (1), 1–29.

Köhler, Reinhard (1986). Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Brockmeyer, Bochum.

Marle, Jaap v. (1985). On the Paradigmatic Dimension of Morphological Creativity. Foris, Dordrecht.

Marle, Jaap v. (1988). Betekenis als factor bij produktiviteitsverandering (iets over de deverbale categorieën op *-lijk* en *-baar*). Spektator 17, 341–359.

Matthews, Peter Hugo (1974). Morphology. An introduction to the theory of word structure. Cambridge University Press, London.

Paivio, Allan, Yuille, John C. and Madigan, Stephan (1968). Concreteness, imagery, and meaningness values for 925 nouns. Journal of Experimental Psychology Monograph.

Reder, Lynne M., Anderson, John R. and Bjork, Robert A. (1974). A semantic interpretation of encoding specificity. Journal of Experimental Psychology 102, 648–656.

Scarborough, Don L., Cortese, Charles and Scarborough, Hollis S. (1977). Frequency and repetition effects in lexical memory. Journal of Experimental Psychology: Human Perception and Performance 3, 1–17.

Schultink, Henk (1962). De Morfologische Valentie van het Ongelede Adjectief in Modern Nederlands [The morphological valency of the simplex adjective in modern Dutch]. Van Goor & Zonen, Den Haag.

Sereno, Joan and Jongman, Allard (1997). Processing of English inflectional morphology. Memory and Cognition 25, 425–437.

Appendix

Table 1: **Base word classes of -heid formations**: f: number of types; $meanf$: mean log token frequency; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level ($*$: 0.05; $**$: 0.01); SEMI: doubtful morphologically complex words; MONO: monomorphemic words; SY: synthetic compounds; COMP: compounds consisting of two nouns and a possible linking morpheme; PART: present participle.

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
on-	264	3.44	812	0.0818	182.12	12.93	6.33	**	0.0007
-ig	255	3.90	528	0.0532	118.42	10.59	12.90	**	0.0002
-(e)lijk	188	4.85	335	0.0338	75.13	8.52	13.25	**	0.0001
-baar	91	3.27	181	0.0182	40.60	6.31	7.98	**	0.0011
-(e)loos	71	3.92	157	0.0158	35.21	5.89	6.08	**	0.0003
-erig	46	2.69	130	0.0131	29.16	5.36	3.14	*	0.0049
-zaam	30	4.36	32	0.0032	7.18	2.67	8.53	**	0.0004
-achtig	28	2.32	251	0.0253	56.29	7.41	-3.82	**	0.0072
-s	18	3.77	111	0.0112	24.90	4.96	-1.39		0.0004
-matig	9	4.55	17	0.0017	3.81	1.95	2.66		0.0002
SEMI	313	3.91	2015	0.2030	451.93	18.98	-7.32	**	0.0002
MONO	311	5.54	593	0.0597	133.00	11.18	15.92	**	0.0000
SY	249	3.06	1169	0.1178	262.19	15.21	-0.87		0.0018
PART	246	4.26	1270	0.1280	284.84	15.76	-2.46		0.0001
COMP	91	2.98	1184	0.1193	265.55	15.29	-11.41	**	0.0012

$n = \sum f = 2226$

Table 2: **Compound constituents**: f: number of types; $meanf$: mean log token frequency; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level ($*$: 0.05; $**$: 0.01); MONO: monomorphemic words; SEMI: doubtful morphologically complex words; DER: derived words; COMP: compounds consisting of two nouns and a possible linking morpheme; SY: synthetic compounds; O: remaining words not belonging to any of the other classes.

a. Dutch left compound constituents

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$
MONO	2592	4.37	5180	0.0952	592.24	23.15	86.39	**
SEMI	1567	3.59	10647	0.1957	1217.29	31.29	11.18	**
DER	1335	3.23	9911	0.1822	1133.14	30.44	6.63	**
COMP	658	2.36	28168	0.5178	3220.50	39.41	-65.02	**
SY	45	2.63	970	0.0178	110.90	10.44	-6.31	**
O	23	2.51	876	0.0161	100.15	9.93	-7.77	**

$n = \sum f = 6220$

b. Dutch right compound constituents

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$
MONO	2342	4.37	5180	0.0952	502.36	21.32	86.29	**
SEMI	1288	3.59	10647	0.1957	1032.55	28.82	8.86	**
DER	1184	3.23	9911	0.1822	961.17	28.04	7.95	**
COMP	428	2.36	28168	0.5178	2731.73	36.30	-63.47	**
SY	20	2.63	970	0.0178	94.07	9.61	-7.71	**
O	14	2.51	876	0.0161	84.95	9.14	-7.76	**

$n = \sum f = 5276$

c. German left compound constituents

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$
MONO	1534	3.17	4355	0.2101	539.07	20.64	48.21	**
DER	579	2.44	5849	0.2822	724.00	22.80	-6.36	**
SEMI	283	2.53	1964	0.0947	243.11	14.83	2.69	*
COMP	155	1.89	7638	0.3685	945.45	24.44	-32.35	**
O	14	1.95	779	0.0376	96.43	9.63	-8.56	**
SY	1	2.03	145	0.0070	17.95	4.22	-4.01	**

$n = \sum f = 2566$

d. German right compound constituents

class	f	$meanf$	$fcel$	p	E	s	Z	$sign$
MONO	1401	3.17	4355	0.2101	478.78	19.45	47.42	**
DER	546	2.44	5849	0.2822	643.02	21.48	-4.52	**
SEMI	191	2.53	1964	0.0947	215.92	13.98	-1.78	
COMP	104	1.89	7638	0.3685	839.70	23.03	-31.95	**
O	33	1.95	779	0.0376	85.64	9.08	-5.80	**
SY	4	2.03	145	0.0070	15.94	3.98	-3.00	**

$n = \sum f = 2279$

Table 3: **Derivation classes in Dutch left compound constituents**: $f1$: number of types; $meanf$: mean log token frequency; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level ($*$: 0.05; $**$: 0.01).

class	$f1$	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
-ing	551	3.72	1986	0.0365	227.06	14.79	21.900913	**	0.0003
-atie	156	3.64	373	0.0069	42.65	6.51	17.417801	**	0.0003
-er	127	3.22	881	0.0162	100.73	9.95	2.639323		0.0007
-heid	83	2.86	1759	0.0323	201.11	13.95	-8.466550	**	0.0013
-ie	48	3.58	292	0.0054	33.38	5.76	2.536264		0.0003
-iteit	30	3.64	159	0.0029	18.18	4.26	2.776615		0.0004
-te	22	4.27	67	0.0012	7.66	2.77	5.184279	**	0.0001
-tie	20	4.10	51	0.0009	5.83	2.41	5.870515	**	0.0001
-sel	18	3.35	118	0.0022	13.49	3.67	1.228884		0.0005
-schap	16	3.55	113	0.0021	12.92	3.59	0.857924		0.0005
-ier	16	3.59	41	0.0008	4.69	2.16	5.226879	**	0.0005
-aat	16	3.67	56	0.0010	6.40	2.53	3.794897	**	0.0003
-eur	16	3.63	53	0.0010	6.06	2.46	4.040119	**	0.0002
ge-	15	2.69	468	0.0086	53.51	7.28	-5.287044	**	0.0008
-aar	14	3.29	128	0.0024	14.63	3.82	-0.166053		0.0011
-age	13	3.64	35	0.0006	4.00	2.00	4.499729	**	0.0000
-ant	12	3.25	63	0.0012	7.20	2.68	1.788444		0.0011
-ling	11	3.39	92	0.0017	10.52	3.24	0.148577		0.0007
-ent	11	4.74	28	0.0005	3.20	1.79	4.359850	**	0.0002
-ement	11	3.96	27	0.0005	3.09	1.76	4.504902	**	0.0003
-st	9	5.99	24	0.0004	2.74	1.66	3.777509	**	0.0000
-iek	8	3.62	83	0.0015	9.49	3.08	-0.483909		0.0003
-nis	8	4.64	32	0.0006	3.66	1.91	2.270370		0.0001
-erij	7	2.61	169	0.0031	19.32	4.39	-2.807586		0.0033
on-	7	3.77	62	0.0011	7.09	2.66	-0.033289		0.0006

Table 3 (continued)

class	$f1$	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
-ist	7	2.95	90	0.0017	10.29	3.21	-1.03		0.0013
-in	7	3.50	49	0.0009	5.60	2.37	0.59		0.0002
-ster	3	2.42	139	0.0026	15.89	3.98	-3.24	*	0.0045
-or	3	2.90	52	0.0010	5.95	2.44	-1.21		0.0010
-uur	3	3.72	20	0.0004	2.29	1.51	0.47		0.0005
-dom	3	4.28	15	0.0003	1.71	1.31	0.98		0.0000
-aal	3	3.84	7	0.0001	0.80	0.89	2.46		0.0000
-es	2	3.00	53	0.0010	6.06	2.46	-1.65		0.0019
-erie	2	2.91	10	0.0002	1.14	1.07	0.80		0.0036
-ateur	0	3.05	16	0.0003	1.83	1.35	-1.35		0.0029

$n = \sum f1 = 1278$

Table 4: **Derivation classes in Dutch right compound constituents**: $f2$: number of types; $meanf$: mean log token frequency; $fcel$: number of class members in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level ($*$: 0.05; $**$: 0.01).

class	$f2$	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
-ing	354	3.72	1986	0.0365	192.60	13.62	11.85	**	0.0003
-er	172	3.22	881	0.0162	85.44	9.17	9.44	**	0.0007
-heid	79	2.86	1759	0.0323	170.59	12.85	-7.13	**	0.0013
-atie	73	3.64	373	0.0069	36.17	5.99	6.14	**	0.0003
-ie	65	3.58	292	0.0054	28.32	5.31	6.91	**	0.0003
ge-	53	2.69	468	0.0086	45.39	6.71	1.13		0.0008
-sel	33	3.35	118	0.0022	11.44	3.38	6.38	**	0.0005

Table 4 (continued)

class	$f2$	$meanf$	$fcel$	p	E	s	Z	$sign$	$prod$
-aar	29	3.29	128	0.0024	12.41	3.52	4.71	**	0.0011
-iek	26	3.62	83	0.0015	8.05	2.83	6.33	**	0.0003
-te	25	4.27	67	0.0012	6.50	2.55	7.26	**	0.0001
-ster	21	2.42	139	0.0026	13.48	3.67	2.05		0.0045
-aat	21	3.67	56	0.0010	5.43	2.33	6.68	**	0.0003
-schap	16	3.55	113	0.0021	10.96	3.31	1.52		0.0005
-tie	16	4.10	51	0.0009	4.95	2.22	4.97	**	0.0001
-iteit	15	3.64	159	0.0029	15.42	3.92	-0.11		0.0004
-ent	14	4.74	28	0.0005	2.72	1.65	6.85	**	0.0002
-eur	14	3.63	53	0.0010	5.14	2.27	3.91	**	0.0002
-erij	13	2.61	169	0.0031	16.39	4.04	-0.84		0.0033
-st	13	5.99	24	0.0004	2.33	1.53	7.00	**	0.0000
-or	11	2.90	52	0.0010	5.04	2.24	2.65		0.0010
-ant	10	3.25	63	0.0012	6.11	2.47	1.57		0.0011
-nis	10	4.64	32	0.0006	3.10	1.76	3.92	**	0.0001
-ist	8	2.95	90	0.0017	8.73	2.95	-0.25		0.0013
-ement	8	3.96	27	0.0005	2.62	1.62	3.33	*	0.0003
-age	8	3.64	35	0.0006	3.39	1.84	2.50		0.0000
-es	7	3.00	53	0.0010	5.14	2.27	0.82		0.0019
-ier	6	3.59	41	0.0008	3.98	1.99	1.02		0.0005
-uur	6	3.72	20	0.0004	1.94	1.39	2.92		0.0005
on-	4	3.77	62	0.0011	6.01	2.45	-0.82		0.0006
-ateur	4	3.05	16	0.0003	1.55	1.25	1.97		0.0029
-dom	4	4.28	15	0.0003	1.45	1.21	2.11		0.0000
-ling	3	3.39	92	0.0017	8.92	2.98	-1.98		0.0007
-in	3	3.50	49	0.0009	4.75	2.18	-0.80		0.0002
-erie	1	2.91	10	0.0002	0.97	0.98	0.03		0.0036
-aal	1	3.84	7	0.0001	0.68	0.82	0.39		0.0000

$n = \sum f2 = 1146$

Table 5: **Length of left Dutch compound constituents**: f: number of types; $meanf$: mean log token frequency; $fcel$: number of types in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level ($*$: 0.05; $**$: 0.01).

a. Morphemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	2591	4.37	5192	0.0954	593.61	23.17	86.20	**
2	1936	2.69	28674	0.5271	3278.35	39.38	-34.09	**
3	117	2.24	9474	0.1741	1083.18	29.91	-32.30	**
4	5	2.48	408	0.0075	46.65	6.80	-6.12	**

$n = \sum f = 4649$

b. Phonemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	8	5.91	11	0.0002	1.26	1.12	6.01	**
2	63	4.44	138	0.0025	15.78	3.97	11.90	**
3	615	4.47	1204	0.0221	137.66	11.60	41.14	**
4	790	4.49	1539	0.0283	175.96	13.08	46.96	**
5	680	4.00	1804	0.0332	206.25	14.12	33.55	**
6	575	3.32	2809	0.0516	321.16	17.45	14.54	**
7	561	3.01	4491	0.0826	513.46	21.70	2.19	
8	493	2.88	5579	0.1025	637.86	23.93	-6.05	**
9	365	2.70	5452	0.1002	623.34	23.68	-10.91	**
10	220	2.57	4697	0.0863	537.02	22.15	-14.31	**
11	125	2.45	3798	0.0698	434.23	20.10	-15.39	**
12	73	2.33	3144	0.0578	359.46	18.40	-15.57	**
13	42	2.28	2563	0.0471	293.03	16.71	-15.02	**
14	14	2.17	1946	0.0358	222.49	14.65	-14.23	**
15	15	2.11	1465	0.0269	167.50	12.77	-11.94	**
16	4	1.99	1073	0.0197	122.68	10.97	-10.82	**
17	4	2.06	760	0.0140	86.89	9.26	-8.96	**
18	1	2.04	516	0.0096	59.00	7.64	-7.59	**
19	1	1.96	330	0.0061	37.73	6.12	-6.00	**

$n = \sum f = 4649$

Table 6: **Length of Dutch right compound constituents**: f: number of types; $meanf$: mean log token frequency; $fcel$: number of types in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level ($*$: 0.05; $**$: 0.01).

a. Morphemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	2343	4.37	5192	0.0954	503.52	21.34	86.19	**
2	1564	2.69	28674	0.5271	2780.80	36.26	-33.55	**
3	74	2.24	9474	0.1741	918.79	27.55	-30.67	**
4	3	2.48	408	0.0075	39.57	6.27	-5.84	**

$n = \sum f = 3984$

b. Phonemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	3	5.91	11	0.0002	1.07	1.03	1.87	
2	46	4.44	138	0.0025	13.38	3.65	8.93	**
3	588	4.47	1204	0.0221	116.76	10.69	44.10	**
4	775	4.49	1539	0.0283	149.25	12.04	51.96	**
5	659	4.00	1804	0.0332	174.95	13.01	37.22	**
6	500	3.32	2809	0.0516	272.42	16.07	14.16	**
7	409	3.01	4491	0.0826	435.54	19.99	-1.33	
8	366	2.88	5579	0.1025	541.05	22.04	-7.94	**
9	279	2.70	5452	0.1002	528.73	21.81	-11.45	**
10	173	2.57	4697	0.0863	455.51	20.40	-13.85	**
11	84	2.45	3798	0.0698	368.33	18.51	-15.36	**
12	44	2.33	3144	0.0578	304.90	16.95	-15.39	**
13	22	2.28	2563	0.0471	248.56	15.39	-14.72	**
14	12	2.17	1946	0.0358	188.72	13.49	-13.10	**
15	11	2.11	1465	0.0269	142.08	11.76	-11.15	**
16	5	1.99	1073	0.0197	104.06	10.10	-9.81	**
17	4	2.06	760	0.0140	73.70	8.52	-8.18	**
18	1	2.04	516	0.0095	50.04	7.04	-6.97	**
19	2	1.96	330	0.0061	32.00	5.64	-5.32	**
21	1	1.90	133	0.0024	12.90	3.59	-3.32	**

$n = \sum f = 3984$

Table 7: **Length of base words of -heid formations**: f: number of types; $meanf$: mean log token frequency; $fcel$: number of types in CELEX; p: probability of a word being a member of the class; E: expected number of types; s: standard deviation; Z: Z-score; $sign$: Bonferroni-adjusted significance level ($*$: 0.05; $**$: 0.01).

a. Morphemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
1	311	5.54	593	0.0597	133.00	11.18	15.92	**
2	1107	3.44	4921	0.4958	1103.69	23.59	0.14	
3	247	3.05	1109	0.1117	248.73	14.86	-0.12	
4	2	2.55	17	0.0017	3.81	1.95	-0.93	

$n = \sum f = 1667$

b. Phonemic length

length	f	$meanf$	$fcel$	p	E	s	Z	$sign$
2	6	6.25	21	0.0021	4.71	2.17	0.60	
3	112	5.86	186	0.0187	41.72	6.40	10.99	**
4	118	5.38	218	0.0220	48.89	6.92	9.99	**
5	141	4.47	323	0.0325	72.44	8.37	8.19	**
6	140	3.81	498	0.0502	111.69	10.30	2.75	*
7	189	3.71	720	0.0725	161.48	12.24	2.25	
8	265	3.51	1049	0.1057	235.27	14.51	2.05	
9	243	3.38	1040	0.1048	233.25	14.45	0.67	
10	211	3.30	928	0.0935	208.13	13.74	0.21	
11	125	3.11	673	0.0678	150.94	11.86	-2.19	
12	74	2.30	442	0.0445	99.13	9.73	-2.58	
13	22	3.06	227	0.0229	50.91	7.05	-4.10	**
14	11	2.83	145	0.0146	32.52	5.66	-3.80	**
15	6	2.80	83	0.0084	18.62	4.30	-2.94	*
16	4	2.38	40	0.0040	8.97	2.99	-1.66	

$n = \sum f = 1667$