Productivity in context: a case study of a Dutch suffix*

R. HARALD BAAYEN and ANNEKE NEIJT

Abstract

The Dutch suffix -heid, *like* -ness *in English, forms abstract nouns from adjectives. In this paper, we explore the hypothesis that* -heid *gives rise to two kinds of abstract nouns: on the one hand, nouns referring to concepts, and on the other hand, nouns referring to states of affairs. An examination of a corpus of newspaper Dutch reveals that the referential function of* -heid *is typical for the lowest-frequency words, most of which are neologisms. Conversely, its conceptual function is found predominantly among the highest-frequency words. Detailed investigations of the use of these two sorts of words in context showed that the high-frequency words tend to be less well anchored in their context than the low-frequency words, and that they pattern more as independent units. Our data argue against the view that productive word formation goes hand in hand with the absence of any storage of full forms in the mental lexicon. Instead, we claim that high-frequency formations with the productive suffix* -heid *are available in the mental lexicon, whereas low-frequency words and neologisms are produced and understood by rule.*

1. Introduction

Concepts can be expressed in various ways. In English, for example, many are expressed by means of monomorphemic words such as *human* and *speed*. Other concepts such as *childhood* and *grandmother* happen to be expressed by means of morphologically complex words, while yet others require phrasal expressions: *blue heron*, *Chinese chequers*, *high chair*. Comparing languages, one further finds that these may express very similar concepts in different ways. In Dutch, the words for *grandmother*, *blue heron*, and *Chinese chequers* are all monomorphemic (*oma*, *reiger*, *halma*), whereas *speed* is expressed by means of a complex noun,

Linguistics 35 (1997), 565–587 0024–3949/97/0035–0565

snelheid 'quickness'. Clearly, culturally well entrenched concepts such as 'speed', 'high chair', and 'grandmother' are part of our lexical inventory irrespective of their morphological structure.

At the same time, it is well known that many morphologically complex words do not express concepts other than those expressed by their base word. Particularly in the domain of verbal inflectional morphology, it is clear that *walks* refers to exactly the same kind of activity as *walk*, and that only the number and person marking is modified by suffixation of *-s*. Thus morphological structure may fulfill two different functions: concept formation, and syntactic functions such as agreement. Interestingly, these two functions can also be observed in the domain of derivational morphology, and even for one and the same derivational affix. Consider examples (1) and (2).

(1)	werkloosheid	'unemployment'
	snelheid	'speed, velocity'
	vrijheid	'freedom'
	waarheid	'truth'
(2)	gesitueerdheid	'situatedness'
	kwakkeligheid	'ailinglyness'
	onregeerbaarheid	'ungovernableness'
	pretentieloosheid	'pretentionlessness'

The nouns in (1) denote well-known concepts. For the nouns in (2), all of which occurred in issues of the Dutch newspaper *Trouw* in 1994, it seems less likely that their primary function is to express new concepts. Comparing the Dutch nouns with their English counterparts, we find that the English nouns in (1) are either monomorphemic or formed by less productive or unproductive suffixes, while in Dutch both the nouns in (1) and the nouns in (2) are coined by means of a very productive suffix, *-heid*, the functional equivalent of *-ness* in English. The examples in (3) illustrate that a similar asymmetry holds for English *-ness* and its translations in Dutch.

(3)	business	'zaak, bedrijf'
	illness	'ziekte'
	consciousness	'bewustzijn'
	happiness	'geluk'
(4)	wrongheadedness	'eigenzinnigheid'
	tenderheartedness	'teerhartigheid'
	stand-offishness	'hooghartigheid'
	disorderliness	'wanordelijkheid'

The examples in (3) require translation equivalents in Dutch that are simplex or formed by less productive or unproductive combinations. Conversely, the examples in (4) mirror those in (2) in that the English formations in *-ness* translate straightforwardly with *-heid* in Dutch.

All words listed in the left-hand columns of (1)–(4) are formally completely regular. Except for *business*, these formations are also semantically compositional, in the sense that both the base and the suffix clearly contribute to the meaning of the whole word. Nevertheless, the meanings of the words in (1) and (3) display a higher degree of semantic richness than the words listed in (2) and (4). For instance, *werkloosheid* does not simply denote a state of idleness, it refers to the social circumstances of being unemployed. Similarly, *snelheid* does not necessarily refer to one's being quick, it also and often denotes the physical concept of velocity.

The examples in (1)–(4) illustrate extreme positions on a cline ranging from conceptual denotation to a function that appears to be more similar to the function of inflectional morphology: reference tracking. Inflectional endings establish how words should be linked up within sentences. Pronominal reference fulfills a similar function for the participants in sentences and discourse. Reference tracking for states of affairs seems to be one of the functions of suffixes such as *-ness* and *-heid* in cases such as (2) and (4). Consider examples (5) and (6).

(5) Althans volgens Goslinga, en hij wijst ons op het 'meest gezag-hebbend' onderzoek. Maar welke instantie bepaalt de *gezaghebbendheid* van onderzoeken? (*Trouw*, September 16, 1993).
'That is, according to Goslinga, and he refers us to the "most authoritative" research. But which agency determines the authoritativeness of the research?'

(6) In het boek wordt veel en vooral 's nachts gedronken en gefilosofeerd en dus nauwelijks gevreeën. ... De *liederlijkheid* eist echter haar tol (*Trouw*, September 16, 1993).
'In the book there is a lot of drinking and philosophizing, especially at night, and hence hardly any lovemaking. ... The *dissoluteness*, however, has its price.'

The suffix *-heid* in *gezaghebbendheid* in (5) does not build a new concept from *gezaghebbend*, but refers to the property of being authoritative. Similarly, *liederlijkheid* in (6) refers to the state of affairs described in the preceding discourse. Thus *-heid* in *gezaghebbendheid* and *liederlijkheid* does not have as its primary function the formation of a new concept; instead its primary function is to refer to the property or the state denoted by the base word to which it is attached.

The distinction between the referential and conceptual functions of suffixes such as *-heid* is intuitively clear, and we are not the first to call attention to this distinction (see Kastovsky 1986). The goal of our study is to inquire whether this distinction is reflected in the context of use of formations with this suffix. With respect to its context of use, we expect that *-heid* words for well-established concepts display lesser degrees of contextual embedding than words in *-heid* with primarily a referential function. Degrees of contextual embedding might become apparent in the preceding and following semantically related words, as well as in the use of anaphoric modifiers such as possessive pronouns. We also expect the distinction between the referential and conceptual functions of *-heid* to be reflected in its productivity. The words for established concepts are formally regular, but semantically often idiosyncratic. From a processing point of view, the meanings of such words are stored in the mental lexicon. Conversely, the words with primarily a referential function appear to be both formally and semantically fully regular. It is for these words that comprehension and production are most likely to require processing by rule.

2. Contextual characteristics of *-heid* formations

In our comparison of the examples in (1) with those in (2), we observed two differences. First, the examples in (1) are more conceptlike than those in (2). Second, the English translation equivalents of the examples in (2) all make use of the suffix *-ness*, while those in (1) either are monomorphemic or make use of less productive or unproductive affixes. There is a third difference between these two sets that we have not yet mentioned. The examples in (1) all concern high-frequency words, those in (2) are words with a very low frequency of use. This correlation between frequency of use and concepthood is reminiscent of the correlation between frequency and irregularity (irregular forms are typically found in the higher frequency ranges) and the inverse correlation between frequency and productivity (large numbers of very low-frequency words are characteristic for productive word-formation rules). This suggests that independent concepts are most likely to appear among the highest-frequency formations, whereas the more productive use of *-heid* and especially its referential function might be primarily instantiated among the lowest-frequency words.

In order to investigate potential differences between conceptual and referential use of *-heid*, we have exploited this contrast between the highest-frequency formations and the lowest-frequency formations, the so-called hapax legomena, the words that occur once only in a given text or corpus. We selected 20 occurrences of 15 of the highest-frequency formations in *-heid* as well as 300 hapax legomena from a corpus of 85 issues of the Dutch newspaper *Trouw*. For each occurrence, we investigated its use in a context consisting of the ten preceding and the ten following lines. Nine graduate students in Dutch and the present authors independently analyzed the resulting 600 contexts such that each context was investigated at least twice. These analyses revealed a number of consistent differences between the high-frequency formations and the hapax legomena.

2.1. *Contextual anchoring*

If high-frequency words are more conceptlike than hapax legomena and as such relatively independent of context in their use, we predict that they are less well anchored in their context than hapax legomena. We operationalized the notion of textual anchoring in two ways. First, a complex word can be anchored in its context via its base word, which might appear by itself or as the constituent of other complex words. We will refer to this kind of contextual embedding, illustrated in (7) for *snelheid* 'speed' and its base *snel* 'fast', as morphological anchoring.

(7) Bruguera was verrast door de *snelheid* van de Hagenaar. "Richard was ongelooflijk *snel* aan het net."
'Bruguera was surprised by the speed of the player from The Hague. "Richard was surprisingly fast at the net"' (*Trouw*, April 11, 1994).

Second, formations in *-heid* might be semantically anchored in their context other than via their base words.

(8) Slechts Rousseau leek het niets te deren, zo makkelijk rolde hij naar de opzienbarende 2.07,51 uur, de achtste tijd ooit op de klassieke afstand gelopen. ... Ten Kate ... laat een bulderend gelach los als hij de snelheid van Rousseau op de laatste kilometer verneemt (*Trouw*, April 18, 1994).
'Only Rousseau seemed unaffected, he rolled so easily to the remarkable 2,07.51 hours, the eighth time ever run on the classical distance. ... Ten Kate ... bursts out laughing when he learns about the speed of Rousseau on the last kilometer.'

In (8), the notion of speed has already been topicalized by the expression 'the remarkable 2,07.51 hours', to which *snelheid* refers. Other examples of semantic anchoring are *zonder haren* 'without hairs' preceding *kaalheid* 'baldness' and *secularisatie* 'secularization' in the context of *buitenkerkelijkheid*, 'outside-churchly-ness'.

Our hypothesis is that in general the hapax legomena should reveal higher degrees of contextual anchoring than the high-frequency formations, as the latter are less dependent on context for their interpretation. This hypothesis applies straightforwardly to thematic anchoring, but special care is required for morphological anchoring. In general, higher-frequency complex words tend to be derived from higher-frequency base words. Furthermore, higher-frequency base words tend to give rise to more and to relatively high-frequency complex words. Consequently, the a priori probability that a complex word will occur in the vicinity of a morphologically related word is much higher for the high-frequency words than for the hapax legomena. Counts based on the CELEX lexical database show that for our sets of high-frequency words and hapax legomena the respective probabilities of occurrence of a morphologically related word are 0.00070 and 0.00038 respectively. In other words, for high-frequency words in *-heid* the probability of finding a morphologically related word in the context is approximately twice that for the hapax legomena, irrespective of the possible effect of morphological anchoring that we are interested in.

We estimated these probabilities as follows. For the hapax legomena and the high-frequency words separately, we counted the number of tokens of words that contain the corresponding base words. For the 15 high-frequency words, we counted a total of 444,781 such word tokens, excluding the tokens of the 15 *-heid* words themselves. Each type has on average 444,781/15 = 29,652.07 morphologically related words according to the CELEX lexical database (Baayen et al. 1993), which is based on a corpus of 42,380,000 words. We can therefore estimate the probability of such a related word by the sample relative frequency, 29,652.07/42,380,000 = 0.00070. For the 296 hapax legomena, each type has on average 4,794,542/296 = 16,197.78 morphologically related tokens in the corpus, which amounts to an estimated probability of 16,197.78/42,380,000 = 0.00038.

Thus it appears that two opposing forces are at play. On the one hand, we may expect the high-frequency formations to reveal more morphological anchoring, simply because their base words are more frequent. On the other hand, our hypothesis is that hapax legomena require more anchoring in general, and hence also more morphological anchoring than high-frequency words. Since we do not know what the balance of these two forces might be, we cannot formulate a prediction concerning possible differences in the amount of morphological anchoring. An additional complicating factor is that probably hapax legomena and high-frequency formations will both reveal more morphological anchoring than expected under chance conditions. After all, we are dealing with cohesive texts, and in cohesive texts words will not be uniformly distributed. The minimum that we may expect to find, at least if our hypothesis is correct, is a difference in the probabilities for the observed numbers of contexts with morphological anchoring given a null hypothesis of a uniform (random) distribution.

Table 1 summarizes the number of contexts with morphological anchoring (and possibly thematic anchoring), with only thematic anchoring but no morphological anchoring, and with no anchoring at all, for the high-frequency nouns and for the hapax legomena. Figure 1 presents the corresponding bar chart. Contexts with no anchoring are in the majority. Overall, anchoring occurs for only 22% of the contexts in our sample. Nevertheless, there are significant differences in the behavior of the sets of high-frequency words and the hapax legomena ($X^2(2) = 16.57, p < 0.001$). The most important difference concerns the amount of thematic anchoring. As expected, we count significantly more contexts with thematic anchoring for the hapax legomena (13.5%) than for the high-frequency words (4.3%, $p < 0.001$, one-tailed proportions test). This higher degree of thematic anchoring suggests that hapax legomena are more dependent on their context than are high-frequency words.

Turning to the amount of morphological anchoring recorded, we find that the number of contexts with morphological anchoring is slightly higher for the high-frequency words (15.5%) than for the hapax legomena (11.5%). This difference, however, is not significant ($p > 0.10$, using a two-tailed proportions test as we have no a priori prediction concerning

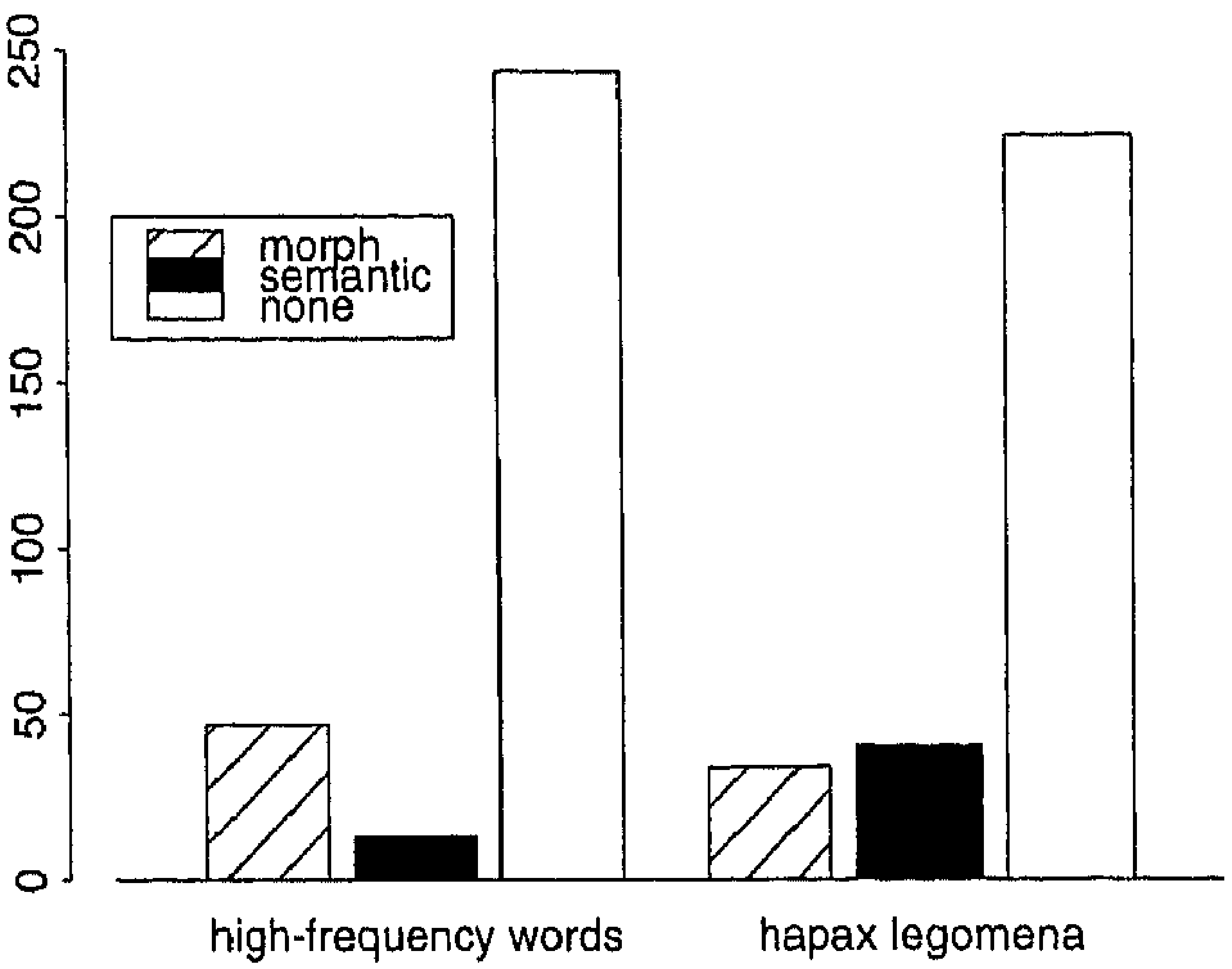
Table 1. *Count of contexts with and without anchoring for high-frequency words and hapax legomena in* -heid

	High-frequency nouns		Hapax legomena		
	no.	(%)	no.	(%)	Total
Morphological[a]	47	(15.5)	34	(11.5)	81
Semantic[b]	13	(4.3)	40	(13.5)	53
None	243	(80.2)	222	(76.0)	465
Total	303	(100.0)	296	(100.0)	599

$X^2_{(2)} = 16.57, p < 0.001.$

a. Morphological anchoring: formal anchoring through the base word.

b. Semantic anchoring: semantic-thematic anchoring (other than via morphological anchoring) with full inter-rater agreement.



morph: formal morphological anchoring through the stem
semantic: semantic-thematic embedding (without morphological support) with full interrater agreement

Figure 1. *Bar plot for the counts of kinds of contextual anchoring for hapax legomena and high-frequency words in* -heid

the direction of a possible difference in morphological anchoring). Apparently, the opposing forces of frequency and context-dependence are more or less balanced. A closer examination of our data reveals that the absence of a significant difference between the proportions of morphological anchoring for high-frequency words and hapax legomena is due to a high degree of morphological anchoring that masks the difference that one would expect under chance conditions. In other words, without morphological anchoring, the proportions of contexts with morphologically related words of the high-frequency formations on the one hand and the hapax legomena on the other would have been significantly different in favor of the former.

To see this, we can proceed as follows. First, we estimate the probability P that a given context contains at least one morphological anchor. Using the binomial distribution, these probabilities are estimated by

$$P = 1 - (1 - p)^n,$$

with p the probability of morphologically related words and n the (average) number of words in the context. For the high-frequency words,

$$P = 1 - (1 - 0.0007)^{175} = 0.1153;$$

for the hapax legomena, we have

$$P = 1 - (1 - 0.00038)^{189} = 0.0697.$$

Next, again using the binomial model, we estimate the expected number of contexts with at least one instance of morphological anchoring by $N \cdot P$, with N the number of contexts. For the high-frequency formations, we thus expect $303 * 0.115 = 34.93$ such contexts; for the hapax legomena, the expected count is $296 * 0.0697 = 20.63$. The corresponding standard deviations ($\sqrt{NP(1-P)}$) are 5.559 and 4.381 respectively. Finally, we use the normal approximation to the binomial distribution to calculate the probability that the observed large number of contexts with morphological anchoring is due to chance. For the high-frequency words, we find that $Z = (47 - 34.93)/5.55908 = 2.171$, and $Z = (34 - 20.63)/4.381 = 3.052$ for the hapax legomena. The corresponding one-tailed probabilities are 0.015 and 0.001 respectively. Thus, we observe a significant effect of anchoring for both high-frequency and low-frequency words, but the observed number of contexts with morphological anchoring for the hapax legomena is less likely (by a factor 10) to have arisen by change than the number of such contexts for the high-frequency formations.

Figure 2 plots the percentages of semantic-thematic anchoring (upper panel) and morphological anchoring (bottom panel) for the 15 high-frequency words of our study as well as the overall percentage for the hapax legomena considered jointly. The plots have to be interpreted with care, as the counts for the individual high-frequency words are quite low (see Table 2), and many of the observable differences are statistically not reliable. Nevertheless, these plots reveal the kind of variation in our data. With respect to semantic-thematic anchoring, we see that there is only one word with a higher percentage, *schoonheid* 'beauty'. Six high-frequency words do not reveal any semantic-thematic anchoring in our data, and eight show lower percentages than the hapax legomena. This plot shows that the overall differences between the hapax legomena and the high-frequency formations do not arise due to extreme values for particular words, and that it is robust across words.

Turning to the plot for morphological anchoring, the reader will observe that there is substantial variation among the high-frequency words. Nine words have a higher degree of morphological anchoring than the hapax legomena, while for six words, this percentage is lower. Clearly, morphological anchoring is not typical for the hapax legomena in the same way as semantic-thematic anchoring. Given the greater statistical likelihood for higher-frequency formations to have more morphological anchoring, the finding that nevertheless roughly one-third of

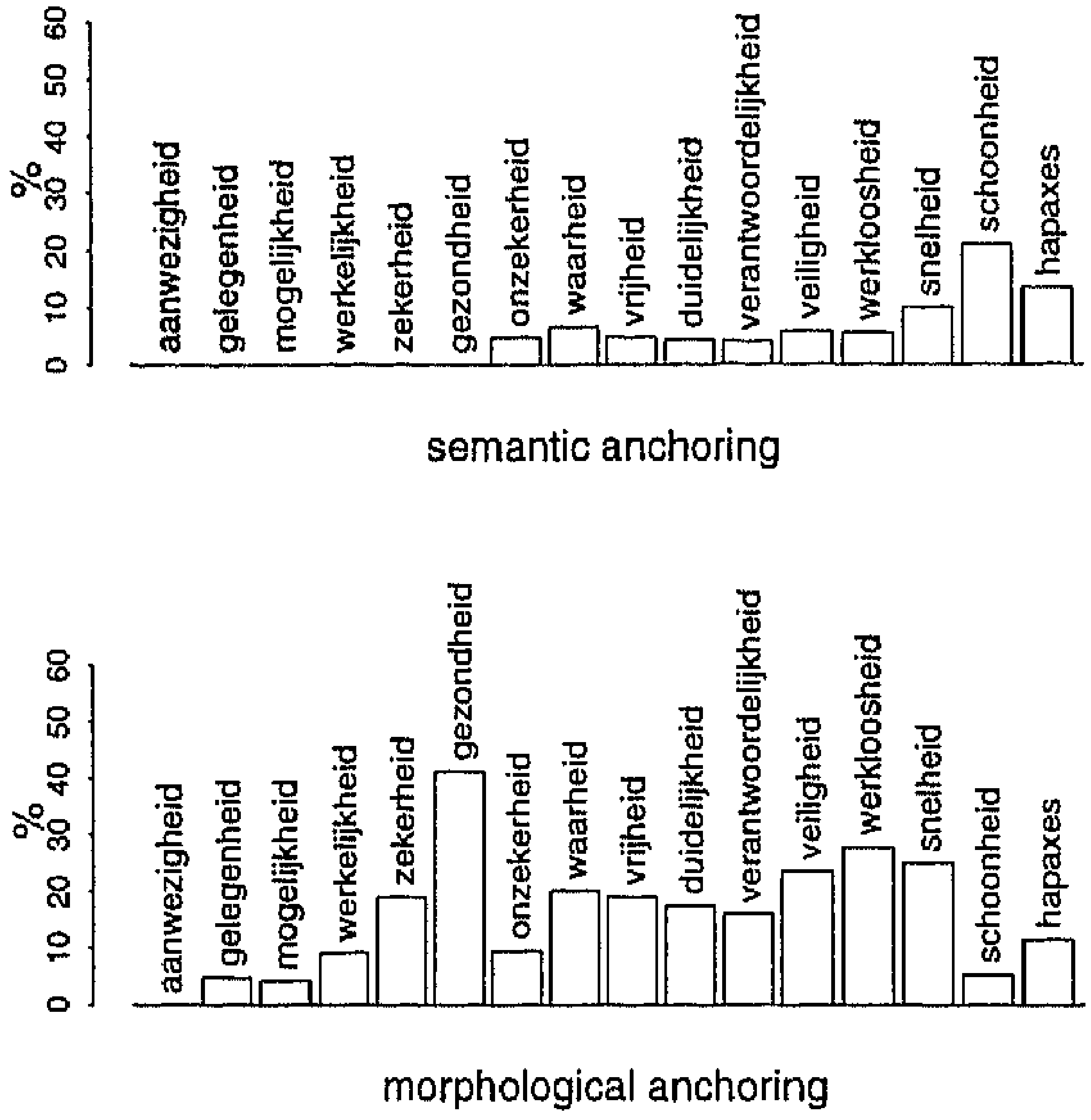


Figure 2. *Bar plot of the percentage of contexts with semantic-thematic anchoring (upper panel) and morphological anchoring (lower panel) for each of the 15 high-frequency words individually as well as for the combined set of all hapax legomena*

Table 2. *Count of contexts with morphological anchoring (morph) and with semantic-thematic anchoring (sem), as well as the total number of contexts for the 15 high-frequency words individually and for the set of hapax legomena considered jointly*

	sem	morph	total
aanwezigheid	0	0	19
gelegenheid	0	1	21
mogelijkheid	0	1	24
werkelijkheid	0	2	22
zekerheid	0	4	21
gezondheid	0	7	17
onzekerheid	1	2	21
waarheid	1	3	15
vrijheid	1	4	21
duidelijkheid	1	4	23
verantwoordelijkheid	1	4	25
veiligheid	1	4	17
werkloosheid	1	5	18
snelheid	2	5	20
schoonheid	4	1	19
hapax legomena	40	34	296

the high-frequency words show the same or a lower degree of morphological anchoring is surprising.

An additional factor that might lead to a greater degree of morphological anchoring for at least some high-frequency words is topicality. Words such as *gezondheid* 'health' and *werkloosheid* 'unemployment' denote topics that are regularly discussed in some detail in our newspaper. By contrast, formations such as *aanwezigheid* 'presence' and *mogelijkheid* 'possibility' are much less probable as the topic of an article, and it seems to us that most of our hapax legomena are similarly nontopical in nature. If correct, this line of reasoning adds topicality to the set of dimensions on which hapax legomena and high-frequency words may differ.

Three of the high-frequency words are to a greater or lesser extent semantically opaque: *schoonheid* 'beauty' (from *schoon*, 'clean, beautiful'), *gelegenheid* 'occasion' (from *gelegen*, 'situated, convenient'), and *werkelijkheid* 'reality' (from *werkelijk*, 'really, actual'). Might semantic opacity be correlated with the extent of anchoring? With respect to semantic-thematic anchoring, *schoonheid* reveals the highest degree of anchoring in our data, but on the other hand *gelegenheid* and *werkelijkheid* have the lowest possible score, 0. With respect to morphological anchoring, the opaque formations do not appear with high scores, but there are transparent formations that reveal similarly low scores (e.g. *aanwezigheid* 'presence' and *mogelijkheid* 'possibility'). As far as we can see, then, anchoring is not confounded with semantic transparency.

In sum, we find more semantic-thematic anchoring for the hapax legomena than for the high-frequency words. Our counts suggest that morphological anchoring is more common for the high-frequency words than for the hapax legomena. At the same time, the possible higher topicality of high-frequency words and the greater a priori likelihood of morphological anchoring for high-frequency words suggest that the observed extent of morphological anchoring for the hapax legomena is surprisingly high. We therefore conclude that our hypothesis that hapax legomena are characterized by a higher degree of contextual anchoring than high-frequency words is supported by our investigation of the use of formations in *-heid* in our newspaper corpus.

2.2. *Pre-text and post-text*

If the observed differences between high-frequency and low-frequency formations are linked with differences in contextual anchoring, then we would also expect a difference in contextual preparation. More specifically, we predict that, to the extent that anchoring is preparational in nature, the cases of anchoring should appear primarily in the preceding context (henceforth pre-text) rather than in the following context (henceforth post-text). Conversely, if our counts are independent of contextual preparation, then we should find roughly equal amounts of anchoring in pre-text and post-text.

To obtain some insight into the distribution of sorts of anchoring for the high-frequency words and the hapax legomena, we have counted all individual instances of anchoring for the various formations. Note that this procedure differs from that used in the preceding section, where each context was counted only once. This restrictive way of counting made it possible to estimate the various probabilities for the amount of anchoring that we have exploited. The counts for which we have opted here, non-restrictive overall counts of all the markup assigned by our raters, provide a complementary window on the pattern in our data: all instances of anchoring are now evaluated.

Table 3 and Figure 3 summarize our counts. As expected, we observe more overall anchoring in the pre-text (515) than in the post-text (141). In a loglinear analysis, position — in pre-text or post-text — is a significant main effect ($F(1, 2) = 824.42, p < 0.01$). This supports our hypothesis that morphologically complex words are morphologically or thematically prepared in the preceding context.

Interestingly, a difference in the counts can be observed for the two sorts of anchoring. There is very little morphological anchoring in the post-text (38 instances versus 278 in the pre-text), compared to what we find for semantic-thematic anchoring (103 instances in the post-text, versus 237 in the pre-text). This interaction is significant ($F(1,2) = 158.44$,

Table 3. *Amount of anchoring (cumulative over raters) as a function of position (in the pre-text versus in the post-text), frequency (high-frequency formations versus hapax legomena), and kind of anchoring (morphological versus semantic-thematic)*

	High-frequency nouns		Hapax legomena	
	pre-text	post-text	pre-text	post-text
Morphological	183	25	95	13
Semantic	95	45	142	58
None	390		352	

Loglinear model (conditional on the presence of anchoring):
position $p < 0.001$.
position : anchoring $p < 0.01$.
anchoring : frequency $p < 0.01$.

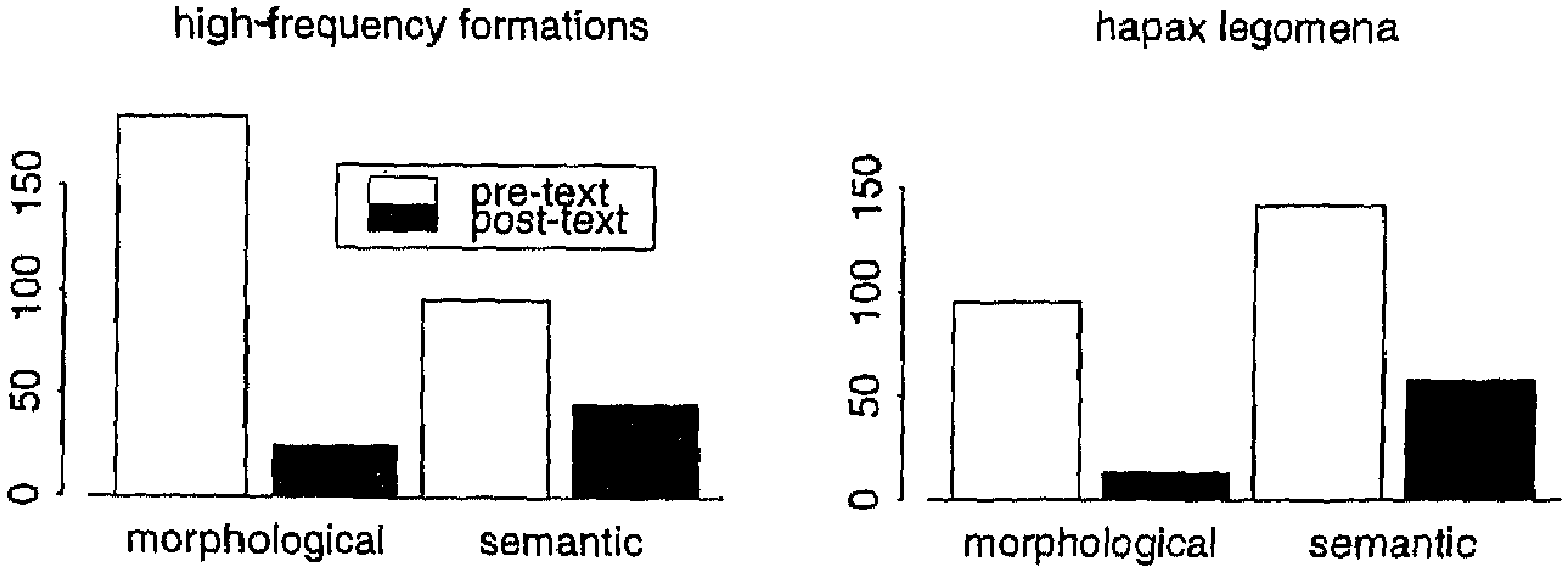


Figure 3. *Degree of anchoring for high-frequency formations and hapax legomena in* -heid, *subcategorized for morphologically motivated contexts ("morphological") versus semantic-thematic contexts ("semantic"), and for anchoring in the preceding ("pre-text") versus the following ("post-text") discourse*

$p < 0.01$). Possibly, the stylistic convention to avoid repeated use of the same word or morpheme is responsible for this difference. The *-heid* formations display large amounts of morphological anchoring in the pre-text. We suspect that, following the complex word, the likelihood of again using a word from the same morphological family quickly diminishes due to stylistic "saturation." To avoid stylistic saturation, writers are more likely to turn to semantically related words when further developing the topic under discussion. This may underlie the observed higher amount of semantic-thematic anchoring compared to morphological anchoring in the post-text.

We also find more morphological anchoring (208) than semantic-thematic anchoring (140) for the high-frequency words, while for the hapax legomena, we observe the reverse: more semantic-thematic anchoring (200) than morphological anchoring (108). In the loglinear analysis, this interaction is also significant ($F(3,2) = 71.01$, $p < 0.02$).[1] This interaction is fully in line with the counts presented in the preceding section, where we similarly observed that semantic-thematic anchoring occurred more often with the hapax legomena. But while the number of contexts with morphological anchoring for the high-frequency formations in these counts was not significantly higher than that for the hapax legomena, the number of instances of morphological anchoring in these contexts analyzed here is significantly higher for the high-frequency formations (208) than for the hapax legomena (108) ($p < 0.001$, proportions test). The higher number of individual instances of morphological anchoring (compared to the number of contexts with at least one instance of morphological anchoring) is to be expected on the basis of our finding that high-frequency words have a substantially higher probability of occurring in the neighborhood of a morphologically related word.

In sum, our counts show that anchoring typically occurs in the pre-text and is less often found in the post-text. This suggests that the use of morphologically complex words in *-heid* is contextually prepared. In addition, these counts, in which we have considered all occurrences of anchoring, further support our initial counts in showing that semantic-thematic anchoring prevails for the hapax legomena. In the following sections we present additional case studies that provide further support for our claims.

2.3. *Occurrences in titles*

In our *Trouw* corpus, there are positions where words are unlikely to be contextually anchored. For instance, the titles of articles, the titles of radio and television programs, and the descriptions of words in crossword puzzles do not lend themselves to contextual anchoring, as there is no pre-text and in the latter cases no post-text either.

For these positions in the text our hypothesis predicts that high-frequency formations should occur more often than hapax legomena. Table 4 and Figure 4 summarize our counts for 20 journal issues. Not surprisingly, the majority of formations in *-heid* do not occur in titles or crossword puzzles. Interestingly, of the 16 formations that do occur in these positions, 13 are high-frequency words and 3 are hapax legomena. The proportion of high-frequency words (13/161 = 9.3%) is significantly

Table 4. *Number of instances of high-frequency words and hapax legomena in* -heid *in 20 randomly selected journal issues*

	High-frequency nouns		Hapax legomena		
	no.	(%)	no.	(%)	Total
In title[a]	13	(9.3)	3	(2.0)	16
Not in title[b]	148	(90.7)	140	(98.0)	288
Total	161	(100.0)	143	(100.0)	304

$X^2_{(1)} = 5.91$, $p < 0.02$ (with continuity correction).

a. In title: appearance in the title of an article, in a crossword puzzle, or in television or radio announcements.

b. Not in title: occurrence in the body of an article.

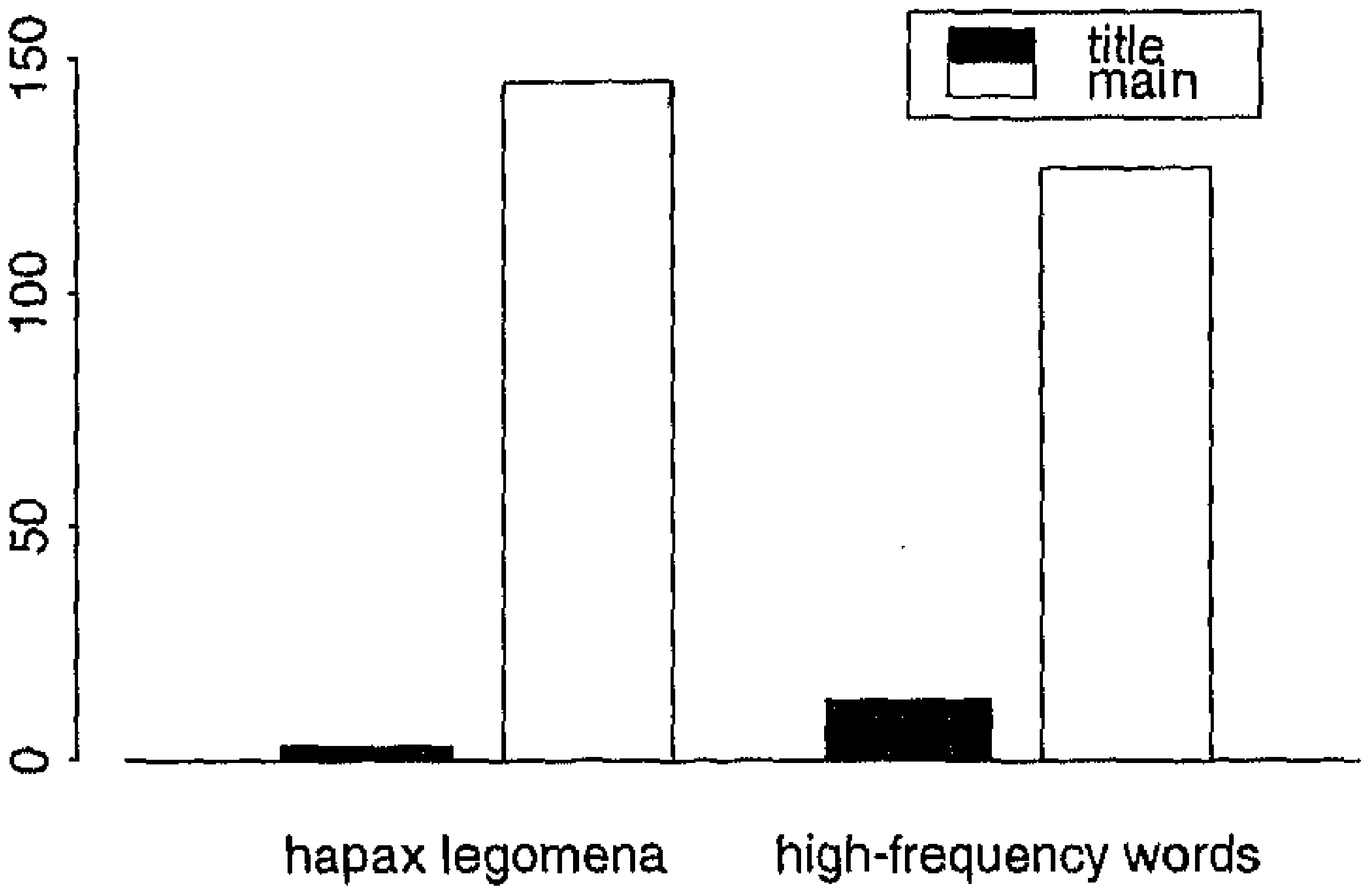


Figure 4. *Bar plot for the numbers of hapax legomena and high-frequency formations that appear in titles, crossword puzzles, and television program announcements ("title"), and for the numbers of occurrences in the main body of the text ("main")*

higher ($p < 0.001$) than the corresponding proportion of hapax legomena (3/143 = 2.0%).

Although these counts support our hypothesis that high-frequency words in *-heid* are less dependent on context than hapax legomena, they should be interpreted with caution. Another factor might be at play. It is possible that higher-frequency words, which tend to be shorter than low-frequency words, are more appropriate for titles as they will tend to occupy less journal space. The same might hold for words in *-heid*, so that the preponderance of high-frequency formations probably cannot be attributed only to their relative context-independence.

2.4. *Rule priming?*

If the hapax legomena are more often created by rule than retrieved from memory compared to the high-frequency formations, we expect to find evidence that once used, the rule for coining words in *-heid* is "primed" and might be used again. A stylistic construction that lends itself particularly well to "rule priming" is parataxis, as in *donkerheid en opgeslotenheid* 'darkness and confinedness' and *verslagenheid en stomheid* 'dejectedness and dumbfoundedness'.

Table 5 and Figure 5, based on counts in 32 issues of *Trouw*, show that such parallel use of words in *-heid* is relatively rare but that, as expected, it is significantly more common for hapax legomena than for high-frequency words ($p < 0.03$, one-tailed proportions test).

Table 5. *Rule priming for* -heid *measured by the number of contexts with different words in* -heid *occurring in coordination (counts based on a sample of 32 newspaper issues)*

	High-frequency nouns		Hapax legomena	
	no.	(%)	no.	(%)
Rule priming	3	(1.7)	11	(6.4)
No rule priming	170	(98.3)	162	(93.6)
Total	173	(100.0)	173	(100.0)

$Z = 1.877$, $p < 0.03$, one-tailed proportions test.

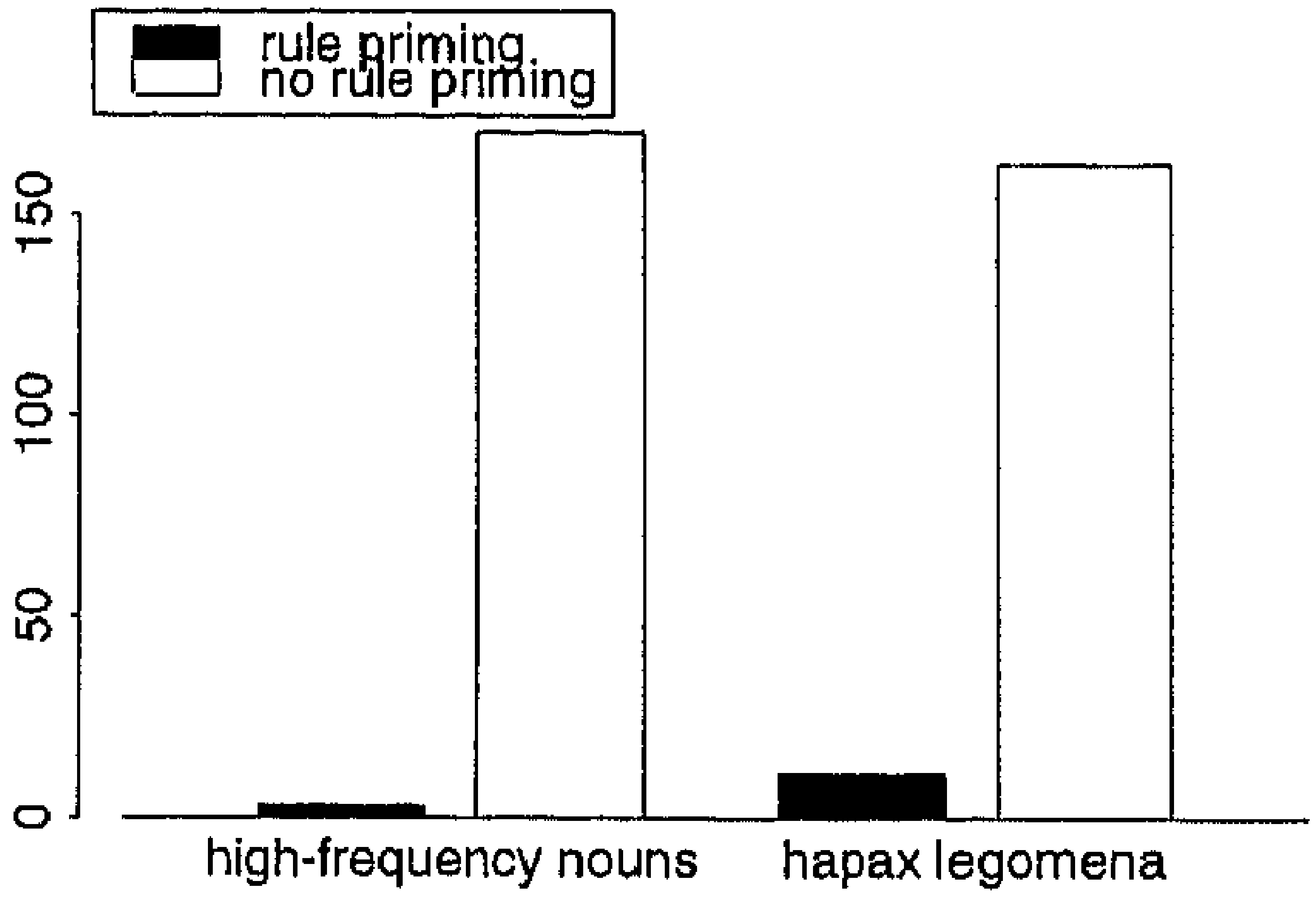


Figure 5. *Counts of* -heid *formation with and without "rule priming" for hapax legomena and high-frequency words*

2.5. *Possessive pronouns*

As a final check of our hypothesis, we investigated whether explicit contextual anchoring by means of premodifying possessives occurs more often for hapax legomena than for high-frequency formations. Premodifying possessives are of special interest as the possessive pronoun provides its antecedent as the argument for the adjective underlying the *-heid* formation. For instance, in

(9) Janet went home. Her lightheartedness was visible to everyone.

her provides its antecedent *Janet* as a grammaticalized anchor for the predicate *lighthearted.* We would therefore expect that the hapax legomena, which according to our hypothesis are more dependent on anchoring, should benefit from occurring in conjunction with a possessive pronoun. Counts based on 15 newspaper issues indeed revealed the

Table 6. *Contextual anchoring for high-frequency words and hapax legomena in* -heid *by means of possessive pronouns (counts based on 15 issues of the newspaper* Trouw*)*

	High-frequency nouns	Hapax legomena	Total
Possessive	8	25	33
No possessive	201	189	390
Total	209	214	423

$X^2_{(1)} = 9.01, p < 0.005$ (with continuity correction).

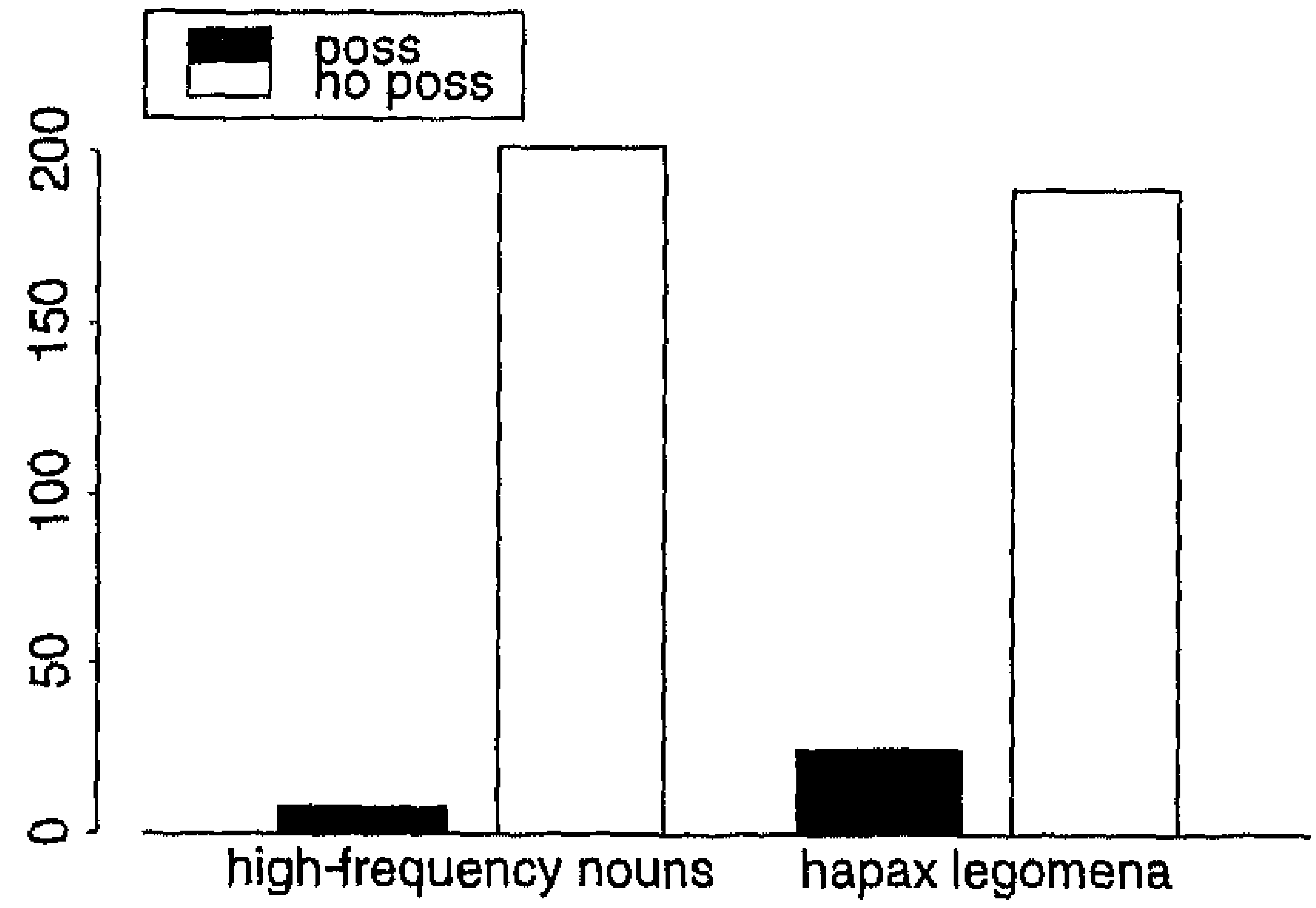


Figure 6. *Bar plot for the numbers of* -heid *formations modified and not modified by a possessive pronoun for high-frequency formations and hapax legomena*

expected difference between the two sets of words (see Table 6 and Figure 6). Possessive pronouns occur significantly more often preceding hapax legomena (25 instances, 12%) than preceding high-frequency formations (8 instances, 4%) ($p < 0.001$, proportions test).

3. Text types

So far we have explored the differences between high-frequency formations in *-heid* and hapax legomena with this suffix in terms of various kinds of contextual anchoring. The counts presented in the previous section all support our hypothesis that the high-frequency formations represent concepts just as monomorphemic words do, while the hapax legomena display a greater degree of referential anchoring. But if high-frequency words denote specific concepts, then their use might be more restricted to particular text types than use of the hapax legomena. Our to their denotational specificity, they might be appropriate only for discussing specialized topics. For instance, *werkloosheid* 'unemployment' is more likely to occur in the sections in our newspaper on politics and economics than in the sections on sports and arts. Similarly, *schoonheid* 'beauty' is appropriate primarily in contexts of artistic appraisal and hence unlikely to appear outside the arts sections. What we may expect for our 15 high-frequency words, then, is that they pattern unevenly across the text types in our newspaper. Because the hapax legomena generally do not denote specialized concepts, one might expect them to occur more uniformly across these text types.

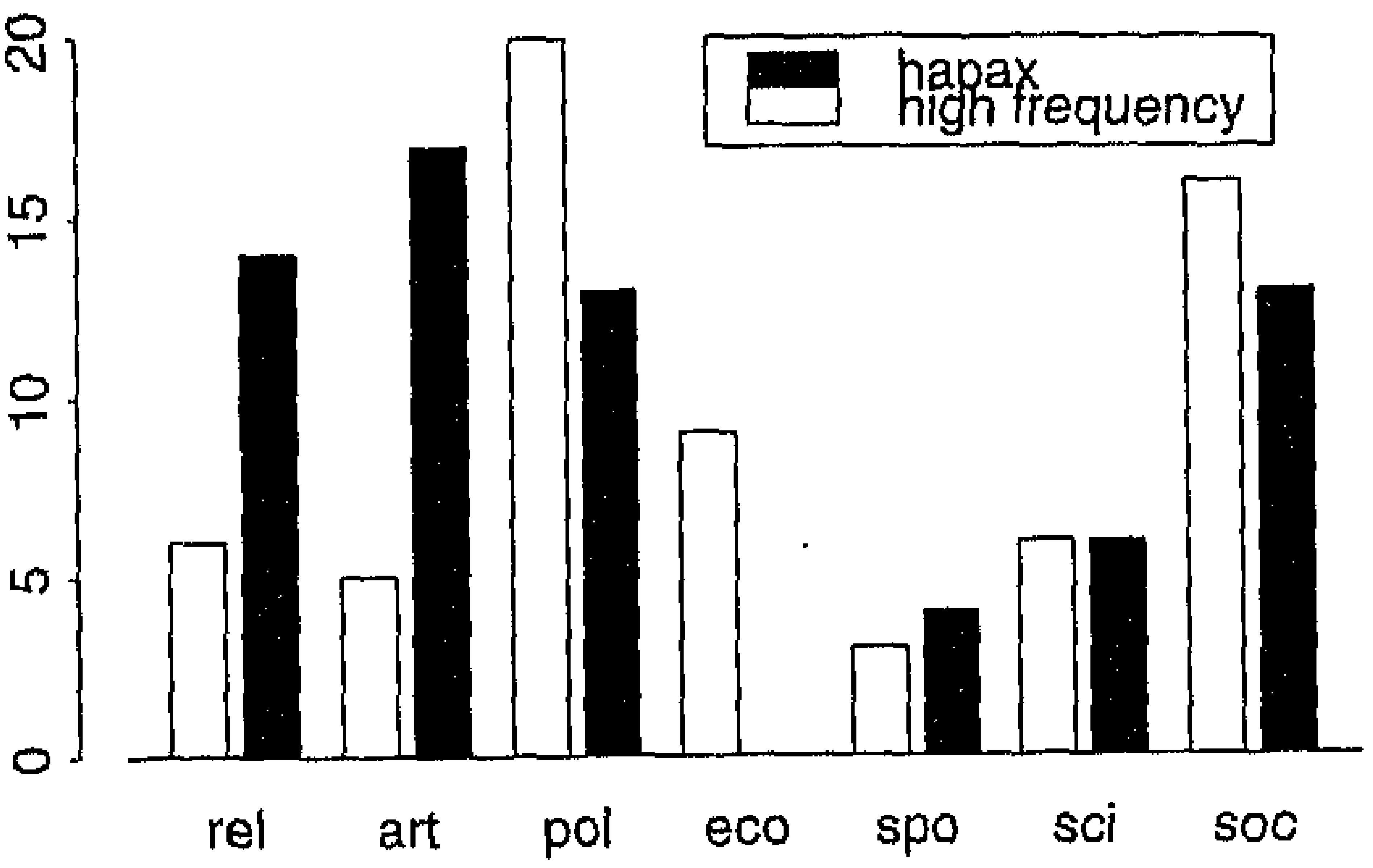
Table 7 and Figure 7 show that these expectations are only partially supported when we count the occurrences of high-frequency words and hapax legomena in *-heid* in the seven main text types in a random subset of 17 issues of our newspaper (religion, art, politics, economics, sports, science, and society). As expected, the high-frequency words reveal a nonuniform distribution. They occur primarily in the sections on politics, economics, and society, but quite infrequently in the sports sections. This uneven pattern supports our intuition that the specialized semantics of these high-frequency formations limits their use to specific topics.

Interestingly, the hapax legomena show a similar range of variation. The arts and economics sections represent the extremes in our data. We find no hapax legomena in *-heid* at all in the sections on economics. Conversely, they appear most frequently in the arts sections. This variation cannot be attributed to the semantics of individual words as such, as in the case of the high-frequency formations. After all, we are dealing with 67 different words that have only the suffix *-heid* in common.

Table 7. *Register differences in the use of hapax legomena and high-frequency formations in* -heid *in 17 issues of the Dutch newspaper* Trouw

	High-frequency nouns	Hapax legomena	Total
Religion	6	14	20
Art	5	17	22
Politics	20	13	33
Economy	9	0	9
Sport	3	4	7
Science and book reviews	6	6	12
Society	16	13	29
Total	65	67	132

$X^2_{(6)} = 20.68$, $p < 0.005$.
Fisher exact test of independence: $p < 0.001$.



rel: religion; art: art; pol: politics; spo: sport; sci: science; soc: society

Figure 7. *Register differences in the use of hapax legomena and high-frequency formations in* -heid

Apparently, some text types are more likely to require innovative use of *-heid* than others — the productivity of *-heid* appears to be codetermined by text type.[2]

More detailed studies are required to chart the precise interdependencies between text type, frequency, and productivity. Crucial for the central hypothesis of the present paper is the finding that the hapax legomena and the high-frequency words reveal substantially different patterns across text types. This supports our intuition that the set of formations with the suffix *-heid* is not a homogeneous one, and that the high-frequency words and the lowest-frequency words may have different semantics and different degrees of pragmatic usefulness as a function of text type.

4. General discussion

In this study we have investigated the semantics of the Dutch suffix *-heid*, which, like *-ness* in English, coins abstract nouns from adjectives. We have advanced the hypothesis that *-heid* has two distinguishable semantic functions. On the one hand, it creates abstract concepts such as *snelheid* 'speed'. On the other hand, *-heid* may serve the function of referring to the quality or state expressed by the base word. In this use, its primary function is to establish referential links between states of affairs in the unfolding discourse in the same way that pronouns establish anaphoric or cataphoric links between discourse participants. We have linked this hypothesis of the dual function of *-heid* with the further claim that it is primarily the highest-frequency formations that denote abstract concepts, and that the referential function of *-heid* is instantiated foremost among the hapax legomena. To substantiate this claim, we have investigated the use of *-heid* in a corpus of Dutch newspaper issues. We have shown that, compared to the hapax legomena, the high-frequency words evidence a higher degree of context-independence. They reveal lesser degrees of contextual anchoring, they have a higher likelihood of appearing in titles and a lower probability of occurring in paratactic constructions that may have been brought about by rule priming, and they occur relatively seldom with possessive pronouns. In addition, the two sets of formations reveal different distributions across text types. We interpret these findings as supporting our hypothesis. The highest-frequency words predominantly express abstract concepts that are relatively independent of their contexts. Conversely, the referential function of *-heid* is more prominently realized among the hapax legomena, which as a result reveal more contextual anchoring.

We view the referential and conceptual functions as two distinct components of the semantics of *-heid*. The frequency domain emerges as a scalar dimension with poles at which the two functions dissociate quantitatively. Note that for any particular word in *-heid*, both functions can be realized. In fact, the two functions can be realized simultaneously. However, as the frequency of a given word increases, the likelihood increases likewise that the conceptual function of *-heid* is the more prominent one. Crucially, our materials do not suggest that the difference between the high-frequency words and the hapax legomena is primarily driven by semantic opacity. Opaque and transparent high-frequency words show a similar pattern compared to the hapax legomena.

Our finding that for one and the same affix the semantics may shift from primarily conceptual to more referential as a function of diminishing frequency of use has important consequences for theories of storage and computation in the mental lexicon. First consider language comprehension. Upon hearing or reading a fully regular complex word such as *snelheid*, the meaning ‘speed’ should be retrieved rather than the interpretation referring to a state of ‘quickness’. This can be modelled by assuming that a form-based access representation for *snelheid* is available that provides a pointer to its specialized conceptual semantics. For very low-frequency formations such as *onregeerbaarheid* ‘ungovernableness’, full-form representations in the access system are probably not available. These forms are recognized on the basis of their constituent morphemes, and hence only the compositional referential meaning is available. (In a parallel dual route model, such as developed in Baayen, Burani, et al. [1997], the recognition route that exploits the morphological structure of a complex word operates in parallel with the recognition route that uses full-form representations to achieve lexical access. With respect to high-frequency words such as *snelheid*, this model predicts that the conceptual meaning is the first to become available, but that its referential meaning also becomes available, albeit at a slightly later point in time.)

Next consider language production. We assume that speakers are familiar with the concept of 'speed'. In other words, we take it for granted that speakers do not construct the concept 'speed' anew for each instance of use, but that they have available in memory a representation for this concept. Once the concept for 'speed' has been activated, the appropriate word form encoding this concept has to be selected. Speakers of Dutch know that the concept of 'speed' is encoded by two morphemes, *snel* and *-heid*, in the linear sequence *snel + heid* (see Roelofs 1997 for evidence for morphological structure in language production). For the expression of the corresponding referential semantics speakers can opt to select *snel* or a near-synonym such as *vlug* in combination with *-heid*. Crucially, speakers do not provide *vlugheid* when they want to express 'speed'. Thus, in both speech production and language perception, the mental lexicon contains the information that links the concept 'speed' with the forms *snelheid* (comprehension) and *snel* and *-heid* (production). Even though *snelheid* is fully regular with respect to its form, its slightly specialized meaning induces storage of both the concept and the specific forms encoding the concept in comprehension and production.

The Dutch suffix *-heid* is among the most productive derivational suffixes of Dutch. Most formations in *-heid* are probably not stored in the mental lexicon. Storage is least likely for the hapax legomena, which comprise at least half the total number of different types, and is also unlikely for the many other words with a low frequency of use. But for the highest-frequency words, words that denote concepts that are part of the conceptual stock of present-day Dutch, storage in the mental lexicon along the lines sketched above is probably necessary. Without wanting to claim that these highest-frequency words are irregular, we can say that their specialized conceptual semantics sets them apart from the large numbers of productively coined ephemeral formations in the lowest-frequency ranges. In the present study we have shown that there are significant differences in the way the high-frequency words and the hapax legomena are anchored in their context. We have shown that the words that express specialized concepts require less contextual anchoring than the hapax legomena, formations that have a more referential function and require more contextual anchoring. This corpus-derived distributional evidence is in line with the psycholinguistic evidence that suggests that high-frequency fully regular words can be stored in the mental lexicon (Baayen, Burani, et al. 1997; Baayen, Dijkstra, et al. forthcoming).

Our approach to accounting for productive word formation in the mental lexicon differs from the way in which Anshen and Aronoff (1988, 1997) propose to model the difference between unproductive and productive word formation in the mental lexicon. Using experimental, statistical, and historical evidence, Anshen and Aronoff argue that storage in the mental lexicon takes place only for words containing unproductive affixes. Words falling into productive morphological categories, they claim, are not stored in the mental lexicon. Our data on the use of *-heid* in context provide a counterexample to their claim that storage and a lack of productivity, and likewise the absence of storage and productivity, are two sides of the same coin. To our mind, productivity is a cline, with at the one extreme fully unproductive categories comprising formations that are all stored in the mental lexicon (for instance, *-th* as in *strength* in English), and with categories comprising formations that are always produced or understood by rule (such as verbal inflections) at the other extreme, and, especially in the domain of derivational morphology, with a large variety of intermediate positions. Whenever regular morphology involves concept formation with lexical specialization, storage of meaning and concomitantly (often minimal) storage of the forms expressing these meanings takes place.

Studies of morphological productivity have generally focused on word structure, and little attention has been paid to the use of complex words in context. Kastovsky (1986) is the only study known to us that explicitly calls attention to what we have called the referential function of morphology. His examples, carefully brought together from personal observation in spoken and written English, necessarily remain somewhat anecdotical in nature. The availability of large corpora has created the possibility of studying the various functions of word formation in greater detail. We have shown, albeit for a single Dutch suffix only, that the referential and conceptual functions can indeed be distinguished in text corpora and that methods of corpus linguistics can profitably be used to complement the experimental methods of psycholinguistics and the structural methods of theoretical linguistics.

Received 5 August 1996
Revised version received
8 November 1996

Max Planck Institute for
Psycholinguistics, Nijmegen
University of Nijmegen

Notes

* We would like to thank our students, Lisa de Groot, Joyce Karreman, Wendy van Krimpen, Jan Michielsen, Tamara Römer, Marianne Top, Peter Van den Broeck, Nicole Verbakel, and Femke van der Weide, for their contributions to this study. We would also like to thank Robert Schreuder and Antoinette Renouf for their comments on an earlier version of this paper. Correspondence address: Max-Planck-Institut für Psycholinguistik, Postbus 310, NL-6500 AH Nijmegen, The Netherlands. E-mail: baayen@mpi.nl.

[1] No other main effects or interactions appear in the most parsimonious model obtained by stepwise model selection.

[2] For language variation as a function of text type and register, see Biber (1995), and for morphological productivity as a function of register, see Baayen (1994).

References

Anshen, Frank; and Aronoff, Mark (1988). Producing morphologically complex words. *Linguistics* 26, 63–72.

—; and Aronoff, Mark (1997). Morphology in real time. In *Yearbook of Morphology 1996*, Geert E. Booij and Jaap van Marle (eds.), 9–12. Dordrecht: Kluwer.

Baayen, R. Harald (1994). Derivational productivity and text typology. *Journal of Quantitative Linguistics* 1, 16–34.

—; Burani, Cristina; and Schreuder, Robert (1997). Effects of semantic markedness in the processing of regular nominal singulars and plurals in Italian. In *Yearbook of Morphology 1996*, Geert E. Booij and Jaap van Marle (eds.), 13–34. Dordrecht: Kluwer.

—; Dijkstra, Ton; and Schreuder, Robert (forthcoming). Singulars and plurals in Dutch: evidence for a parallel dual route model. *Journal of Memory and Language*.

—; Piepenbrock, Richard; and van Rijn, Hedderik (1993). *The CELEX Lexical Database* (CD-ROM). Philadelphia: Linguistic Data Consortium.

Biber, Douglas (1995). *Dimensions of Register Variation*. Cambridge: Cambridge University Press.

Kastovsky, Dieter (1986). Productivity in word formation. *Linguistics* 24, 585–600.

Roelofs, Ardi (1997). Morpheme frequency in speech production: testing WEAVER. In *Yearbook of Morphology 1996*, Geert E. Booij and Jaap van Marle (eds.), 135–154. Dordrecht: Kluwer.