# Lexical frequency in morphology: is everything relative?[1]

JENNIFER HAY

*Abstract*

*While it is widely assumed that high-frequency morphologically complex forms tend to display characteristics of noncompositionality, models of morphological processing do not predict a direct relationship between absolute frequency and decomposition. Rather, they predict a relationship between decomposition and the relative frequency of the derived form and the base. This paper argues that such a relative frequency effect does, indeed, exist.*

*First, the results of a simple experiment demonstrate that subjects perceive derived forms that are more frequent than their bases to be significantly less complex than matched counterparts that are less frequent than their bases. And second, dictionary calculations reveal that derived forms that are more frequent than their bases are significantly more likely to display symptoms of semantic drift than derived forms containing higher-frequency bases. High-frequency forms, however, are no more prone to semantic drift than low-frequency forms.*

*These results provide evidence that it is relative frequency, rather than absolute frequency, that affects the decomposability of morphologically complex words. A low-frequency form is likely to be nontransparent if it is composed of even-lower-frequency parts. And a high-frequency form may be highly decomposable if the base word it contains is higher frequency still.*

## 1. Introduction

Many researchers have argued that lexical frequency affects morphological decomposition. High-frequency forms, it is claimed, tend to be accessed whole, are not easily decomposed, and so do not contribute to the productivity of the affixes they contain. However, when we closely examine current models of morphological processing, we find that they

*Linguistics* 39–6 (2001), 1041–1070

0024–3949/01/0039–1041

predict an effect of relative frequency, rather than absolute frequency. That is, they predict that derived forms that are more frequent than their base should be less decomposable than derived forms that are less frequent than their base, regardless of the absolute frequency of the derived form.

This paper sets out to resolve this apparent incongruity. Does relative frequency or absolute frequency affect morphological decomposition? Section 2 briefly reviews the literature, which places a large emphasis on absolute frequency, assuming high-frequency forms to behave differently from low-frequency forms. Section 3 then outlines current models of morphological processing, showing how they actually predict an effect of relative frequency.

Section 4 describes a simple experiment that demonstrates that a clear effect of relative frequency does, in fact, exist. The nature of the relationship between relative and absolute frequency is explored in section 5, and then section 6 describes a study of the relationship between frequency and semantic transparency. Using dictionary definitions to index degree of semantic transparency, we find that the relative frequency of the derived form and the base is a good predictor of semantic transparency. The absolute frequency of the derived form, however, appears to be unrelated to semantic transparency.

This result has consequences for models of morphological access, and also for the methodology used to investigate these issues. These consequences are discussed in section 7.

## 2. Absolute frequency and morphological decomposition

High-frequency forms, it has been argued, are not highly decomposable and display low levels of semantic transparency (Modor 1992; Baayen 1992, 1994; Baayen et al. 1997; Bybee 1988, 1995a, and others). Bybee, for example, makes strong claims about the relationship between the frequency of the derived form and degree of compositionality:

> There is a universal tendency for morphological irregularity to be to the highest frequency forms of a language (Bybee 1995a: 235).

This, claims Bybee, is true of both phonological irregularity and semantic irregularity. She claims (1985, 1995a, 1995b) that frequent derived forms diverge both phonologically and semantically from their bases and have a tendency to become autonomous. The vast majority of work bearing on this issue actually deals with inflectional morphology. High-frequency



inflected forms have been shown to be stored, while low-frequency forms are believed (by many) to be derived by rule (see, e.g., Stemberger and MacWhinney 1986, 1988; Losiewicz 1992; Alegre and Gordon 1999a).

One piece of evidence relating the claim to derivational affixation comes from Pagliuca (1976, cited in Bybee 1995a), who shows that both phonological and semantic transparency in *pre-* affixation can be related to lexical frequency.

The belief that semantic transparency and the frequency of the derived form are linked is so widely held that it is sometimes stated as fact, without examples or references. Baayen (1993), for example, when justifying a methodological choice, starts out, "Since transparency is inversely correlated with frequency ..." (Baayen 1993: 203).

Because Baayen (1992, 1993) assumes a link between frequency and opacity, he argues that there is a connection between high token frequency and lack of productivity — frequent words are not accessed via the components they contain and so do not contribute to the productivity of the related affix. Any word-formation process that is characterized by only high-frequency tokens will therefore be an unproductive one.

This apparent link between transparency and frequency is often explained in terms of memory constraints. Baayen and Lieber (1997), for example, explain that

> A high frequency of use guarantees that their opaque reading can be retained in memory. Thus it is only to be expected that opaque formations show a tendency to appear in the highest ranges of the frequency spectrum (Baayen and Lieber 1997: 283).

Note, however, that monomorphemic words are inherently opaque, and that we are able to retain the meanings of relatively low-frequency monomorphemic words in memory without apparent difficulty. The problem with retaining opaque complex words in memory relates to the fact that there is a competing analysis. If the components of that competing analysis are of relatively lower frequency, then retaining an opaque meaning for a complex word in memory should provide no obstacle. It is only when the competing analysis is a strong and competitive one that problems may start to arise.[2] Noncompositionality should be a viable option for complex words that are more frequent than their parts, regardless of the absolute frequency of the derived form.

Indeed, we will see in the next section that despite the fact that a link between high frequency and noncompositionality is widely assumed, the existence of such a link does not directly follow from current models of morphological processing.

## 3. Models of morphological processing

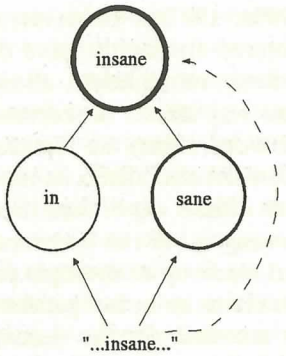
Models of morphological processing must make some assumptions about the role of decomposition. Do we decompose affixed words upon encountering them, breaking them down into their parts in order to access lexical entries associated with their component morphemes? Or do we access affixed words as wholes, accessing an independent lexical entry? Some researchers have argued that there is no decomposition during access (e.g. Butterworth 1983), and others have claimed that there is a prelexical stage of compulsory morphological decomposition (e.g. Taft 1985). But most current models are mixed — allowing for both a decomposed access route, and a direct-access, nondecomposed route. In many models, the two routes compete, and, in any given encounter with a word, either the decomposed route or the direct route will win (Wurm 1997; Frauenfelder and Schreuder 1992; Baayen 1992; Caramazza et al. 1988). Direct competition does not necessarily take place, however. Baayen and Schreuder have recently argued that the two routes may interactively converge on the correct meaning representation (Schreuder and Baayen 1995; Baayen and Schreuder 1999, 2000). In this model too, there will necessarily be some forms for which the decomposed route dominates access, and others in which the direct whole-word representation is primarily responsible for access.

Figure 1 shows an idealization of a dual-route model, in which two routes race to access the lexical entry for *insane*. The solid lines represent the decomposed route, and the broken line indicates the direct-access route.

One factor that will affect the speed of each of the routes is lexical frequency. Frequent words are more easily and quickly retrieved (Connine et al. 1993; Grosjean 1980; Balota and Chumbley 1984). Nodes associated with frequent words (or morphemes) will therefore be accessed more quickly than nodes associated with infrequent items. The CELEX lexical database (Baayen et al. 1995) lists *sane* with a frequency of 149/17.4 million, and *insane* with a frequency of 258/17.4 million. The relatively higher frequency of *insane* relative to *sane* is represented in Figure 1 by a greater line thickness.

When we consider the two routes in Figure 1 it is clear that the whole-word route has an advantage. The higher relative frequency of *insane* speeds the whole-word route, relative to the decomposed route. *Insane* can be compared with a word like *infirm*. *Infirm* is fairly infrequent (27/17.4 million), and, importantly, its base (*firm*) is highly frequent (715/17.4 million). As such, we predict the decomposed route should have a strong advantage over the whole-word access route.






*Key:*
- solid lines indicate the decomposed route
- broken line indicates the direct route
- line width of each node indicates the resting activation level: *insane* is more frequent than *sane*

Figure 1. *Schematized dual-route model*

Importantly, because words compete, the absolute frequency of the derived form is not so important as its frequency relative to the base form with which it is competing. While I have illustrated this prediction with a schematized dual-route "fight to the death" model, the same prediction follows from all models in which both decomposition and whole-word access are available options (regardless whether they compete or converge), or in which the presence of the base word can be variably salient.

Implicit in much of the literature outlined above, and also in my discussion below, is an assumed link between morphological decomposition and semantic transparency. When we consider the workings of the dual-route model sketched here, it becomes clear that the two are necessarily related. This link arises because the way in which we tend to access a word must have long-term consequences for that word's representation.

In order to successfully model the fact that high-frequency morphologically simple words have different characteristics from low-frequency simple words, we must assume that accessing a word leaves its traces on the lexicon. Recognizing an item affects that item's representation and increases the probability that it will be successfully recognized in the future. Some models capture this process by raising the resting activation level of the relevant lexical entry (see, e.g., Norris et al. 2000; McClelland and Elman 1986). Other models, based on exemplars, assume that identifying a word involves adding a new exemplar to the appropriate exemplar

cloud (e.g. Johnson 1997a, 1997b). However, in order to capture the fact that words encountered frequently have different properties from words encountered relatively infrequently, all models must assume that accessing a word in some way affects the representation of that word.

If accessing a simple word affects its representation, then it follows that accessing a complex word also affects its representation. And if there are two ways to access an affixed word, then there are two ways to affect its representation. Accessing it via the decomposed route reinforces its status as an affixed word made up of multiple parts. Accessing it via the direct route reinforces its status as an independent entity. Thus, we expect words that tend to be accessed via the decomposed route to remain tightly constrained by the characteristics and meaning of the base word, and to remain robustly semantically transparent. Words that tend to be accessed via the direct route have the potential to become liberated and to take on characteristics wholly independent of the base word. Such words are prone to semantic drift and may display low levels of semantic transparency.

We therefore predict derived words that are more frequent than the bases they contain (characterized by direct access) to be less robustly semantically transparent than words that are less frequent than their bases (characterized by decomposition).

There therefore appears to be a marked incongruity between the type of frequency effect predicted by models of processing (a RELATIVE frequency effect), and that assumed to exist in the literature outlined in the previous section (an ABSOLUTE frequency effect). The next section describes an experiment designed to test whether an effect of relative frequency does, in fact, exist.

## 4. Relative frequency and morphological complexity

Do words that are more frequent than their embedded bases appear more easily decomposable than words that are not more frequent than their embedded bases? In this section I describe a simple experiment, which asked subjects this question directly.

### 4.1. *Materials and methodology*

This experiment simply presented subjects with pairs of words and asked them to provide intuitions about which member of the pair is more easily decomposable. Thirty-four pairs of words were constructed — 17 prefixed pairs, and 17 suffixed pairs. Members of word pairs shared affixes and were matched for the probability of the junctural phonotactics, the stress pattern and syllable count, and the surface frequency of the derived form.[3] One member of each pair was more frequent than the base, and one member was less frequent. Frequency information was obtained from the CELEX lexical database. The prefixed and suffixed word pairs are shown in Tables 1 and 2 respectively. The A members of each pair are more frequent than the bases they contain, so are predicted to be rated less complex than the B members of each pair, which are less frequent than the bases they contain. Note that in the suffixed stimuli the pair *agility–fragility* is included. Unlike the other words in column A, *agility* is NOT more frequent than *agile* — they are of roughly equal frequency. However, *fragile* is considerably more frequent than *fragility*, and so the pair should still exhibit the expected contrast. They were included so as to obtain a reasonably sized dataset, while still meeting the considerable restrictions imposed by controlling for various phonological factors.

The pairs were counterbalanced for order of presentation and randomized together with 30 filler word pairs. The fillers paired together pseudo-affixed and affixed words. Example filler pairs include *defend–dethrone*, *indignant–inexact*, *family–busily* and *adjective–protective*. The complete set of stimuli consisted of 64 word pairs.



Table 1. *Prefixed stimuli*

| word A | freq. | base freq. | word B | freq. | base freq. |
|---|---|---|---|---|---|
| refurbish | 33 | 1 | rekindle | 22 | 41 |
| inaudible | 292 | 100 | inadequate | 399 | 540 |
| incongruous | 55 | 3 | invulnerable | 23 | 400 |
| uncanny | 89 | 20 | uncommon | 114 | 3376 |
| unleash | 65 | 16 | unscrew | 44 | 187 |
| immutable | 40 | 4 | immoderate | 6 | 223 |
| unobtrusive | 42 | 17 | unaffected | 54 | 169 |
| entwine | 32 | 27 | enshrine | 44 | 98 |
| immortal | 112 | 53 | immoral | 94 | 143 |
| illegible | 14 | 10 | illiberal | 11 | 55 |
| intractable | 45 | 12 | impractical | 47 | 1228 |
| uncouth | 34 | 2 | unkind | 72 | 390 |
| impatient | 227 | 114 | imperfect | 50 | 1131 |
| revamp | 13 | 4 | retool | 10 | 800 |
| inanimate | 34 | 4 | inaccurate | 53 | 377 |
| reiterate | 47 | 0 | reorganise | 61 | 1118 |
| immobile | 55 | 11 | immodest | 13 | 521 |

Table 2. *Suffixed stimuli*

| word A | freq. | base freq. | word B | freq. | base freq. |
|---|---|---|---|---|---|
| diagonally | 36 | 29 | eternally | 58 | 355 |
| abasement | 6 | 2 | enticement | 3 | 64 |
| meekly | 47 | 41 | bleakly | 22 | 196 |
| swiftly | 268 | 221 | softly | 440 | 1464 |
| diligently | 35 | 31 | arrogantly | 17 | 116 |
| rueful | 14 | 9 | woeful | 14 | 68 |
| respiration | 39 | 4 | adoration | 49 | 218 |
| alignment | 57 | 44 | adornment | 41 | 75 |
| equally | 1303 | 1084 | generally | 1663 | 4624 |
| hapless | 22 | 13 | topless | 27 | 3089 |
| listless | 42 | 19 | tasteless | 30 | 402 |
| frequently | 1036 | 396 | recently | 1676 | 1814 |
| exactly | 2535 | 532 | directly | 1278 | 1472 |
| agility | 34 | 38 | fragility | 36 | 207 |
| slimy | 61 | 35 | creamy | 74 | 540 |
| virility | 41 | 31 | sterility | 36 | 121 |
| scruffy | 42 | 7 | puffy | 48 | 159 |

For each pair, subjects were asked to indicate which member of the pair they considered more "complex." The exact instructions were as follows.

This is an experiment about complex words.

A complex word is a word which can be broken down into smaller, meaningful, units.

In English, for example, the word *writer* can be broken down into two units: *write* and *-er*. *-er* is a unit which occurs at the end of many English words. In *writer*, *-er* has been added to the word *write* to make a new, more complex word *writer*. We call a word which has been made out of smaller units in this way a complex word.

*Rewrite* is another example of a complex word in English. It can be broken down into *re-* and *write*.

Words which are not complex are called simple words. Here are some examples of simple words in English: *yellow*, *sing*, *table*. It is impossible to break down the word *table* into smaller units. *Table* is not complex.

In this experiment, you will be presented with pairs of complex words, and asked to decide which one you think is MORE complex.

For example *happiness* is very complex — it can be easily broken down into *happy* and *-ness*. *Business*, however, is not quite so complex. While it is possible to break *business* down into *busy* and *-ness*, it does not seem completely natural to do so. *Business* is complex, but not as complex as *happiness*.



Another example of a complex word is *dishorn*. Even though you may never have heard the word *dishorn* before, you can understand its meaning, because it can be broken down into *dis-* and *horn*. *Discard* is also complex — it can be broken down into *dis-* and *card*. But *discard* does not seem as complex as *dishorn*. We do not need to break *discard* into its parts in order to understand its meaning, and, in fact, it seems slightly unnatural to do so.

For each pair of words below, please read both words silently to yourself, and then circle the word you think is more complex. It is very important that you provide an answer for every pair, even if you are not certain of your answer. Just follow your intuition, and provide your best guess.

The experiment was completed using pen and paper, and subjects worked at their own pace. Twenty Northwestern University undergraduate students completed the task, in fulfilment of a course experimental requirement.

### 4.2. *Results*

Subjects who did not consistently distinguish between the pseudo-affixed and affixed filler pairs were not included in the analysis. Any subject who did not provide the same answer (in either direction) for at least twenty of the thirty fillers was excluded.

Sixteen subjects were therefore analyzed. Of these, two interpreted "complex" in the opposite manner from that intended. This could be seen from their consistent behavior on the filler items (i.e. they rated *family* more complex than *busily*, *adjective* more complex than *protective* and so on). This consistent behavior indicates that their confusion was a terminological one, rather than a conceptual one, and so their data was included, with their answers reversed.

Forms that were more frequent than the bases they contained were consistently rated less complex than their counterparts, which were less frequent than the bases they contained. This is true both for suffixed forms (wilcoxon, by subjects: $p < 0.005$, by items: $p < 0.002$), and prefixed forms (wilcoxon, by subjects: $p < 0.002$, by items: $p < 0.01$).

Among prefixed pairs, 65% of responses favored the form for which the base was more frequent than the whole. Only 35% of responses judged forms that were more frequent than their bases to be more complex than their matched counterpart. The tendency for suffixed forms was of roughly equal strength (66% to 34%).

This provides strong evidence that the frequency of the base form is involved in facilitating decomposability. When the base is more frequent than the whole, the word is easily and readily decomposable. However,

when the derived form is more frequent than the base it contains, it is more difficult to decompose and appears to be less complex. The relative frequency effect predicted by models of morphological processing does, indeed, exist.

## 5. Relative frequency distributions in affixed words

The above simple experiment demonstrated a clear effect of relative frequency upon morphological decomposition. This raises the question of the degree to which previous observations about high-frequency forms are actually artifacts of relative frequency. We can get an initial indication of this by considering the degree to which absolute frequency and relative frequency are correlated.

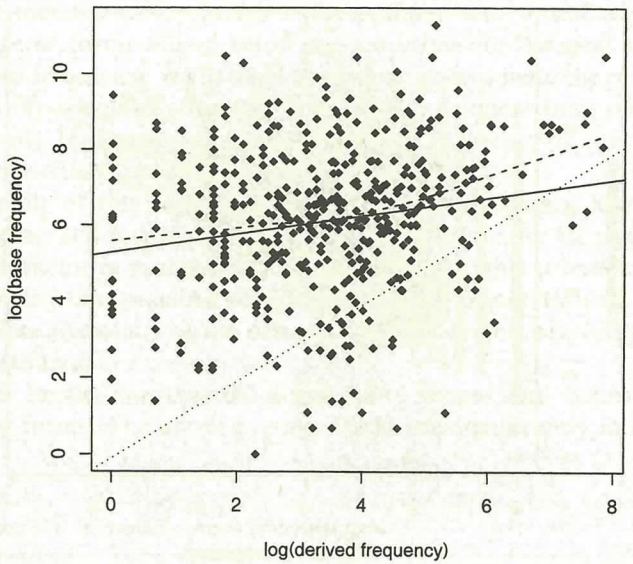
A small dataset was created in order to investigate this question. Nine consonant-final prefixes and five consonant-initial suffixes were chosen, and all forms listed in CELEX as containing these affixes were extracted. CELEX codes words as affixed only when there is a corresponding root entry. That is, words like *transfer* are not coded as prefixed, but rather as containing "obscure morphology." The CELEX entries provide frequency information for the entire form, as well as a morphological decomposition, which allowed for the automatic retrieval of the frequency of the base. Only affixed forms with monomorphemic bases were included, and duplicate entries were deleted. In cases of multiple entries, only the most frequent was included, in an attempt to approximate the frequency of the form in its most usual, or "default," use. This led to a corpus consisting of 515 prefixed and 2028 suffixed words.

Harwood and Wright (1956) have pointed out that in general derived forms (or in their terminology *resultants*) tend to be less frequent than their bases (the *underlying* forms). This can be confirmed examining the data in Figures 2 and 3, which show the relationship of derived frequency to base frequency, for the set of prefixed and suffixed words, respectively.

The dotted lines through the graphs indicate the line at which derived frequency and base frequency are equal. Points falling below this line represent forms for which the derived frequency is greater than the frequency of the base. Points falling above it represent forms for which the derived form is of lower frequency than the base it contains. The solid line represents a regression line fit through the data, and the dashed line represents a nonparametric scatterplot smoother (Cleveland 1979).

These graphs reinforce Harwood and Wright's (1956) argument that derived forms tend to be less frequent than the bases — many more points fall above the dotted x = y line, than below it. The points below





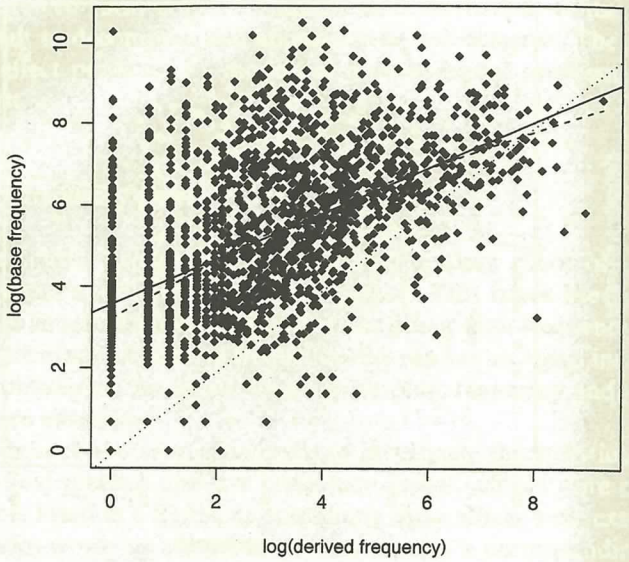

*Key:*
- solid line shows the result of a linear regression ($r^2 = 0.04$, $p < 0.001$)
- dashed line shows a nonparametric scatterplot smoother (Cleveland 1979) fit through the data
- dotted line shows the x = y line
- the correlation is significant both by parametric (pearson's $r = 0.2$, $p < 0.001$) and nonparametric (spearman's $r = 0.21$, $p < 0.001$) measures

Figure 2. *Log derived frequency and base frequency, for 515 English prefixed forms*

the x = y line represent forms that are more frequent than their bases. There is a sense, then, in which these forms have escaped — and become liberated from the properties of the base. It is these forms that should tend to be accessed via a whole-word representation and are not robustly decomposable.

Examination of Figures 2 and 3 reveals that for both prefixes and suffixes, a positive and significant correlation holds between the frequency of a base form, and the frequency of a derived form that contains it. More-frequent bases tend to be associated with more-frequent derivatives. In general, how often a (transparent) derived form is deployed in speech is likely to be a partial function of the frequency of the form upon which it is based.

Importantly, for both suffixes and prefixes the frequency of the derived form and the base are positively correlated on a slope that falls consistently ABOVE the line at which the derived form and the base are




*Key:*
- solid line shows the result of a linear regression ($r^2 = 0.28$, $p < 0.001$)
- dashed line shows a nonparametric scatterplot smoother (Cleveland 1979) fit through the data
- dotted line shows the x = y line
- the correlation is significant both by parametric (pearson's $r = 0.53$, $p < 0.001$) and nonparametric (spearman's $r = 0.56$, $p < 0.001$) measures

Figure 3. *Log derived frequency and base frequency, for 2028 English suffixed forms*

equal. Given the frequency of any base, then, the frequency of a derived form containing that base is partially predictable, and, in all cases, the prediction is that the derived form will be somewhat less frequent than the base.

Examination of the data represented in these figures provides some insight into the degree to which the absolute frequency of the derived form is correlated with the relative frequency of the derived form and the base. The chances of a high-frequency derived form (toward the right of each graph) being more frequent than its base are much higher than the chances of a low-frequency derived form being more frequent than its base. The forms occurring at the very left of each graph have a listed lexical frequency of 0 (treated here as 1, so as to facilitate taking the log). For such forms, it is obviously impossible for the base to be less frequent. The degree of opportunity for the derived form to be more frequent than the base increases as we move to the right of the graph.



To see that this is so, Tables 3 (for prefixes) and 4 (suffixes) list the percentage of forms falling below the x = y line for different values of the derived frequency. When the derived frequency is high, the proportion of forms for which the derived form is more frequent than the base is substantially higher than when the derived frequency is low. This is particularly true of the suffixed forms (Table 4).

The result of this tendency is that absolute frequency and relative frequency are not independent of one another. If relative frequency is an important factor in morphological decomposition, this correlation raises the possibility that previous observations regarding correlations between the absolute frequency of the derived form and decomposability may in fact be artifactual.

Table 5 breaks the prefixed forms into above- and below-average frequency forms. The above-average forms are significantly more likely

Table 3. *Number and percentage of prefixed forms with derived form more frequent than the base, broken down by frequency range of derived form*

| | derived > base | total forms | percentage |
|---|---|---|---|
| log(der) < 2 | 5 | 156 | 3 |
| 2 < log(der) < 4 | 24 | 196 | 12 |
| 4 < log(der) < 6 | 21 | 133 | 16 |
| 6 < log(der) < 8 | 3 | 19 | 16 |

Table 4. *Number and percentage of suffixed forms with derived form more frequent than the base, broken down by frequency range of derived form*

| | derived > base | total forms | percentage |
|---|---|---|---|
| log(der) < 2 | 13 | 914 | 1 |
| 2 < log(der) < 4 | 53 | 633 | 8 |
| 4 < log(der) < 6 | 39 | 355 | 11 |
| 6 < log(der) < 8 | 43 | 115 | 38 |
| 8 < log(der) | 9 | 11 | 82 |

Table 5. *Number of prefixed forms of above/below-average frequency, with derived form more/less frequent than the base (chi-square = 7.12, df = 1, p < 0.01)*

| | above-average freq. (%) | below-average freq. (%) |
|---|---|---|
| derived > base | 19 (36) | 88 (19) |
| derived < base | 34 (64) | 373 (81) |

to be more frequent than their base than the below-average forms (chi-square = 7.12, df = 1, $p < 0.01$).

This tendency is stronger still in suffixed forms, as shown in Table 6 (chi-square = 132.4, df = 1, $p < 0.001$).

Suffixed forms differ from prefixed forms in an important way. The onset of the derived form and the base are simultaneous in suffixed forms, whereas for prefixed forms the onset of the derived form is temporally prior to the onset of the base. Because the onset of the derived form precedes the onset of the base in prefixed forms, the derived form has a natural advantage in perception. If it is more frequent than the base, this reinforces the advantage, making a decomposed route unlikely. In suffixed forms, however, the temporal onset of the derived form and the base is identical. Furthermore, the offset of the base is reached before the offset of the entire derived form. Suffixed forms, then, do not afford the whole-form access route the same natural advantage over the decomposed route that prefixed forms do.

For both prefixed and suffixed forms, decomposition and whole-route access are possible, and so in both cases we expect to see an interaction between the frequency of the derived form and the base: the more frequent the derived form is relative to the base, the more likely a whole-word representation and access strategy will be. However, because the whole-word route does not have quite the same advantage in suffixed forms, one might reasonably expect the derived form to need to be relatively more frequent than the base in suffixed forms than in prefixed forms before signs of noncompositionality can be detected. Furthermore, we should expect prefixed words to become liberated from their bases at a higher rate and so be more likely to drop below the x = y line than suffixed words.

In light of these difference between prefixes and suffixes, it is worth carefully comparing Figures 2 and 3. Comparing the two figures reveals some interesting differences. Recall that the temporal nature of speech predicts forms containing prefixes to generally be less decomposed than suffixes. Some evidence for this comes from comparing the proportion of forms falling below the dotted x = y line for the two graphs. If we

Table 6. *Number of suffixed forms of above/below-average frequency, with derived form more/less frequent than the base (chi-square = 132.4, df = 1, $p < 0.001$)*

| | above-average freq. (%) | below-average freq. (%) |
|---|---|---|
| derived > base | 74 (46) | 222 (12) |
| derived < base | 87 (54) | 1619 (88) |



assume a link between relative frequency and decompositionality, then a larger proportion of points below this line would indicate a higher proportion of noncompositionality. In the case of a fully productive affix creating highly decomposable forms, we should expect points to gather fairly tightly around the regression line, and uniformly above the point at which base frequency and derived frequency are equal. Comparing Figures 2 and 3, note that a larger proportion of points falls below the x = y line for prefixes than suffixes. For prefixes, 53 of 515 forms fall below the line — 10%. For suffixes, a slightly smaller proportion of forms falls below the line — 158 of 2028 forms, or 8%.

This lends some support to the hypothesis that prefixed forms, in general, tend to be less decomposed. The correlation between base frequency and derived frequency is also much stronger for suffixes — the correlation for prefixes is extremely low. This, too, can be taken as indicative of the higher rates of decompositionality among suffixes than prefixes.

As argued above, we might expect the derived form to need to be relatively MORE frequent than the base in suffixed forms than in prefixed forms before signs of noncompositionality can be detected. That is, the relevant threshold of relative frequency may differ for prefixes and suffixes. In absolute terms, a derived prefixed form may not necessarily be more frequent than its base, before it is frequent enough to win out in access. And likewise, a derived suffix form that is slightly more frequent than the base it contains may still not be quite frequent enough to escape the bias afforded the parsing route.

Despite the fact that the relevant threshold may be slightly different for prefixes and suffixes, we still expect relative frequency to be related to decompositionality in both prefixed and suffixed forms. And, for the purposes of investigating this hypothesis, drawing a line at derived frequency = base frequency suffices as a first approximation at dividing the data.

Now then, we turn to a closer investigation of affixed forms, in order to determine whether the forms that have fallen below the dotted x = y line in Figures 2 and 3 display different properties than forms that fall above the line. In particular, we turn to an investigation of the relationship between lexical frequency and semantic transparency.

## 6. Lexical frequency and semantic transparency

This section investigates degree of semantic drift as a function of the frequency characteristics of the derived form and the base.

Two entries for words taken from the Websters 1913 Unabridged English Dictionary are shown below. Some tags relating to pronunciation have been removed.

**Dishorn** <def>To deprive of horns; <as>as, to <ex>dishorn</ex> cattle</as>.</def>

**Dislocate** <def>To displace; to put out of its proper place. Especially, of a bone: To remove from its normal connections with a neighboring bone; to put out of joint; to move from its socket; to disjoint; <as>as, to <ex>dislocate</ex> your bones</as>.</def>

The word *dishorn* is maximally semantically transparent. It has neither shifted nor proliferated in meaning. That it has not proliferated can be evidenced by the small number of definitions associated with it. One index of the fact that it has not shifted is the explicit presence of the base word *horn* in the definition. *Dislocate*, on the other hand, is not quite so transparent in meaning. Thus, not only is it associated with more different meanings than *dishorn* is, but the base word *locate* is not explicitly invoked in the definition.

If a prefixed word is highly transparent, then it should be easily defined by referring to the base word. We can therefore take the absence of the base word from the definition of an affixed word to be meaningful — a clear sign of semantic drift. If forms that are more frequent than their bases are more likely to undergo semantic drift than forms that are less frequent than their bases, then the former set should be less likely to mention their base word in the definition than the latter set. The availability of Websters 1913 dictionary in ascii form allowed for the automatic retrieval of definitions for the words in the dataset described in the previous section (515 prefixed and 2028 suffixed words — all with monomorphemic bases).

A short Perl script was written to test whether affixed forms in this dataset that are more frequent than their bases are less semantically transparent than those that are less frequent than the bases they contain. For each word in the dataset, Websters was consulted and the relevant entry was retrieved. The affix was stripped from the head word, and then the definitions were consulted to see if they contained the base word. If the word was found, a match was returned, and if not, some simple transformations were performed on the definition words (such as the removal of "s" endings, to catch plurals), and the definition was consulted again. If no entry was found for a CELEX prefixed word in the Websters dictionary, it was assumed to be transparent and was tagged as if the base was present in its definition. This set constituted an extremely small proportion of the words.



All forms that were tagged as not listing their base in the definition were then checked by hand. This process caught some bases that were not identified by the simple heuristics used in the Perl script. The script did not attempt any modifications on the base words. The word *dissatisfy*, for example, included *satisfied* in the definition, which was not caught by the script but was subsequently tagged by hand as a match. In addition, some definitions referred to other definitions that contained the base word. The definition for *Encrust*, for example, was simply "To incrust, see *Incrust*." And the definition for *Incrust* does contain the base word *crust*.

Only inflectional variants of the base words were accepted. Two exceptions involve definitions for the prefixes *dis-* and *un-*. Bases prefixed with *un-* were admitted for head words prefixed in *dis*, and vice versa. (1), for example, contains the entry for *unjoin*.

(1) **unjoin:** To disjoin.

This was coded as if the base word was mentioned in the definition. Words defined in this way were clearly semantically transparent. The results for the prefixed and suffixed datasets will be discussed separately.

### 6.1. *Semantic transparency of prefixed forms*

Table 7 shows the number of words for which the base was mentioned in the definition, for prefixed words with bases more/less frequent than their derived forms.

Relative frequency is related to semantic drift. Words for which the derived form is more frequent than the base are significantly less likely to mention their base in their definition than words for which the derived form is less frequent than the base (chi-square = 11.68, $p < 0.001$).

Table 7. *Number of forms with base absent/present in definition, for prefixed forms with the derived form more/less frequent than the base it contains (chi-square = 11.68, df = 1, p < 0.001)*

| | der > base (%) | base > der (%) |
|---|---|---|
| base absent | 20 (38) | 79 (17) |
| base present | 33 (62) | 382 (83) |

As discussed above, the literature assumes a strong relationship between the frequency of the derived form and semantic transparency. Table 8 shows the collapsed distribution of above-average-frequency and below-average-frequency derived forms for which the base is present or absent in the definition.

Above-average-frequency prefixed forms are no more likely than below-average-frequency forms to mention their bases explicitly in their definitions. The absolute frequency of the derived form does not appear to be relevant to semantic drift (chi-square = 0.23, n.s.).

Here, then, we have clear evidence that the relative frequency of the derived form and its base is relevant to semantic transparency of prefixed forms. The effect observed here cannot be an artifact of absolute frequency of the derived form — as the absolute frequency of the derived form appears to have absolutely no effect on the observed phenomena. The evidence involving prefixed forms is clear. In the following section we turn to an investigation of the relationship of lexical frequency to the semantic transparency of suffixed forms in English. Do the same patterns hold?

### 6.2. *Semantic transparency of suffixed forms*

The set of 2028 suffixed words was examined in order to assess the influence of relative frequency upon semantic drift in suffixed forms. We hypothesize that derived forms that are more frequent than their bases should be less likely to explicitly invoke their base in their definition than derived forms that are less frequent than their bases. The same procedure was followed as that described above for prefixed words.

The relevant distribution is shown in Table 9. The predicted pattern is present. Forms for which the derived form is more frequent than the base are significantly less likely to mention their base in their definition than forms for which the base is more frequent (chi-square = 6.63, df = 1, p < 0.01).

Table 8. *Number of forms with base absent/present in definition, for prefixed forms with the derived form having above/below average lexical frequency (chi-square = 0.23, n.s.)*

| | above-average freq. (%) | below-average freq. (%) |
|---|---|---|
| base absent | 23 (21) | 76 (19) |
| base present | 85 (79) | 331 (81) |



Table 9. *Number of forms with base absent/present in definition, for suffixed forms with the derived form more/less frequent than the base it contains (chi-square = 6.63, df = 1, p < 0.01)*

| | der > base (%) | base > der (%) |
|---|---|---|
| base absent | 25 (16) | 169 (9) |
| base present | 133 (84) | 1675 (91) |

This indicates that the pattern observed with prefixed forms is also present for suffixed words. The same dataset is given in Table 10, this time broken down by absolute frequency. Numbers of forms for which the base was present or absent in the definition are given, for forms that are of above-average or below-average lexical frequency. The average lexical frequency for derived forms in the dataset was 122.2 (per 17.9 million).

There is a tendency for frequent derived forms to mention their bases in their definitions less often than infrequent derived forms. However, this tendency does not reach significance. Thus the significant relative-frequency effect observed above is not an artifact of a strong effect of absolute frequency. This reinforces the claim that there is a relationship between relative frequency and semantic transparency in suffixed forms in English, and that relative frequency is more directly relevant to decomposition than absolute lexical frequency.

## 7. Discussion

### 7.1. *Lexical frequency, decomposition, and transparency*

The literature widely assumes a direct link between high lexical frequency, nondecomposability, and nontransparency. However, such a direct link is in fact not predicted by models of morphological processing. Rather, processing models appear to predict a link between nondecomposition

Table 10. *Number of forms with base absent/present in definition, for suffixed forms with the derived form having above/below-average lexical frequency (chi-square = 3.58, df = 1, n.s. [p < 0.06])*

| | above-average freq. (%) | below-average freq. (%) |
|---|---|---|
| base absent | 38 (13) | 158 (9) |
| base present | 258 (87) | 1574 (91) |

and relative frequency — the frequency of the derived form relative to its base.

This paper has provided experimental evidence that a relative-frequency effect does, in fact, exist — subjects consistently rate derived forms that are more frequent than their bases as appearing less complex than matched counterparts that are less frequent than the bases they contain. A careful investigation of the relationship between derived frequencies and base frequencies reveals that relative frequency and the frequency of the derived form are highly correlated. This raises the possibility that previously observed effects of the frequency of the derived form may in fact be artifactual.

In order to investigate this question further we turned to an analysis of semantic transparency. This dictionary study revealed that semantic transparency is much better predicted by relative frequency than by absolute frequency. The lack of direct relation between high frequency and nontransparency conflicts with widespread assumptions in the literature and should lead us to seriously reconsider arguments based on this supposed link. The assumed link between transparency and frequency, for example, has been used as a basis for model building. Baayen (1993), in defining a measure of productivity, limits the measure to take into account only those forms falling below a certain frequency threshold $\theta$.

The low choice of $\theta$ in the present paper is motivated by the constraint that only semantically transparent complex words contribute to the activation level A. Since transparency is inversely correlated with frequency, higher values of $\theta$ would lead to the inclusion of opaque and less transparent forms in the frequency counts. In the absence of indications in the CELEX database of the (degree of) semantic transparency of complex words, and in the absence of principled methods by means of which degrees of transparency and their effect on processing can be properly evaluated, the research strategy adopted here is to concentrate on that frequency range where complex words are most likely to be transparent (Baayen 1993: 203).

If the threshold is set low enough, then Baayen may be right that nontransparent forms will be excluded from the calculation. For a very-low-frequency form, the chances of the base being even lower frequency are small. However, setting such a threshold low also excludes a large number of higher-frequency words that are nonetheless transparent. The results of this paper suggest that the most efficient frequency-based method to exclude nontransparent forms from a dataset would be to exclude all forms that are more frequent than the bases they contain.

One caveat is that the results presented here are based exclusively on bimorphemic words — prefixed or suffixed words with monomorphemic bases. The degree to which these results transfer to multiply complex words remains to be tested.



### 7.2. *Methodological consequences*

The results presented here provide evidence that relative frequency of the derived form and the base is related to decomposability, for both prefixed and suffixed forms. This result has consequences, not only for models of morphological access and representation, but also for methodology used to investigate these issues.

There is a large experimental literature dedicated to investigating whether affixed words are accessed whole, or in decomposed form. Comparisons are made between regular and irregular affixes, inflections versus derivations, level one versus level two derivations, phonologically transparent versus nontransparent affixes, prefixes and suffixes, and various other contrasts, to attempt to establish in which cases decomposition takes place. Much of the literature deals with orthographic input. And almost all of it either controls for or manipulates lexical frequency.

There are numerous paradigms used to test for decomposition, but two of them dominate the literature — repetition priming, and response latency measurement.

In the priming literature, an attempt is generally made to control for lexical frequency, and priming effects are investigated. Does a derived form prime its base? Does a base prime related derived forms? And if so, to what degree? Results are mixed, but as a whole, the body of literature certainly demonstrates that priming takes place in some cases. Delimiting those cases has proved to be a difficult task.

I will not give an exhaustive review of the literature here but will describe some typical results, so as to give its general flavor. Many affixed forms do tend to prime the bases they contain. Some argue that this priming effect is stronger for inflectional affixes than derivational affixes (e.g. Stanners et al. 1979 for English; Feldman 1994 for Serbian). Others have demonstrated equivalent priming for inflections and derivations (Burani and Laudanna 1992 for Italian; Napps 1989 and Raveh and Rueckl 2000 for English).

Among derivational affixes, some experiments tend to suggest that priming is the same regardless of phonological transparency (Marslen-Wilson and Zhou 1999), whereas others suggest that more priming occurs when the base is phonologically transparent (Tsapkini et al. 1999). Auditory presentation of derived suffixed forms does not facilitate recognition of other suffixed forms sharing the same base. Presentation

of *sanely*, for example, does not facilitate *sanity* (Marslen-Wilson et al. 1994; Feldman and Soltano 1999; Marslen-Wilson and Zhou 1999). There may be a modality difference here, with priming occurring between derived suffixed forms when presented visually, and no apparent effect cross-modally (Feldman and Soltano 1999). However, priming appears to occur between prefixed and suffixed forms in all modalities (Feldman and Soltano 1999). Productive derivational affixes also appear to prime themselves — *darkness*, for example, facilitates recognition of *toughness* (Marslen-Wilson, Ford, and Zhou 1997, cited in Marslen-Wilson and Zhou 1999). Finally, low-frequency derived forms appear to prime their stems more than high-frequency derived forms do (Meunier and Segui 1999b).

Among inflected forms, regular forms appear to display full priming (Stanners et al. 1979; Napps 1989; Marslen-Wilson et al. 1993). Experimenters have variously claimed that English irregular inflections display no priming (Kempley and Morton 1982), reduced priming (Stanners et al. 1979), and full priming (Fowler et al. 1985). Marslen-Wilson et al. (1993) argue that while no priming occurs for irregulars such as *burn–burnt*, inhibition actually occurs when the inflection involves a vowel change in the stem. That is *gave* inhibits recognition of *give*.

In the literature on response latencies, lexical frequency is typically manipulated in a lexical decision task. If, for example, derived forms with high-frequency bases are recognized more quickly than derived forms with low-frequency bases, this would provide some evidence for the role of the base in access, that is, for decomposition. If, on the other hand, high-frequency derived forms are recognized faster than low-frequency derived forms, this is thought to provide evidence for a whole-word access strategy.

When base frequencies are kept constant, some regular inflections show effects of surface frequency upon response latencies. Sereno and Jongman (1997), for example, demonstrate an effect of surface frequency upon recognition of English regular plurals. Bertram et al. (2000) suggest that this may be due to the fact that *-s* is homonymal in English, which would slow down processing in decomposition. They present similar homonymal examples from Dutch and Finnish, which display surface-frequency effects. Burani and Caramazza (1987) argue that reaction times for suffixed derived words in Italian are related to surface frequency. Meunier and Segni (1999a, 1999b) demonstrate that some suffixed forms in French display surface-frequency effects and some do not and argue that this depends on the word's position in the morphological family — forms compete with one another based on surface frequency. Vannest and Boland (1999) find a surface-frequency effect for forms in *-ity* and *-ation*, but none for forms suffixed in *-less*. They argue that this may reflect a difference between level 1 and level 2 affixes, but, in a second experiment involving more level 2 affixes, they find little evidence to support this claim.



When stem frequency is manipulated and surface frequency is kept constant, Bertram et al. (2000) argue that stem-frequency effects are absent for affixes that are homonymal but may otherwise be present. Burani and Caramazza (1987) show that they are present for suffixed derived words in Italian. Alegre and Gordon (1999b) argue that there is a frequency effect, with stem-frequency effects emerging when the surface frequency is below six occurrences per million (although Baayen et al. i.p. argue that this result is an artifact of the small corpus used). Meunier and Segui (1999b) have demonstrated that there is an effect of stem frequency for suffixed words with both high and low surface frequency. However, Vannest and Boland (1999) show that the effect is present for some affixes but not others and suggest that it will always be absent for level 1 affixes. Bradley (1979) found similarly variable results. The results of Cole et al. (1989) indicate that the cumulative root frequency is relevant for the processing of suffixed forms, but not for prefixed forms. This disjunction is supported by the results of Beauvillain and Segui (1992), who demonstrate an effect of surface frequency for prefixed forms but not suffixed forms. (However, note the results outlined above, which claim there is a surface-frequency effect for at least some suffixed forms.) And Taft and Forster (1975) and Taft (1979) argue that there is a stem-frequency effect for prefixed forms, even when they have bound stems.

As should be apparent, there are conflicting and confusing results in both the repetition priming and the response latency literatures. Many of the experiments are conducted with an implicit assumption that words containing the same affix are a largely homogeneous set. This makes the results difficult to interpret in relation to the word-specific effects described in this paper and may go some way toward explaining the variable results.

While it is apparent that the base sometimes plays a role in the perception of derived forms, exactly what that role is, and in what cases it is relevant, is not at all clear. And the finding that the RELATIVE frequency of the derived form and the base may affect the access strategy makes the results of many of the studies reported above particularly difficult to interpret. Most of the studies either control or manipulate surface and/or base frequency, because there is an implicit assumption that these factors may affect response times. However, if the relationship of these factors to each other affects the access strategy, then controlling

one or both factors can have radically different consequences depending on the level at which they are controlled RELATIVE TO ONE ANOTHER.

Consider, for example, a hypothetical response latency experiment where we are controlling base frequency and manipulating surface frequency, in order to determine whether there are whole-word representations for words derived with a given affix. Assume that we have (miraculously) managed to control all of the base words such that they all have frequencies 300/million in English. Now, what do we expect to happen when we manipulate surface frequency? Well, if the relative frequency of the base and the derived form has an important effect on representation and access, then it will depend on the frequency range within which we manipulate surface frequency. If we manipulate our derived forms such that they all range between 400–1000/million in English, then we expect that all should have robust whole-word representations (as they are more frequent than the bases they contain). Then the more-frequent derived forms will be accessed more quickly than the less-frequent derived forms, and a significant effect of surface frequency should arise. Now, what if we manipulate our derived forms such that they all range between 50–200/million? In this case, all the derived forms are less frequent than the bases they contain. If the affix is phonologically transparent and is productive, then such forms may well be accessed via their component parts. As such, because the bases are of roughly the same frequency, there should be little difference in the response latencies for these forms. In a third possible scenario, we might manipulate the surface frequencies such that they range from 100 per million to 600 per million, thus straddling the level at which base frequency is controlled. In such a case we expect variable behavior. If the whole-word route is significantly faster than the decomposed route, then perhaps the more frequent words in our set will be accessed faster. If so, a misleading frequency effect will arise. Words with high surface frequency will be accessed faster than words with low surface frequency, but it would be incorrect to conclude that words with this affix are all accessed as whole words, and that it is the frequency of these words that is directly responsible for the observed difference.

It should be clear that, if relative frequency is relevant to access, controlling both factors, or manipulating one of them, should have radically different effects depending on the relative frequency ranges in which they are controlled. However, such information is rarely reported in the literature. When frequency characteristics are reported, they tend to take the form of means, or ranges. Vannest and Boland (1999), for example, report that they matched words for surface frequency and manipulated base frequency; and they report means for each of the two



sets, making it impossible to deduce their overall profile in terms of relative frequency. Other experimenters also report the range within which their stimuli were controlled. For example, in their priming studies, Raveh and Rueckl (2000) report that their base-form targets had a mean frequency of 27 (median 12, range 1–200); their inflected primes had a mean frequency of 27 (median 14, range 0–172); and the derivational primes had a mean frequency of 18 (median 8, range 0–89) (Raveh and Rueckl 2000: 112). As such, the relative frequency of the affixed form and the base must have displayed variation within each condition, but we cannot know whether the derived form was more often more frequent than the base in one condition than the other.

In some cases, the base frequencies differ markedly across conditions. For example Sonnenstuhl et al. (1999) investigate cross-modal priming of German regular and irregular inflections. They controlled stem frequency within conditions but were not able to do this across conditions.

> ... since nouns which take *-s* plurals tend to have lower frequencies in German than nouns that take *-er* plurals (CELEX mean lemma frequencies are 52 for *-s* plurals and 563 for *-er* plurals), the lemma frequencies of the nouns we used for *-er* plurals were higher (CELEX mean frequency 538) than those of the nouns with *-s* plurals (CELEX mean lemma frequency 43). We would expect this difference to lead to shorter lexical decision times for nouns that take *-er* plurals in each of the three experimental conditions, especially in the unprimed control condition. Morphological priming effects, however, are not determined by directly comparing *-s* and *-er* plurals, but are measured within target sets, that is by comparing the same targets in the experimental, identity and control condition. As this is done separately for *-s* plurals and for *-er* plurals, the different lemma frequencies mentioned above should not affect the priming results (Sonnenstuhl et al. 1999: 215–216).

The argument they put forward stands if lexical frequency's only role is to affect speed of access. However, if lexical frequency, and in particular, the relative frequency of the derived form and the base, affects the actual form of the representation and the preferred access strategy, then such an experimental approach is problematic. The difference between the two conditions described above presumably means that the relative frequency relationships in the two conditions also differ, though the frequency of the inflected forms was not controlled, so we cannot know this for sure.

In short, relative lexical frequency is not taken into account in current experimental work on this topic. And while we are often told that experimenters "controlled for lexical frequency," the exact relationships that hold vary across conditions, experiments, and experimenters and are often impossible to deduce. Of course this last generalization is not

without the occassional exception — for a recent example see Bertram, Schreuder, and Baayen (2000), who carefully and fully document the frequency characteristics of their stimuli.

In general, relative lexical frequency is not currently controlled for in experiments on morphological access and representation. This failure to take account of relative-frequency relations may well account for some of the conflicting results reported in the literature.

## 8. Conclusion

It is widely believed that high-frequency morphologically complex forms tend to display characteristics of noncompositionality. Bybee even goes so far as to call this a universal (Bybee 1995a: 235). This paper has demonstrated that relative frequency matters more than absolute frequency. A low-frequency form is likely to be nontransparent if it is composed of even-lower-frequency parts. And a high-frequency form may be highly decomposable if the base word it contains is higher frequency still. This finding resolves an incongruity between previous assumptions in the literature and the predictions of processing models and provides evidence in support of a time-tested adage: Everything Is Relative.

*Received 9 January 2001*
*Revised version received*
*10 July 2001*

*University of Canterbury*

## Notes

1. The work reported in this paper forms part of my Northwestern University dissertation (Hay 2000), which was completed under the excellent guidance of Janet Pierrehumbert. It has also greatly benefited from the comments of Harald Baayen, and one anonymous reviewer. Correspondence address: Department of Linguistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. E-mail: j.hay@ling.canterbury,ac.nz.
2. These problems may not be of the obvious kind, however. See Schreuder et al. (i.p.), who provide evidence that opaque low-frequency words with high-frequency constituents are processed in a way that the opaque reading is in fact the first to become available. They explain this result by positing integration nodes, which "serve the purpose of redirecting the activation of the constituents to the correct opaque meaning." These results suggest that the processing system may have effective strategies for dealing with opaque words with higher-frequency constituents. Nonetheless, they also provide evidence that the meanings of the component constituents do become available, suggesting that opaque words with higher-frequency constituents may still be disadvantaged relative to opaque words with lower-frequency constituents, due to the effects of these later, strong competitors.
3. Note that morphological family size is not controlled for in this data set (see, e.g., Schreuder and Baayen 1997; De Jong et al. 2000). Post-hoc analysis of the results reveals that morphological family size does play some role in predicting behavior in this task; however, none of the results reported here are artifacts of lack of control for family size.



## References

Alegre, Maria; and Gordon, Peter (1999a). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40, 41–61.

—; and Gordon, Peter (1999b). Rule-based versus associative processes in derivational morphology. *Brain and Language* 68, 347–354.

Baayen, R. Harald (1992). Quantitative aspects of morphological productivity. In *Yearbook of Morphology 1991*, Geert Booij and Jaap van Marle (eds.), 109–150. Dordrecht: Kluwer Academic.

—(1993). On frequency, transparency and productivity. In *Yearbook of Morphology 1992*, Geert Booij and Jaap van Marle (eds.), 181–208. Kluwer Academic.

—(1994). Productivity in language production. *Language and Cognitive Processes* 9(3), 447–469.

—; and Lieber, Rochelle (1997). Word frequency distributions and lexical semantics. *Computers and the Humanities* 30, 281–291.

—; Lieber, Rochelle; and Schreuder, Robert (1997). The morphological complexity of simplex nouns. *Linguistics* 35, 861–877.

—; Piepenbrock, Richard; and Gulikens, Leon (1995). The CELEX lexical database (release 2), cd-rom. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

—; and Schreuder, Robert (1999). War and peace: morphemes and full forms in a non-interactive activation parallel dual-route model. *Brain and Language* 68, 27–32.

—; and Schreuder, Robert (2000). Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society*, Series A: *Mathematical, Physical and Engineering Sciences* 358, 1–3.

—; Schreuder, Robert; De Jong, Nivja; and Krott, Andrea (i.p.). Dutch inflection: the rules that prove the exception. In *Storage and Computation in the Language Faculty*, Sieb Nooteboom, Fred Weerman, and Frank Wijnen (eds.). Dordrecht: Kluwer Academic.

Balota, David A.; and Chumbley, James I. (1984). Are lexical decisions good measures of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance* 10, 340–357.

Beauvillain, Cecile; and Segui, Juan (1992). Representation and processing of morphological information. In *Orthography, Phonology, Morphology, and Meaning*, Ram Frost and Leonard Katz (eds.), chapter 19, 377–388. Amsterdam: Elsevier.

Bertram, Raymond; Laine, Matti; Baayen, R. Harald; Schreuder, Robert; and Hyöna, Jukka (2000). Affixal homonymy triggers full-form storage, even with inflected words, even in a morphologically rich language. *Cognition* 74, B13–B25.

—; Schreuder, Robert; and Baayen, R. Harald (2000). The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26(2), 489–511.

Bradley, Diane (1979). Lexical representation of derivational relations. In *Juncture*, Mark Aronoff and Mary-Louise Kean (eds.). Cambridge, MA: MIT Press.

Burani, Cristina; and Caramazza, Alfonso (1987). Representation and processing of derived words. *Language and Cognitive Processes* 2, 217–227.

—; and Laudanna, Alessandro (1992). Units of representation for derived words in the lexicon. In *Orthography, Phonology, Morphology, and Meaning*, Ram Frost and Leonard Katz (eds.), chapter 18, 361–376. Amsterdam: Elsevier.

Butterworth, Brian (1983). Lexical representation. In *Language Production*, vol. 2, Brian Butterworth (ed.), 257–294. London: Academic Press.

Bybee, Joan (1985). *Morphology: A Study of the Relation Between Meaning and Form*. Amsterdam: Benjamins.

—(1988). Morphology as lexical organization. *Theoretical Morphology: Approaches in Modern Linguistics*, Michael Hammond and Michael Noonan (eds.), 119–142. San Diego: Academic Press.

—(1995a). Diachronic and typological properties of morphology and their implications for representation. In *Morphological Aspects of Language Processing*, Laurie Beth Feldman (ed.), 225–246. Hillsdale, NJ: Erlbaum.

—(1995b). Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5), 425–455.

Caramazza, Alfonso; Laudanna, Alessandro; and Romani, Cristina (1988). Lexical access and inflectional morphology. *Cognition* 28, 297–332.

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

Cole, P.; Beauvillan, C.; and Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words: a differential frequency effect. *Journal of Memory and Language* 28, 1–13.

Connine, Cynthia M.; Titone, Debra; and Wang, Jian (1993). Auditory word recognition: extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition* 19(1), 81–94.

De Jong, Nivja H.; Schreuder, Robert; and Baayen, R. Harald (2000). The morphological family size effect and morphology. *Language and Cognitive Processes* 15, 329–365.

Feldman, Laurie Beth (1994). Beyond orthography and phonology: differences between inflections and derivations. *Journal of Memory and Language* 33, 442–470.

—; and Soltano, Emily G. (1999). Morphological priming: the role of prime duration, semantic transparency, and affix position. *Brain and Language* 68, 33–39.

Fowler, Carol A.; Napps, Shirley E.; and Feldman, Laurie Beth (1985). Relations among regular and irregular morphologically related words in the lexicon as revealed by repetition priming. *Memory and Cognition* 13, 241–255.

Frauenfelder, Uli H.; and Schreuder, Robert (1992). Constraining psycholinguistic models of morphological processing and representation: the role of productivity. In *Yearbook of Morphology 1991*, Geert Booij and Jaap van Marle (eds.), 165–185. Dordrecht: Kluwer Academic.

Grosjean, Francois (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics* 45, 189–195.

Harwood, F. W.; and Wright, Alison M. (1956). Statistical study of English word formation. *Language* 32(2), 260–273.

Hay, Jennifer (2000). Causes and consequences of word structure. Unpublished Ph.D. thesis, Northwestern University.

Johnson, Keith (1997a). The auditory/perceptual basis for speech segmentation. In *Ohio State University Working Papers in Linguistics: Papers from the Linguistics Laboratory*, vol. 50, Kim Ainsworth-Darnell and Mariapaola D'Imperio (eds.). Columbus: Ohio State University.



—(1997b). Speech perception without speaker normalization. In *Talker Variability in Speech Processing*, K. Johnson and J. Mullenix (eds.). San Diego: Academic Press.

Kempley, S.; and Morton, J. (1982). The effects of priming with regularly and irregularly related words in auditory word recognition. *British Journal of Psychology* 73, 441–454.

Losiewicz, Beth L. (1992). The effect of frequency on linguistic morphology. Unpublished Ph.D. thesis, University of Texas at Austin.

Marslen-Wilson, William; Ford, Michael; and Zhou, Xiaolin (1997). The combinatorial lexicon: priming derivational affixes. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

—; Hare, Mary; and Older, Lianne (1993). Inflectional morphology and phonological regularity in the English mental lexicon. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Princeton, NJ: Erlbaum.

—; Tyler, Lorraine K.; Waksler, Rachelle; and Older, Lianne (1994). Morphology and meaning in the English mental lexicon. *Psychological Review* 101(1), 3–33.

—; and Zhou, Xiaolin (1999). Abstractness, allomorphy, and lexical architecture. *Language and Cognitive Processes* 14(4), 321–352.

—; Zhou, Xiaolin; and Ford, Michael (1997). Morphology, modality, and lexical architecture. In *Yearbook of Morphology 1996*, Geert Booij and Jaap van Marle (eds.), 117–134. Dordrecht: Kluwer Academic.

McClelland, James L.; and Elman, Jeffrey L. (1986). The TRACE model of speech perception. *Cognitive Psychology* 18, 1–86.

Meunier, Fanny; and Segui, Juan (1999a). Frequency effects in auditory word recognition: the case of suffixed words. *Journal of Memory and Language* 41, 327–344.

—; and Segui, Juan (1999b). Morphological priming effect: the role of surface frequency. *Brain and Language* 68, 54–60.

Modor, Carol Lynn (1992). Productivity and categorization in morphological classes. Unpublished Ph.D. thesis, SUNY at Buffalo.

Napps, Shirley E. (1989). Morphemic relationships in the lexicon: are they distinct from semantic and formal relationships? *Memory and Cognition* 17(6), 729–739.

Norris, Dennis; McQueen, James M.; and Cutler, Anne (2000). Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences* 23(3).

Pagliuca, William (1976). Pre-fixing. Unpublished manuscript, SUNY at Buffalo.

Raveh, Michal; and Rueckl, Jay G. (2000). Equivalent effects of inflected and derived primes: long-term morphological priming in fragment completion and lexical decision. *Journal of Memory and Language* 42, 103–119.

Schreuder, Robert; and Baayen, R. Harald (1995). Modeling morphological processing. In *Morphological Aspects of Language Processing*, Laurie Beth Feldman (ed.), 131–156. Hillsdale, NJ: Erlbaum.

—; and Baayen, R. Harald (1997). How complex simplex words can be. *Journal of Memory and Language* 37, 118–139.

—; Burani, Cristina; and Baayen, R. Harald (i.p.). Parsing and semantic opacity. In *Morphology and the Mental Lexicon*, Egbert Assink and Dominiek Sandra (eds.). Dordrecht: Kluwer Academic.

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

Sonnenstuhl, Ingrid; Eisenbeiss, Sonja; and Clahsen, Harald (1999). Morphological priming in the German mental lexicon. *Cognition* 72, 203–236.

Stanners, R. F.; Neiser, J. J.; Hernon, W. P.; and Hall, R. (1979). Memory representations for morphologically related words. *Journal of Verbal Learning and Verbal Behavior* 18, 399–412.

Stemberger, Joseph Paul; and MacWhinney, Brian (1986). Frequency and the storage of regularly inflected forms. *Memory and Cognition* 14, 17–26.
—; and MacWhinney, Brian (1988). Are inflected forms stored in the lexicon? In *Theoretical Morphology: Approaches in Modern Linguistics*, Michael Hammond and Michael Noonan (eds.). San Diego: Academic Press.
Taft, Marcus (1979). Recognition of affixed words and the word frequency effect. *Memory and Cognition* 7(4), 263–272.
—(1985). The decoding of words in lexical access: a review of the morphographic approach. In *Reading Research: Advances in Theory and Practice*, vol. 5, Derek Besner, T. Gary Waller, and G. Ernest Mackinnon (eds.). London: Academic Press.
—; and Forster, Kenneth (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior* 14, 638–647.
Tsapkini, Kyrana; Kehayia, Eva; and Jarema, Gonia (1999). Does phonological change play a role in the recognition of derived forms across modalities? *Brain and Language* 68, 318–323.
Vannest, Jennifer; and Boland, Julie E. (1999). Lexical morphology and lexical access. *Brain and Language* 6, 324–332.
Wurm, Lee (1997). Auditory processing of prefixed English words is both continuous and decompositional. *Journal of Memory and Language* 37, 438–461.

# Syntactic definiteness in the grammar of Modern Hebrew*

GABI DANON

*Abstract*

*Definiteness has often been assumed to play a role in syntax, most notably in relation to various "definiteness effects" and case alternations (Belletti 1988; De Hoop 1992; and many others). The question whether this involves a semantic property that is relevant in syntax, or an independent syntactic representation of definiteness, remains to a large extent unanswered. This paper shows that, on the one hand, Hebrew provides independent evidence for assuming a definiteness feature in syntax; and on the other hand, this formal definiteness does not simply correlate with semantic definiteness, and there is no simple one-to-one mapping between the two kinds of definiteness. The second part of this paper focuses on the Hebrew object marker* et, *which appears only in front of DPs having the syntactic definiteness feature. I argue that* et *fulfills a requirement for structural case that Hebrew verbs cannot assign, and that this requirement is related to the representation of definiteness as a formal feature and not to any semantic property. In this light I consider Belletti's (1988) theory of abstract partitive and show that Hebrew object marking seems to provide evidence against it.*

## 1. Formal definiteness features

Definiteness in natural language is usually seen as a semantic or pragmatic property of noun phrases. Over the years, however, definiteness has also been discussed in the syntactic literature as well. Definiteness FEATURES, taken as formal features that are marked on certain lexical entries and play a role in syntactic processes, have often been either explicitly proposed or implicitly assumed (for Hebrew, see Hazout 1990; Siloni 1997; Borer 1998, and others). But as opposed to phi features such as number and gender, whose morphological realizations in many languages are clear and which trigger purely syntactic agreement phenomena, the

*Linguistics* 39–6 (2001), 1071–1116
0024–3949/01/0039–1071