Computers and the Humanities **30:** 281–291, 1997.

Word Frequency Distributions and Lexical Semantics

R. Harald Baayen[1] * and Rochelle Lieber[2] †
[1]*Max Planck Institute for Psycholinguistics, Nijmegen 6525 XD, The Netherlands;* [2]*University of New Hampshire, Durham, NH 03824, USA; e-mail: baayen@mpi.nl, rl@christa.unh.edu*

Key words: word frequency distributions, lexical conceptual structure, lognormality, bimodal, density estimation

Abstract

This paper addresses the relation between meaning, lexical productivity, and frequency of use. Using density estimation as a visualization tool, we show that differences in semantic structure can be reflected in probability density functions estimated for word frequency distributions. We call attention to an example of a bimodal density, and suggest that bimodality arises when distributions of well-entrenched lexical items, which appear to be lognormal, are mixed with distributions of productively created nonce formations.

1. Introduction

Words have widely varying frequencies of use. The distributions of these frequencies provide a rich source of information that is exploited in quantitative stylistics (Brunet, 1978; Holmes, 1994; Hubert and Labbe, 1988; Muller, 1977), lexicography (Martin 1983, 1988), and linguistics (Anshen and Aronoff, 1988; Baayen, 1992, 1993, 1994; Harwood and Wright, 1956; Koehler, 1986). For example, it has been found that when particular aspects of meaning are consistently expressed by overt affixes, differently shaped word distributions can be observed; the word frequency distribution of nouns in the English suffix *-ness* (*goodness*) is differently shaped than that of nouns in *-ee* (*escapee*). Presumably it is the combination of form and meaning which crucially underlies these differences.

In this paper we explore two related questions concerning the relation between morphology, lexical semantics, and frequency distribution. First, some affixes express a number of related but slightly different meanings. Do these different meanings bring about differences in the frequency distributions of subsets of words expressing these meanings? Second, we ask whether it might also be possible to observe differences between frequency distributions for sets of semantically different conversion verbs, verbs that are not formally marked as different from monomorphemic words. Are differences in meaning between sets of noun to verb conversion verbs (*to saddle, to boss*) reflected in their frequencies of use?

2. Word Frequency Distributions

To answer these questions, we need a way to summarize frequency distributions. One way to do so is to consider the so-called frequency spectrum, which summarizes for each frequency of use f the number of types n_f with that frequency. Most frequency spectra are characterized by the property that the numbers of types decrease with increasing frequency of use. For instance, the number of hapax legomena (items of frequency 1), n_1, generally is much larger than the number of dis legomena (items of frequency 2), n_2. In turn, n_2 is almost always larger than n_3, although the difference in magnitude may be less than for n_1 and n_2.

This is shown in the upper left panel of Figure 1, which plots n_f against f for the morphological category of words in the English suffix *-ness*. The frequency counts underlying this graph are taken from the

* Harald Baayen is staff member of the Max Planck Institute for Psycholinguistics.

† Rochelle Lieber is professor of English at the University of New Hampshire, Durham.

CELEX lexical database (Baayen, Piepenbrock, and Van Rijn, 1993). (This database lists the frequencies of use of English words in the 18 million wordform Cobuild corpus (Renouf, 1987) and of Dutch words in a 42 million wordform corpus compiled by the Institute for Dutch Lexicography in Leiden. CELEX is the source for all word frequency distributions discussed in this paper.) For ease of presentation, the most frequent word in *-ness, business*, which occurs 4234 times, has not been plotted. Note that the largest numbers of types are concentrated at the very left edge of this plot, and that there is a long tail of higher frequencies that are instantiated by only one type.

Models for word frequency distributions assume that the frequency spectrum is best approximated by a monotonically decreasing continuous function. All the models that we are aware of (see Chitashvili and Baayen, 1993, for a review) focus on the head of the frequency spectrum, n_1, n_2, ..., n_k, and the total number of different types $V = \sum_{f=1}^{f\,max} n_f$, to set the parameters of this decreasing function. This seems reasonable enough, given that most types occur with the lower frequencies of use. However, approximation by a monotonically decreasing continuous function implies that higher frequency types should occur more sparsely with increasing frequency of use. For the higher frequencies of use, the continuous function approximating the frequency spectrum predicts that n_f should become less than one. This is impossible for empirical frequency distributions, which are discrete. But the approximation will still be valid if incidentally frequencies of use will be represented by one particular type. But as n_f approaches zero, the rate at which incidental frequencies of use are realized should also approach zero. The upper left panel of Figure 1 suggests informally that this is indeed the case. Although the hypothesis of increasing sparseness in the high-frequency tail of the frequency spectrum is valid in general, we shall see that there are distributions for which it is not strictly correct.

In order to study the full frequency spectrum, including its high-frequency tail, we need a better visualization technique than the one used in the upper left panel of Figure 1. The technique we have used is density estimation on the basis of the logarithmically transformed frequency spectrum. The logarithmic transformation allows us to pack all attested frequencies of use into a limited interval, while preserving n_f as a decreasing function of f. At the same time, the logarithmic transform is well-motivated from a psycholinguistic point of view, as various studies have shown that frequency of use is perceived logarithmically. (Carroll, 1967, 1970; Rubenstein and Pollack, 1963; Scarborough, Cortese, and Scarborough, 1977; Shapiro, 1969) The result of the logarithmic transformation is shown in the upper right panel of Figure 1, which now covers all words in *-ness*, including *business*. The decrease of n_f for increasing f is clearly visible for the highest values of f. Although almost all word types have very low frequencies of use, this graph reveals that there is a non-negligible number of types that have a frequency of use in the middle range around $\log f = 4.0$. This frequency range is fairly densely populated by words with similar frequencies of use. Even though the values of n_f in this range are quite low, the numbers of types within successive intervals of frequencies of use may well be more similar and may decrease less rapidly than the logarithmic plot of the frequency spectrum suggests.

The appropriate technique for testing for this possibility is density estimation. Density estimation can be viewed as a method to improve on histograms. Histograms are known to be unreliable, because the shape of the distribution that they suggest is heavily dependent on how the first, leftmost bin (or bar) of the histogram is positioned, as well as on the width of the bins themselves. Even for fixed bin width, the choice of origin may change a distribution that first appeared to be unimodal into a bimodal or even a trimodal distribution (Haerdle, 1991). Since the choice of origin may determine what a distribution will look like, it is important to find a method that is independent of the particular origin chosen. The solution adopted in density estimation is to average over histograms with shifted origins. The resulting average histogram is a curve representing the probability density function of the distribution which, for normally distributed random variables, has the well-known bell-shaped form.

The bottom panel of Figure 1 plots the probability density functions for the English suffixes *-ness* and *-ity*. First consider the curve for *-ness*. Not surprisingly, it has its mode around zero. It also clearly shows that there is a reasonable probability of observing words with a log frequency around 4.0 that is much larger than the upper right panel of Figure 1 would lead one to believe. Note that the estimated probability density function is non-zero for log frequencies smaller than zero, which means that it assigns non-zero probabilities to frequencies of use less than one. This is a direct consequence of modeling a discrete distribution by means of a continuous function. This is no problem

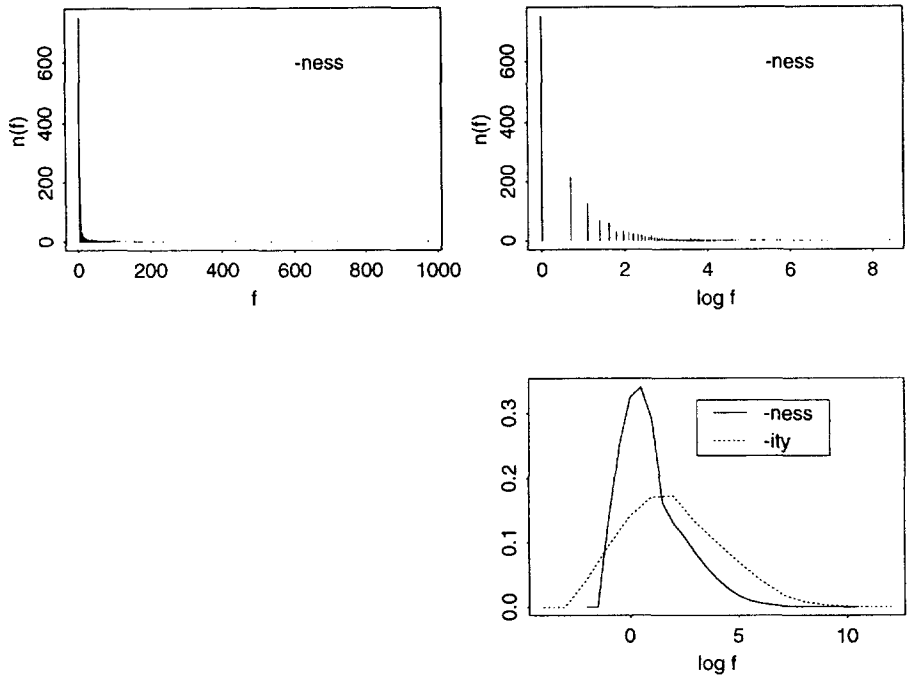

Figure 1. Summarizing the frequency spectrum. The upper left panel plots the number of words $n(f)$ in *-ness* with frequency f against f. The upper right panel plots the same distribution, but now on a logarithmic frequency scale. The lower panel shows the probability density functions of *-ness* (solid line) and its less productive near synonym *-ity* (dotted line).

for our purposes, since the area under the curve to the left of the line $x = 0$, the Y-axis, is part of the area to the left of the line $x = 0 + \epsilon$ that represents the probability of the hapax legomena. (The bin width used in density estimation determines the value of ϵ.)

The bottom panel of Figure 1 also shows the probability density function of the English suffix *-ity*. This suffix is known to be less productive than its near synonym *-ness* (Aronoff, 1976; Anshen and Aronoff, 1988), which is reflected in the graph by a marked reduction in the extent to which its distribution is skewed to the left. In other words, the distribution of *-ity* is characterized by relatively large numbers of words with high frequencies of use, words that are firmly entrenched in the mental lexicons of the language users, and by relatively few instances of low-frequency words, words that require rule-based processing.

Probability density functions not only highlight differences in productivity, but they also make it possible to visualize semantic factors. One well-known semantic factor in word frequency studies is semantic transparency (Baayen, 1992; Koehler, 1986). Semantically transparent words can be fully understood given the meaning of the affix and the meaning of the base. For instance, the meaning of the result nominalization *judgment* is a transparent function of the meaning of the verb *judge* and the semantics of this particular nominalizing suffix. By contrast, the word *department* contains the verb *depart*, but the meaning of *depart* no longer participates in the meaning of *department*. Semantically opaque formations such as *department* show a strong trend to appear among the highest frequency words in a given morphological category. In the case of *-ness*, for instance, the highest frequency formation, *business*, is no longer transparently derived from *busy*, a fact which is also reflected in its pronunciation, in which the /i/ of *busy* is dropped. Opaque formations fall outside the morphological category proper, and their morphological structure no longer plays a role in lexical processing (Marslen-Wilson, Tyler, Waksler, and Older, 1994). Such formations have become individual words, similar to monomorphemic words. Since their meaning cannot be obtained from the meanings of their parts, their use fully depends on retrieval from memory. 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 strong tendency to appear in the highest ranges of the frequency spectrum.

This is shown in the upper left panel of Figure 2, which plots three probability density functions. The solid line represents all formations with the Dutch suffix *ont-*, including both transparent and opaque words. The dashed line represents the subset of opaque formations. As expected, this essentially unimodal distribution reveals a substantial shift to the right. Almost all opaque words have a high frequency of use. The dotted line plots the probability function for the transparent words with the prefix *ont-*. This curve reveals a slight shift to the lower frequency ranges, compared to the density function of all *ont-* formations. In the next section, we will investigate the category of transparent *ont-* formations in some more detail, as it appears to be a bi-modal instead of a uni-modal distribution.

3. The Dutch Prefix Ont-

In this section we investigate why the prefix *ont-* shows up with a bi-modal distribution. Since the higher frequency range is more densely populated than expected, we are led to wonder whether we are dealing with a single morphological category. Might we instead be observing a mixture of two morphological categories? To answer this question, we need to introduce some basic facts concerning the semantics of *ont-*.

This prefix *ont-* attaches to adjectives, nouns, and verbs. A number of denominal verbs in *ont-* are illustrated in (1).

(1) a. REVERSATIVES
ontspelden 'unpin'
ontinkten 'de-ink'
ontkleuren 'decolorize'
ontwortel 'uproot'

b. NONREVERSATIVES
ontsnavelen 'de-beak'
ontvlezen 'strip the flesh off'
onthoofden 'behead'
ontluizen 'delouse'

De Vries (1975) holds the category of denominal *ont-* forms to be a productive one. We distinguish the items in (1a) from those in (1b) on the following basis. The items in (1a) are reversative in the sense that they denote the reversal of an action denoted by the stem; significantly the stem might here be interpreted as a verb that has been derived from a noun by conversion, although De Vries does not do so. We will have more to say about this below. We label the items in (1b) as nonreversative because, although they denote the removal of the object denoted by the base, they cannot be construed as the reversal of an action.

There are three kinds of deverbal formations with *ont-*. The first group is characterized by an inchoative reading, as shown in (2).

(2) INCHOATIVES
ontvlammen 'catch fire'
ontkiemen 'germinate'
ontwaken 'awake'
ontspringen 'well up from'

The second group expresses separation,

(3) SEPARATIVES
ontroven 'rob'
ontduiken 'evade'
ontglippen 'slip, to get away'
ontkomen 'escape'

and the third the reversal of the action denoted by the verb:

(4) REVERSATIVES
ontmagnetiseer 'demagnetize'
ontmythologiseer 'demythologize'
ontkoppelen 'uncouple'
ontspannen 'slacken, to relax'

According to De Vries (1975), all three categories of deverbal *ont-* formations are unproductive. Examples of de-adjectival formations in *ont-* are given in (5).

(5) *ontheilig* 'desecrate'
ontzondig 'remove sin'
ontmenselijk 'dehumanize'

This category is quite small (there are only 27 formations in our database) and unproductive.

In Lieber and Baayen (1993) we argue that the prefix *ont-* contributes a unitary conceptual frame in which the base word is incorporated. Making use of Jackendoff's (1990) theory of lexical conceptual structure (LCS), in which verb meanings are analyzed as hierarchical organizations of a number of semantic primitives such as causation (CAUSE), inchoation (INCH), movement (GO) and existence or location (BE), we propose the following LCS for *ont-*:

(6) [CAUSE ([ ], [INCH [BE ([ ],
[AT-END-OF [FROM [ON ([ ])]]])]]])].

In (6), the AT-END-OF function modifies the prepositional functions FROM and ON to express complete removal. Loosely paraphrased, (6) states that some

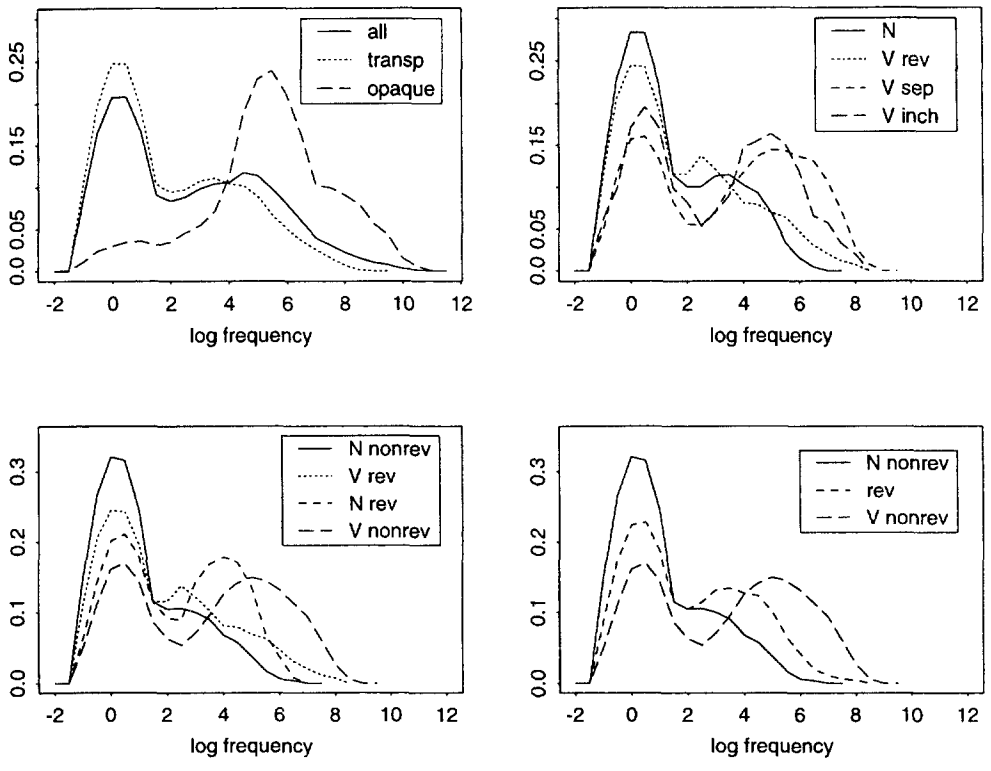

Figure 2. Probability density functions (pdf) for the Dutch prefix *ont-*. The upper left panel plots the complete distribution (*all*), together with its partitions of transparent (*transp*) and opaque (*opaque*) formations. The upper right panel plots the pdf for denominal *ont-* (*N*), and the reversative (*V rev*), separative (*V sep*), and inchoative (*V inch*) deverbal forms. The lower right panel shows the pdf for reversative (*rev*) and non-reversative (*nonrev*) deverbal (*V*) and denominal (*N*) formations. The panel at the bottom right presents the pdf of denominal non-reversatives (*N nonrev*), deverbal non-reversatives (*V nonrev*), and denominal or deverbal reversatives (*rev*).

external agent causes a particular object, property, or activity to be completely removed from some other object, property, or state. For instance, in *de gevangene ontluizen*, 'to delouse the prisoner',

(7) [CAUSE ([ ], [INCH [BE ([luizen],
[AT-END-OF [FROM [ON ([the prisoner])]]])]])],

ont- specifies the removal of lice from the prisoner. In *de achtervolgers ontlopen*, 'to outrun or escape one's pursuers',

(8) [INCH [BE ([lopen ([ ], [ ])],
[AT-END-OF [FROM [ON ([de achtervolgers])]]])]]

the activity of running away results in being out of reach of the pursuers. Note that the LCS of *ontlopen* does not contain the CAUSE function. This is a property that most inchoatives and separatives have in common. The reversatives, by contrast, are always causative.

Can the deverbal and denominal formations really be brought together in a single morphological category? In particular the more complex LCSs required for the deverbal formations suggest that perhaps deverbal verbs constitute a category of their own. To answer this question, we can investigate the probability density functions belonging to the four subcategories of verbs outlined above (for the moment we count all of the denominals as constituting a single category). This is shown in the upper right panel of Figure 2.

The two dashed functions represent the classes of inchoative and separative verbs. They reveal a pattern that suggests either a very high, or a very low frequency of use. Most of the low-frequency formations are rather strange, and appear to be either old-fashioned or stylistically quite marked. In the latter case, they were probably created by analogy to the high-frequency verbs. These high-frequency verbs, by contrast, are all well-know items of the Dutch vocabulary. Interestingly, the functions of the inchoatives and separatives are quite similar, and a Kolmogorov-Smirnov two-sample test (Siegel, 1956) confirms that they are statistically indistinguishable ($p > 0.50$). This suggests that they constitute a single class. The same conclusion can be

reached on the basis of the semantics of inchoation and separation. The inchoative verbs denote change of state, the separative verbs change of position. Both lack the sense of reversal characteristic of the deverbal reversative verbs, and tend not to contain the CAUSE function in their LCS. What we find is that similarity in meaning is reflected in similarities in frequencies of use without overt marking.

The upper left-hand panel of Figure 2 also plots the density functions of the denominal verbs (solid line) and deverbal reversative verbs (dotted line). These two functions are also quite similar, and again this visual impression is confirmed by a Kolmogorov two-sample test ($p > 0.40$). They also differ significantly from the inchoative and separative classes ($p < 0.02$ for both comparisons). The relevant question to ask here is why the denominal *ont-* forms should appear to pattern with the deverbal *ont-* forms. After all, De Vries suggests that the former are productive and the latter unproductive. We believe that the answer to this question lies in the difference we have alluded to above in the two kinds of denominal *ont-* forms. As we have pointed out there are many formations in *ont-* that are ambiguous with respect to the category of their base. For instance, we can analyze a verb such as *ontzadelen*, 'unsaddle', as a denominal formation, meaning 'remove the saddle from'. Alternatively, we might argue that *ontzadelen* is deverbal, meaning 'reverse the action of saddling', since there is a verb *zadelen*, 'to saddle'. De Vries argues that these ambiguous cases should be analyzed as denominal, precisely because he believes that all other deverbal prefixation with *ont-* is not productive. In order to tease apart the similarities in patterning, we therefore partition the class of denominal formations into two sets, unambiguous denominals which we call nonreversatives, and ambiguous denominals which we call reversatives, which we can compare to deverbal reversatives. The resulting probability density functions are shown in the lower left panel of Figure 2. The inchoative and separative deverbal formations have been merged into a single category, and are represented by a long dashed line. The unambiguously denominal formations are plotted with a solid line. The denominal and deverbal reversatives are represented by a dotted and a short dashed line respectively. Although the denominal reversatives reveal a somewhat higher density for the higher frequency ranges, the two distributions do not differ significantly ($p > 0.10$). This leaves us with three distinct classes: denominal nonreversatives, reversatives (both denominal and deverbal), and deverbal non-reversatives. These classes are shown in the bottom right panel of Figure 2. Their distributions are all significantly different. Higher frequencies of use are more probable for the deverbal non-reversatives than for the reversatives ($p < 0.01$) or the denominal non-reversatives ($p < 0.001$), and the reversatives similarly reveal somewhat higher probabilities for high frequencies of use than the denominal non-reversatives ($p < 0.025$).

The bottom right panel of Figure 2 suggests that the non-reversative denominals constitute the most productive class, that the non-reversative verbs are unproductive, and that the reversatives are semi-productive. What semantic properties of these classes give rise to this pattern of results? First consider the non-reversative denominals, essentially verbs of removal, which have the LCS (see (7))

(9) [CAUSE ([ ], [INCH [BE ([BASE NOUN],
[AT-END-OF [FROM [ON ([ ])]]])]]]).

The derivation of nonreversative verbs from nouns is relatively straightforward. The base noun is the argument of the BE function, the Theme that is moved from some position by the Agent. Reversative verbs such as *ontzadelen* would have the LCS in (10) if we treat them as deverbal, but the LCS in (11) if they were denominal.

(10) [CAUSE ([ ]$^{\alpha}$, [INCH [BE
([CAUSE ([ ]$^{\beta}$, [INCH [BE ([SADDLE],
[AT$_{d}$ ([ ]$^{\gamma}$)])]])],
[AT-END-OF [FROM [AT$_{d}$ ([γ])]]])]]).

(11) [CAUSE ([ ], [INCH [BE ([SADDLE],
[AT-END-OF [FROM [AT$_{d}$ ([ ])]]])]]]).

The LCSs given in (10) and (11) share an important property, namely, that a Theme appears as the argument of BE in a causative construction. This holds not only for verbs such as *ontzadelen*, for which a strictly denominal LCS as in (11) is a real possibility, but also for unambiguously deverbal reversative verbs such as *ontpolitiseren*, 'depoliticize'. Of the LCSs in (10) and (11), the more complex structure given in (10) has the advantage that it makes explicit that the reversal of a placement action is involved.

If we assume that (10) is the correct semantic representation for *ontzadelen*, we are also able to understand why the reversative verbs tend to be slightly less productive than the non-reversative denominal verbs. First, the LCS of the reversative verbs is more complex. A greater semantic complexity requires more complex processing in production and perception, which makes it less likely that it will be used for coining ephemeral nonce formations. Second, since reversal verbs,

by their very nature, pair up with verbs of putting into position, they are more likely to denote culturally well-established actions than non-reversative denominal verbs. The latter kind of formation is more innovative, but at the same time more marginal and incidental. This allows non-reversative *ont-* to attach to a wider range of base words. In this light, the larger numbers of non-reversative denominal nonce formations is to be expected.

Why are the deverbal non-reversatives even less productive than the reversative verbs? Both classes have complex LCSs, due to the incorporation of a verb. Therefore, the complexity of the LCS as such cannot be the determining factor. Nevertheless, there are two important differences. Many deverbal non-reversatives lack the outer CAUSE function. This holds for all inchoatives, and for a large number of separatives. More importantly, even in the absence of CAUSE , many separatives are transitive, but their direct object is not the Theme but the Source, the object or place from which separation takes place, as shown for *ontvluchten*, 'flee, escape from' in (8), repeated here as (12) for convenience.

(12) [INCH [BE ([GO ([ ], [ ])],
[AT-END-OF [FROM [ON ([ ])]]])]]

Since it normally is the Theme that appears in the object position, the separatives are exceptional. In the same vein, the pure inchoatives are also exceptional, as they are intransitives instead of transitives. The unproductiveness of the deverbal non-reversatives follows immediately from their exceptional nature.

4. Noun to Verb Conversion

The most productive class of *ont-* verbs, the denominal verbs of removal (denominal nonreversatives), have CAUSE in the outermost layer of their LCS, and incorporate their nominal base as Theme. Is there any evidence in other domains of the grammar that this semantic configuration, which we will henceforth refer to as 'theme causation' is an optimal condition for productive verb formation? In this section, we turn to noun to verb conversion in English and Dutch, and we will adduce some evidence that this is indeed the case.

Noun to verb conversion is semantically much more heterogeneous than prefixation with *ont-*. There are two main classes of conversion verbs which may be distinguished in Jackendovian terms by the presence or absence of a thematic tier in LCS. In the class of verbs with a thematic tier are verbs such as *cork, plant*, and *color* which specify movement, location or existence of a Theme, the argument of the semantic functions GO and BE . Verbs such as *boss* and *puzzle*, by contrast, denote actions that cannot be adequately described by means of GO or BE . Instead, they are pure activity verbs which we will refer to as DO verbs. In Jackendovian terms, these verbs have an action tier in their LCSs, but not a thematic tier.

English, and to a lesser extent, Dutch, makes heavy use of conversion. There are 1619 different conversion verbs in our database of English, and 688 in our database of Dutch, compared to 275 formations with *ont-*. Since conversion imposes no constraints on what kind of actions or states it can denote, it is a useful tool in communication. At the same time, it is precisely its semantic freedom which renders the interpretation of novel conversion verbs heavily context dependent. While the often highly idiosyncratic meanings of higher frequency conversion verbs are stored in memory, the interpretation of novel conversion verbs out of context is often impossible, given the range of possible meanings that might have been intended. Does this imply that context alone is relevant for the interpretation of novel compounds, as argued by Clark and Clark (1979), or are semantic factors also relevant? Without denying the primary role of the pragmatics of conversational interaction in fixing the interpretation of many conversion verbs, we will discuss some evidence that suggests that in addition two semantic factors may also be operative.

First, within the class of DO verbs, we find a subset of verbs derived from personal names and characterizations, such as *to boss* and *to Houdini*. Persons are volitional agents that are unlikely to undergo inchoation, caused movement, or to appear as the object of some action. For instance, *to boss* is very unlikely to mean 'cause one's boss to leave'. Personal names in conversion verbs specify the way in which an action is performed. Such verbs constitute a subclass of the DO verbs that we will gloss with ACT LIKE. If it is indeed the case that personal names provide a good cue to what the conversion verb should mean, 'act in a way typical for', then we may expect a somewhat reduced dependency on memory, and hence a distribution with more low frequency words than the class of other DO verbs. This expectation is born out by the upper panels of Figure 3, which plot the probability density functions of the ACT LIKE and the other DO verbs for Dutch (left) and English (right). The solid and dashed horizontal lines show the means of these

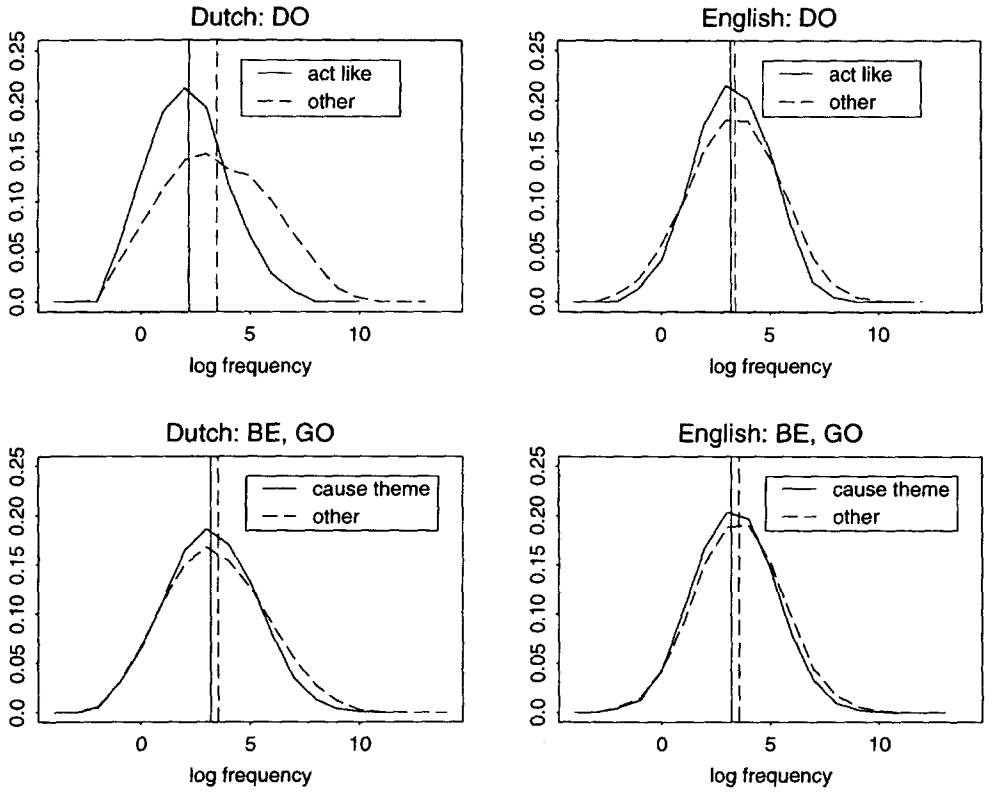

Figure 3. Probability density functions (pdf) for semantic classes of Dutch and English conversion verbs. The upper two panels show, for Dutch (left) and English (right), the pdf for two classes of DO verbs (ACT LIKE versus other kinds of DO verbs). The lower two panels show the pdf for two classes of verbs with a thematic tier (verbs of theme causation versus other GO or BE verbs). The vertical lines highlight the mean log frequencies of these four classes.

distributions. In both cases, the observed difference is statistically reliable ($p < 0.001$ for Dutch, $p < 0.05$ for English, one-tailed Kolmogorov-Smirnov tests).

Second, consider the class of thematic conversion verbs, verbs expressing location, movement, and inchoation. We have seen that among the *ont-* formations the highest degree of productivity occurs for the non-reversative denominals which incorporate the base noun as the Theme in a causative frame. The bottom two panels of Figure 3 reveal that the subset of causative verbs for which the base noun is the argument of GO or BE is characterized by a probability density function that is slightly shifted to the left, compared to the remaining verbs with a thematic tier. This suggests that incorporation of the base noun as Theme in a causative frame is a central semantic operation that, when not marked overtly as in Dutch by means of prefixes such as *ont*, still retains some functionality in noun to verb conversion.

All distributions displayed in Figure 3 are unimodal and much less skewed than the distributions we have seen thus far. Some curves even approximate the probability density function of the normal distribution. Is this a special property of conversion verbs? This is a real possibility but, unfortunately, our data do not allow this conclusion to be drawn, as the CELEX database does not list all conversion verbs and their frequencies that occur in the corpora it has surveyed. It lists all conversion verbs that appear in the machine readable dictionaries it consulted, together with the frequencies with which these conversion verbs appear in the INL corpus (for Dutch) and the Cobuild corpus (for English). Future analyses based on the complete frequency range, as we were able to carry out for *ont-*, may well reveal more formations in the lower frequency ranges than the present data suggest. A reanalysis of the *ont-* data, but now conditioning on occurrence in a dictionary in the same way as for the conversion data, reveals the same overall pattern of results: the same semantic classes ensue, and their word frequency distributions are again reliably different. However, the shape of the associated probability

density functions are, due to the absence of most hapax legomena, much more similar and unimodal. This suggests informally that the semantic effects observed for noun to verb conversion will generalize to the unconditional complete frequency distribution, but that the shapes of the distributions will be changed by the transition from the dictionary-conditioned distributions to the complete distributions.

We are still left with the question why the probability density functions shown in Figure 3 resemble the bell-shaped curve of the normal distribution. Carroll (1967) has suggested that all word frequency distributions should be normal after a logarithmic transformation. His hypothesis of the lognormality of word frequency distributions is clearly wrong for affixes such as *-ness* and *ont-*, but it appears to be correct for our – truncated – conversion distributions. Why? Possibly, it is the truncation of the lower frequency ranges itself that is crucial here. This truncation removes all ephemeral nonce formation from consideration, and leaves us with a more or less well-established stock of words, words the language user knows by heart. In general, their interpretation can proceed without calling upon morphological rules or pragmatic interactional conventions. This line of reasoning leads to the hypothesis that the known, conventionalized part of a morphological category is lognormally distributed.

Interestingly, this hypothesis is supported by the shape of the probability density function of the category of English monomorphemic nouns, the solid curve shown in Figure 4. Monomorphemic nouns constitute a strictly finite set that cannot be extended by rule. Their interpretation is fixed, and they are part of the basic lexical stock of English. The dotted line in Figure 4 plots the probability density function of a random sample with the same mean and variance as our sample of monomorphemic nouns. The two curves are highly similar, although the distribution of observed frequencies is slightly skewed to the left. This is supported by the coefficients of skewedness for these distributions: 24.597 for the random sample, 49.307 for the sample of simplex nouns. That the two distributions are not strictly identical is supported by a Kolmogorov-Smirnov two-sample test ($p = 0.0158, n = 4588$; all negative log frequencies in the random sample were collapsed with the zero log frequency before the Kolmogorov-Smirnov test was applied). Nevertheless, it is clear that the distribution of simplex nouns is approaching lognormality.

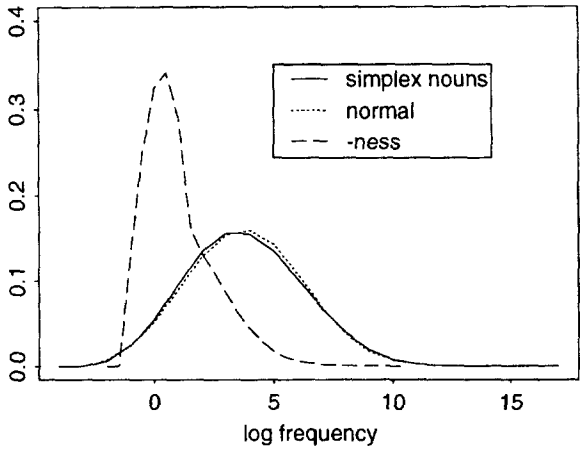

Figure 4. Probability density function (pdf) for English nouns in *-ness* (dashed line) and English monomorphemic nouns (solid line). The dotted line presents the pdf of a normally distributed random variable with the same mean and standard deviation as the monomorphemic nouns.

This result leads us to hypothesize that the frequency distributions of the subclasses of *ont-* formations are mixed distributions, consisting of an asymptotically lognormal distribution of 'entrenched' vocabulary in the higher frequency ranges combined with a leftwards skewed distribution at the lowest frequency range. The appearance of a higher than expected density of types in the log frequency range $[3, 5]$ in the distribution of words in *-ness*, shown in Figure 1 and repeated in Figure 4, might also indicate that we are observing a composite combination of innovative use of productive morphology and conventional use of well-established lexical items.

Interestingly, our hypothesis concerning the category of *ont-* verbs as a mixed distribution also explains why our attempts to fit its word frequency distribution using the statistical models developed by Sichel (1986), Orlov (1983), and Carroll (1967) were completely unsuccessful. All these models are based on the assumption that uni-modal densities are involved. Given the bi-modal probability density functions we have observed for *ont-*, these failures are only to be expected.

5. Conclusions

We have seen that differences in lexical conceptual structure are sometimes mirrored in frequency distributions. Why should frequency and semantics be correlated? For opaque complex words, the answer hinges

on the impossibility of rule-based comprehension or production. Therefore, opaque words fully depend on lexical storage. Storage in memory, in turn, is guaranteed only for words that are used frequently enough. Hence, opaque formations tend to have high frequencies of use.

Conversely, as complex words approach the transparent core meaning of the morphological category, speakers are more sure how the corresponding word formation rule should be applied (Aronoff, 1976). This allows them to apply the rule to a much broader range of base words than would be possible under conditions of uncertainty. Consequently, larger numbers of rare words can appear in the frequency distributions of the more transparent classes.

Perhaps the most important finding in this study is that the probability density function of a word frequency distribution can be bi-modal. We have taken this to indicate that word frequency distributions are mixtures of two distributions. One distribution belongs to the 'rote' class. Its members are the known conventionalized words of the morphological category. The other distribution belongs to the 'rule' class, and contains the rule-based productive lexical innovations. Highly productive affixes such as *-ness* give rise to extreme distributions which display only minimal evidence for the 'rote' class. Conversely, the category of monomorphemic nouns exemplifies the other extreme, with no evidence at all for the 'rule class'. Our dictionary-conditioned classes of conversion verbs fall in line with the monomorphemic nouns, suggesting that they too are heavily dependent on 'rote'. The categories of verbs with *ont-* instantiate intermediate positions, with on the one hand rule-governed neologisms and on the other hand rote-governed use of well-established words. Our analysis of monomorphemic nouns suggests that 'rote' distributions are asympotically lognormal. The 'rule' classes are likely to be Poisson distributed, but at present this is a conjecture only. Future research will have to show whether or not it is correct.

References

Anshen, F. and M. Aronoff. "Producing Morphologically Complex Words." *Linguistics*, 26 (1988), 641–655.

Aronoff, M. *Word Formation in Generative Grammar*. Cambridge, Mass.: MIT Press, 1976.

Baayen, R. H. "Quantitative Aspects of Morphological Productivity." In *Yearbook of Morphology 1991*. Eds. G. E. Booij and J. Van Marle. Dordrecht: Kluwer Academic Publishers, 1992, pp. 109–149.

Baayen, R. H. "Statistical Models for Word Frequency Distributions: A Linguistic Evaluation." *Computers and the Humanities*, 26 (1993), 347–363.

Baayen, R. H. "Derivational Productivity and Text Typology." *Journal of Quantitative Linguistics*, 1 (1994), 16–34.

Baayen, R. H. and R. Lieber. "Productivity and English Derivation: A Corpus-Based Study." *Linguistics*, 29 (1991), 801–843.

Baayen, R. H., R. Piepenbrock, and H. Van Rijn. *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania, 1993.

Brunet, E. *Le Vocabulaire de Jean Giraudoux*. Genève: Slatkine, 1978.

Carroll, J. B. "On Sampling from a Lognormal Model of Word Frequency Distribution." In *Computational Analysis of Present-Day American English*. Eds. H. Kučera and W. N. Francis. Providence: Brown University Press, 1967, pp. 406–424.

Carroll, J. B. "An Alternative to Juilland's Usage Coefficient for Lexical Frequencies, and a Proposal for a Standard Frequency Index (SFI)." *Computer Studies in the Humanities and Verbal Behavior*, 3 (1970), 61–65.

Chitashvili, R. J. and R. H. Baayen. "Word Frequency Distributions." In *Quantitative Text Analysis*. Eds. G. Altmann and L. Hřebíček. Trier: Wissenschaftlicher Verlag Trier, 1993, pp. 54–135.

Clark, H. H. and E. V. Clark. "When Nouns Surface as Verbs." *Language*, 55 (1979), 567–811.

De Vries, J. W. *Lexicale Morfologie van het Werkwoord in Modern Nederlands*. Leiden: Universitaire Pers, 1975.

Haerdle, W. *Smoothing Techniques With Implementation in S*. Berlin: Springer, 1991.

Harwood, F. W. and A. M. Wright. "Statistical Study of English Word Formation." *Language*, 32 (1956), 260–273.

Holmes, D. I. "Authorship Attribution." *Computers and the Humanities*, 28 (1994), 87–106.

Hubert, P. and D. Labbe. "Un Modèle de Partition du Vocabulaire." In *Etudes sur la Richesse et les Structures Lexicales*. Eds. D. Labbe, P. Thoiron and D. Serant. Paris: Slatkine-Champion, 1988, pp. 93–114.

Jackendoff, R. *Semantic Structures*. Cambridge, Mass.: The MIT Press, 1990.

Koehler, R. *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer, 1986.

Lieber, R. and R. H. Baayen. "Verbal Prefixes in Dutch: A Study in Lexical Conceptual Structure." In *Yearbook of Morphology 1993*. Eds. G. E. Booij and J. Van Marle. Dordrecht: Kluwer Academic Publishers, 1993, pp. 51–78.

Marslen-Wilson, W., L. K. Tyler, R. Waksler and L. Older. "Morphology and Meaning in the English Mental Lexicon." *Psychological Review*, 101 (1994), 3–33.

Martin, W. "On the Construction of a Basic Vocabulary." In *Proceedings of the 6th International Conference on Computers and the Humanities*. Eds. S. Burton and D. Short. Comp. Science Press, 1983, pp. 410–414.

Martin, W. "Lexical Frequency." In *Distributions Spatiales et Temporelles, Constellations des Manuscrits: Etudes de Variation Linguistique Offertes à Anthonij Dees à l'Occasion de son 60me Anniversaire*. Ed. K. van Reenen-Stein. Amsterdam: Benjamins, 1988, pp. 139–152.

Muller, C. *Principes et Méthodes de Statistique Lexicale*. Paris: Hachette, 1977.

Orlov, J. K. "Dynamik der Häufigkeitsstrukturen." In *Studies on Zipf's Law*. Eds. H. Guiter and M. V. Arapov. Bochum: Brockmeyer, 1983, pp. 116–153.

Renouf, A. "Corpus Development." In *Looking Up: An Account of the Cobuild Project in Lexical Computing*. Ed. J. M. Sinclair. London: Collins, 1987, pp. 1–40.

Rubenstein, H. and I. Pollack. "Word Predictability and Intelligibility." *Journal of Verbal Learning and Verbal Behavior*, 2 (1963), 147–158.

Scarborough, D. L., C. Cortese and H. S. Scarborough. "Frequency and Repetition Effects in Lexical Memory." *Journal of Experimental Psychology: Human Perception and Performance*, 3 (1977), 1–17.

Shapiro, B. J. "The Subjective Estimation of Word Frequency." *Journal of Verbal Learning and Verbal Behavior*, 8 (1969), 248–251.

Sichel, H. S. "Word Frequency Distribution and Type-Token Characteristics." *Mathematical Scientist*, 11 (1986), 45–72.

Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw Hill, 1956.