Parsing and Productivity[1]

Jennifer Hay and Harald Baayen

University of Canterbury,
Interfaculty Research Unit for Language and Speech, University of Nijmegen
September 20, 2001

1 Introduction

It has often been argued that the (type or token) frequency of an affix in the lexicon cannot be used to predict the degree to which that affix is productive. Affix type frequency refers to the number of different words which contain an affix, token frequency refers to the summed lexical frequency of those words. The observation that neither of these counts relates straightforwardly to productivity, raises difficult questions about the source of different degrees of productivity, making the nature of morphological productivity one of the "central mysteries of word-formation" (Aronoff 1976:35). If productivity does not arise as a function of frequency, then where does it come from?

This paper argues that frequency and productivity are, in fact, intimately linked. Type and token frequency in the lexicon are not good predictors of productivity. But frequency counts of *decomposed* forms in the lexicon *can* predict the degree to which an affix is likely to be productive. The problem with doing a straightforward frequency count of forms containing an affix, is that not all affixed forms contain the affix to the same degree. Some affixed words are highly affixed, and are highly decomposable (e.g. *tasteless*). Other affixed words appear more opaque, and tend to be characterised by whole word access, rather than parsing (e.g. *listless*). We argue that the former set facilitate productivity much more strongly than the latter set.

Decomposed forms in the lexicon arise from parsing in perception. By coming to a clear understanding of the types of factors which tend to lead to parsing in perception, then, we can predict the degree to which an affix is represented by decomposed forms in the lexicon, and so (we argue), the degree to which it is likely to exhibit productivity. Thus, we argue that there is a strong relationship between parsing in perception,

[1]We are indebted to Andrew Carstairs-McCarthy, Wolfgang Dressler, Anke Luedeling, Janet Pierrehumbert, Ingo Plag and Robin Schafer, whose comments have greatly improved the quality and coherence of this paper. Remaining errors or incoherencies are the sole responsibilities of the authors.

and morphological productivity. Increased rates of parsing lead straightforwardly to increased productivity.

One model which has posited a link between parsing and productivity is Baayen's (1993) dual processing race model of morphological access. In this model, there are two routes for processing – the direct route (in which a word is accessed whole), and the parsing route (in which it is accessed via its parts). Whether or not a specific morphologically complex form is accessed via its parts is determined by the token frequency of that form – if it is above a certain threshold of frequency, then the direct route will win, and the word will be accessed whole. If it is below that same threshold of frequency, the parsing route will win, and the word will be accessed via its parts. In order for an affix to remain productive, words containing that affix must be parsed sufficiently often that the resting activation level of that affix remains high. In this way, the model implicitly ties productivity to decomposition in perception. High rates of decomposition should ensure the productivity of an affix. Conversely, an affix which is represented by many words which are characterized by the direct route is unlikely to be productive.

In an attempt to explicitly model such a link between productivity and parsing, and to provide a formal explanatory account of productivity, Baayen proposes a psychologically motivated measure $\mathcal{A}$, which is intended to approximate the resting activation level of a given affix. This measure is the number of words containing the affix which occur below a certain frequency threshold, each weighted by their frequency of occurrence. This measure is distinct from other proposed productivity measures, in that it attempts not only to measure degree of productivity, but also to explain it. The general idea behind the approach is that low frequency types require parsing, and so protect the activation levels of the affixes against decay. Such an approach is attractive, because it provides the potential for a psychologically plausible account of the emergence of productivity. Productivity emerges as a result of parsing.

The power of the statistic $\mathcal{A}$ is weakened, however, by its reliance on the assumption that low frequency forms require parsing, and high frequency forms do not. Recent results have demonstrated that the absolute frequency of a derived form is not straightforwardly related to parsing. More relevant is the nature of the frequency *relations* between the derived form and the base.

Hay (in press) distinguishes between derived forms which are more frequent than the bases they contain (e.g. *illegible* is more frequent than *legible*), and derived forms which are less frequent than their bases (e.g. *illiberal* is less frequent than *liberal*). Derived forms which are more frequent than their bases (e.g. *illegible*) are more prone to whole word access, regardless of the absolute frequency of the derived form (Hay 2000, in press). Thus,

low frequency forms may be accessed directly if their base is of even lower frequency. And high frequency forms may be parsed if the base is higher frequency still. Relative frequency matters.

In order to properly explore the relationship between productivity and parsing, then, we need to find a more accurate heuristic for distinguishing between those words which are prone to parsing, and those which are likely to be accessed whole. And this heuristic must involve relative frequency, rather than absolute frequency.

Hay's (in press) division between derived forms which are more frequent than the bases they contain, and derived forms which are less frequent was a first approximation, and is almost certainly overly simplistic. While the relative frequency of the base and the derived form is clearly important, what is less clear is the exact location of the relevant threshold for parsing. Exactly how frequent does the base form need to be, relative to the derived form, in order to facilitate parsing?

This paper has two primary functions. First, it sets out to refine the notion of relative frequency as a potential heuristic for assessing parsability. What kinds of frequency relations between affixed words and their bases tend to facilitate parsing?

Second, it uses this enhanced understanding of parsing to conduct a systematic investigation into the relationship between parsing and productivity. To what extent can productivity in production be linked to parsing in perception? We demonstrate that there is a strong link. The more often words containing a given affix are parsed during perception, the more productive that affix will be.

We begin, in section 2 with an investigation into affix-specific characteristics in the lexicon. We show that the relationship between the token frequency of base forms, and the token frequency of their related derived forms differs considerably across different affixes. We argue that this variation is linguistically significant, and can provide insight into degrees of decomposition of words containing different affixes in the lexicon.

In section 3 we motivate the location of an approximate "parsing line" (a threshold in the space relating base frequency, and the frequency of corresponding derived forms), above which an affixed form is likely to be decomposed. Having motivated such a parsing line, we are able, for any given affix, to estimate which words containing that affix are likely to be highly decomposable (those falling well above the line), and those which are likely to be non-decomposable and characterised by whole word access (those falling well below the line).

Having come to a more sophisticated understanding of the role of base form frequency and derived form frequency in morphological decomposition, we are then able, in section 4, to turn our attention to the relationship between parsing and morphological productiv-

ity. In this section we demonstrate that the distributional properties of words with respect to the parsing line are statistically related to an affix's productivity.

We conclude, in section 5, by arguing that the results, taken as a whole, provide strong evidence of a robust link between productivity and parsing.

2 Affix-specific characteristics in the lexicon

Hay (in press) shows that for both prefixes and suffixes, the token frequency of the base form (base frequency) and the token frequency of the derived form (derived frequency) are significantly correlated. This is an intuitively reasonable result. Higher frequency base words spawn higher frequency derived words. Extremely low frequency (relatively useless) words, are less likely to spawn derived words, and are particularly unlikely to have high frequency derivatives. Hay's discussion collapses across a set of English affixes – that is, she doesn't investigate whether the observed correlation between base and derived frequency holds for all affixes, or whether it holds to the same *degree* for all affixes. This section describes an investigation of 80 affixes of English, in which we find that the nature of this correlation in fact varies markedly across different affixes. We argue that this variation is linguistically significant. Section 2.1 describes the corpus on which these calculations are based.

2.1 The Data Set

The calculations in this paper are based on a set of words extracted from the CELEX Lexical Database (Baayen, Piepenbrock, and Gullikers, 1995), which is based on an early version of the Cobuild corpus (Renouf, 1987) that contained some 18 million words. The English database in CELEX provides the morphological segmentation for a great many complex words: all the words in the LDOCE machine-readable dictionary, as well as all words in the Cobuild corpus down to a frequency threshold of 15 occurrences per 18 million. We will refer to this list as the segmentation list. It also provides a separate, unanalyzed list of all character strings occurring in the Cobuild corpus, together with their frequency of occurrence in the corpus. We will refer to this list as the string list.

We began with extracting all prefixes and suffixes that appear in the parses in the segmentation list, and vetted each affix for its synchronic plausibility. All bimorphemic words which contained the resultant affixes, and their corresponding monomorphemic base word were then extracted from the segmentation list together with their frequency of occurrence. Any affix which was not represented by at least ten such words was then

discarded. This process resulted in a list of 54 suffixes and 26 prefixes — 80 affixes total.

We chose to work only with affixed words with monomorphemic bases, as they present the simplest possible case. We leave the investigation of complex words with multimorphemic bases to future work. Such complex words may well behave differently, and their investigation is many times more complicated because it requires grappling with the degree of decomposability of the base words.[2]

Because a crucial part of our investigation was to involve the measurement of the productivity of the affixes we were anxious that our materials contained representative tokens and frequency counts. It was especially important that this was true of the lower frequency range – the part of the word frequency distribution which dominates the calculation of $\mathcal{P}$ (the category conditioned degree of productivity – see Baayen 1989, 1992). The segmentation list in CELEX is unsatisfactory in this respect, because it misses an important set of low frequency words. That is, it omits any complex word which appears in Cobuild with a frequency below 15 per 18 million, and which is not listed in the LDOCE dictionary. A consequence of this is that, especially for the more productive affixes, we are missing at least half of the word types that actually occur in the Cobuild corpus. Word frequency distributions are characterised by large numbers of words with very low frequencies of occurrence (Baayen 2001). Thus, for many affixes (and particularly for productive affixes), a very large proportion of the words containing that affix are low frequency.

In order to minimise the extent of this problem, we worked with the string list provided by CELEX, which includes all of the word-forms present in the Cobuild corpus, including a great many misspelled words, hyphenated forms, numbers, and combinations of numbers and letter sequences. We attempted to automatically extract all affixed words from this list which did not appear in the segmentation list.

For each of the prefixes in our set, each word in the file was tested to see whether it was a candidate prefixed form. For example, for the prefix *pre-*, we identified all words which began with this string, and stripped the prefix from those words. The remaining string was looked up in the segmentation list. If the string is listed there as a monomorphemic word, then the entire word is included as part of our set of words prefixed with *pre-*. Automatically conducting such a procedure creates the risk of spurious parses being

[2]E.g. understanding the behaviour of *carelessness* may require coming to an understanding of the role of how the token frequency of *carelessness* relates to the frequency of *careless*, how the frequency of *careless* relates to the frequency of *care*, whether the relationship between the frequency of *carelessness* and *care* is relevant, and whether the relationship between each of these relationships plays any role. As understanding the role of the frequency relationship of *care* to *careless* is complex enough, we begin by restricting ourselves to monomorphemic bases, and leave these related questions for future work.

accepted as candidate prefixed forms (e.g. if *aid* were listed in the string list, it might qualify as a potential affixed word, prefixed with *a-*, because *id* is in the segmentation list.) Spurious parses were therefore limited by the stipulation that the base word must be at least three characters long. Of course, some spurious parses still resulted from this process, but they formed a distinct minority of the total number of forms.

A similar process was conducted for the suffixes. For suffixes beginning in vowels, the truncated "base" string was looked up in the segmentation list, and also the same string with "e" added to enable identification of words such as, e.g., *writer*. In some cases, part of speech was also specified. For instance, we specified that the base of "er" must not be an adjective, in order to avoid including comparatives like *bigger*.

Clearly an algorithm would have to be very complicated in order to account for all possible allomorphy, and we did not attempt to do this. Rather we found that with the very simple heuristics outlined above, we were able to substantially increase the data-sets for the affixes in which we were interested, and so greatly increase our confidence in the range of forms represented in the lower frequency ranges. Thus, the statistics reported throughout this paper are based on the affixed forms with monomorphemic bases as available in the segmentation list, supplemented with the forms extracted from the string list using the process outlined above. In cases in which CELEX includes multiple entries for the same word, the frequency counts of these entries were summed together.

2.2 Correlating base and derived frequency

Hay (2000, in press) reports an overall significant correlation between the token frequency of base forms and their corresponding derived forms, for both prefixes and suffixes. Does such a correlation hold to the same degree (or at all) for individual affixes?

For each of the 80 affixes in our dataset, we performed a robust regression of base frequency on derived frequency using least trimmed squares regression (Rousseeuw and Leroy 1987). This is a form of regression which is particularly robust to outliers – i.e. it attempts to find the best fit in the data (here – the nature of the relationship between base frequency and derived frequency), without being unduly influenced by individual points which do not conform to the pattern of the majority of the data. We chose to use robust regression so we could be confident of the returned results, without examining each of the 80 resultant plots in detail to check for the undue influence of outlier points.

In this regression analyses we used log frequency, rather than raw frequency. We chose to use log frequency for two reasons. First, there is evidence that humans process frequency information in a logarithmic manner – with differences amongst lower

frequencies appearing more salient than equivalent differences amongst higher frequencies (e.g. the difference between 10 and 20 is more important / salient than the difference between 1010 and 1020). Second, by taking the log, we can make the distribution of frequency counts more closely approximate the assumptions required for linear regression techniques such as the one used here (Baayen 2001).

When we regress base frequency on derived frequency, we find that affixes vary considerably in the nature of this relationship. Figure 1 shows example plots for four different affixes, using the same ranges for the two axes for all panels. For three of the four affixes (*-ness, -ism*, and *-ment*), a positive and significant correlation holds between derived frequency and base frequency. For the fourth, (*-ry*, as in *cabinetry, forestry* etc.), there is no significant correlation. The dotted lines represent $x = y$, the position on the graph where base frequency and derived frequency are equivalent. The solid lines on the graphs represent the least trimmed squares regression lines.

A brief inspection of these four graphs reveals that affixes can vary in at least three ways. They can vary in the degree of correlation between base and derived frequency, in the steepness of the resultant slope, and in the intercept of the line (the location where the line crosses the vertical axis). Thus, Hay's calculation demonstrating a significant correlation between base and derived frequency is likely to have included a fair amount of inter-affix variability. What might we be able to learn about an affix based on such factors?

From a production perspective, a significant correlation between base frequency and derived frequency can be seen as a sign of productivity and of a high degree of semantic regularity. If an affix is truly productive and regular, then we may expect that the usefulness of a derived form should be directly predictable from the usefulness of the base word. Put differently — more frequent base words will be more easily available as bases of affixation, and so will create more frequent derived words. The less regular and predictable the relationship is between derived words containing a particular affix, and the base words they contain, the less predictable should be the relationship between the frequency of the derived word and the frequency of the base. Thus, we interpret the result in Figure 1 that *-ness, -ism* and *-ment* show a significant correlation and *-ry* does not, as evidence that the former three affixes are more synchronically productive and / or semantically regular than *-ry*. The affixes investigated are listed in the appendix, together with relevant statistics. The significance level of the correlation between base and derived frequency for each affix is given in the column labelled *prob*.

Interestingly, of the 80 affixes we investigated, 50% of suffixes (27 / 54) and 31% of prefixes (8 / 26) show a significant correlation between base and derived frequency. This pro-

vides evidence that suffixes tend to be more decomposable and lead to more semantically transparent forms than prefixed forms. This difference is likely to relate to the temporal nature of speech perception, which leads bases to be encountered before affixes for suffixed words, but not prefixed words. This is likely to facilitate decomposition for suffixes, but whole word access for prefixed words (see, e.g. Cutler, Hawkins and Gilligan 1985; Segui and Zubizarreta 1985; Hay 2000).

From a perception perspective, the distribution of points in graphs like those shown in Figure 1 is vital. Hay (2000, in press) has argued that derived forms which are more frequent than the bases they contain tend to be accessed whole, whereas derived forms which are less frequent than their bases are more likely to be parsed. This result contradicts longstanding ideas regarding the primacy of the absolute frequency of the derived form. Contrary to claims regarding the non-decomposability of high-frequency forms, Hay argues that derived forms which are less frequent than their bases are prone to decomposition *regardless of absolute frequency*. For any given affix, the more words for which the base word is more frequent than the derived form, the more words will be decomposed during access.

One way to assess the proportion of types for any given affix which are prone to whole word access, then, would be to note the proportion of types falling below the $x = y$ line, i.e., the proportion of types for which the derived form is more frequent than the base. The $x = y$ line appears as a dashed line in the graphs in Figure 1. Hay provides several types of evidence that forms falling below such a line have different properties than those that fall above it – namely, properties associated with non-decomposition. As an indication of how various words are positioned relative to the x=y line, here are some examples of words represented on the graph for the affix *-ment*. Words falling well above the line for *-ment* include *arrestment* and *dazzlement*. Words falling well below the line include *government* and *pavement*. *Argument* and *assessment* fall approximately on the line.

A glance at Figure 1 should make clear that different affixes have different proportions of words which fall below this line. For example, a greater proportion of words containing *-ment* are more frequent than their bases (17.3%) than words containing *-ness* (0.6%), indicating that a much greater proportion of *-ness* forms are regularly parsed and highly decomposable than *-ment* forms.

The slope and the intercept of the regression between base and derived frequency therefore become extremely important from a perception point of view. Consider *-ness* and *-ism*, for example. Both have roughly the same intercept, but *-ness* has a much steeper slope. The consequence of this is that fewer points fall below the $x = y$ line for *-ness* than for *-ism*. Similarly, *-ness* and *-ment* have roughly the same slope, but the intercept for *-ness*

is higher. A consequence of this is that fewer points fall below the $x = y$ line for *-ness* than *-ment*. In sum, a higher intercept and a steeper slope are both likely to contribute to higher rates of decomposition during speech perception.

We will return to these points in detail later in the paper, but first, we explore the nature of the $x = y$ line in some detail. The use of the line $x = y$ is in fact fairly arbitrary, an extremely approximate "parsing line" chosen in the absence of a principled manner to determine the true location of such a line. Hay provides evidence that the relative frequency of the derived form and the base affects the decomposability of an affixed word. What remains unclear is the relevant threshold for this ratio. In the following section we locate its approximate vicinity using empirical means.

3 Locating the Parsing Line

In order to ascertain the location of the parsing line defining the balance of storage and computation for derived words given the frequency of the word itself (its derived frequency) and the frequency of its base (its base frequency), we have made use of Matcheck (Baayen, Schreuder, & Sproat, 2000; Baayen & Schreuder, 2000), a psycholinguistic model for morphological parsing. In what follows, we first present a brief outline of how Matcheck works. We continue with a validation study of Matcheck using experimental data on the Dutch suffix *-heid* (*-ness* in English). Finally, we discuss how the approximate location of the "parsing" line can be ascertained.

Matcheck is a model for morphological segmentation that implements the idea that words and affixes in the mental lexicon compete with each other for recognition upon presentation of a given target word. How strong a competitor is depends on its frequency and its similarity to the target word. The competition process in Matcheck differs from the competition process in a model such as ShortList (Norris, 1994) in that, instead of yielding a single final representation, different segmentations of the input become available over time. In fact, Matcheck can be seen as a tool for ranking segmentations (including the segmentation consisting of just the derived form itself) according to their psychological likelihood of being perceived. Baayen & Schreuder (2000) show that, at least for Dutch, Matcheck is quite successful at ranking correct segmentations before incorrect segmentations. For instance, for the 200 randomly selected complex words of length 5-12 reported in their study, 194 emerge with a correct segmentation at the very first timestep in the model that a segmentation becomes available. Less than half of these parses (94 out of 200) are due to the derived form being recognized before the constituents. It is important to realize that Matcheck does not make use of subcategorisation information of any kind.

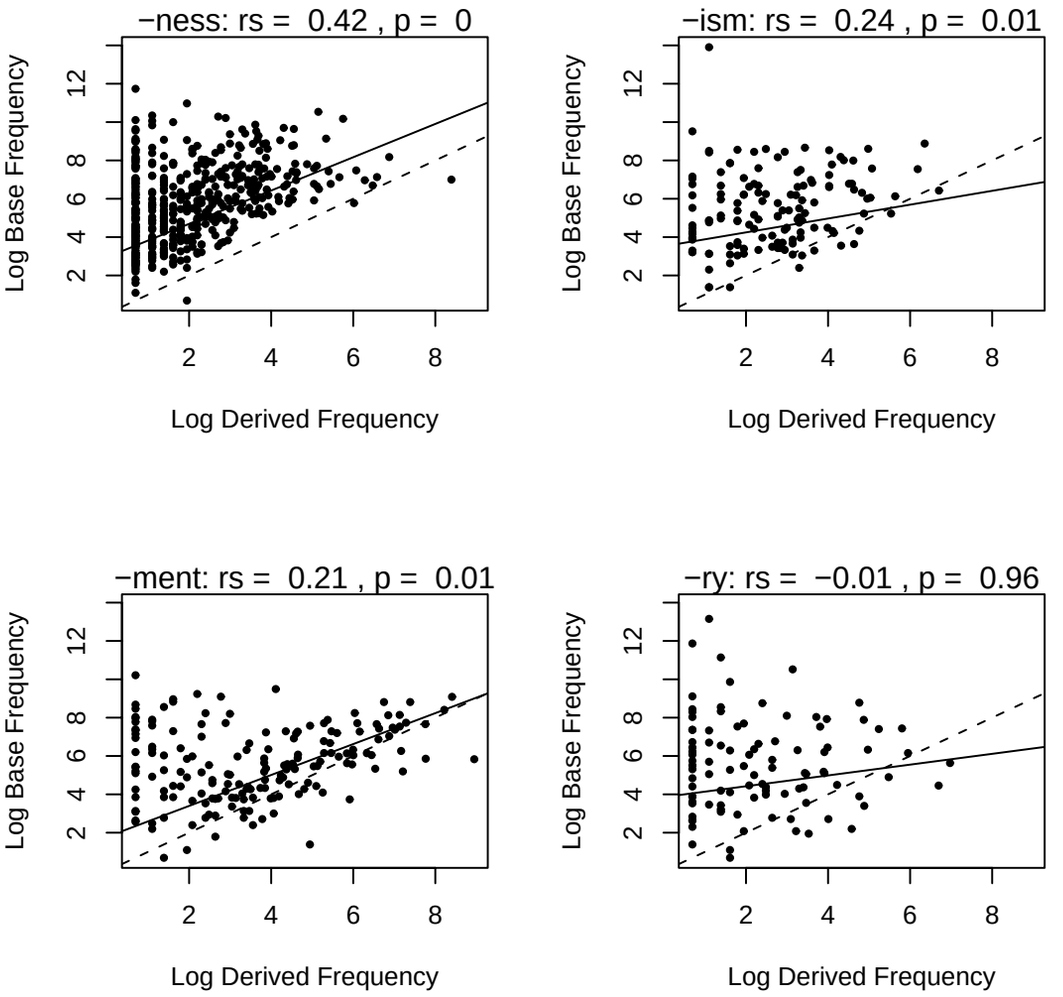

Figure 1: The relation between log derived frequency and log base frequency for four affixes. Solid lines represent least trimmed squares regression lines (Rousseeuw and Leroy 1987). Dashed lines represent the x=y line. rs=non-parametric correlation (Spearman's rho)

It is a simple purely bottom-up model that can make use of only word frequency and form similarity. What makes Matcheck attractive for the purposes of the present study is that it assigns to any complex word two moments in time: t_D, the moment in model time that the derived form has become available, and t_P, the moment in time that the first segmentation spanning the target word has become available. Thanks to these two times, we can gauge the relative importance of the direct route and the parsing route for any given target word. This property makes Matcheck a promising tool for estimating the approximate position of the parsing line.

Section 3.2 describes how we used Matcheck to estimate the location of the parsing line. But first, in section 3.1 we describe a short validation study of Matcheck, which provides evidence in favour of the appropriateness of using Matcheck for this task.

3.1 Validation of Matcheck

Just how accurate is Matcheck at estimating of the points in time at which the derived form and a correct parse become available? In order to provide some empirical validation for Matcheck's timepoint estimates, we compared its predictions with the response latencies obtained for roughly a hundred formations in the Dutch suffix *-heid* (equivalent to *-ness*, in English). These response latencies come from a visual lexical decision task, as reported in Bertram, Baayen, & Schreuder (1999) and Baayen, Schreuder, Bertram, & Tweedie (1999). We used exactly the same Matcheck parameter settings as those used in the corpus-based case study of Baayen & Schreuder (2000).

We find that for forms affixed in *-heid*, the "derived form time" t_D and the "parse time" t_P returned by Matcheck are both highly correlated with subjects' reaction times (RT) (derived form time–RT: 0.73; parse time–RT: 0.48). The derived form time and the parse time also emerge as highly correlated with log derived frequency and log base frequency. As such, the correlations with the response latencies of the logarithmically transformed frequency measures (which correlate much better with response latencies than raw frequency counts) are virtually identical to the correlations of t_D and t_P with the response latencies (log derived frequency–RT: -0.72; log base frequency–RT: -0.48). One may therefore view Matcheck as a psycholinguist's black box accomplishing a log transform such that the correct parsings have a high probability of being the first to become available in model time.

The question that we now have to address is whether Matcheck is more than an algebraic alternative to taking the logarithm of the frequency measures.

As a first attempt to answer this question, we can consider to what extent the timestep

at which the first segmentation becomes available, $\min(t_D, t_P)$, predicts the actually observed response times. Taking the minimum in this way amounts to assuming a morphological race model in which the first access route to win the race determines the response time.

We also include the morphological family size of the base word in the model (the number of derived words and compounds in which that base adjective occurs as a constituent), because there is evidence that it strongly influences response times (Schreuder & Baayen, 1997; Bertram, Baayen, & Schreuder, 1999; De Jong, Schreuder, & Baayen, 2000). So how well does the combination of $\min(t_D, t_P)$ (the timestep of the first analysis returned by Matcheck), and morphological family size predict actual response times to words suffixed in *-heid*? A regression analysis suggests that both $\min(t_D, t_P)$ and the morphological family size (henceforth abbreviated as Vf) are significant predictors of the response latencies, and together account for 37% of the variance.[3]

We could explain more of the variance in the response latencies with a model that integrates the two sources of information, the parse information and the derived form information, *without* taking the minimum. In this way, we can account for 63% of the variance.[4]

However, upon closer examination, it turns out that such a model is still not quite correct. Let us denote the set of words for which $t_D < t_P$ by F: This is the set of words for which Matcheck predicts that the direct route is the first to deliver a complete spanning of the input, namely, the derived form. Let us likewise denote the set of words for which the parsing route wins by P. When we inspect the role of t_D, t_P, and Vf for these two subsets separately, we find that they have different predictive properties. In the case of subset F, only t_D (or, equivalently, log derived frequency) emerges as significant.[5] Turning to set P, we find that t_D and Vf, but not t_P are significant.[6] The same pattern of results is obtained when subsets of the F and P sets are selected such that the correlations between family size and derived frequency are removed. Expressed in terms of derived frequency and

[3] $F(1, 83) = 30.7, p < 0.0001$ for $\min(t_D, t_P)$, $F(1, 83) = 17.1, p < 0.0001$ for Vf. The regression equation of this model is:
$RT = 603.6 + 3.8 * \min(t_D, t_P) - 24.9 * \log \text{Vf}$.
Note that this equation specifies that reaction times (RT) become longer if the moment in time at which the first segmentation becomes available is later, and that a larger family size (Vf) leads to shorter RTs.

[4] The regression model:
$RT = 380.6 + 10.4 * t_D + 2.3 * t_P - 14.1 * \log \text{Vf}$,
accounts for 63% of the variance ($F(1, 82) = 120.0, p < 0.0001$ for t_D, $F(1, 82) = 13.3, p < 0.001$ for t_P, $F(1, 82) = 8.6, p < 0.005$ for Vf). Collinearity due to a high correlation of family size and base frequency (r = 0.71) may inflate the separate roles of these two factors, however.

[5] $F(1, 23) = 53.4, p < 0.0001$ for t_D; $F < 1$ for t_P as well as for Vf.

[6] $F(1, 55) = 79.2, p < 0.0001$ for t_D; $F(1, 55) = 7.7, p < 0.01$ for Vf; $F < 1$ for t_P.

family size, we now have the following model:

$$\begin{aligned} \text{F: RT} &= 288 + 16.8 * \log(\text{derivedfreq}) \\ \text{P: RT} &= 389 + 11.6 * \log(\text{derivedfreq}) - 14.1 * \log(\text{Vf}). \end{aligned} \quad (1)$$

The response latencies predicted by (1) are plotted against the observed response latencies in the upper left panel of Figure 2. Words of the P set (for which the parsing route was first to complete) are represented by solid points, and words of the F set (for which the derived form was first to complete) by open points. The correlation of the predicted and the observed response latencies is 0.80, we account for 64% of the variance in the data.

Two things are surprising about this model. The first is the absence of an effect for base frequency (or, equivalently, parse time t_P), for the P set (the words which Matcheck predicts that the parsing route is the first to complete). And the second is the finding that the effect of family size is restricted to the P set.

In order to understand the apparent lack of family size effect amongst the forms for which the direct route is first to complete, it is important to realize that Matcheck simulates a form-based component of visual lexical processing, and that the family size effect is a semantic effect that can be understood in terms of activation spreading to semantically related words along lines of morphological similarity.

Consider what happens when the direct route is the first to provide a complete spanning of the target word, say, *snel-heid*, "quickness", i.e., "speed". Once the derived form has become available, the corresponding meaning is activated, apparently without activation spreading into the morphological family of its base, "snel". In other words, family members such as *ver-snel-en* ("to increase speed") and *snel-weg* ("freeway") are not co-activated when *snel-heid* has been recognized by means of the direct route. A similar observation has been made by De Jong, Feldman, Schreuder, Pastizzo, and Baayen (2001), who report the absence of base frequency and family size effects combined with the presence of a derived frequency effect and a positional constituent token frequency effect for Dutch and English compounds. Considered jointly, these results lead to the hypothesis that the semantic activation for words recognized primarily by means of the direct route is restricted predominantly to the meaning of the target word itself.

Now consider the case in which the parsing route wins the race. The present experimental data on *-heid* suggest that in this case activation spreads into the morphological family. This makes sense, as initially the comprehension system knows only that it is dealing with a stem that has to be combined with some affix. By allowing the morphological family members to become co-activated, all and only the possibly relevant candidates

are made more available for the processes which combine the stem with the derivational suffix and which compute the meaning of their combination.

In fact, since the derived form is one of the family members of the base, it will be activated more quickly when the base has a large morphological family. This is because it is embedded in a larger network of morphologically related words, and the morphologically related words spread activation to each other. This may explain why log derived frequency remains such a strong predictor of the response latencies even for the words in the P set. We think that derived frequency and family size may conspire to mask an effect of the timestep itself at which the base word itself becomes available, leading to the absence of a measurable effect of base frequency and t_P.

What have we learned from this validation study of Matcheck? Firstly, we have seen that this model does a reasonable job in predicting for which words the direct route is the predominant access route, and for which words the parsing route is demonstrably active. By dividing affixes into approximately these two sets (as identified by Matcheck), and modelling them separately, we significantly increase our ability to accurately predict subjects' response times. For further validation for inflected words in Dutch and German, the reader is referred to Baayen (to appear). Secondly, we have seen that the parse times as predicted by Matcheck cannot be used as predictors of response latencies, probably because subsequent semantic processes mask the potential effect of base frequency. But certainly for the purposes of the present study, we can conclude that Matcheck is a useful tool for distinguishing between words that are likely to be accessed through parsing and subsequent semantic processes, and words that are accessed through their derived forms.

3.2 Locating the Parsing Line with Matcheck

Where in the plane spanned by log derived frequency and log base frequency can we find the words that are prone to parsing and the words that are likely to be accessed directly? The answer is provided by the upper right panel of Figure 2. Each point in this figure represents a Dutch word affixed with the suffix *heid*. The X axis shows the log derived frequency of the derived word, and the Y axis shows the log frequency of the base. Filled circles represent words for which the parsing route completes first in the Matcheck simulation (i.e, the P words). Open circles represent the F words – the words for which the derived form completes first. Interestingly, the two sets are quite distinct, and can be separated reasonably well by a straight line, as shown. This parsing line acts as a discriminant and, crucially, it has a positive slope, exactly as expected given the results of Hay (2000). It is not the absolute frequencies but the *relation* between derived frequency

and base frequency that is crucial for the balance of storage and computation.

The bottom panel of Figure 2 shows the predictions of Matcheck for the bimorphemic words in the English suffix *-ness* represented in our data set, using the same parameter settings but replacing the Dutch input lexicon by an English lexicon. Note that the parsing line is very similar to the one obtained for *-heid*. Approximately the same line also emerged when we ran Matcheck for several different affixes in English, although the exact location of the parsing line depends to some extent on the length of the suffix (see Laudanna & Burani, 1995, and Baayen, to appear, for the relevance of affix length for lexical processing). In what follows, we will ignore the role of affix length and use the parsing line of *-ness*[7] (as determined by a grid search minimizing the misclassification rate of the F and P sets) as a first pass at a more sophisticated way of estimating the balance of storage and computation for English derivational morphology than the arbitrary threshold θ proposed by Baayen (1993), and the x=y line used by Hay (in press). We also ignore possible differences between prefixes and suffixes, and treat them as identical for the purposes of this paper, while acknowledging that future work may well identify different characteristics for parsing lines associated with prefixes than those associated with suffixes.

Note that the parsing line identified here is somewhat higher than x=y. Thus, the number of forms which fall below this parsing line (i.e. are prone to whole word access) is rather higher than was predicted by the number of forms which fall below the x=y line. This is to be expected, because the x=y division in effect weighs up the effort involved in retrieving the base against the effort in retrieving the derived form. It does not take into consideration the added task of retrieving the affix and any subsequent calculations which may be associated with parsing, both of which add to the effort involved in successfully decomposing a word into its parts. Thus, if a derived word and its base are of equal frequency, the direct route is likely to have an advantage in terms of access time over the parsing route. This is reflected in the fact that the parsing line located using Matcheck is substantially higher than x=y. It appears that the derived form can be somewhat less frequent than the base word, and still have a good chance at dominating the "race", i.e. can still be robustly associated with whole-word access.

The remainder of the paper takes this parsing line as a given, and explores what we can learn about different affixes, based on how words containing those affixes are distributed relative to the parsing line. However, it is important to emphasise that we do not regard the parsing line as an absolute. That is, it is not the case that all words falling

[7] This parsing line has a slope of .76, and an intercept of 3.76

above the line are definitely and fully decomposed, and all words falling below the line are definitely and completely accessed via their derived forms. On the contrary, we regard decomposition as a continuum (see Hay 2000), and assume that both parsing and whole-word access are likely to play some role in the access of most affixed words — and may interactively converge on the correct meaning representation (Baayen and Schreuder 2000). The parsing line provides us with a valuable tool for assessing the *relative contribution* of parsing and direct access in the representation of words containing the affixes we investigate. Thus, for words which fall well above the parsing line (e.g. *dazzlement*), parsing is likely to play a fairly dominant role, and for words falling well below the parsing line (e.g. *government*) direct access is likely to play a more dominant role. The closer a word is to the line, the more equal becomes the balance between computation and storage. Thus, by making a fairly crude division between words which fall above the parsing line, and those that fall below it, we are able to approximately identify those words for which parsing is likely to play a substantial role. Viewed differently, we can isolate those words for which the base word is substantially and saliently present inside the derived word.

4 Parsing and Productivity

4.1 Parsing Ratios

Given the parsing line, we can calculate, for any given affix, what proportion of words containing this affix fall above it. These words are the words that are likely to be parsed or for which the parsing process contributes measurably to lexical access. We will henceforth refer to this proportion as the *parsing ratio*. This statistic provides information about how robust the affix is, from the point of view of perception. Affixes which are represented only by words which fall above the line are likely to be extremely robust. Due to the involvement of the parser, the formal and semantic relationship between the constituents remains transparent, and hence the syntactic and semantic functions of the affix. If one assumes (as we do) that productivity arises as a result of decomposition in the lexicon, then we should expect such affixes to have a high probability of being productive.

Affixes which are represented only by words which fall below the line, on the other hand, have very little probability of gaining an independent robust representation. Words containing such affixes are seldom decomposed during access – the whole-word access route is dominant. We predict that such affixes are extremely unlikely to be productive, as they are seldom needed in comprehension.

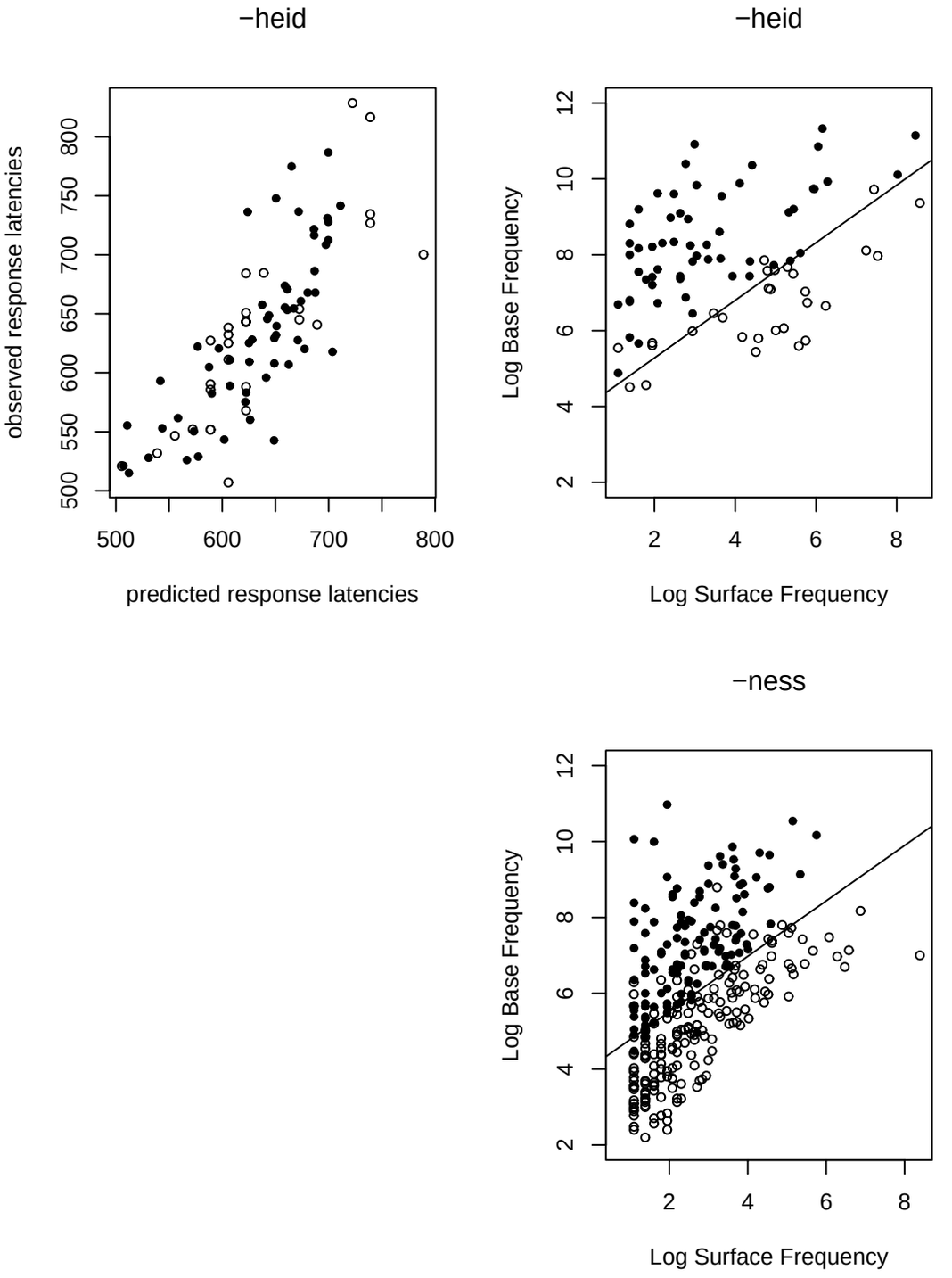

Figure 2: Left Panel: The correlation between the model times produced by MATCHECK and observed response latencies for the Dutch suffix *-heid*. Upper Right Panel: The relation between log derived frequency and log base frequency for *-heid*. Bottom Right Panel: The relation between log derived frequency and log base frequency for *-ness*. Solid points represent forms for which the parsing route is the first to produce a complete spanning in MATCHECK. Open points represent forms for which the derived form is the first to provide a complete spanning. The lines in the right panels optimally divide the words that are parsed from those that are recognized on the basis of their derived forms.

We can distinguish two types of parsing ratio – the *type parsing ratio,* and the *token parsing ratio.* The type parsing ratio can be calculated by establishing, for a given affix, what proportion of types (i.e. distinct words) containing that affix fall above the parsing line. The token parsing ratio can be calculated by establishing what proportion of all tokens containg that affix fall above the parsing line.

So how do the type parsing ratio and the token parsing ratio relate to morphological productivity? Is there a tight link, as the discussion above predicts? We can test the hypothesized link between parsing and productivity by examining the corpus of 80 English affixes described in section 2.1. How does the parsing ratio for each affix relate to its productivity?

Productivity is multifaceted, and so can be assessed in several different ways. Baayen (1994) proposes that productivity be assessed with respect to three different measures which tap into different components of productivity: $\mathcal{P}$, $\mathcal{P}^*$ and V. The category conditioned degree of productivity, $\mathcal{P}$, assesses the likelihood, given we are encountering a word containing a certain affix, of that word representing a new type. It is calculated by the total number of hapaxes (forms containing the affix which are represented just once in the corpus) as a proportion of all tokens containing the affix (N). $\mathcal{P}^*$ is the hapax conditioned degree of productivity. It expresses the probability that, if we are encountering an entirely new word, that word will contain the affix in question. It is measured by calculating what proportion of all hapaxes in the corpus are associated with that affix. Finally, in addition to these two statistics, productivity is affected by the type frequency V, which allows us to gauge how many different words the affix in question has been responsible for. Taken together, these three statistics assess the overall productivity of an affix.

Our calculations reveal that, of these three components of probability, parsing ratios are best correlated with the category conditioned degree of productivity, $\mathcal{P}$. Consider Figure 3, which is based on our set of 80 affixes – every point represents an affix. The top left panel relates the token parsing ratio to the log number of tokens representing an affix (N). N forms the denominator in the calculation of $\mathcal{P}$ – that is – it is inversely related to the level of productivity of an affix. Affixes towards the left of the graph are represented by a small number of tokens. Affixes towards the right of the graph have a high token frequency. The Y-axis represents the proportion of those tokens which are parsed, i.e., which fall above the parsing line. The solid line represents a non-parametric scatterplot smoother. The graph shows that there is a significant inverse relationship between token frequency, and the proportion of tokens which are parsed. This inverse relationship between the token parsing ratio and token frequency provides support for the hypothesis that parsing and productivity are linked.

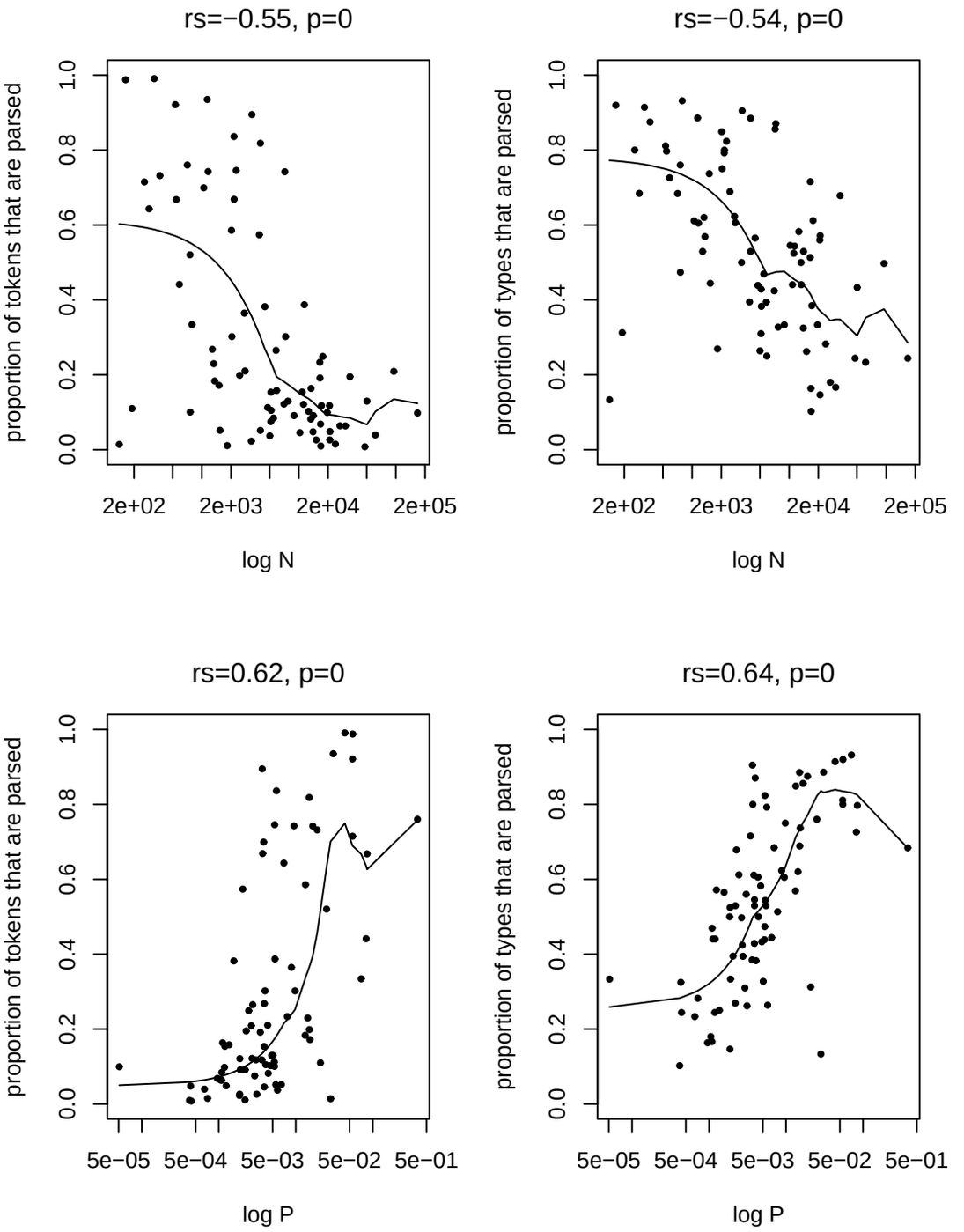

Figure 3: The relation between log N (upper panels) and $\mathcal{P}$ (lower panels) and the proportions of tokens (left) and types (right) that are parsed. Every point represents a single affix. The lines represent a non-parametric scatterplot smoother (Cleveland, 1979) fit through the points. rs=non-parametric correlation (Spearman's rho)

That such a relationship should hold between token frequency and parsing also becomes clear when we consider that high frequency forms are significantly more likely to be more frequent than their bases than low frequency forms are (Hay 2000). Thus, the more high frequency forms an affix has, the larger the number of non-parsed forms is likely to be. The top right graph demonstrates that such a relationship holds, regardless of whether we consider the proportion of types or the proportion of tokens which is being parsed. This shows how token frequency is related to the type parsing ratio of an affix. Affixes represented by high token frequency contain a smaller proportion of both types and tokens which are actually decomposed during access. Paradoxically, this leads to the generalization that the more often we encounter an affix in running speech, the less likely we are to parse words containing it.

These results demonstrate that productivity cannot be construed as arising as a simple function of the number of times one has encountered an affix. On the contrary – the more often you encounter an affix, the less likely you are to decompose forms containing it, and so the less productive that affix is likely to be.

The bottom panels show how parsing ratios relate to category-conditioned productivity $\mathcal{P}$, itself a partial function of N (Baayen 1989, 1992, Baayen and Lieber 1991). It is a measure which has met some success in measuring degree of productivity for different affixes. If an affix is highly productive, then new forms will be constantly coined with that affix, and so a corpus should contain many forms which are encountered just once. For non-productive affixes, however, there is a strictly finite set of words which contain that affix, so, for a large corpus, the chances are high that those words will have been encountered more than once. $\mathcal{P}$ is calculated as the number of hapaxes (V1) containing the affix, as a proportion of all tokens containing the affix.

For any given affix, both the proportion of tokens which are parsed (bottom left graph) and also the proportion of types which are parsed (bottom right) are significantly related to the log productivity of that affix.[8] Productivity is related — in a statistically well-behaved way — to parsing.

We see this relationship as causal. Productivity arises from decomposed forms in the lexicon. And decomposed forms in the lexicon arise from parsing. Thus, we expect those affixes which are represented by words which are likely to be parsed, to be associated with a certain level of productivity.

We have shown figures for both proportion of types and proportion of tokens, to show that the relationship between parsing and productivity is robust, regardless of the per-

[8]We take the logarithm of $\mathcal{P}$ because this makes the structure in the data more visible to the eye.

spective from which it is viewed. Both types and tokens are likely to play important roles. A certain proportion of types is required in order for generalisation to take place. That is, if just one word containing a particular affix is likely to be parsed, then, regardless of how frequent that word is, speakers are unlikely to generalise on the basis of one word, and so the affix is unlikely to become productive. So type frequency is important (see also Pierrehumbert, 2001). Token frequency is also important, because once speakers have generalised from a set of words and an affix has a robust representation, then words with a high token frequency (which are nonetheless parsed), serve to regularly activate (and so strengthen) the affix's representation. Thus, both a high type-parsing ratio and a high token-parsing ratio influence the degree to which an affix will be productive.

Why, of the three components of productivity, do parsing ratios best correlate with $\mathcal{P}$? Recall that $\mathcal{P}$ is the category conditioned degree of productivity. It tells us, given that we are encountering a word containing a particular affix, the probability that the word has been productively coined – i.e. the probability that the speaker or writer did not retrieve a stored representation for that word, but produced it from its constituents. The calculation definitionally limits the domain of comparison to a single affix, and does not compare it, either implicitly or explicitly, to the behaviour of any other affix. The parsing ratio does exactly the same thing — it tells us the category conditioned degree of parsing. It is the perceptual counterpart to $\mathcal{P}$. It tells us, given that a listener is encountering a word containing a particular affix, the probability that the word will be decomposed during access. Summing up, the results in this section demonstrate that the category conditioned degree of parsing is a statistically reliable predictor of the category conditioned degree of productivity.

4.2 The Intercept

When we regress base frequency on derived frequency, the intercept of the resulting line displays a fair amount of variation across affixes. When we consider the graphs in Figure 1 for example, the intercept for *-ment* (i.e. the place where it crosses the Y axis) is 1.25, and the intercept for *-ism* is 3.53. A consequence of this difference is that fewer points fall below the x=y line (and also, the more empirically motivated parsing line – not shown in figure 1) for *-ism* than for *-ment*.

From a production perspective, the intercept (regardless of the slope of the line) indicates how frequent a base word needs to be before it is likely to spawn an affixed word. Affixes with high intercepts are affixes for which the base words are fairly high frequency relative to the derived forms. For such affixes, a base word needs to be fairly frequent

before it is likely to produce an affixed word. A high intercept is likely to indicate that the category of the base word is more "useful" – i.e., more frequently employed, than the category of the derived word. Affixes with low intercepts are affixes for which the base words tend to be of lower frequency relative to the derived words. This type of profile might be expected for affixes for which the category created by the affix is highly useful, and so leads to relatively frequently used words.

From a perception perspective, a high intercept reflects an overall pattern in which base frequencies tend to be high relative to derived frequencies. That is, it reflects a distribution in which many words are prone to parsing, and very few are prone to whole word access. A low intercept, on the other hand, would reflect a distribution in which a larger proportion of forms fall below the parsing line. Such a distribution has a larger proportion of forms which are prone to whole word access.

Taking the claims about production and perception together, we reach a surprising conclusion — but one which is nonetheless borne out by the empirical facts. The less useful an affix is (in terms of the *degree* of use of the words it creates – not in terms of how many different words it could potentially create), the more likely it is to be parsed, and so the more productive it is likely to be. Relatively useless affixes remain productive because their derived forms remain low frequency relative to the frequency of the base words. This leads to high rates of parsing, and so to a robust representation of the affix.

Evidence for the existence of a relationship between the intercept and productivity can be seen in the graphs in Figure 4. These graphs are based on robust regression lines fit through individual affixes (a small subset of which were shown in Figure 1). Each point in the graphs in Figure 4 represents a single affix. The X-axis of the graphs shows the intercept for robust regression lines fit through derived and base frequency for each individual affix. Of the 80 affixes, 44% show a significant correlation between derived frequency and base frequency. All 80 are shown here, however, as the intercept is relevant (both from a production and a perception perspective), regardless of the significance of the slope of the line with which it is associated. For those with a non-significant slope, the intercept merely reflects the average base frequency.

The left panel of Figure 4 shows a significant relationship between the intercept, and productivity as measured by $\mathcal{P}$. Affixes which return high intercept values when base frequency is regressed on derived frequency, show significantly higher levels of productivity. The right panel shows the relationship between the value of the intercept, and the total number of tokens containing the affix. Affixes with high token frequency (towards the top of the graph) are more likely to be represented by high frequency words, which fall below the parsing line, and so are prone to whole word access. When we regress base

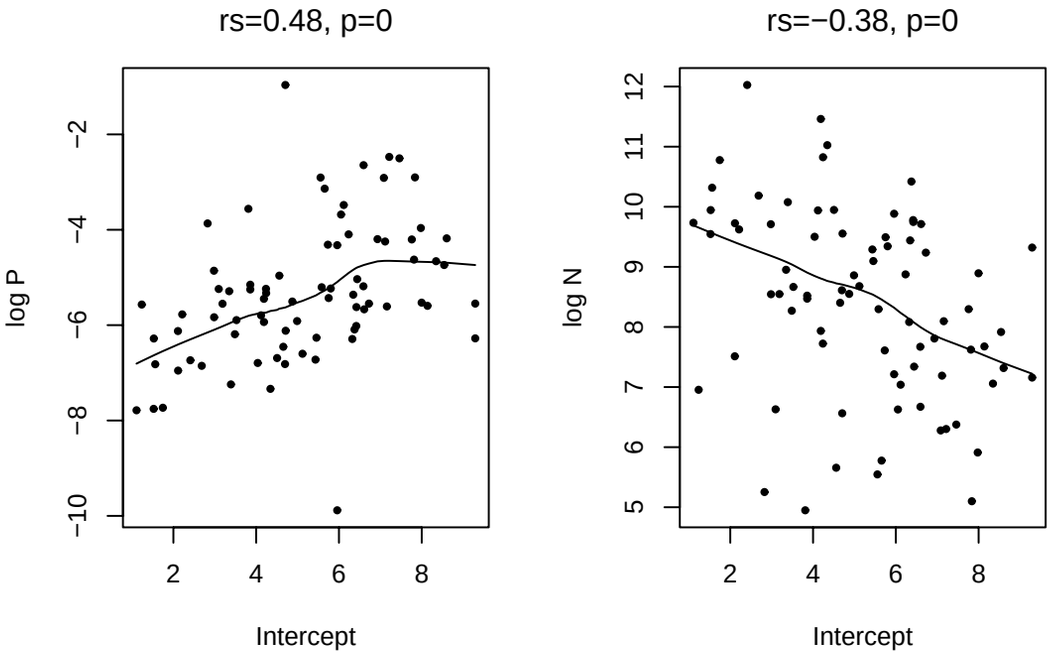

Figure 4: The relation between the Intercepts of the by-affix linear models regressing base frequency on derived frequency, and log $\mathcal{P}$ (left panel) and log N (right panel). Every point represents an affix. The lines represent a non-parametric scatterplot smoother (Cleveland, 1979) fit through the points. rs=non-parametric correlation (Spearman's rho).

frequency on derived frequency, large numbers of high frequency words will lower the value of the intercept by pulling the regression line closer to the X-axis.

The results shown in Figure 4, then, further demonstrate that the relationship between base frequency and derived frequency for a given individual affix profoundly influences that affix's degree of productivity.

4.3 Estimating the Activation Level

In section 4.1 we showed that there is a significant relationship between the proportion of forms that are parsed (the parsing ratio) and the productivity of the associated affix. When we limit the domain of analysis to individual affixes, the proportion of forms which is parsed is an extremely good predictor of the proportion of forms which will be productively coined.

However, the proportion of forms is not necessarily the best indicator of the overall activation level of the affix. We can more accurately compare the different degrees of activation different affixes receive by comparing the actual number of forms for each affix that are parsed. In terms of perception, there is a sense in which the forms which are not parsed do little or nothing to contribute to the activation level of the affix. Rather, the degree to which the affix is activated can be assessed by calculating the total number of forms containing that affix which are characterized by decomposition.

When we investigate the relationship of the total number of forms parsed (i.e. falling above the parsing line) to the various measures of productivity, we find a very interesting result. This number is not a good predictor of the category conditioned degree of productivity $\mathcal{P}$. It is, however, extremely highly correlated with the other two aspects of productivity — the type frequency V, and the hapax conditioned degree of productivity $\mathcal{P}^*$.

Recall that the hapax conditioned degree of productivity $\mathcal{P}*$ is calculated as the total number of hapaxes representing the affix in question, as a proportion of all hapaxes in the corpus. Because we are working with a single corpus here, the total number of hapaxes in the corpus is a constant. The number of hapaxes representing any given affix ($V1$), then, can be viewed as a measure of that affix's hapax conditioned degree of productivity $\mathcal{P}^*$.

Figure 5 shows how the total number of forms parsed is related to the number of types an affix has (V), and the number of hapaxes that represent it ($V1$). Each point represents a single affix. In the top panels, the X-axis shows the log number of types of words represented by the affix. The Y-axis shows the log number of tokens (left panel) and types (right panel) containing that affix which are likely to be parsed (ie. which fall above the

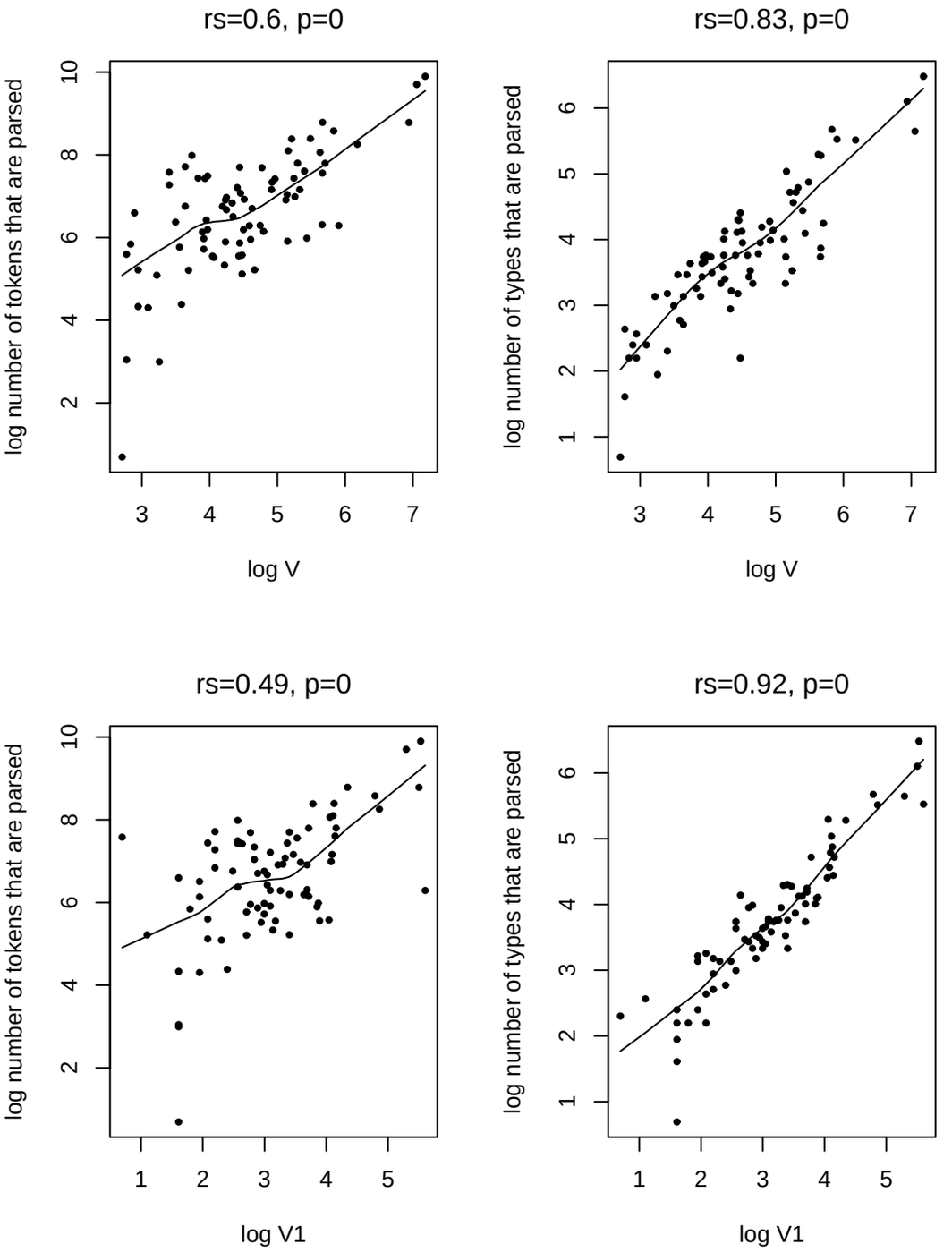

Figure 5: The relation between log V (upper panels) and $V1$ (lower panels) and the log number of tokens (left) and types (right) that are parsed. Every point represents an affix. The lines represent a non-parametric scatterplot smoother (Cleveland, 1979) fit through the points. rs=non-parametric correlation (Spearman's rho).

parsing line). What we see is that affixes which have a large number of types associated with them, have a larger number of both tokens (left panel) and types (right panel) which are prone to parsing. However, the relation between the number of types, and number of types or tokens parsed is fairly trivial. The larger the number of different words that exist, the larger the number which is likely to be parsed.

The bottom panel shows that for any given affix, the number of tokens parsed (left) and types parsed (right) can be predicted by the total number of hapaxes containing that affix — i.e. the number of new words which are coined with that affix in a given period, and a measure of the hapax conditioned degree of productivity.

In the left panel, we see a statistical relationship between the number of hapaxes, and the log number of tokens that are parsed. This relationship is by no means trivial. Hapaxes contribute extremely minimally to overall token counts, and so there is no a priori reason we should expect the number of hapaxes to correlate with the total number of tokens which are parsed. Yet we *do* see this relationship, and the reason we see it (we suggest), is because there is a causal relationship between parsing and productivity. The larger the number of tokens that is parsed, the more activated and robust the representation of the affix is, and so the more available it becomes for the formation of new forms.

Finally, in the bottom right panel we see an extremely tight correlation between number of hapaxes and the number of types which are parsed. One long-standing puzzle in the study of productivity is that there is no reliable correlation between the productivity of an affix and the number of different words in which it can be found. Type frequency, it has often been claimed, cannot be related to productivity (Dressler 1997, Anshen and Aronoff 1999).

> Apparently, a language can have in its lexicon a fairly large number of words from which one could potentially analogize to a productive pattern without any consequent productivity. (Anshen and Aronoff, 1999:25).

Anshen and Aronoff see this as a "formidable obstacle" to those who argue for quantitative analogy over rules. It certainly appears to be the case that type frequency alone can not predict productivity. What is crucially missing from any analysis focussing on type frequency alone is any information about how decomposable the types are. Not all words contribute equally to the productivity of an affix. In particular, words which tend to be accessed directly, and for which decomposition plays no effective role, do not contribute significantly to the productivity of the affixes they contain. By finding a principled way of approximately identifying that subset of words, for any given affix, which are prone to

parsing, we are able to demonstrate an astonishing level of correlation between the size of that subset, and the frequency with which new words are likely to be formed.

The number of forms parsed, we claim, corresponds to the overall activation level of an affix. And the activation level of an affix is directly related to that affix's productivity — as measured both by the overall likelihood of encountering a new word containing that affix, and its degree of generalisability.

4.4 The significance of the slope

When we more closely examine many of the results described above, we tend to find a marked difference in the behaviour of affixes which display a significant correlation between base frequency and derived frequency, and those that do not. Figure 6, for example, repeats the top left graph from Figure 3. It shows the relationship between token frequency and the token parsing ratio — the proportion of tokens that are parsed. Each point represents an affix. Filled points represent affixes for which there is no significant correlation between base frequency and derived frequency. Unfilled points represent affixes for which a significant correlation holds. Separate lines are fitted through each of the two sets of affixes. These lines show that, regardless of whether there is a significant correlation between base and derived frequency, a strong relationship holds between total token frequency (N) and token parsing ratios. Affixes which have higher token frequency have a smaller proportion of tokens which are actually parsed.

The difference between these two lines tells us something about the implications of a significant correlation between base and derived frequency. Recall that in section 2.2 we argued that significant correlations should tend to hold for affixes which display high rates of semantic transparency and/or are highly productive. Figure 6 shows that those affixes with significant correlations tend to have higher parsing rates compared to affixes with non-significant correlations with the same token frequency N. In other words, given two affixes with the same token frequency, the affix with the significant correlation between base and derived frequency will display a higher parsing ratio.

What Figure 6 shows, then, is that although it is detrimental for the productivity of an affix to have a large value of N, it is less detrimental when the affix has a significant slope, i.e., when the words it occurs tend to be transparent to their base. In other words, of two affixes with similar token parsing ratios, the one maintaining a significant correlation between derived frequency and base frequency is used more often: Purely in terms of intensity of use, it is the more productive affix.

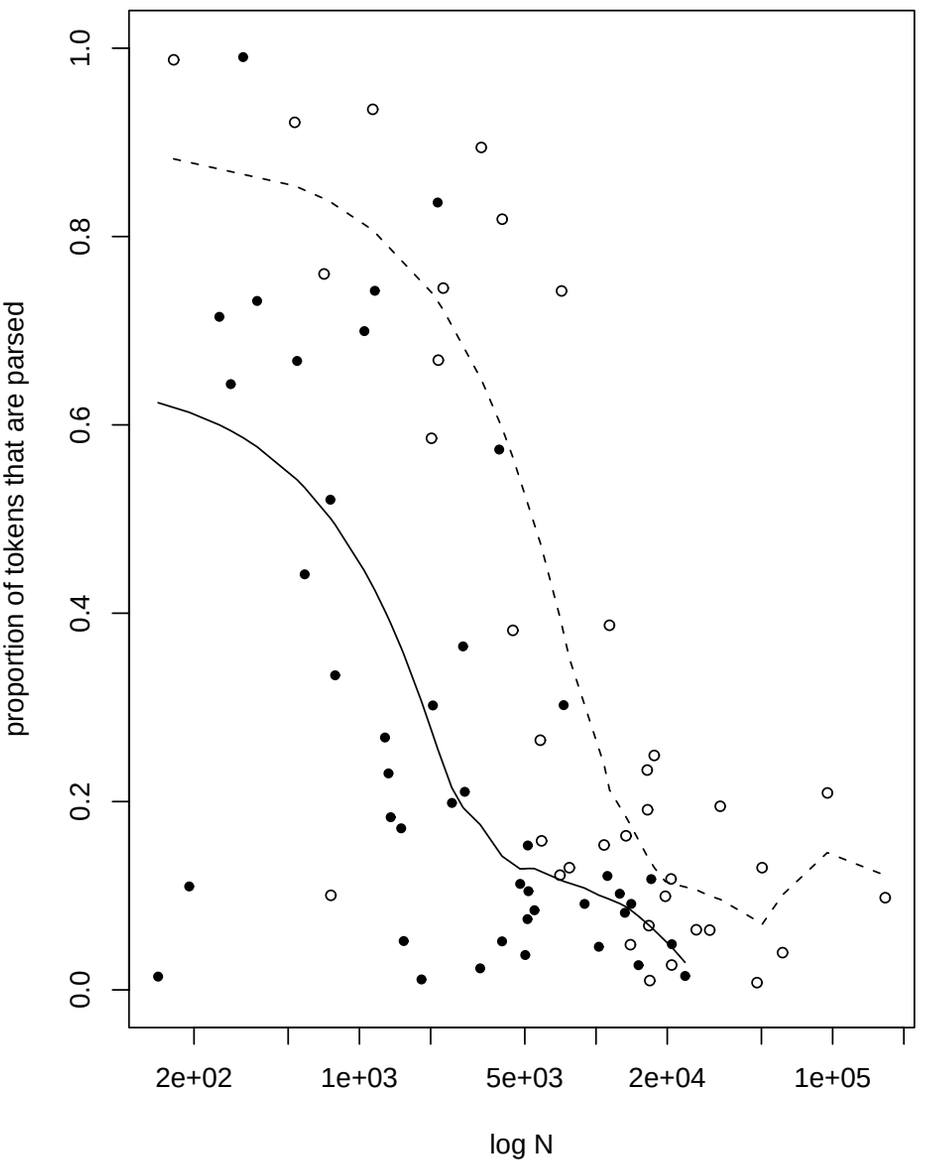

Figure 6: The relation between log N and the proportion of tokens that are parsed. Unfilled points represent affixes with a significant correlation between base frequency and derived frequency. Filled points represent affixes with no such correlation. The dotted line represents a non-parametric scatterplot smoother for the affixes in the former set, the solid line represents a similar smoother for affixes in the latter set.

5 Discussion

We have identified a critical ratio of base and derived frequency above which derived forms are likely to be parsed. This critical ratio is defined by the parsing line shown in Figure 2. We have motivated this parsing line both theoretically and empirically, albeit as yet only on the basis of a case study of a single affix. We have demonstrated that the distributional properties of words with respect to this parsing line are statistically related to an affix's productivity.

These findings enable us to considerably refine the psychologically motivated measure $\mathcal{A}$ developed by Baayen (1993). Baayen proposes to set an activation level at a certain frequency threshold — θ. What $\mathcal{A}$ measures is the number of types of a certain category occurring below this frequency threshold θ, each weighted by their frequency. This measure is distinct from the other productivity measures he proposes, in that it attempts not only to measure degree of productivity, but also to explain it. The general idea behind the approach is that low frequency types require parsing, and so protect the activation levels of the affixes against decay. The choice of the frequency threshold is not straightforward, as Baayen himself notes. He chooses a fairly low threshold, explaining:

> The low choice of θ ... ...is motivated by the constraint that only semantically transparent complex words contribute to the activation level $\mathcal{A}$. Since transparency is inversely correlated with frequency, higher values of θ 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)

Recent work, however, has demonstrated that frequency by itself is not a good predictor of degree of semantic transparency (Hay in press). High frequency forms are not significantly more likely to display signs of semantic drift than low frequency forms. The *relative* frequency of the derived form and the base, however, is a good predictor, both for prefixes and suffixes. Thus, while drawing a line at a given frequency threshold may have been a reasonable first approximation, it does not reliably distinguish between opaque and transparent forms, as Baayen had intended it to. Note that, in the plane spanned

by derived frequency and base frequency, a frequency threshold such as θ amounts to a vertical parsing line at a fixed position of derived frequency for all affixes.

Conversely, the parsing line motivated in section 3, and the results in subsequent sections showing how this parsing line leads to various correlational patterns relating to different facets of productivity, suggest that there is a strong psycholinguistic motivation for using this parsing line as critical diagnostic rather than using a simple frequency threshold.

We have shown that it is possible to estimate the frequency with which an affix is activated, by considering the relationship between base and derived frequency for words containing that affix, and how this distribution relates to the parsing line. This allows us not only to estimate which words containing a particular affix are likely to be parsed, but also to accommodate the insight that the *relation* between derived frequency and base frequency is a crucial factor in language comprehension.

This increased understanding of the nature of morphological parsing has made it possible to come to grips with various aspects of morphological productivity. Given any particular affix – the likelihood that it will be parsed during access is statistically predictive of $\mathcal{P}$, the likelihood of a word containing that affix having been productively coined. And $\mathcal{P}*$ – the likelihood, given all productively coined words, that a coined word will contain the affix of interest, is a function of the frequency of activation of that affix – as measured by the number of forms containing the affix which tend to be accessed via parsing.

Independently of the parsing line, we have shown that the nature of the correlation between base frequency and derived frequency varies markedly across individual affixes. This variation is linguistically significant, and can be related to affixes' parsability and productivity.

Taken together, the results in this paper provide evidence of a strong link between productivity and parsing. The frequency with which an affix is activated during processing directly affects the degree to which it is productive.

References

Anshen, F. and Aronoff, M.: 1999, Using dictionaries to study the mental lexicon, *Brain and Language* **68**, 16–26.

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

Baayen, H.: 1989, *A Corpus-based approach to morphological productivity: Statistical analysis and psycholinguistic interpretation*, PhD thesis, Vrije Universiteit, Amsterdam.

Baayen, H.: 1992, Quantitative aspects of morphological productivity, *in* G. Booij and J. van Marle (eds), *Yearbook of Morphology 1991*, Kluwer Academic Publishers, Dordrecht, pp. 109–150.

Baayen, H.: 1993, On frequency, transparency and productivity, *in* G. Booij and J. van Marle (eds), *Yearbook of Morphology 1992*, Kluwer Academic Publishers, pp. 181–208.

Baayen, H.: to appear, Probability in morphology, *in* Bod, R., Hay, J. and S. Jannedy (eds), *Probability Theory in Linguistics*, MIT Press, Cambridge, MA.

Baayen, H. and Lieber, R.: 1991, Productivity and English derivation: A corpus based study, *Linguistics* **29**, 801–843.

Baayen, R. H.: 1994, Productivity in language production, *Language and Cognitive Processes* **9**(3), 447–469.

Baayen, R. H.: 2001, *Word Frequency Distributions*, Kluwer Academic Publishers.

Baayen, R. H. and Schreuder, R.: 2000, Towards a psycholinguistic computational model for morphological parsing, *Philosophical Transactions of the Royal Society (Series A: Mathematical, Physical and Engineering Sciences* **358**, 1–13.

Baayen, R. H., Schreuder, R. and Sproat, R.: 2000, Morphology in the mental lexicon: a computational model for visual word recognition, *in* F. van Eynde and D. Gibbon (eds), *Lexicon Development for Speech and Language Processing*, Kluwer Academic Publishers, pp. 267–291.

Baayen, R. H., Schreuder, R., Bertram, R. and Tweedie, F.: 1999, The semantic functions of the dutch suffix *-heid*: evidence from lexicography, lexical statistics, and psycholinguistics, pp. 1–29.

Baayen, R.H, Piepenbrock, R. and Gulikens, L.: 1995, The CELEX lexical database (release 2) cd-rom., Philadelphia, PA: Linguistic Data Consortium, Univesity of Pennsylvania (Distributor).

Bertram, R., Baayen, R. H. and Schreuder, R.: 2000, Effects of family size for complex words, *Journal of Memory and Language* **42**, 390–405.

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

Cutler, A., Hawkins, J. A., and Gilligan, G.: 1985, The suffixing preference: a processing explanation, *Linguistics* **23**, 723–758.

De Jong, N. H., Feldman, L., Schreuder, R., Pastizzo M., and Baayen, H.: 2001, The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects, *Brain and Language*.

De Jong, N. H., Schreuder, R. and Baayen, R. H.: 2000, The morphological family size effect and morphology, *Language and Cognitive Processes* **15**, 329–365.

Dressler, W. U.: 1997, On productivity and potentiality in inflectional morphology, *CLASNET Working Papers* **7**, 2–22.

Hay, J.: 2000, *Causes and Consequences of Word Structure*, PhD thesis, Northwestern University.

Hay, J.: in press, Lexical frequency in morphology: Is everything relative?, *Linguistics*.

Laudanna, A. and Burani, C.: 1995, Distributional properties of derivational affixes: Implications for processing, *in* L. B. Feldman (ed.), *Morphological Aspects of Language Processing*, Lawrence Erlbaum Associates, Hillsdale, N.J., pp. 345–364.

Norris, D. G.: 1994, Shortlist: A connectionist model of continuous speech recognition, *Cognition* **52**, 189–234.

Pierrehumbert, J.: 2001, Why phonological constraints are so coarse-grained, *Language and Cognitive Processes*.

Renouf, A.: 1987, Corpus development, *in* J. Sinclair (ed.), *Looking up: An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*, Collins, pp. 1–40.

Rousseeuw, P. and Leroy, A.: 1987, *Robust regression and outlier detection*, John Wiley and Sons, New York.

Schreuder, R. and Baayen, R. H.: 1997, How complex simplex words can be, *Journal of Memory and Language* **37**, 118–139.

Segui, J. and Zubizarreta, M.-L.: 1985, Mental representation of morphologically complex words and lexical access, *Linguistics* **23**, 759–774.

(Jennifer Hay)
Department of Linguistics
University of Canterbury
Private Bag 4800
Christchurch
New Zealand
email: j.hay@ling.canterbury.ac.nz

(Harald Baayen)
Max-Planck-Institut für Psycholinguistik
Wundtlaan 1
6525 XD Nijmegen
The Netherlands
email: baayen@mpi.nl

Appendix

The following tables list all of the affixes investigated, together with the major calculations discussed. Note that figures given here are rounded to two or three decimal places, for ease of presentation. However the statistics described in this paper were all performed on figures accurate to 6 decimal places.

affix	cor	prob	int	V1	$\mathcal{P}$	V	types-P	type-PR,	tokens-P	token-PR
anti	0.11	0.312	7.46	48	0.082	84	61	0.73	259	0.44
be	0.19	0.078	4.50	26	0.001	91	52	0.57	1017	0.05
con	0.05	0.702	3.19	20	0.004	70	30	0.43	790	0.15
counter	0.29	0.033	7.09	29	0.054	53	43	0.81	491	0.92
cross	0.23	0.186	5.66	14	0.043	35	32	0.91	320	0.99
de	-0.02	0.810	6.73	40	0.004	121	66	0.55	469	0.05
dis	0.23	0.013	4.04	15	0.001	118	52	0.44	2187	0.16
em	0.13	0.612	9.29	5	0.004	17	9	0.53	344	0.27
en	0.13	0.612	9.29	21	0.002	82	43	0.52	1350	0.12
fore	0.39	0.005	4.24	12	0.005	51	42	0.82	1683	0.75
im	0.13	0.363	4.70	6	0.001	49	23	0.47	463	0.08
in	-0.09	0.206	5.75	58	0.004	192	96	0.5	1084	0.08
inter	0.10	0.475	7.81	20	0.010	52	39	0.75	617	0.30
mid	0.31	0.010	6.11	35	0.031	70	62	0.89	1065	0.94
mis	0.26	0.062	6.59	12	0.006	53	42	0.79	1791	0.84
non	0.16	0.147	6.59	56	0.071	88	82	0.93	264	0.33
out	0.11	0.299	7.80	29	0.004	85	74	0.87	2204	0.30
over	0.23	0.002	7.76	60	0.015	174	154	0.89	3283	0.82
pre	0.03	0.818	6.93	37	0.015	90	62	0.69	488	0.20
re	0.21	0.000	6.38	76	0.002	289	196	0.68	6542	0.20
self	0.04	0.803	6.05	19	0.025	50	38	0.76	393	0.52
sub	-0.16	0.243	8.60	23	0.015	57	42	0.74	258	0.17
super	-0.02	0.875	7.22	46	0.084	69	55	0.80	364	0.67
trans	0.08	0.754	1.24	4	0.004	18	11	0.61	733	0.7
un	0.16	0.012	5.80	61	0.005	241	131	0.54	4417	0.39
under	0.46	0.000	5.73	27	0.013	86	73	0.85	1182	0.59

Table 1: Prefixes investigated.
cor: non-parametric correlation between base and derived frequency (Spearman's rho).
prob: significance level of the above correlation.
int: intercept returned from robust regression of base frequency on derived frequency.
V1: number of hapaxes.
$\mathcal{P}$: Productivity, as measured by the number of hapaxes, as a proportion of total token frequency.
V: number of distinct words containing the affix.
types-P: number of distinct words which fall above the parsing line.
type-PR: The type parsing ratio = the proportion of types which fall above the parsing line.
tokens-P: The summed frequency of the words which fall above the parsing line.
token-PR: The token parsing ratio = the proportion of tokens which fall above the parsing line.

affix	cor	prob	int	V1	$\mathcal{P}$	V	types-P	type-PR,	tokens-P	token-PR
able	0.13	0.031	6.61	57	0.003	278	199	0.72	3161	0.19
age	0.09	0.287	4.71	31	0.002	136	72	0.53	1285	0.09
al	0.34	0.000	4.34	40	0.001	300	70	0.23	2430	0.04
an	0.11	0.109	2.21	47	0.003	229	60	0.26	398	0.03
ance	0.27	0.019	1.52	6	0.000	77	25	0.32	671	0.05
ant	0.14	0.159	5.46	17	0.002	102	34	0.33	816	0.09
ary	0.39	0.001	5.12	8	0.001	76	19	0.25	931	0.16
ate	0.09	0.361	2.99	15	0.003	100	31	0.31	385	0.07
ation	0.53	0.000	2.68	28	0.001	189	34	0.18	1695	0.06
dom	0.30	0.175	6.32	6	0.002	22	11	0.50	74	0.02
ee	0.04	0.761	5.58	22	0.005	68	36	0.53	207	0.05
eer	0.46	0.047	3.09	4	0.005	19	9	0.47	76	0.10
en	0.16	0.020	4.12	63	0.003	200	112	0.56	2443	0.12
ence	0.45	0.000	1.11	7	0.000	88	9	0.10	167	0.01
ent	0.17	0.119	3.39	17	0.001	85	24	0.28	354	0.01
er	0.35	1.000	4.19	251	0.003	1313	653	0.50	19872	0.21
ery	0.16	0.083	4.87	21	0.004	115	44	0.38	542	0.10
ese	0.20	0.324	2.11	4	0.002	26	7	0.27	20	0.01
ess	-0.10	0.461	5.96	18	0.013	58	33	0.57	249	0.18
ette	0.05	0.788	6.44	10	0.006	36	16	0.44	80	0.05
fold	0.48	0.017	7.83	9	0.055	25	23	0.92	162	0.99
ful	0.16	0.035	6.42	43	0.002	183	112	0.61	4391	0.25
hood	0.51	0.005	8.14	8	0.004	30	24	0.80	1441	0.67
ian	0.11	0.275	3.86	29	0.006	106	28	0.26	185	0.04
ic	0.28	0.000	1.53	39	0.002	287	42	0.15	550	0.03
ier	0.39	0.102	4.56	2	0.007	19	13	0.68	184	0.64
ify	0.53	0.000	4.66	7	0.002	46	26	0.57	1701	0.38
ish	0.06	0.426	6.34	59	0.005	206	120	0.58	1286	0.10
ism	0.24	0.005	3.53	16	0.003	137	54	0.39	1540	0.27
ist	0.38	0.000	3.35	39	0.005	168	55	0.33	1001	0.13
itis	-0.30	0.283	3.81	4	0.028	15	2	0.13	2	0.01
ity	0.55	0.000	1.56	33	0.001	288	48	0.17	1916	0.06
ive	0.27	0.027	4.99	19	0.003	66	28	0.42	857	0.12
ize	0.23	0.006	5.43	13	0.001	143	63	0.44	1664	0.15
less	0.32	0.000	6.23	119	0.017	340	291	0.86	5313	0.74
let	0.19	0.198	7.12	19	0.014	50	31	0.62	305	0.23
like	0.24	0.000	4.71	270	0.381	367	251	0.68	539	0.76
ling	-0.21	0.358	7.75	0	0.000	21	13	0.62	110	0.10
ly	0.51	0.000	2.41	198	0.001	1158	283	0.24	16347	0.10

affix	cor	prob	int	V1	$\mathcal{P}$	V	types-P	type-PR,	tokens-P	token-PR
ment	0.21	0.005	1.75	21	0.000	172	42	0.24	370	0.01
most	-0.04	0.874	7.98	7	0.019	16	14	0.88	270	0.73
ness	0.42	0.000	2.98	128	0.008	483	248	0.51	3845	0.23
oid	-0.16	0.556	2.83	4	0.021	16	5	0.31	21	0.11
or	0.00	0.972	6.42	62	0.004	221	85	0.38	2013	0.12
ory	0.26	0.111	3.49	8	0.002	38	15	0.40	2239	0.57
ous	0.41	0.000	2.11	16	0.001	171	28	0.16	1142	0.07
proof	0.27	0.097	5.56	14	0.055	40	32	0.80	183	0.71
ry	-0.01	0.960	3.86	25	0.005	98	43	0.44	537	0.11
ship	0.06	0.615	8.54	24	0.009	69	43	0.62	1000	0.36
some	0.14	0.387	8.35	11	0.009	38	23	0.61	862	0.74
ster	0.06	0.743	4.19	12	0.004	33	20	0.61	586	0.21
th	0.53	0.003	5.96	1	0.000	30	10	0.33	1959	0.10
ward	0.38	0.014	7.16	12	0.004	42	38	0.90	2932	0.89
y	0.16	0.000	4.24	244	0.005	1032	447	0.43	6522	0.13

Table 2: Suffixes investigated.
cor: non-parametric correlation between base and derived frequency (Spearman's rho).
prob: significance level of the above correlation.
int: intercept returned from robust regression of base frequency on derived frequency.
V1: number of hapaxes.
$\mathcal{P}$: Productivity, as measured by the number of hapaxes, as a proportion of total token frequency.
V: number of distinct words containing the affix.
types-P: number of distinct words which fall above the parsing line.
type-PR: The type parsing ratio = the proportion of types which fall above the parsing line.
tokens-P: The summed frequency of the words which fall above the parsing line.
token-PR: The token parsing ratio = the proportion of tokens which fall above the parsing line.