Putting the bits together: An information theoretical perspective on morphological processing

Fermín Moscoso del Prado Martín[a,b,*] Aleksandar Kostić[c]

R. Harald Baayen[a]

[a]*Max Planck Inst. for Psycholinguistics & Univ. of Nijmegen, The Netherlands*

[b]*MRC–Cognition and Brain Sciences Unit, Cambridge, U.K.*

[c]*Lab. for Experimental Psychology, Univ. of Belgrade, Serbia and Montenegro*

Abstract

In this study we introduce an information-theoretical formulation of the emergence of type-based and token-based effects in morphological processing. We describe a probabilistic measure of the informational complexity of a word, its information residual, which encompasses the combined influences of the amount of information contained by the target word and the amount of information carried by its nested morphological paradigms. By means of re-analyses of previously published data on Dutch words we show that the information residual outperforms the combination of traditional token-based and type-based counts in predicting response latencies in visual lexical decision, and at the same time provides a parsimonious account of inflectional, derivational, and compounding processes.

Key words: Inflection, Derivation, Compound, Morphology, Information Theory

Introduction

In the present study, we develop a concise mathematical formulation of the morphological complexity of a word: its information residual. This measure is inspired by the approach to processing of inflected morphology proposed by Kostić (Kostić, 1991, 1995, 2003; Kostić, Marković, & Baucal, in press). Our single measure outperforms in terms of explained variance the combined effects on lexical decision latencies of a wide range of apparently contradicting measures that had been described in the literature. More importantly, it provides us with a parsimonious treatment of monomorphemic, polymorphemic, and compound words, where the apparent processing differences between them, arise naturally as consequences of the probabilistic distributions found in different types of morphological paradigms. This constitutes a new tool that allows to investigate in detail the consequences for lexical processing of the hyerarchical structure of morphological paradigms, that had been largely overlooked in the literature.

Several token-based counts (i.e., counting the number of occurrences of a word, a base form, or a root in a sufficiently large corpus), are reported to affect response latencies in the visual lexical decision task. Surface frequency, the number of times that each particular inflectional variant of a word appears in a corpus, is negatively correlated with response latencies to monomorphemic words in visual lexical decision (Taft, 1979; Whaley, 1978). Also base frequency, that is, the summed frequency of all inflected variants of a word, has been shown to correlate positively with response latencies, even after partialling out the effect of surface frequency (Baayen, Dijkstra, & Schreuder, 1997; Taft, 1979). Similarly, Hay (2001) showed that the logarithm of the ratio between the surface frequency and the base frequency of a word, its inflectional ratio, correlates negatively with response latencies.

* Corresponding author. MRC–Cognition and Brain Sciences Unit, 15 Chaucer Road, CB2 2EF Cambridge, U.K.

Email address: fm01@mrc-cbu.cam.ac.uk (Fermín Moscoso del Prado Martín).

Finally, the summed base frequency of all words derived from the same stem, its cumulative root frequency, has also been reported to have a negative correlation with response latencies after partialling out the effects of surface and base frequency (Colé, Beauvillain, & Seguí, 1989; Taft, 1979).

In contrast to these token-based counts, Schreuder and Baayen (1997) introduced a type-based measure that has an independent effect on responses to visual lexical decision in Dutch. The morphological family size of a word is the number of other polymorphemic words and compounds in which it appears as a constituent, independently of their frequencies of occurrence. For instance, the morphological family of the word *fear* contains the words *fearful*, *fearfully*, *fearfulness*, *fearless*, *fearlessly*, *fearlessness*, *fearsome*, and *godfearing*, according to the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), so the family size of *fear* equals 9. A facilitatory effect of family size has also been documented in a range of languages other than Dutch (Baayen, Lieber, & Schreuder, 1997; Ford, Marslen-Wilson, & Davis, in press; Lüdeling & De Jong, 2001; Moscoso del Prado Martín, 2003; Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, to appear), and appears to arise at the level of semantic processing (cf., De Jong, 2002). Table 1 provides a summary of the different frequency counts which have been shown to influence visual lexical decision latencies.

Morphological family size is highly correlated with cumulative root frequency. In general, the more family members a word has, the higher their summed frequency will be. However, Schreuder and Baayen (1997) report that, when the effect of family size is controlled for, there is no effect of cumulative root frequency. However, a re-analysis of their experimental results by Baayen, Tweedie, and Schreuder (2002) using a more sensitive linear mixed effect model (Pinheiro & Bates, 2000) revealed a small inhibitory effect of cumulative root frequency after having partialled out the facilitatory effect of family size. [1]

[1] Note here that, in their original studies, Baayen and colleagues referred to family frequency,

Table 1

Summary of the morphological variables from which we derive the information residual.

Variable	Description	Example	Type- or token-based	Correlations with RT reported in the literature
Surface Frequency	Number of times that a word appears in a corpus.	$F(\text{“page”})$	token-based	negative
Base Frequency	Sum of the surface frequencies of all inflectional variants of a word.	$F(\text{page}) = F(\text{“page”}) + F(\text{“pages”})$	token-based	negative
Inflectional Ratio	Quotient between surface and base frequency of a word.	$\rho_I = \frac{F(\text{“move”})}{F(\text{move})}$	token-based	negative
Cumulative Root Frequency	Summed based frequencies of all words sharing a stem	$Nf(\text{page}) = F(\text{page}) + F(\text{paginate}) + F(\text{pagination}) +$ $+F(\text{frontpage}) + F(\text{title-page}) + F(\text{page-marker})$	token-based	negative/positive
Morphological Family Size	Number of different words that contain the same stem (excluding inflectional variants)	$Vf(\text{page}) = \lvert\text{page, paginate, pagination,}$ $\text{, frontpage, page-marker, title-page}\rvert = 6$	type-based	negative
Positional Family Size	Number of compounds containing the same left or right constituent.	$Vf_{left}(\text{page}) = \lvert\text{page-marker}\rvert = 1$ $Vf_{right}(\text{page}) = \lvert\text{title-page, frontpage}\rvert = 2$	type-based	negative
Positional Cumulative Root Frequency	Summed frequency of compounds containing the same left or right constituent.	$Nf_{left}(\text{page}) = F(\text{page-marker})$ $Nf_{right}(\text{page}) = F(\text{title-page}) + F(\text{frontpage})$	token-based	negative

Kostić (Kostić, 1991, 1995, 2003; Kostić et al., in press) addressed the relevance of token- and type-based counts for inflectional processing using an information theoretical framework. In a series of lexical decision experiments with Serbian inflected nouns, he demonstrated that the amount of information carried by an inflected noun form is inversely proportional to its processing latency. The amount of information, however, was not derived from the token counts alone but from the ratio of the surface frequency to the type count of syntactic functions and meanings carried by an inflected form within a given paradigm (e.g., feminine singular nouns). For instance, the Serbian noun *voda* (*water*) has the inflectional variants *"voda"*,[2] *"vode"*, *"vodi"*, *"vodu"*, *"vodom"*, and *"vodama"*, each occurring with a particular frequency in a large corpus. One can estimate the probability of each of the inflected forms (e.g., *"vode"* within the inflectional paradigm of *voda*), by calculating its inflectional ratio ($p(\text{“vode”}|\text{voda}) \simeq F(\text{“vode”})/F(\text{voda})$), where $F(\text{voda}) = F(\text{“voda”}) + \ldots + F(\text{“vodama”})$. Each inflected form of *voda* is used to express different combinations of cases and numbers, for instance, *"vode"* can be used to express the genitive singular and the nominative or accusative plurals of *voda*. In turn, each Serbian case can be used in a variety of syntactic and semantic contexts, which leads to some ambiguity in the interpretation of an isolated inflected form. If we denote by $N_s(w)$ the number of syntactic functions and meanings expressed by a form w, then, the average probability of each form within an inflectional paradigm would be the quotient between the probability of the inflected form, and the number of functions it can convey, in the case of *"vode"* the estimated average probability of each particular usage would be $p(\text{“vode”}_i|\text{voda}) = p(\text{“vode”}|\text{voda})/N_s(\text{“}vode''})$. Kostić showed that the response latencies to each inflected form (w) belonging to an inflectional paradigm ($\mathcal{P}$) is inversely

which is the cumulative root frequency minus the base frequency of a word. However, since the frequency of the word is partialled out in all analyses, we will consider cumulative root frequency as equivalent to family frequency. In the analyses, we report the best result from using either of the counts, referring to both of them as cumulative root frequency.

[2] Throughout this paper we will use double quotes to refer to surface forms.

proportional to:

$$I_m(w) = -\log_2 \frac{p(w_i|\mathcal{P})}{\sum_{v \in \mathcal{P}} p(v_i|\mathcal{P})} = \log_2 \sum_{v \in \mathcal{P}} p(v_i|\mathcal{P}) - \log_2 p(w_i|\mathcal{P}). \tag{1}$$

This predictor accounted for almost all processing variability of inflected noun forms of all three Serbian grammatical genders. The first term in (1) provides an estimation of the information content of the whole inflectional paradigm, by summing the individual amounts of information of each of the paradigm members, while the second term estimates the amount of information contained by a particular individual form.

The previous results seem to indicate that inflectional and derivational paradigms affect the recognition of a word in different ways. While for inflectional paradigms one should consider the summed frequency of all the members of the paradigm (base frequency), or the information content of the paradigms themselves, both of which are mainly token-based counts, it appears that the influence of derivational paradigms is best quantified by morphological family size, a completely type-based count, with only a small additional effect of token-based cumulative root frequency. This picture is further complicated by compound words. De Jong, Feldman, Schreuder, Pastizzo, and Baayen (2002) reported that the morphological paradigm of a compound is only formed by those compounds that share the right constituent as a right constituent or the left constituent as a left constituent. This claim is based on analyses where they observed that it was the positional family size and positional cumulative root frequency (see Table 1), that is, restricted to the paradigm members in which the right constituent of the compounds appears as a right constituent, or left constituent appears as a left constituent, that best accounted for response latencies.

The question arises whether the two classes of counts (i.e., type-based and token-based) tap into different properties of the cognitive system, or reflect aspects of a single process. In this study, we develop a measure that reconciles the apparently contradictory findings as to the respective predictive powers of each of the counts across various types of words, including monomorphemic words, polymorphemic words, and compound words. In this way, our measure offers a parsimonious treatment of inflectional, derivational, and compounding relations, using a single measure, therefore eliminating the need for different cognitive processes dealing with each of these types of relations.

In what follows, we begin by formulating the information residual measure and discussing its relationship to different counts that have been described in the literature. We continue by evaluating the performance of this measure in predicting the response latencies of three previously published Dutch visual lexical decision experiments, as compared to using the traditional type and token frequency counts. Finally, we conclude with an outline of the implications of this measure for models of lexical processing.

Information Residual of a Word

In this section we provide a general formulation of the information residual of a word. Our formulation will be guided by the different morphological effects (see Table 1) that have been shown to affect response latencies in visual lexical decision.

Amount of Information Contained by a Surface Form

The surface frequency of a word can be expressed by Shannon's amount of information, that is, the minimum number of bits that would be necessary to encode the word in an optimal binary coding of all the words in the lexicon (Shannon, 1948; for an extensive introduction to information theory see, e.g., Cover & Joy, 1991). In this way the *amount of information*$(I_s(w))$ of a surface form w, with a frequency $F(w)$ in a corpus of size N is:

$$I_s(w) = -\log_2 p(w) \simeq -\log_2 \frac{F(w)}{N}, \tag{2}$$

where $p(w) \simeq F(w)/N$ is the probability of encountering w in a corpus. Note that, according to (2), the amount of information is inversely proportional to the log frequency of the word. As it has been established that logarithmic frequency correlates negatively with lexical decision latencies (Rubenstein & Pollack, 1963; Scarborough, Cortese, & Scarborough, 1977; Shapiro, 1969), the amount of information of a word should show a positive correlation with lexical decision latencies. When, as Kostić did in his experiments, we restrict our experimental items to members of a particular paradigm, the second term of (1) is equivalent to the information content of a word as described in (2), with an additional weighting by the number of syntactic functions and meanings. For simplicity reasons, in the present study we will consciously overlook the effect of the number of syntactic functions and meanings (it is not clear how such estimate should be calculated for derivational forms). However, future refinements of our technique should take this factor into account.

Morphological Paradigms

Kostić (2003) showed that the average amount of information of an inflectional paradigm (e.g., feminine nouns) is inversely related to the processing speed of the individual inflected forms that constitute the paradigm – high average amounts of information are paralleled by shorter processing speed per one bit in a given experiment. In order to estimate the amount of information of an inflectional paradigm (i.e., the first term in (1)), instead of the plain sum proposed by Kostić, we propose a more standard informational measure, the entropy (Shannon, 1948) of the paradigm. We can consider the inflectional paradigm of a word to be a random variable whose possible values are the different inflected forms that a base word can take. Hence, we can calculate the entropy of the inflectional paradigm, its *inflectional entropy.* In general, the entropy of a paradigm $\mathcal{P}$ with $V(\mathcal{P})$ members $\{x_1, \ldots, x_{V(\mathcal{P})}\}$, each of which has a probability of occurrence of $p(x|\mathcal{P}) \simeq F(x)/F(\mathcal{P})$, is:

$$H(\mathcal{P}) = -\sum_{x\in\mathcal{P}} p(x|\mathcal{P}) \log_2 p(x|\mathcal{P}) \simeq -\sum_{x\in\mathcal{P}} \frac{F(x)}{F(\mathcal{P})} \log_2 \frac{F(x)}{F(\mathcal{P})}, \tag{3}$$

where $F(\mathcal{P})$ is the base frequency of the inflectional paradigm, and $F(x)$ is the surface frequency of the word. Note that this measure is related to the base frequencies and inflectional ratios (the inflectional entropy of a paradigm is the weighted sum of the inflectional ratios of its members). Therefore, we predict that the inflectional entropy will correlate negatively with response latencies, in line with the negative correlations with response latencies found for base frequency and inflectional ratios.

For example, the inflectional paradigm of the base form *car* consists of the forms "car" and "cars" with surface frequencies $F(\text{“car”})$ and $F(\text{“cars”})$. The base frequency of the inflectional paradigm *car* is $F(\text{car}) = F(\text{“car”}) + F(\text{“cars”})$ and the probabilities of the inflected form being "car" or "cars" within the inflectional paradigm of *car* are $p(\text{“car”}) = F(\text{“car”})/F(\text{car})$ and $p(\text{“cars”}) = F(\text{“cars”})/F(\text{car})$. The entropy of the inflectional paradigm of *car* will then be $H(\text{car}) = -p(\text{“car”}) \log_2 p(\text{“car”}) - p(\text{“cars”}) \log_2 p(\text{“cars”})$.

We can also conceive of derivational paradigms as random variables, for which we can calculate the entropy. Once again, we can estimate the probability of occurrence of each of word inside the paradigm by its derivational ratio: the word's base frequency divided by the summed frequency of all members of the paradigm (the cumulative root frequency), and then we can calculate the entropy according to (3).

As one can deduce from (3), in general, the greater the number of members in a morphological paradigm, the greater the entropy of the paradigm will tend to be. All else being equal, increasing the number of members will decrease their probability of occurrence within the paradigm, thus increasing the number of bits required to represent each of them. In fact, family size gives us an estimate of the maximum entropy situation (i.e., the case when all paradigm members have the same probability of occurrence). Figure 1 compares the number of bits of entropy in the paradigm for all Dutch monomorphemic words in the CELEX Lexical Database (Baayen et al., 1995) in the family size range $[1, 200]$, with the maximum entropy for words with that morphological family size. The fact that both curves follow very similar trajectories ensures that both measures should correlate similarly with reaction times.

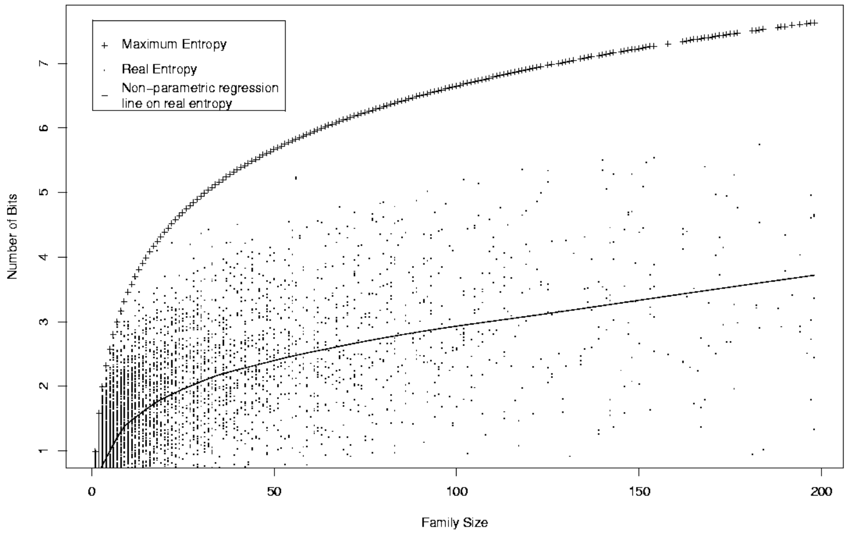

Figure 1. Comparison between real and maximum entropy for all monomorphemic Dutch words in the family size range $[1, 200]$.

To illustrate the way in which the derivational entropy measure also accounts for the inhibitory effects of cumulative root frequency after having partialled out the effect of family size, consider Figure 2. It compares the frequency histograms for the derivational paradigms of the Dutch words *barbaar* (*barbarian*), that includes the words *barbaars* (*barbarous*), *barbaarsheid* (*barbarity*), *barbarisch* (*barbaric*), and *cultuurbarbaar* (*cultural barbarian*), with the paradigm of *faam* (*fame*), including *befaamd* (*famed*), *wereldfaam* (*world fame*), *wijdbefaamd* (*widely famed*), and *befaamdheid* (*'famedness'*). These two words are taken from the dataset where Baayen and colleagues found the inhibitory effect of cumulative root frequency (family frequency). Both words have five members in their derivational paradigms, and both have similar base frequencies (228 for *barbaar* and 218 for *faam*).[3] However, the cumulative root frequency (obtained by summing the frequencies of all the paradigm members) is greater in the paradigm of *faam* (676) than in the paradigm of *barbaar* (566). The difference in the frequencies of the members of the paradigm of *faam* with respect to the frequencies of those of the paradigm of *barbaar* is not constant however. Note that it consists of a very strong increase in the frequency of the most frequent paradigm member, together with smaller increases (in this case even decreases) in the frequencies of the other members. This is consistent with the Zipfian distributions of morphological paradigms. Importantly, the more skewed distribution of frequencies in the paradigm with the greater cumulative root frequency produces a lower entropy value for that paradigm. We can now calculate the entropies of both paradigms according to (3),[4] by substituting in the equation the probability of each item for the quotient of its (base) frequency and the summed frequency of all paradigm members (e.g., $p(\text{barbaar}|[\text{barbaar}]) = 228/566$).[5] Correspondingly, the entropy of the paradigm of *faam* ($H(\text{faam}) = 1.13$) is slightly lower than that of the paradigm of *barbaar* ($H(\text{barbaar}) = 1.36$). As the entropy of the paradigm is negatively correlated with response latencies, the result is that words from the paradigm of *faam* receive less facilitation from their derivational paradigm than the words from the paradigm of *barbaar*, so that their response latencies are longer.

Hierarchical Structure of the Paradigms

Inflectional paradigms are nested within derivational paradigms, which in turn are nested

[3] The frequency counts are based on a corpus of 42 million words.

[4] In order to avoid taking the logarithm of zero in the entropy calculations, we added 0.1 to the frequency of the items that were listed in CELEX as having a frequency of zero. As long as the magnitude added is significantly smaller than one, its actual value does not affect the main effects reported here.

[5] We will use square brackets to refer to derivational paradigms.

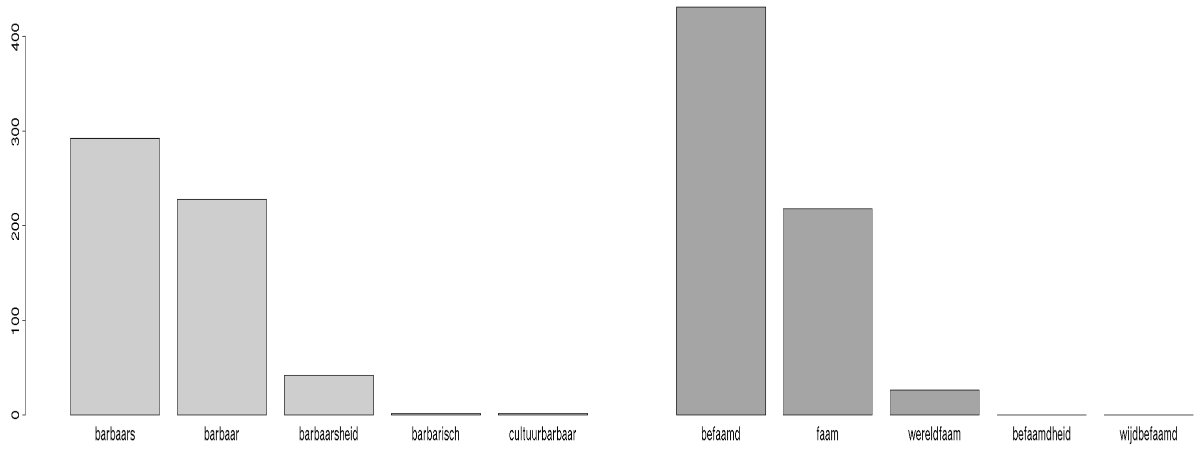

Figure 2. Comparison of the frequencies of the members of the derivational paradigms of the Dutch words *barbaar* and *faam*.

within each other so as to form tree-like structures. This hierarchical structure implies that a single word can belong simultaneously to several morphological paradigms. For instance, Figure 3 shows the morphological paradigms that include the polymorphemic word “thinkers”. “Thinkers” and its singular form “thinker” belong to the inflectional paradigm of *thinker*, which in turn belongs to the derivational paradigm of [thinker]. [thinker] itself is a paradigm within the larger paradigm of [think], which includes other elements such as [rethink] as well as the inflectional paradigm of *think*.

Because morphological paradigms are nested, we need to specify the joint entropy of the paradigms to which a given word belongs. In general, except for compound words, the joint entropy of the morphological paradigms can be calculated as the sum of the entropies of all the morphological paradigms that dominate it in the tree. This is because the paradigms are nested within each other. In this case the joint probability of the nested paradigms is the product of the probabilities of each of them and, in a logarithmic scale, this product converts to a sum. We will refer to the *paradigmatic entropy* ($H_{tot}(w)$) of a word as the joint entropy of all the morphological paradigms that dominate it in the morphological tree.

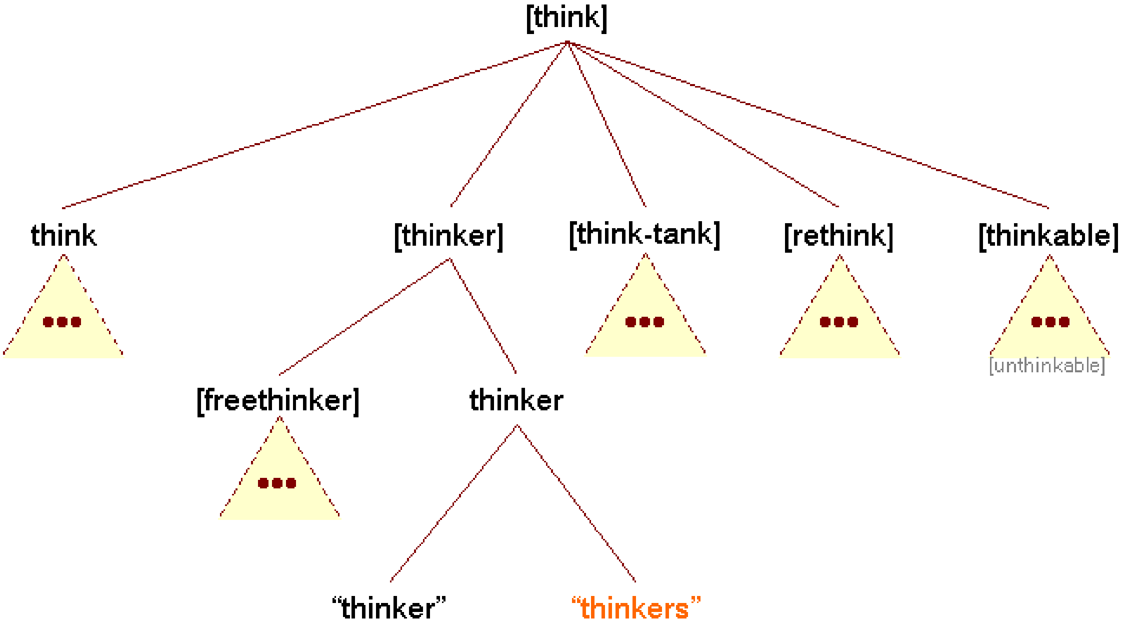

Figure 3. Morphological paradigms for the polymorphemic word *“thinkers”*.

For instance, in the above example, the joint entropy of the paradigms to which “thinkers” belongs can be calculated as the sum of the individual entropies of the paradigms:

$$\begin{aligned} H(\mathit{thinker}, [\text{thinker}], [\text{think}]) &= H([\text{think}]) + H([\text{thinker}] | [\text{think}]) \\ &\quad + H(\text{thinker} | [\text{thinker}], [\text{think}]) \\ &= H(\text{thinker}) + H([\text{thinker}]) + H([\text{think}]). \end{aligned} \tag{4}$$

Compound Words

Figure 4a shows the typical paradigmatic structure of a compound. In this case, to calculate the joint paradigmatic entropies of the levels that dominate [think-tank] is problematic because the compound has two direct ancestors. We cannot calculate the joint entropy of the two paradigms by addition – considering the two paradigms of [think] and [tank] separately and adding them up. The members of the paradigms of [think] and [tank] are mutually exclusive except for [think-tank] itself. This means that we can consider the union of the [think] and [tank] paradigms as a single random variable, as shown in Figure 4b, and then simply calculate the entropy according to (3).

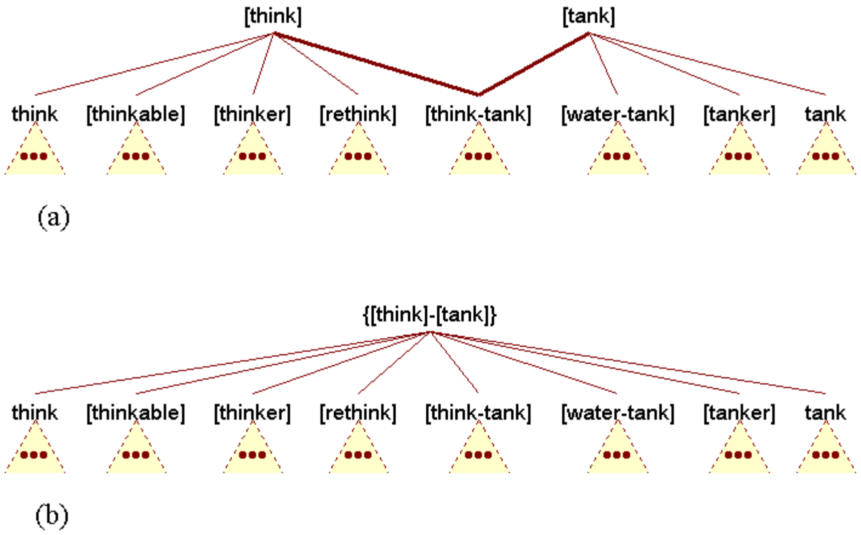

Figure 4. (a) Derivational paradigms of [think] and [tank] (b) Derivational paradigm of $\{[\text{think}], [\text{tank}]\}$.

With respect to the positional restriction on compound families suggested by De Jong et al. (2002), an analysis of the distribution of the cumulative root frequencies of the left and right constituents of the compounds in their study revealed that, in their experimental dataset, left constituents tend to appear as left constituents in other compounds more often than they appear as right constituents in other compounds (paired $t = -7.8203,\ df = 109,\ p < 0.0001$), and vice-versa , right constituents tend to appear as right constituents in other compounds (paired $t = 5.0184,\ df = 111,\ t < 0.0001$). Hence, the positional family effects that they reported may arise from the inner distribution of the families.

Putting the Bits Together

At the beginning of this section, we showed how surface frequency can be described as the amount of information contained by a word, $I_s(w)$. On the one hand, $I_s(w)$ gives us an estimate of how difficult it is to recognize a word by itself. The greater the amount of information contained by the word, the more costly it will be to recognize. On the other hand, the total paradigmatic entropy $H_{tot}(w)$ gives us an estimate of how much support that word receives from the morphological paradigms to which it belongs. As it was done in Kostić's original approach described by (1), both the cost of recognizing a word (I_s) and the support provided by its morphological paradigms (H_{tot}) are measured in logarithmic scale (bits), and we can express the quotient between them and as a difference, *the information residual of a word*:

$$I_R(w) = I_s(w) - H_{tot}(w). \tag{5}$$

Note that we predict that the information residual will be positively correlated with processing latencies. This is a direct consequence of I_s being positively correlated with response latencies, and H_{tot} being negatively correlated with response latencies.

Re-analyses of Previously Published Experiments

In this section, we compare the performance of our information residual measure in predicting response latencies in visual lexical decision experiments on the one hand, to the performance of the traditional measures (surface and base frequency, family size, and cumulative root frequency) on the other.

Methods

Materials: We obtained the three datasets from previously published visual lexical decision experiments on monomorphemic words (Schreuder & Baayen, 1997; Experiment 3), polymorphemic words (Neijt, Schreuder & Baayen, 2003), and compounds (De Jong et al., 2002; Experiment 1b). The experiment on monomorphemic words contrasted morphological family size, controlling for frequency and cumulative root frequency. The experiment on polymorphemic words consisted of Dutch words with different feminine agentive suffixes (such as the *-ess* in English *countess*), and the researchers investigated the effects of the different origins of the suffixes (Latinate or Germanic) on visual word recognition. As this distinction goes beyond the scope of this paper, we will partial out that variance by adding an additional variable "suffix type" into our regression analyses. Finally, the compound word dataset was an experiment in which the researchers used compounds with constant left constituents that were controlled for base frequency and word length in letters (of the right constituent and the compounds), and they contrasted the morphological family size of the right contituents of the compounds.

In order to standardize measures across all experiments, we recalculated all frequency, family size, and cumulative root frequency measures using the morphological parses and word frequency counts available in the CELEX lexical database (Baayen et al., 1995). In order to provide the maximum possible accuracy for items where the CELEX parsing was erroneous according to standard dictionary-based parses, we corrected these parses by hand. We calculated the information residual value for all the words in the three datasets using the method described in the previous section. Before performing any analyses we excluded from the datasets all trials and items with reaction times above or below two and a half standard deviations from the mean (in logarithmic scale).

Procedure: We fitted two multiple regression models (each including a by-participant and a by-item regression) to each of the three datasets. One of the regressions had the logarithms of the traditional counts (word frequency, morphological family size, and cumulative root frequency) as independent variables, while the other had only the information residual as an independent variable. Both regressions had the logarithm of the response latencies as the dependent variable.[6] The regressions on the polymorphemic word dataset (Neijt et al., 2003) had suffix type as an additional independent variable, whose values were the different feminine agentive suffixes present in the experimental items.

Results and Discussion

In all cases, we report sequential analyses of variance on by-participant multi-level regression analyses (Alegre & Gordon, 1999; Baayen et al., 2002; Pinheiro & Bates, 2000; Table 2) and traditional by-item linear regression models (Table 3).

The first three columns of Tables 2 and 3 present the results of the by-participant and by-item analyses on each of the datasets. The upper section of the tables summarizes the analyses using the traditional counts, while the middle section contains the results of the analyses using the information residual measure. In all analyses, the effects are reported in the order in which they were entered into the regressions, so the significance of each of the effects is calculated after partialling out the contribution of the effects reported above it. The signs between brackets (+ or −) represent the direction of the effects (the sign of their coefficient in the models), for the effects that were significant in each of the analyses. The bottom row of the tables compares the amount of variance explained by the regression using the information residual, with the amount of variance explained by the traditional type- and token-based counts. Additionally, in the bottom row of Table 3, we report analyses of variance testing whether the models are significantly different in terms of explained variance (it is not possible to do this on the by-participant multilevel regressions).

[6] We use the logarithm of the response latencies to avoid the large deviations from normality that are present in reaction times, that would violate the sphericity conditions of the regression analyses.

Table 2

Summary of by-participant multi-level regressions. Freq. refers to frequency, Cum. Freq. refers to Cumulative root frequency, Fam. Size to morphological family size, I_R to the information residual, I'_R to the modified information residual, and Suffix to the effects of suffix type. The signs in brackets in front of the effect represent the direction of the effects. Significance codes are: $^+ : p < 0.1$, $^* : p < 0.05$, $^{**} : p < 0.005$, and $^{***} : p < 0.0005$

	Schreuder & Baayen, 1997 – Exp.3	Neijt et al., 2003	De Jong et al., 2003 – Exp.1b	
Traditional	(−) Freq.: $F(1,956) = 562.57^{***}$	(−) Freq.: $F(1,1036) = 187.12^{***}$	(−) Freq.: $F(1,1287) = 97.39^{***}$	
analyses	(−) Fam. Size: $F(1,956) = 129.13^{***}$	(−) Fam. Size.: $F(1,1287) = 39.73^{***}$	(−) Pos. Fam. Size.: $F(1,1287) = 33.69^{***}$	
	(+) Cum. Freq.: $F(1,956) = 68.29^{***}$	Cum. Freq.: $F < 1$	(+) Pos. Cum. Freq.: $F(1,1287) = 5.90^{*}$	
		Suffix: $F(10,1036) = 2.66^{**}$		
Explained variance (r^2)	44%	47%	39%	
I_R analyses	(+) I_R: $F(1,958) = 923.37^{***}$	(+) I_R: $F(1,1037) = 212.25^{***}$	(+) I_R: $F(1,1289) = 60.96^{***}$	(+) I'_R: $F(1,1289) = 168.49^{***}$
		Suffix: $F(10,1037) = 2.90^{**}$		
Explained variance (r^2)	48%	47%	35%	40%
Comparison of models	+4%	0%	−4%	+1%

Table 3

Summary of by-item regressions. Freq. refers to frequency, Cum. Freq. refers to Cumulative root frequency, Fam. Size to morphological family size, I_R to the information residual, I'_R to the modified information residual, and Suffix to the effects of suffix type. The signs in brackets in front of the effect represent the direction of the effects. Significance codes are: $^{+}$: $p < 0.1$, * : $p < 0.05$, ** : $p < 0.005$, and *** : $p < 0.0005$

	Schreuder & Baayen, 1997 – Exp.3	Neijt et al., 2003	De Jong et al., 2003 – Exp.1b	
Traditional	(−) Freq.: $F(1,33) = 18.10^{***}$	(−) Freq.: $F(1,37) = 52.87^{***}$	(−) Freq.: $F(1,108) = 18.57^{***}$	
analyses	(−) Fam. Size: $F(1,33) = 4.40^{*}$	(−) Fam. Size: $F(1,37) = 8.19^{**}$	(−) Pos. Fam. Size: $F(1,108) = 7.54^{**}$	
	Cum. Freq.: $F(1,32) = 2.31$	Cum. Freq: $F < 1$	Pos. Cum. Freq.: $F < 1$	
		Suffix: $F < 1$		
Explained variance	37%	60%	18%	
(adjusted r^2)				
I_R analyses	(+) I_R: $F(1,34) = 32.92^{***}$	(+) I_R: $F(1,38) = 58.09^{***}$	(+) I_R: $F(1,109) = 13.95^{***}$	(+) I'_R: $F(1,109) = 39.26^{***}$
		Suffix: $F < 1$		
Explained variance	48%	59%	11%	26%
(adjusted r^2)				
Comparison of models	+11%, $F(33,34) = 4.80^{*}$	−1%, $F(38,39) = 1.78$	−7%, $F(108,109) = 10.89^{**}$	+8%, $F(108,109) = 9.40^{**}$

Results of the analyses using the traditional counts, reveal a facilitatory effect of frequency in all three experiments. Next, in all datasets, we also find a facilitatory effect of morphological family size. Finally the effect of cumulative root frequency and positional family size only reaches significance in the by-participant regressions (except in the second dataset, where it does not reach significance at all). Interestingly, as we predicted in the previous section, once the facilitatory effect of morphological family size has been partialled out, cumulative root frequency, when present, shows an inhibitory effect.

In contrast to the different effects found across the experiments in the analyses that use the traditional counts, we find that the information residual count has a consistent inhibitory effect in all experiments. The consistency of the information residual effect represents an advantage over the changing effects of the traditional counts. Moreover, if we compare the amount of variance explained by the information residual with the variance explained by the traditional counts, we find that the information residual analyses significantly outperforms the traditional counts in the first dataset, and that its predictive power is not significantly different for the second dataset.

By contrast, in the case of the compounds, the traditional counts clearly outperform the information residual analyses in terms of explained variance. This underperformance for the compounds led us to investigate in more detail the contribution to the effect of each of the components that form the information residual, that is, the information content of the surface form, and the paradigmatic entropies at the different levels. Towards this end, we fitted regression models with log reaction time as the dependent variable and amount of information of the word and the paradigmatic entropies at the different levels as the independent variables. This decomposition showed that the amount of information of the word and the paradigmatic entropies up to and including the joint entropy between the two constituents were showing effects in the predicted direction (inhibitory for the amount of information and facilitatory for the paradigmatic entropies). However, the paradigmatic entropies of the levels in the tree higher than the node where the joint entropy of the compound constituents is calculated showed an inhibitory effect, opposite to the direction we had predicted. This inhibitory effect was significant both in the by-participant ($F(1, 1288) = 60.45, p < 0.0001$) and in the by-item regressions ($F(1, 108) = 12.26, p = 0.0007$). The inhibitory effect of these 'upper' paradigms of a compound suggests that a refinement in our description of the total paradigmatic entropy of a word is required.

The members of the paradigms at the higher tree levels for compounds tend to show much weaker semantic relations to the target than do the members of the more immediate paradigms that are dominated by the level at which the constituents of the compound are joined. The presence of such 'upper' paradigms produces heterogeneity with respect to the meanings of the paradigm members. For instance, in our dataset at these levels we find words that bear very distant semantic relations to the targets, such as *vangen* (*to catch*) in the paradigm of *gevangenispsychiater* (*prison psychiatrist*), or *boter* (*butter*) in the paradigm of *avondboterham* (*evening sandwich*). It has been shown that the effects of morphological paradigms arise as a consequence of the semantic relations that bind the members of a morphological paradigm (cf. De Jong, 2002). Interestingly, Moscoso del Prado Martín (2003) reports an inhibitory effect for semantically distant family members in Hebrew morphological families whose members belong to heterogeneous semantic fields. These findings suggest that the contribution of the morphological paradigms to the I_R measure should take semantic relatedness into account. In our approach we approximate this by dividing the paradigmatic entropy measure into two separate components. On the one hand there is the joint paradigmatic entropy of the paradigms whose members bear very close semantic relations to the target ($H_{related}(w)$). The information in these paradigms provides support for the recognition of the word. On the other hand, the joint entropy of more distant paradigms ($H_{unrelated}(w)$) not only fails to support the recognition of the word, but rather makes it more difficult.

Introducing this change in the information residual measure from (5) we obtain:

$$I_R(w) = I_s(w) - H_{related}(w) + H_{unrelated}(w). \tag{6}$$

We recalculated the value of the information residual of the compounds according to (6). For compounds with at least one polymorphemic constituent, we considered the paradigms above the node where the compounds constituents are joined to be semantically distant, and that node and the levels that it dominates were considered as semantically close. Note here that, for simplicity reasons, we have chosen the 'upper' morphological paradigms to be semantically distant, while the paradigms below the node joining both constituents of the compound are considered semantically close. A more realistic approach would require a less clear-cut boundary, but would instead weight the contribution of each paradigm by an estimate of its semantic relatedness to the target. This is left for future research.

The fourth columns in Tables 2 and 3 report the results of the regression analyses including this modification to the information residual measure. Observe that the information residual measure (labelled I'_R in the tables to distinguish it from the previous value) now significantly outperforms the traditional measures in terms of explained variance.

General Discussion

In this paper we have shown that the effects on visual lexical decision response latencies attributed to frequency counts such as surface frequency and base frequency, inflectional ratio, cumulative root frequency, and morphological family size, can be accounted for in a more parsimonious manner using the information residual of a word. This is a measure of the cost of recognizing a word, considering the decreases and increases in uncertainty contributed by the morphological paradigms to which the word belongs.

The information residual measure performs at least as well as a combination of the traditional counts in predicting response latencies in the visual recognition of monomorphemic words, polymorphemic words, and compounds. Its relevance is that no other single measure can quantify the processing cost of the wide range of different types of words we have studied (monomorphemic, polymorphemic, and compounds). In summary, the use of the information residual measure presents several theoretical and practical advantages. First, on the practical side, the traditional frequency counts are strongly correlated with one another. These high degrees of collinearity pose problems in any regression model that includes several of those counts. More specifically, high collinearity makes regression models unreliable, as it makes it difficult to assess the real magnitude and direction of an effect (Belsley, 1991). By contrast, the present approach combines the traditional counts into a single measure, thereby avoiding the collinearity problem.

On the theoretical side, an important advantage of the information residual measure relative to the traditional counts is that it provides a parsimonious description of inflectional and derivational paradigms. In our approach, the effects on response latencies of the different counts, such as type-based morphological family size for derivational paradigms and token-based base frequency for inflectional paradigms, arise as by-products of the quantitative differences between the statistical distributions within the paradigms. Also, our uniform account of nested paradigm structures offers a straightforward approach that can be applied to languages in which even inflectional paradigms are nested within each other. This is the case in strongly inflectional languages such as those of the Romance and Slavic families, or in agglutinative languages such as Finnish.

Our approach is consistent with frequency counts reflecting informational complexity as proposed by McDonald and Shillcock (2001). They report that a measure they call Contextual Distinctiveness is a better predictor of reaction times than is word frequency. Interestingly, their measure is based on the entropy of the distribution of the different contexts in which a word is used, what we could call its syntagmatic entropy. Our intuition is that this measure is highly related to the first component of the information residual, the amount of information contained by the word. It will be interesting to investigate whether that component can be substituted for contextual distinctiveness in order to express the complexity of recognizing a word in terms of a combination of its paradigmatic and syntagmatic entropies.

The effect of the information residual of a word has important consequences for theories of morphological processing. First, and most obvious, the information residual is a purely probabilistic measure, and reflects the extreme sensitivity of the human language processing system to stochastic factors. A second consequence comes from the markedly hierarchical nature of morphological paradigms. Namely, the effect is not driven by individual relations between pairs of words, but by more structural relations between hierarchically organized paradigms. The third consequence arises from the sensitivity to graded semantic similarity between paradigms, as indicated by the opposite influences of the morphologically closer and more distant paradigms to which a compound word belongs. This is also in accordance with the converging evidence for the graded influence of semantic similarity in inflectional (Baayen & Moscoso del Prado Martín, 2003; Kostić et al., in press; Ramscar, 2002) and derivational processes (e.g., De Jong, 2002; Feldman, Barac-Cikoja, & Kostić, 2002; Feldman, Pastizzo, Soltano, & Francis, in press; Moscoso del Prado Martín, 2003). These properties emerge naturally in distributed connectionist models of lexical processing (e.g., Gaskell & Marslen-Wilson, 1997; Plaut & Booth, 2000; Plaut & Gonnerman, 2000; Seidenberg & Gonnerman, 2000), which view morphological relations as the simultaneous effect of similarity between distributed representations of form and meaning. In this respect, McKay (2003) indicates that information theory is indeed the apropriate tool for analyzing the behavior of a neuronal network. Additionally, Moscoso del Prado Martín (2003) shows how paradigmatic effects of the kind reported in this paper emerge naturally in a connectionist model that is trained to map distributed representations of form into distributed representations of meaning.

In summary, our approach succeeds in extending the ideas of Kostić (Kostić, 1991; 1995; 2003; Kostić et al., in press) from inflectional to derivational morphology and compounds while at the same time using a more standard measure of the amount of information carried by a paradigm. Nevertheless, further refinements of the technique are anticipated to account for the fine-grained sensitivity to semantic relatedness as revealed by the semantic constraints on morphological family size and inflectional morphology. Additionally, as we mentioned before, the contribution of each morphological sub-paradigm to the information residual of a word should ideally be defined as a continuous function of its semantic relatedness and form similarity to the target.

Acknowledgement

The authors are indebted to Jen Hay, Laurie Feldman, Stefan Frank, Stewart Robinson, Rob Schreuder and three anonymous reviewers for helpful comments and suggestions on previous versions of this manuscript.

This work was supported by the Dutch Research Council (NWO) through a Pionier grant to the third author. The first author received additional support from the Medical Research Council (U.K.) and from the European Community under the "Information Society Technologies Programme" (IST-2001-35282).

References

Alegre, M., Gordon, P., 1999. Frequency effects and the representational status of regular inflections. Journal of Memory and Language 40, 41–61.

Baayen, R. H., Lieber, R., Schreuder, R., 1997. The morphological complexity of simplex nouns. Linguistics 35, 861–877.

Baayen, R. H., Moscoso del Prado Martín, F., 2003. Questioning the unquestionable: Semantic density and past-tense formation in three Germanic languages. Manuscript submitted for publication, Max Planck Institute for Psycholinguistics .

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

Baayen, R. H., Tweedie, F. J., Schreuder, R., 2002. The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. Brain and Language 81, 55–65.

Belsley, D. A., 1991. Conditioning Diagnostics: Collinearity and Weak Data in Regression. Wiley, New York.

Colé, P., Beauvillain, C., Segui, J., 1989. On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. Journal of Memory and Language 28, 1–13.

Cover, T. M., Joy, A. T., 1991. Elements of Information Theory. John Wiley & Sons, New York.

De Jong, N. H., 2002. Morphological Families in the Mental Lexicon. MPI Series in Psycholinguistics. Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., Baayen, R. H., 2002. The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects. Brain and Language 81, 555–567.

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

Feldman, L. B., Barac-Cikoja, D., Kostić, A., 2002. Semantic transparency influences morphological processing. Memory and Cognition 30 (4), 629–636.

Feldman, L. B., Pastizzo, M., Soltano, E., Francis, S., in press. Semantic transparency influences morphological processing. Brain and Language .

Ford, M. A., Marslen-Wilson, W. D., Davis, M. H., in press. Morphology and frequency: Contrasting methodologies. In: Baayen, R. H., Schreuder, R. (Eds.), Morphological structure in language processing. Mouton de Gruyter, Berlin.

Gaskell, M. G., Marslen-Wilson, W., 1997. Integrating form and meaning: A distributed model of speech perception. Language and Cognitive Processes 12, 613–656.

Hay, J., 2001. Lexical frequency in morphology: Is everything relative? Linguistics 39, 1041–1070.

Kostić, A., 1991. Informational approach to processing inflected morphology: Standard data reconsidered. Psychological Research 53 (1), 62–70.

Kostić, A., 1995. Informational load constraints on processing inflected morphology. In: Feldman, L. B. (Ed.), Morphological Aspects of Language Processing. Lawrence Erlbaum Inc. Publishers, New Jersey.

Kostić, A., 2003. The effects of the amount of information on processing of inflected morphology. Manuscript submitted for publication, University of Belgrade .

Kostić, A., Marković, T., Baucal, A., in press. Inflectional morphology and word meaning: orthogonal or co-implicative domains? In: Baayen, R. H., Schreuder, R. (Eds.), Morphological structure in language processing. Mouton de Gruyter, Berlin.

Lüdeling, A., De Jong, N. H., 2002. German particle verbs and word-formation. In: Dehé, N., Jackendoff, R., McIntyre, A., Urban, S. (Eds.), Verb-particle explorations. Mouton de Gruyter, Berlin, pp. 315–333.

McDonald, S., Shillcock, R., 2001. Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. Language and Speech 44, 295–323.

McKay, D. J., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, U.K.

Moscoso del Prado Martín, F., 2003. Paradigmatic Effects in Morphological Processing: Computational and cross-linguistic experimental studies. MPI Series in Psycholinguistics. Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., Baayen, R. H., to appear. Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. Journal of Experimental Psychology: Learning, Memory, and Cognition .

Neijt, A., Schreuder, R., Baayen, R. H., 2003. Verpleegsters, ambassadrices, and masseuses. stratum differences in the comprehension of Dutch words with feminine agent suffixes. In: Cornips, L., Fikkert, P. (Eds.), Linguistics in the Netherlands 2003. Benjamins, Amsterdam, pp. 117–127.

Pinheiro, J. C., Bates, D. M., 2000. Mixed-effects models in S and S-PLUS. Statistics and Computing. Springer, New York.

Plaut, D. C., Booth, J. R., 2000. Individual and developmental differences in semantic priming: Empirical and computational support for a single mechanism account of lexical processing. Psychological Review 107, 786–823.

Plaut, D. C., Gonnerman, L. M., 2000. Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? Language and Cognitive Processes 15 (4/5), 445–485.

Ramscar, M., 2002. The role of meaning in inflection: Why the past tense doesn't require a rule. Cognitive Psychology 45, 45–94.

Rubenstein, H., Pollack, I., 1963. Word predictability and intelligibility. Journal of Verbal Learning and Verbal Behavior 2, 147–158.

Scarborough, D. L., Cortese, C., Scarborough, H. S., 1977. Frequency and repetition effects in lexical memory. Journal of Experimental Psychology: Human Perception and Performance 3, 1–17.

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

Seidenberg, M. S., Gonnerman, L. M., 2000. Explaining derivational morphology as the convergence of codes. Trends in Cognitive Sciences 4 (9), 353–361.

Shannon, C. E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423.

Shapiro, B. J., 1969. The subjective estimation of word frequency. Journal of verbal learning and verbal behavior 8, 248–251.

Taft, M., 1979. Recognition of affixed words and the word frequency effect. Memory and Cognition 7, 263–272.

Whaley, C. P., 1978. Word-nonword classification time. Journal of Verbal Language and Verbal Behavior 17, 143–154.