Paradigms bit by bit: an information-theoretic approach to the processing of paradigmatic structure in inflection and derivation

Petar Milin* [a,e], Victor Kuperman [b], Aleksandar Kostić [c,e] and R. Harald Baayen [d]

[a] *Department of Psychology, University of Novi Sad*

[b] *Radboud University Nijmegen*

[c] *Department of Psychology, University of Belgrade*

[d] *Department of Linguistics, University of Alberta*

[e] *Laboratory for Experimental Psychology, University of Belgrade*

[*] Corresponding Author: Petar Milin; Department of Psychology, University of Novi Sad; Dr Zorana Djindjica 2; Novi Sad 21000, Serbia; Phone: +381 (21) 458 948; Fax: +381 (21) 458 948. E-mail: pmilin@ff.ns.ac.yu (Petar Milin); Victor.Kuperman@mpi.nl (Victor Kuperman); akostic@f.bg.ac.yu (Aleksandar Kostić); baayen@ualberta.ca (Harald Baayen).

1 Introduction

Most experimental work on morphological processing has been inspired by syntagmatically oriented theories of word structure. Processing models that assume obligatory morphological decomposition during lexical processing such as proposed by Taft (1994, 1979, 2004) fit well with, for instance, distributed morphology (Halle and Marantz, 1993). The same holds for the dual mechanism model of Pinker (1991, 1999), which claims that regular inflected forms are not available in lexical memory but derived on-line using morphological rules. The processing literature offers extensive discussion of the question to which representational levels morphemic representations should be allocated, and seems, almost universally, to proceed on the assumption that affixes are morphemes in the classical structuralist sense. Work in theoretical morphology questioning the morphemic status of lexical formatives (Hockett, 1954; Aronoff, 1994; Beard, 1995; Matthews, 1974; Anderson, 1992; Blevins, 2003, 2006) has not had any impact on the models psychologists have proposed for processing and representation in the mental lexicon.

In this chapter, we present a survey of a line of research that departs from the theoretical assumptions of mainstream experimental psycholinguistics in that it is very close in spirit to Word and Paradigm morphology. It is becoming increasingly clear that, contrary to the assumptions of the dual mechanism model and other models positing obligatory decomposition into morphemes, that morphologically complex words leave traces in lexical memory.

A central diagnostic for the presence of memory traces in long-term memory has been the word frequency effect. A higher frequency of use allows for shorter processing latencies in both visual and auditory comprehension (cf. Baayen et al., 2003; New et al., 2004; Baayen et al., 2006, etc.), and lower rates of speech errors in production (Stemberger and MacWhinney, 1986). The effect of word frequency tends to be stronger for irregular complex words than for regular complex words, and stronger for derived words than for inflected words. But even for regular inflected words, the effect of prior experience clearly emerges (Baayen et al., 2008b). The ubiqitous effect of word frequency shows that large numbers of complex words are available in the mental lexicon. This fits will with the central tenet of Word and Paradigm morphology that inflected words are available in the lexicon and form the basis for analogical generalization.

In Word and Paradigm morphology, inflected words are organized into paradigms and paradigms into inflectional classes. (In what follows, we will use the term *inflectional paradigm* to refer to the set of inflected variants of a given lexeme, and the term *inflectional class* to refer to a set of lexemes that use the same set of exponents in their inflectional paradigms.) This raises the question of whether there is experimental evidence supporting such a paradigmatic organizational structure for the mental lexicon.

For derivational morphology, work on the morphological family size effect (see, e.g. Moscoso del Prado Martín et al., 2004a) has clarified that how a given word is processed is co-determined by other words in lexical memory to which it is morphologically related. This constitutes evidence for paradigmatic organization in the mental lexicon. From the perspective of inflection, however, morphological families are very heterogeneous, and do not allow words to be grouped into higher-order sets similar to inflectional classes.

In this chapter, we first review a series of recent experimental studies that explore the role of paradigmatic structure for inflected words. We then present new experimental results showing how the principles that structure inflectional paradigmatics can be generalized to subsets of derived words.

The approach to morphological organization and morphological processing that we describe in this chapter differs from both theoretical morphology and mainstream experimental psycholinguistics in that it makes use of central concepts from information theory. A basic insight from information theory that we apply to lexical processing is that the amount of information carried by an event (e.g., a word's inflected variant, an exponent, or an inflectional class) is negatively correlated with the probability of that event, and positively correlated with processing costs (see for a similar approach to syntax Levy, 2008). We believe information theory offers exactly the right tools for studying the processing consequences of paradigmatic relations. The use of these tools does not imply that we think the mental lexicon is organized in terms of optimally coded bit streams. We will remain agnostic about how paradigmatic structure is implemented in the brain. We do believe that the concepts of information science provide us with excellent tools to probe the functional organization of the (mental) lexicon.

We begin this chapter with an introduction to a number of central concepts from information theory and illustrate how these concepts can be applied to the different levels of paradigmatic organization in the (mental) lexicon. We then focus on three key issues: (i) the processing cost of an exponent given its inflectional class, (ii) the processing cost associated with paradigms and inflectional classes, and (iii) the processing cost that arises when the probabilistic distributional properties of paradigms and inflectional classes diverge.

2 Central concepts from information theory

A fundamental insight of information theory is that the amount of information I carried by (linguistic) unit u can be defined as the negative binary logarithm of its probability:

$$I_u = -\log_2 \Pr(u). \qquad (1)$$

Consider someone in the tip-of-the tongue state saying *the eh eh eh eh eh eh key*. The word *eh* has the greatest probability, $6/8$, and is least informative. Its amount of information is $-\log_2(6/8) = 0.415$ bits. The words *the* and *key* have a probability of $1/8$ and the amount of information they carry is 3 bits. In what follows, we assume that lexical units that have a higher information load are more costly to access in long-term memory. Hence, we expect processing costs to be proportional to the amount of information. This is, of course, exactly what the word frequency effect tells us: higher frequency words, which have lower information loads, are processed faster than low-frequency, high-information words.

We estimate probabilities from relative frequencies. By way of illustration, consider the inflected variants of the Serbian feminine noun *planina*, “mountain”. Serbian nouns have six cases and two numbers. Due to syncretism, the twelve combinations of case and number are represented by only 6 distinct inflected variants. These inflected variants are listed in column 1 of the upper part of Table 1. The second column lists the frequencies of these inflected variants in a two-million word corpus of written Serbian.

In what follows, we consider two complementary ways of estimating probabilities from frequencies. The probabilities listed in the third column of Table 1 are obtained by normalizing the frequency counts with respect to a lexeme’s inflectional paradigm (column three). More specifically, the probability $\mathrm{Pr}_\pi(w_e)$ of an inflected variant w_e of lexeme w is estimated in this table as its form-specific frequency F (henceforth *word frequency*) of occurrence, normalized for the sum of the frequencies of all the distinct inflected variants of its lexeme, henceforth *stem frequency*:

$$\mathrm{Pr}_\pi(w_e) = \frac{F(w_e)}{\sum_e F(w_e)}. \qquad (2)$$

The corresponding amounts of information, obtained by applying (1), are listed in column four. Table 1 also lists the frequencies of the six exponents (column 5), calculated by summing the word frequencies of all forms in the corpus with these exponents. The probabilities listed for these exponents (column six) are obtained by normalizing with respect to the summed frequencies of these exponents:

$$\mathrm{Pr}_\pi(e) = \frac{F(e)}{\sum_e F(w_e)}. \qquad (3)$$

The corresponding amount of information is listed in column seven.

The second way in which we can estimate probabilities is by normalizing with respect to the number of tokens N in the corpus. The probability of a lexeme w is then estimated as the sum of the frequencies of its inflected variants, divided by N:

$$\mathrm{Pr}_N(w) = \frac{F(w)}{N} = \frac{\sum_e F(w_e)}{N}. \qquad (4)$$

feminine nouns

Inflected variant	Inflected variant frequency	Inflected variant relative frequency	Information of inflected variant	Exponent frequency	Exponent relative frequency	Information of exponent
	$F(w_e)$	$\Pr_\pi(w_e)$	I_{w_e}	$F(e)$	$\Pr_\pi(e)$	I_e
planin-*a*	169	0.31	1.69	18715	0.26	1.94
planin-*u*	48	0.09	3.47	9918	0.14	2.84
planin-*e*	191	0.35	1.51	27803	0.39	1.36
planin-*i*	88	0.16	2.64	7072	0.1	3.32
planin-*om*	30	0.05	4.32	4265	0.06	4.06
planin-*ama*	26	0.05	4.32	4409	0.06	4.06

masculine nouns

Inflected variant	Inflected variant frequency	Inflected variant relative frequency	Information of inflected variant	Exponent frequency	Exponent relative frequency	Information of exponent
	$F(w_e)$	$\Pr_\pi(w_e)$	I_{w_e}	$F(e)$	$\Pr_\pi(e)$	I_e
prostor-*ø*	153	0.38	1.40	25399	0.35	1.51
prostor-*a*	69	0.17	2.56	18523	0.26	1.94
prostor-*u*	67	0.17	2.56	8409	0.12	3.06
prostor-*om*	15	0.04	4.64	3688	0.05	4.32
prostor-*e*	48	0.12	3.06	5634	0.08	3.64
prostor-*i*	23	0.06	4.06	6772	0.09	3.47
prostor-*ima*	23	0.06	4.06	3169	0.04	4.64

Table 1: Inflected nouns in Serbian. The upper part of the table shows inflected variants for the feminine noun “planina” (*mountain*), the lower part shows the inflected variants of the masculine noun “prostor” (*space*). Columns present frequencies and relative frequencies of respective inflectional paradigm and the class to which it belongs.

In this approach, the probability of an inflected variant can be construed as the joint probability of its lexeme w and its exponent:

$$\begin{aligned} \mathrm{Pr}_N(w_e) &= \mathrm{Pr}(w,e) \\ &= \mathrm{Pr}(e,w) \\ &= \frac{F(w_e)}{N}. \end{aligned} \tag{5}$$

Likewise, the probability $Pr(e)$ of an exponent (e.g., *-a* for nominative singular and genitive plural in Serbian feminine nouns) can be quantified as the relative frequency of occurrence of e in the corpus:

$$\mathrm{Pr}_N(e) = \frac{F(e)}{N}. \tag{6}$$

The probabilities considered thus far are unconditional, a priori, decontextualized probabilities. As exponents appear in the context of stems, we need to consider the conditional probability of an exponent given its lexeme, $Pr(e|w)$. Using Bayes' theorem, we rewrite this probability as:

$$\begin{aligned} \mathrm{Pr}_N(e|w) &= \frac{\mathrm{Pr}_N(e,w)}{\mathrm{Pr}(w)} \\ &= \frac{F(w_e)}{N}\frac{N}{F(w)} \\ &= \frac{F(w_e)}{F(w)} \\ &= \mathrm{Pr}_\pi(w_e). \end{aligned} \tag{7}$$

Likewise, the conditional probability of the lemma given the exponent is defined as:

$$\begin{aligned} \mathrm{Pr}_N(w|e) &= \frac{\mathrm{Pr}_N(w,e)}{\mathrm{Pr}_N(e)} \\ &= \frac{F(w_e)}{N}\frac{N}{F(e)} \\ &= \frac{F(w_e)}{F(e)}. \end{aligned} \tag{8}$$

For each lexical probability we can compute the corresponding amount of information. We allow for the possibility that each source of information may have its own distinct effect on lexical processing by means of positive weights ω_{1-5}:

$$\mathrm{I}_{w_e} = -\omega_1 \log_2 F(w_e) + \omega_1 \log_2 N$$

$$
\begin{aligned}
\mathrm{I}_w &= -\omega_2 \log_2 F(w) + \omega_2 \log_2 N \\
\mathrm{I}_e &= -\omega_3 \log_2 F(e) + \omega_3 \log_2 N \\
\mathrm{I}_{e|w} &= -\omega_4 \log_2 F(w_e) + \omega_4 \log_2 F(w) \\
\mathrm{I}_{w|e} &= -\omega_5 \log_2 F(w_e) + \omega_5 \log_2 F(e).
\end{aligned} \tag{9}
$$

We assume that the cost of retrieving lexical information from long-term memory is proportional to the amount of information retrieved. Hence the cost of processing an inflected word w_e is proportional to at least the amounts of information in (9). More formally, we can express this processing cost (measured experimentally as a reaction time RT) as a linear function:

$$
\begin{aligned}
RT &\propto I_{w_e} + I_w + I_e + I_{e|w} + I_{w|e} \\
&= (\omega_1 + \omega_2 + \omega_3) \log_2 N - (\omega_1 + \omega_4 + \omega_5) \log_2 F(w_e) \\
&- (\omega_2 - \omega_4) \log_2 F(w) - (\omega_3 - \omega_5) \log_2 F(e).
\end{aligned} \tag{10}
$$

There are several predictions for the effects of lexical probabilities on lexical processing that follow directly from (10). First, word frequency $F(w_e)$ will always elicit a facilitatory effect, as all its coefficients have a negative sign in (10). Second, stem frequency may either facilitate or inhibit processing, depending on the relative strengths of the coefficients ω_2 and ω_4. Third, the frequency of the exponent can also either speed up or hinder processing depending on values of ω_3 and ω_5. The first two predictions are supported by the large-scale regression studies reported by Baayen et al. (2008b) and Kuperman et al. (2008).

We now proceeed from basic lexical probabilities that operate at the level of individual inflected words to the quantification of the information carried by inflectional paradigms and inflectional classes. The paradigm of a given lexeme can be associated with a distribution of probabilities $\{\mathrm{Pr}_\pi(w_e)\}$. For *planina* in Table 1, this probability distribution is given in column three. The amount of information carried by its paradigm as a whole is given by the *entropy* of the paradigm's probability distribution:

$$
H = -\sum_e \mathrm{Pr}_\pi(w_e) \log_2(\mathrm{Pr}_\pi(w_e)). \tag{11}
$$

Formally, H is the expected (weighted average) amount of information in a paradigm. The entropy increases with the number of members of the paradigm. It also increases when the probabilities of the members are more similar. For a given number of members, the entropy is maximal when all probabilities are the same. H also represents the average number of binary decisions required to identify a member of the paradigm, i.e., to reduce all uncertainty about which member of the paradigm is at issue, provided that the paradigm is represented by an optimal

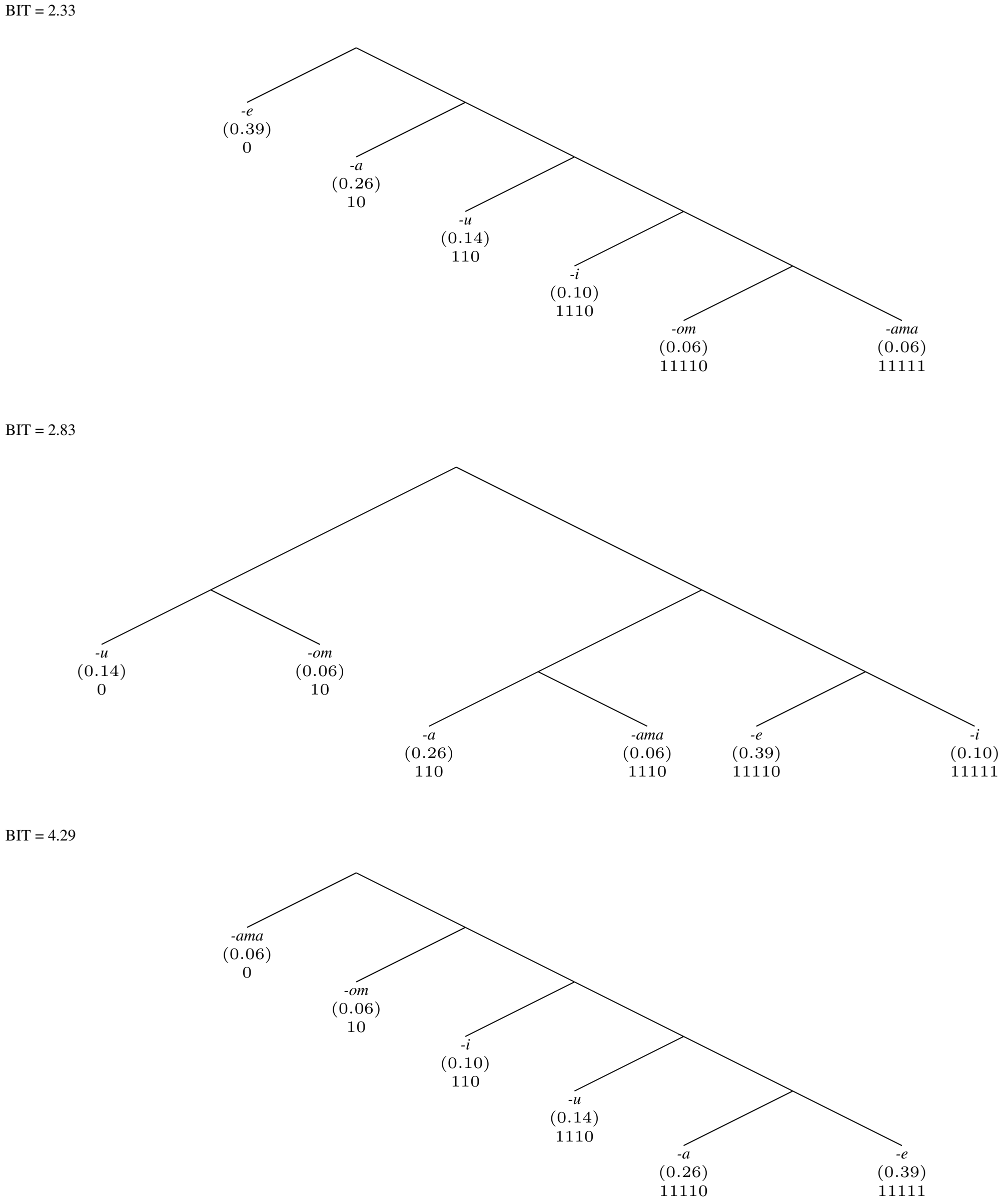

Figure 1: Optimal and non-optimal binary coding schemes for the inflectional class of regular feminine nouns in Serbian.

binary coding. We illustrate the concept of optimal coding in Figure 1 using as an example the inflectional class of regular feminine nouns in Serbian.

The upper panel of Figure 1 shows an optimal binary coding scheme, in which the most probable exponent (*-e*, $\mathrm{Pr}_\pi = 0.39$) occupies the highest leaf node in the tree. The lower the probability of the other exponents, the lower in the tree they are located. Thus, the exponents with the lowest probabilities in the inflectional class, *-om* ($\mathrm{Pr}_\pi = 0.06$) and *-ama* ($\mathrm{Pr}_\pi = 0.06$) are found at the lowest leaf nodes. The second panel of Figure 1 represents another possible coding, which is suboptimal in that some exponents with relatively high probabilities are located below lower-probability exponents in the tree. Finally, the third panel shows the least optimal coding, in which the less probable the exponent is, the *higher* it is positioned in the tree. The average number of binary decisions (the number of bits) required to identify a given paradigm member, i.e., to reach the paradigm member's leaf node when starting at the root node of the tree, is the sum of the products of the number of steps and the members' probabilities. This average is never greater than the entropy of the paradigm $H + 1$ (Ross, 1988). For the upper panel of Figure 1, the average number of binary decisions is 2.33 bits, for the coding in the second panel, it is 2.83, and for the worst coding in the third panel, it is 4.29. In section 4 we will review experimental studies showing that paradigmatic entropies co-determine lexical processing.

Thus far, we have considered probabilities and the corresponding entropy at the level of the inflectional class of regular feminine nouns in Serbian. However, the probability distribution of the inflected variants of a given lexeme may differ substantially from the probability distribution of the exponents at the level of the inflectional class. As a consequence, the corresponding entropies may differ substantially from each other as well. The extent to which these probability distributions differ is quantified by the relative entropy, also known as Kullback-Leibler divergence. By way of example, consider again the Serbian feminine noun *planina* 'mountain' and its inflectional class as shown in Table 1. The third column lists the estimated probabilities for the paradigm, and the sixth column lists the probability distribution of the class. Let P denote the probability distribution of the paradigm, and Q the probability distribution of the inflectional class. The relative entropy can now be introduced as:

$$D(P||Q) = \sum_e \mathrm{Pr}_\pi(w_e) \log_2 \frac{\mathrm{Pr}_\pi(w_e)}{\mathrm{Pr}_\pi(e)}. \tag{12}$$

Relative entropy is also known as *information gain*,

$$\begin{aligned} D(P||Q) &= IG(\mathrm{Pr}_\pi(e|w)||\mathrm{Pr}_\pi(e|c)) \\ &= \sum_e \mathrm{Pr}_\pi(e|w) \log_2 \frac{\mathrm{Pr}_\pi(e|w)}{\mathrm{Pr}_\pi(e|c)} \end{aligned}$$

$$= \sum_e \mathrm{Pr}_\pi(w_e) \log_2 \frac{\mathrm{Pr}_\pi(w_e)}{\mathrm{Pr}_\pi(e)}, \tag{13}$$

as it measures the reduction in our uncertainty about the exponent when going from the situation in which we only know its inflectional class to the situation in which we also know the lexeme. For *planina*, $H = 2.22$, and $D(P||Q) = 0.05$. For the masculine noun listed in the lower half of Table 1, $H = 2.42$ and $D(P||Q) = 0.07$. In both cases, the two distributions are fairly similar, so the relative entropies are small. There is little that the knowledge of *planina* adds to what we already new about regular feminine nouns. If we approximate the probability distribution of *planina* with the probability distribution of its class, we are doing quite well. In what follows, we will refer to relative entropy simply as RE. In section 4.2 we review a recent study demonstrating that RE is yet another information theoretic predictor of lexical processing costs.

In what follows, we will review a series of studies that illustrate how these information theoretic concepts help us to understand paradigmatic organization in the mental lexicon. Section 3 addresses the question of how the probability of an exponent given its inflectional class is reflected in measures of lexical processing costs. Section 4 reviews studies that make use of entropy and relative entropy to gauge lexical processing and paradigmatic organization. Finally, in section 5 we present new experimental results showing how concepts from information theory that have proved useful for understanding inflection can also be made fruitful for understanding derivation.

3 The Structure of Inflectional Classes

The consequence of the amount of information carried by an exponent for lexical processing has been explored in a series of experimental studies on Serbian (Kostić, 1991, 1995; Kostić et al., 2003). A starting point for this line of research is the amount of information carried by an exponent,

$$I_e = -\log_2 \Pr_\pi(e).$$

The problem addressed by Kostić and colleagues is that exponents are not equal with respect to their functional load. Some exponents (given their inflectional class) express only a few functions and meanings, others express many. Table 2 lists the functions and meanings for the exponents of the masculine and regular feminine inflectional class of Serbian. The count of numbers of functions and meanings for a given exponent were taken from an independent comprehensive lexicological survey of Serbian (see also the appendix of Kostić et al. 2003, for a shortlist of functions and meanings). Instead of using just the flat corpus-based

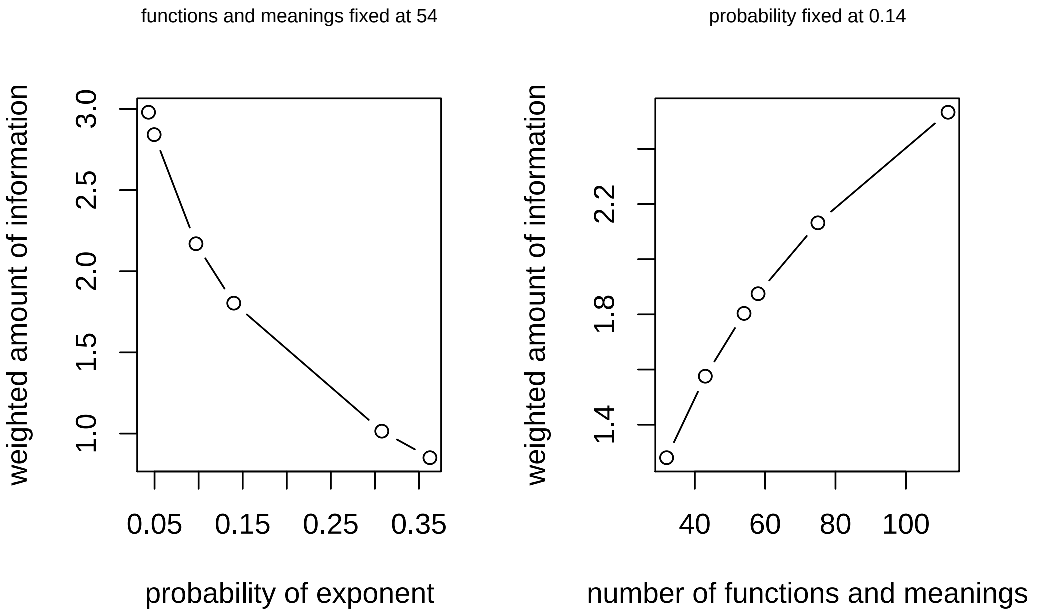

Figure 2: Partial effects of the probability of an exponent and its number of syntactic functions and meanings on the weighted amount of information I'_e.

relative frequencies, Kostić and colleagues propose to weight these probabilities for their functions and meanings. Let R_e denote the number of functions and meanings carried by exponent e. Then the weighted amount of information I'_e can be expressed as follows:

$$I'_e = -\log_2 \left(\frac{\Pr_\pi(e)/R_e}{\sum_e \Pr_\pi(e)/R_e} \right) \tag{14}$$

The ratio $(\Pr_\pi(e)/R_e)$ gives us the average probability per syntactic function/meaning for a given exponent. In order to take the other exponents within the inflectional class into account, this ratio is weighted by the sum of the ratios for each of the exponents (see, e.g., Luce, 1959). The resulting proportion is log-transformed to obtain the corresponding amount of information in bits. The partial effects of probability on the one hand, and the number of functions and meanings on the other, is shown in Figure 2. The weighted information is predicted to decrease with probability, and to increase with the number of functions and meanings. Table 2 lists I'_e for each of the exponents of the masculine and regular feminine inflectional classes.

To assess the predictivity of I'_e, Kostić et al. (2003); Kostić (2008) calculated

masculine nouns				
Exponent	Case and Number	Frequency	Functions and Meanings	Information
∅	nom sg	12.83	3	0.434
a	gen sg/acc sg /gen pl	18.01	109	5.128
u	dat sg /loc sg	4.64	43	5.744
om	ins sg	1.90	32	6.608
e	acc pl	2.21	58	7.243
i	nom pl	3.33	3	2.381
ima	dat pl/loc pl/ins pl	1.49	75	8.186

feminine nouns				
Exponent	Case and Number	Frequency	Functions and Meanings	Information
a	nom sg/gen pl	12.06	54	1.464
u	acc sg	5.48	58	2.705
e	gen sg /nom pl/acc pl	14.20	112	2.280
i	dat sg /loc sg	3.80	43	2.803
om	ins sg	1.94	32	3.346
ama	dat pl/loc pl/ins pl	1.69	75	4.773

Table 2: Exponents, case and number, frequency of the exponent, number of functions and meanings of the exponents, and amount of information carried by the exponents, for masculine nouns (upper table) and regular feminine nouns (lower table).

the mean lexical decision latency for each exponent in a given inflectional class, and investigated whether these mean latencies can be predicted from the weighted amounts of information such as listed in Table 2. The Pearson correlation between the mean latencies and the weighted information scores was highly significant for both masculine and feminine nouns ($R^2 = 0.88$ for masculine nouns, $R^2 = 0.98$ for regular feminine nouns and $R^2 = 0.99$ for irregular feminine nouns). Furthermore, when mean reaction time is regressed on the weighted information load, the slopes of the regression lines are positive. Exponents carrying a greater average amount of information are more difficult to process. In other words, these data show that the average processing cost of an exponent in its inflectional class is very well predicted from its frequency and its functional load as given by (14) and illustrated above in Figure 2.

The probabilities that we considered in these analyses were estimated by summing across all words with a given exponent in a given inflectional class. In this way, the information about the probabilities of the different exponents in the inflectional paradigms of specific words is lost. In order to address the possibility that word-specific probabilities of exponents also co-determine lexical processing, Kostić et al. (2003) first applied the same weighting scheme underlying (14) at the level of individual lexemes, giving a lexeme-specific weighted information I'_{w_e}:

$$I'_{w_e} = -\log_2 \left(\frac{\Pr_\pi(w_e)/R_e}{\sum_e \Pr_\pi(w_e)/R_e} \right). \tag{15}$$

Kostić et al. (2003) then constructed two sets of lexemes (henceforth Inflectional Groups) which contrasted maximally with respect to I'_{w_e}. For each of the two inflectional groups, they then calculated the average value of I'_{w_e} for each of the exponents. Regression analysis showed that these group-averaged amounts of information contributed independently to the model, over and above the general class-based information values I'_e. As before, larger values for the group-averaged amounts of information I'_{w_e} corresponded to longer mean lexical decision latencies.

It is useful to probe the lexeme-specific weighted information (15) with respect to how it relates to the frequential properties of the lexeme and its inflected variants, as well as to the functional ambiguities existing in inflectional paradigms and classes. First consider a simple lower bound for (15):

$$\begin{aligned} I'_{w_e} &= -\log_2 \left(\frac{\Pr_\pi(w_e)/R_e}{\sum_e \Pr_\pi(w_e)/R_{w_e}} \right) \\ &= -\log_2 \frac{\Pr_\pi(w_e)}{R_e} + \log_2 \sum_e \frac{\Pr_\pi(w_e)}{R_e} \\ &\geq -\log_2 \Pr_\pi(w_e) + \log_2 R_e + \log_2 \prod_e \frac{\Pr_\pi(w_e)}{R_e} \\ &\geq -\log_2 \Pr_\pi(w_e) + \log_2 R_e + \sum_e \log_2 \frac{\Pr_\pi(w_e)}{R_e} \\ &\geq \log_2 R_e - \sum_e \log_2 R_e - \log_2 \Pr_\pi(w_e) + \sum_e \log_2 \Pr_\pi w_e. \end{aligned} \tag{16}$$

The third term is the amount of information carried by the inflected variant, I_{w_e}, see (2), and $\sum_j \log_2 \Pr_\pi w_j$ is a measure of the lexeme's stem frequency, evaluated by summing the log frequencies of its inflected variants rather than by summing the bare frequencies of its inflected variants. At the level of the inflected variant, then, the amount of information (15) incorporates two well-known frequency effects that have been studied extensively in the processing literature. The word frequency effect ($-\log_2 \Pr_\pi(w_e)$) is facilitatory, as expected. By contrast, the stem frequency effect ($\sum_e \log_2 \mathrm{Pr}_\pi w_e$) is predicted to be inhibitory. However, both frequency effects are complemented by measures gauging ambiguity. Ambiguity of the given exponent is harmful, whereas ambiguity in the rest of the paradigm is facilitatory. Thus, the stem frequency effect emerges from this model as a composite effect with an inhibitory and a facilitatory component. This may help explain why stem frequency effects are often much less robustly attested in experimental data (see, e.g., Baayen et al., 2008b) compared to word frequency effects.

In order to evaluate how well the lower bound given in (16) approximates the original measure given in (15), we examined for the two inflectional groups for regular feminine nouns the exponent frequency, the group average functions and meanings, information values, and mean reaction times, as listed in Table 3 (data from Kostić et al., 2003). As a consequence, the terms in (16) represent the ambiguity of the exponent, the joint ambiguity of all exponents, the word frequency effect of the inflected variant, and the stem frequency effect of its lexeme.

For the data in Table 3, we first carried out a linear regression analysis with RT as dependent variable and I' and Inflectional Group as predictors. The R^2 for this model was 0.863. We then carried out a linear regression analysis, but now with as predictors the two measures that figure in the lower bound of the amount of information: exponent frequency and the number of functions and meanings of the exponent R. The R^2 of this model was 0.830. Furthermore, the effect of the number of functions and meanings was inhibitory ($\hat{\beta} = 27.5, t(8) = 2.512, p = 0.0362$) and the effect of exponent frequency was facilitatory ($\hat{\beta} = -5.2, t(8) = -5.813, p = 0.0004$) as expected given (16). In other words, the two variables that according to (16) should capture a substantial proportion of the variance explained by the amount of information I', indeed succeed in doing so: 0.830 is 96% of 0.863.

The lower bound estimate in (16) is a simplification of the full model I'_{w_e} defined by (15). Because the simplification allows us to separate the word and stem frequency effects, it clarifies that these two frequency effects are given the same overall weight. There is evidence, however, that stem frequency has a much more modest weight than word frequency (Baayen et al., 2008b), and may even have a different functional form. This suggests that it may be preferable to rewrite (15) as:

$$I'_{w_e} = -\log_2 \left(\frac{\omega_1 \, \mathrm{Pr}_\pi(w_e)/R_e}{\omega_2 \, \sum_e \mathrm{Pr}_\pi(w_e)/R_e} \right), \tag{17}$$

with separate weights ω for numerator and denominator.

On the other hand, at the level of a given class the lower bound estimate in (17) reduces to the exponent frequency and the overall class frequency. The exponent frequency can be translated into affix frequency, for which Baayen et al. (2008b)

Exponent	Exponent frequency	R	I'	Inflectional Group	RT
a	12.06	3.99	1.46	high	674
e	14.20	4.72	2.28	high	687
i	3.80	3.76	2.80	high	685
u	5.48	4.06	2.71	high	693
om	1.94	3.47	3.35	high	718
ama	1.69	4.32	4.77	high	744
a	12.06	3.99	1.46	low	687
e	14.20	4.72	2.28	low	685
i	3.80	3.76	2.80	low	730
u	5.48	4.06	2.71	low	712
om	1.94	3.47	3.35	low	722
ama	1.69	4.32	4.77	low	746

Table 3: Mean reaction times in visual lexical decision (RT), exponent frequency, number of functions and meanings of the exponent (R), amount of information (I), and Inflectional Group (high versus low by-word amount of information) for the Exponents of the regular feminine declension class.

confirmed a significant facilitatory effect. However, it is presently unclear how class frequency could be generalized and gauged with derivations. Inflectional classes are well contained and it is easy to count-out their overall frequencies. However, within and between derivational classes there are no clear partitions of the lexical space and while inflected words belong to only one inflectional class, any given base word may participate in several derivations. We shall address the issue of relations between base words and their derivatives in co-determining lexical processing in great detail in section 5.

It is also useful to rewrite (14) along similar lines as we did for (15). In this case, the lower bound for the amount of information can be written as the sum of two conditional probabilities. First consider the probability of exponent e given its inflectional class c:

$$\begin{aligned} \Pr(e|c) &= \frac{\Pr(e,c)}{\Pr(c)} \\ &= \frac{\Pr(e)}{\Pr(c)}. \end{aligned}$$

(Note that the probability of an exponent is defined strictly with respect to its inflectional class. We never sum frequencies of exponents across inflectional classes.) The information corresponding to this conditional probability is

$$\begin{aligned}
I_{e|c} &= -\log_2 \frac{\Pr(e)}{\Pr(c)} \\
&= -\log_2 \Pr(e) + \log_2 \Pr(c) \\
&= -\log_2 \Pr(e) + \log_2 \sum_j \Pr(e_j) \\
&\geq -\log_2 \Pr(e) + \log_2 \prod_j \Pr(e_j) \\
&\geq -\log_2 \Pr(e) + \sum_j \log_2 \Pr(e_j) \\
&= I'_{e|c} \qquad (18)
\end{aligned}$$

Note that $I'_{e|c}$ is a lower bound of $I_{e|c}$.

Next, let R_e denote the number of functions and meanings of exponent e in class c, and let R_c denote the total count of functions and meanings within the class. The conditional probability of the functions and meanings of exponent e given its class c is

$$\begin{aligned}
\Pr(R_e|R_c) &= \frac{\Pr(R_e, R_c)}{\Pr(R_c)} \\
&= \frac{\Pr(R_e)}{\Pr(R_c)} \\
&= \frac{R_e}{R_c}
\end{aligned}$$

and the corresponding information is therefore

$$\begin{aligned}
I_{R_e|R_c} &= -\log_2 \frac{R_e}{R_c} \\
&= -\log_2 R_e + \log_2 R_c \\
&= -\log_2 R_e + \log_2 \sum_j R_j \\
&\leq -\log_2 R_e + \log_2 \prod_j R_j \\
&\leq -\log_2 R_e + \sum_j \log_2 R_j \\
&= I'_{R_e|R_c} \qquad (19)
\end{aligned}$$

Here, $I'_{R_e|R_c}$ is an upper bound of $I_{R_e|R_c}$.

Taking into account that $I'_{e|c}$ is a lower bound of $I_{e|c}$, and that $I'_{R_i|R_c}$ is an upper bound of $I_{R_i|R_c}$, we can now approximate (14) as follows:

$$\begin{aligned} I_{w_e} &\approx \log_2 R_e - \sum_j \log_2 R_j - \log_2 \mathrm{Pr}_\pi w_e + \sum_j \log_2 \mathrm{Pr}_\pi w_j \\ &\approx -I'_{R_e|R_c} + I'_{e|c}. \end{aligned} \tag{20}$$

In other words, the amount of information as defined in (14) is related to the sum of two conditional probabilities: (i) the probability of the exponent given its class, and (ii) the probability of the ambiguity of the exponent given the ambiguity in its class. The partial effects of these two conditional probabilities are shown in Figure 3. As expected, the partial effects are very similar to those shown in Figure 2.

At this point, the question arises why $I'_{R_e|R_c}$ appears with a negative sign in (20). To answer this question, we need to consider the function of exponents in their classes: to differentiate between the functions and meanings an inflected form can have in the discourse. Now consider the case in which $R_e \to R_c$. The more the functions expressed by exponent e become similar to the universe of functions and meanings carried by the inflectional class, the less distinctive the exponent becomes. In other words, an exponent is more successful as a distinctive functional unit of the language when $|R_c - R_e|$ is large. If so, the corresponding amount of information is small, and processing is fast. By contrast, an exponent for which $I_{R_e|R_c}$ is large is dysfunctional, and therefore harder to process, leading to longer processing latencies.

4 The information structure of paradigms

4.1 Entropy

Thus far, we have considered the processing load of an inflected form given its paradigm, or an exponent given its inflectional class. Moscoso del Prado Martín et al. (2004b) added a new dimension to the experimental study of paradigmatics by considering the cost of the complexity of a paradigm as such, gauged by means of the entropy measure H. We illustrate the difference between Kostić's approach and the one developed by Moscoso del Prado and his colleagues by means of Figure 1 shown above. Ignoring the weighting for numbers of functions and meanings, Kostić's measure simplifies to $-\log_2(\mathrm{Pr}_\pi(e))$, which reflects the number of steps from the root node to the leaf node of the exponent e in an optimal binary coding scheme (see the upper panel; for numbers of nodes that are integer powers of two, the $-\log_2(\mathrm{Pr}_\pi(e))$ is exactly equal to the number of steps). However, this measure is insensitive to the size and configuration of the tree. To

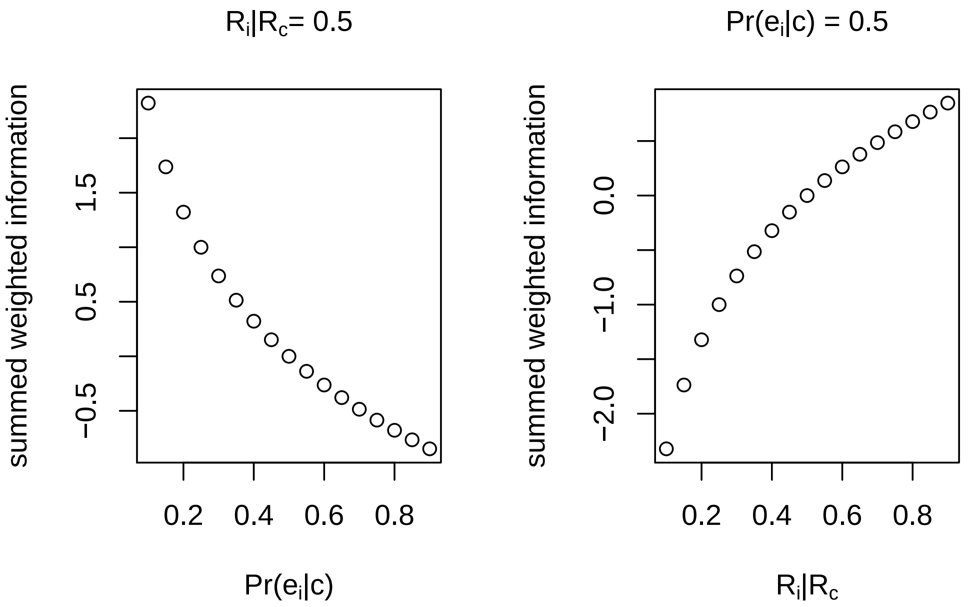

Figure 3: The left panel shows the partial effect of the information carried by the probability of the exponent given its class $I'_{e|c}$. The right panel shows the partial effect of the information carried by the proportion of the number of functions and meanings conditioned on the total number of functions and meanings for the class $I'_{R_e|R_c}$. Both partial effects are calibrated for the other effect evaluated at 0.5, and are calculated straightforwardly from (20).

capture these aspects of the tree, we can make use of the entropy measure. The entropy, which is the same for each and every member of the paradigm, quantifies the expected number of steps from the root to a leaf node.

Moscoso del Prado Martín et al. (2004b) applied the entropy measure to Dutch paradigms, but used a much broader definition of paradigms that extended the concept of the morphological family. Table 4 shows the words listed in CELEX that contain *neighbour* as a constituent. The left two columns list the morphological family as defined by Schreuder and Baayen (1997), the middle columns list the inflected variants that were found for two of the members of the family, and the rightmost columns list the set that merges the family members with the inflected variants. Moscoso del Prado and colleagues calculated the entropy over this merged set, and proposed this entropy as an enhanced measure for capturing the morphological family size effect. They pointed out that when all family members are equiprobable, the entropy of the family reduces to the log of the number

morphological family		inflectional paradigms		merged paradigms	
word	F	word	F	word	F
neighbour	901	*neighbour*	343	*neighbour*	343
neighbourhood	407	*neighbours*	558	*neighbours*	558
neighbouring	203			*neighbourhood*	386
neighbourliness	3	*neighbourhood*	386	*neighbourhoods*	21
neighbourly	14	*neighbourhoods*	21	*neighbouring*	203
				neighbourliness	3
				neighbourly	14

Table 4: Morphological family and inflectional paradigms for *neighbor*.

of family members. Since it is exactly this log-transformed count that emerged as predictive for processing latencies, the entropy of the family can be viewed as a principled way of weighting family members for their token frequency.

Moscoso del Prado and colleagues combined this generalized entropy measure with the amount of information carried by a word (inflected or uninflected) as estimated from its relative frequency to obtain what they called the information residual:

$$I_R = I_w - H = \log N - \log_2 F_w - H. \tag{21}$$

This information residual performed well in a series of post-hoc analyses of processing of Dutch complex words.

By bringing several measures together in a single predictor, I_R, stem frequency and entropy receive exactly the same regression weight:

$$\begin{aligned} RT &\propto \beta_0 + \beta_1 I_R \\ &= \beta_0 + \beta_1 (I_w - H) \\ & \quad \beta_0 - \beta_1 \log_2 F_w - \beta_1 H. \end{aligned} \tag{22}$$

However, subsequent work (Baayen et al., 2006) suggests that frequency, the entropy calculated over the morphological family while excluding inflected variants, and the entropy of the paradigms of individual lexemes should be allowed to have different importance (i.e, different β weights). Their study examined a wide range of lexical predictors for simple English nouns and verbs, and observed independent effects of inflectional entropy (henceforth H_i) across both the visual lexical decision and word naming tasks. An effect of derivational entropy (henceforth H_d) was present only in the visual lexical decision task. Here, it emerged with a U-shaped curve, indicating the presence of some inhibition for words with very information-rich families. In their study of the lexical processing of 8486 complex words in English, Baayen et al. (2008b) also observed an independent facilitatory effect of inflectional entropy, side by side with a facilitatory effect of the family size of the lexeme.

These results suggest that, when considered in terms of optimal binary coding schemes, inflected words and lexemes should not be brought together in one encompassing binary tree. Instead, lexemes form one tree, and each lexeme then comes with its own separate disjoint tree for its inflected variants.

Inflectional paradigms in languages such as Dutch and English are trivially simple compared to the paradigms one finds in morphologically rich languages. This raises the question to what extent entropy measures inform us about the processing complexity of more substantive paradigmatic structure. We address this issue for nominal paradigms in Serbian.

4.2 Relative entropy

When the inflectional entropy is computed for a given lexeme, it provides an estimate for the complexity of this lexeme's inflectional paradigm. This measure, however, does not take into account the complexity of the inflectional class, and the extent to which the probability distribution of a lexeme's paradigm diverges from the probability distribution of its inflectional class. We could consider bringing the entropy of the inflectional class into our model, but this class entropy would be the same for all lexemes in the class. Hence, it would not be much more informative than a plain name for that class (for example, Latin declension I, or Serbian declension III). Therefore, Milin et al. (2008) considered the simultaneous influence of paradigms and classes on the processing of inflected nouns in Serbian by means of relative entropy, RE.

Milin et al. (2008) investigated whether relative entropy is predictive for lexical processing in visual lexical decision using masculine and feminine nouns with the case endings *-om, -u* and *-e*. A mixed-effects analysis with word frequency and stem frequency, bigram frequency, number of orthographic neighbors and entropy as covariates revealed an independent inhibitory effect of RE, as shown in the lower right panel of Figure 4. Comparison with the other significant partial effects in the model shows that the magnitude of the effect of RE is comparable to that of stem frequency and orthographic neighborhood size. However, the effect of the entropy did not reach significance ($p > 0.15$).

What this experiment shows is that it is neither the probability distribution of the inflected variants in a word's paradigm, nor the probability distribution in its inflectional class considered separately that are at issue, but rather the divergence between the two distributions. The greater this divergence, the longer the response latencies. A similar pattern was observed for the accuracy measure as well: the greater the divergence of the probability distribution of the paradigm from the probability distribution of the class, the more errors were made.

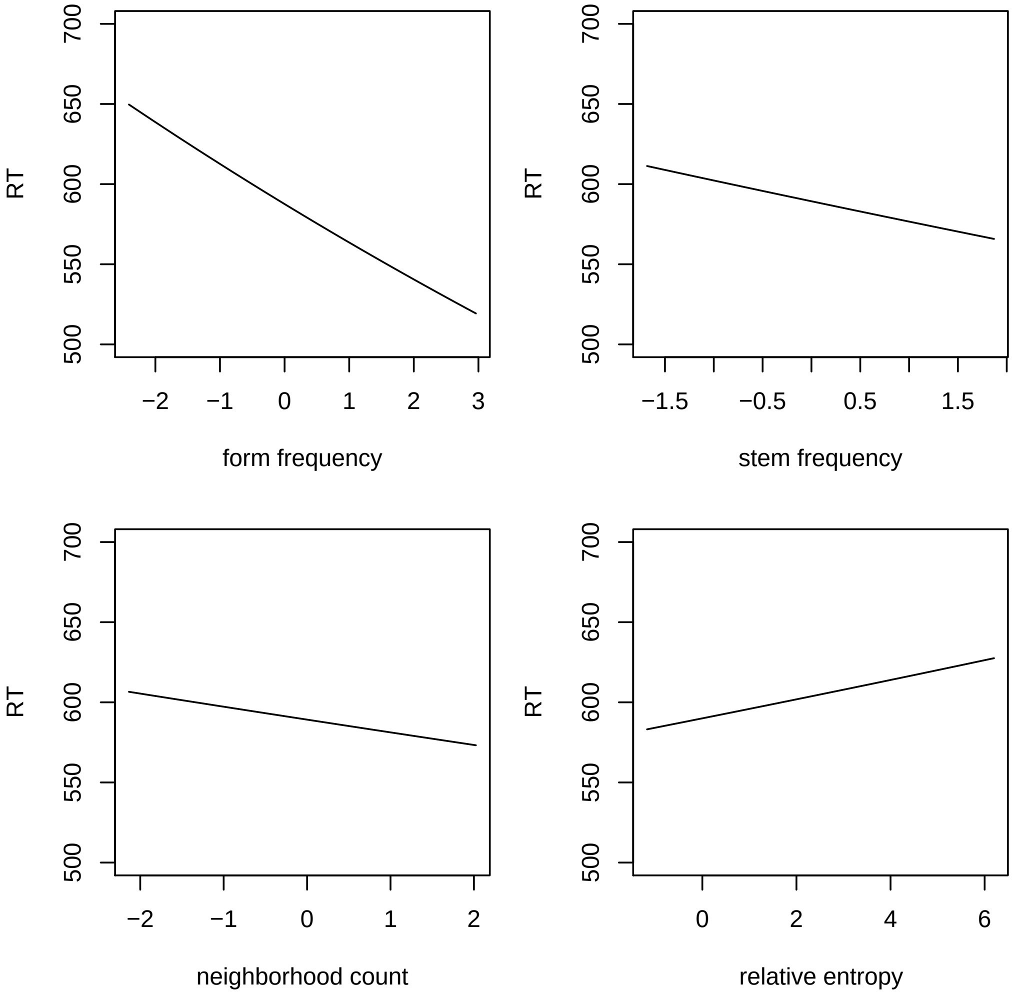

Figure 4: Partial effects of distributional predictors for the response latencies in visual lexical decision to Serbian nouns (Milin et al., 2008).

From the perspective of cognitive psychology, these results are interesting in that they provide further evidence for the importance of structured lexical connectivity. From the perspective of linguistic morphology, they support the theoretical concepts of paradigms and inflectional classes. Combined with the presence of a strong effect of the word frequency, an effect that is much stronger than the effect of the word's stem (compare the upper panels in Figure 4), these results provide strong support for Word and Paradigm morphology (Matthews, 1974; Blevins, 2003, 2006) and for exemplar-based approaches to lexical processing in general (see, e.g., Baayen, 2003).

5 Paradigmatic structure in derivation

In languages such as Dutch or English, morphological families consist predominantly of compounds. As a consequence, the family size effect (cf., Schreuder and Baayen, 1997) is driven almost exclusively by lexical connectivity between compounds. Little is known about the role of derived words. The problem here is that a given base word combines with only a handful of derivational affixes at best. Counts of the number of different prefixes and suffixes that English monomorphemic base words combine with, based on the English section of the CELEX lexical database (Baayen et al., 1995), illustrate that 60% English monomorphemic base words combine with only one affix. Table 5 shows a steep decrease (a Zipfian distribution) in the number of derivational affixes that are attested for a given base word. The verbs *act* and *play* are exceptional in combining with 11 different affixes. The maximum family size in English, 187, observed for *man*, is an order of magnitude larger. With such small numbers of derived family members, it becomes very difficult to gauge the role of a strictly derivational family size count in lexical processing.

Derived words, however, enter into more systematic relations than most compounds, even when we take into account that the meaning of a compound is predictable from its constituents to a much greater extent than has traditionally been assumed (Gagné and Shoben, 1997; Gagné, 2001). For instance, derived adjectives with the prefix *un-* systematically express negation. Taking this fact into account, we asked ourselves whether such systematic relations between base words and their derivatives co-determine lexical processing. As a first step towards an answer, we introduce two simple concepts: the mini-paradigm and the mini-class. Here, the term mini-paradigm refers to pairs of base words and their derivatives. Thus, *kind* and *unkind* form a mini-paradigm, and so do *clear* and *clearly*. In the same line, the term mini-class refers to the set of mini-paradigms sharing the same derivational affix. All pairs of base words and the corresponding *un-* derivatives constitute the mini-class of: *kind - unkind, true - untrue, pleasant - unpleasant*,

Number of affixes	Count of base words
1	3449
2	1391
3	516
4	202
5	105
6	31
7	13
8	11
9	2
10	3
11	2

Table 5: The number of monomorphemic base words that can attach the given number of affixes (prefixes or suffixes) when forming bi-morphemic derived words.

	simple base	complex base
-*able*	70	0
-*er* (comparative)	98	0
-*er* (deverbal)	240	24
-*ly* (adverbial)	21	355
-*ness* (complex base)	0	65
-*ness* (simple base)	152	0
-*est* (superlative)	95	0
un-	18	111

Table 6: Affixes in the study based on latencies extracted from the English Lexicon Project, cross-classified by the complexity of their base words.

etc. Mini-paradigms and mini-classes approximate inflectional paradigms and inflectional classes in the sense that the semantic relations within the pairs tend to be more consistent and transparent than in general morphological families or in families of derived words with different prefixes and suffixes.

In what follows, we therefore investigate whether the measures of entropy and relative entropy are significant predictors for lexical processing when applied to mini-paradigms and mini-classes.

5.1 Materials

We selected six suffixes and one prefix, for which we extracted all formations listed in the CELEX lexical database and for which latencies were also available in the English Lexicon Project (Balota et al., 2007) for both the derived word and its base. The resulting counts of formations are available in Table 6, cross-classified by whether the base word is simple or complex. For all words, we extracted from CELEX their frequency of occurrence, their length in letters, the number of synsets for the base as listed in WordNet (Miller, 1990; Beckwith et al., 1991, and studied by Baayen et al., 2006), the family size of the base (calculated from the morphological parses in CELEX), and their frequency in the demographic subcorpus of conversational English in the British National Corpus (Burnard, 1995). From the English Lexicon Project, we added the by-item mean naming latencies and the by-item mean lexical decision latencies.

For each pair of base and derivative, we calculated its entropy and its relative entropy. For the derived words, the entropy of the mini-paradigm was calculated on the basis of the relative frequencies of the derivative and its base word (e.g., for *kind* and *unkind*, the relative frequencies are $72/(72 + 390)$ and $390/(72 + 390)$). For the base words, we distinguished between base words with only one derivative, and base words with two or more derivatives. For base words with a single derivative, the procedure for estimating the entropy was the same as for derived words. For base words with more than one derivative, the problem arises how to calculate entropies. Selection of a single derivative seems arbitrary. Taking all derivations linked with a given base word into account is possible, but then the mini-class distribution would contain the maximum number of 11 relative frequencies (see Table 5), most of which would be zero for almost all words. words would have a much smaller number of non-zero relative frequencies. We therefore opted for taking only two relative frequencies into account when calculating the entropy: the frequency of the base itself, and the summed frequency of all its derivatives.

The probability distribution for a given mini-class was obtained by summing the frequencies of all base words in the class on the one hand, and all derivatives in the class on the other hand. The resulting frequencies were then transformed into relative frequencies. These relative frequencies then served as the Q distribution (also known as the reference distribution) for the calculation of the relative entropy.

In the following analyses, frequency measures, family size, number of synsets, and response latencies were log-transformed to eliminate the adverse effect of outliers on the model fit.

5.2 Derived words

We investigated the predictivity of the entropy and relative entropy measures for word naming and lexical decision latencies to the derived words. For that, we applied linear mixed-effects modeling (Baayen et al., 2008a; Bates, 2005, 2006; Baayen, 2008), with Task (lexical decision versus naming) as a fixed-effect factor, and with the set of relevant covariates including length, base frequency, word frequency, spoken word frequency, number of synsets in WordNet, morphological family size, entropy and relative entropy. Word and affix were considered as random effects.

For the covariates, we investigated whether nonlinearity was present. This turned out to be the case only for word length. We also observed interactions of Task with word frequency and spoken word frequency, with length (only the quadratic term), and with entropy and relative entropy. Finally, we considered whether by-word or by-affix random slopes were required. It turned out that by-affix random slopes were necessary only for the two entropy measures.

Inspection of the coefficients for the entropy measures in the resulting model revealed that entropy and relative entropy had positive coefficients of similar magnitude ($H : 0.034, \hat{\sigma} = 0.025$; $RE : 0.058, \hat{\sigma} = 0.016$), with small differences across the two tasks. In word naming, the effect of entropy was slightly larger, while the effect of relative entropy was fractionally smaller (H in naming: $0.034 + 0.041$; RE in naming: $0.058 - 0.014$).

These observations invite a simplification of the regression model. Let β_0 denote the coefficient for the intercept, and let β_1 and β_2 denote the coefficients for entropy and relative entropy respectively. Given that β_1 and β_2 are very similar, we can proceed as follows:

$$\begin{aligned} \beta_0 + \beta_1 H + \beta_2 RE &\approx \beta_0 + \beta_1 H + \beta_1 RE \\ &= \beta_0 + \beta_1 (H + RE). \end{aligned} \tag{23}$$

Interestingly, the sum of entropy and relative entropy is equal to another information theoretical measure, the *cross entropy* (CE) (Manning and Schütze, 1999; Cover and Thomas, 1991). Applied to the present data, we have that

$$\begin{aligned} CE &= H + RE = \\ &= -\sum_L \mathrm{Pr}_\pi(w_L) \log_2(\mathrm{Pr}_\pi(w_L) + RE \\ &= -\sum_L \mathrm{Pr}_\pi(w_L) \log_2(\mathrm{Pr}_\pi(w_L) + \sum_L \mathrm{Pr}_\pi(w_L) \log_2 \frac{\mathrm{Pr}_\pi(w_L)}{\mathrm{Pr}_\pi(c_L)} \\ &= -\sum_L \mathrm{Pr}_\pi(w_L) \log_2(\mathrm{Pr}_\pi(c_L)). \end{aligned} \tag{24}$$

In (24), L indexes the base and derived lexemes for mini-paradigms, and the sets of base words and derived words for the mini-class. Thus, $\Pr_{\pi}(w_L)$ denotes the probability of a base or derived lexeme in its mini-paradigm, and $\Pr_{\pi}(c_L)$ denotes the corresponding probability in the mini-class. Technically, the cross entropy between the probability distribution of the mini-paradigm and the probability distribution of the mini-class measures the average number of bits needed to identify a form from the set of possible forms in the mini-paradigm, if a coding scheme is used based on the reference probability distribution $\Pr_{\pi} c_e$ of the mini-class, rather than the "true" distribution $\Pr_{\pi} w_e$ of the mini-paradigm. More informally, we can interpret the cross entropy as gauging the average amount of information in the mini-paradigm, corrected for the departure from the prior reference distribution of the corresponding mini-class.

We therefore replaced entropy H and relative entropy RE as predictors in our regression model by a single predictor, the cross entropy CE, and refitted the model to the data. After removal of outliers and refitting, we obtained the model summarized in Table 7 and visualized in Figure 5. The standard deviation of the by-word random intercepts was 0.0637, the standard deviation for the by-affix random intercepts was 0.0399, the standard deviation for the by-affix random slopes for cross entropy was 0.0277, and the standard deviation for the residual error was 0.0663. All random slopes and random intercepts were supported by likelihood ratio tests (all p-values < 0.0001).

	Estimate	Lower	Upper	P
Intercept	6.6679	6.5830	6.7607	0.0001
Task=naming	-0.1419	-0.2158	-0.0688	0.0001
length (linear)	0.0056	-0.0109	0.0228	0.5162
length (quadratic)	0.0012	0.0004	0.0020	0.0034
word frequency	-0.0382	-0.0428	-0.0333	0.0001
spoken frequency	-0.0183	-0.0245	-0.0117	0.0001
synset count	-0.0277	-0.0339	-0.0212	0.0001
cross entropy	0.0565	0.0164	0.0937	0.0076
Task=naming: word frequency	0.0067	0.0022	0.0112	0.0036
Task=naming:length (linear)	0.0132	-0.0025	0.0283	0.0914
Task=naming:length (quadratic)	-0.0011	-0.0019	-0.0003	0.0026
Task=naming:spoken frequency	0.0124	0.0062	0.0186	0.0001

Table 7: Partial effects of the predictors for the visual lexical decision and naming latencies to derived words. The reference level for Task is lexical decision. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

With respect to the control variables, we note that word length was a strongly nonlinear (positively accelerated) predictor for especially lexical decision, with longer lengths eliciting elongated response latencies. The word frequency effect was similar for both tasks, albeit slightly stronger for lexical decision. Similarly, the spoken word frequency added facilitation specifically for lexical decision. The effect of number of synonyms, as gauged with the help of the synset count, was facilitatory and the same across the two tasks. The effect of cross entropy was inhibitory, and also did not differ across tasks. Its effect size (roughly 100 ms) exceeds that of the spoken frequency effect and that of the number of meanings. Interestingly, the model with cross entropy as predictor provides an equally tight fit to the data as the model with entropy and relative entropy as predictors, even though the latter model had two additional parameters (a beta coefficient for a second entropy measure, and a random-effects standard deviation for by-item slopes for the second entropy measure): the log likelihood of the simpler model with cross entropy was 2364, while for the more complex model with entropy and relative entropy it was 2362 (a greater log likelihood implies a better fit). From this, we conclude that the relevant entropy measure for understanding the role of paradigmatic complexity during lexical processing of derived words is the cross entropy measure.

The synset measure in our data estimates the number of meanings that a base word has (e.g., *bank* as a part of the river and a financial institution). Generally, the meaning of a derivative builds on only one of the meanings of its base word (e.g., *embank*). The lower the number of synsets, the tighter we may expect the relationship between the base and its derivatives to be. The synset measure does not interact with cross entropy, nor does it substantially affect the estimate of its slope. To further rule out potential semantic confounds, we also considered a semantic measure that specifically gauges the semantic similarity between a given derived word and its base. The measure that we used is the LSA score for the distance between the derived word and its base in co-occurrence space (Landauer and Dumais, 1997), using the software available at `http://lsa.colorado.edu`. For the subset of our mini-paradigms, the LSA scores elicited a significant facilitatory effect on lexical decision latencies ($\hat{\beta} = -0.1196, p = 0.0001$). As for the synset measure, there was no significant effect for word naming. Crucially, the measure of cross entropy retained significance also when the pairwise semantic similarity between base and derived word in mini-paradigms has been taken into account.

The presence of random slopes for cross entropy in this model indicates that the effect of cross entropy varied with mini-class. Table 8 lists the individual slopes for the different mini-classes that we considered. Slopes range from 0.097 for superlative *-est* to 0.004 for *-ness* formations derived from simple base words.

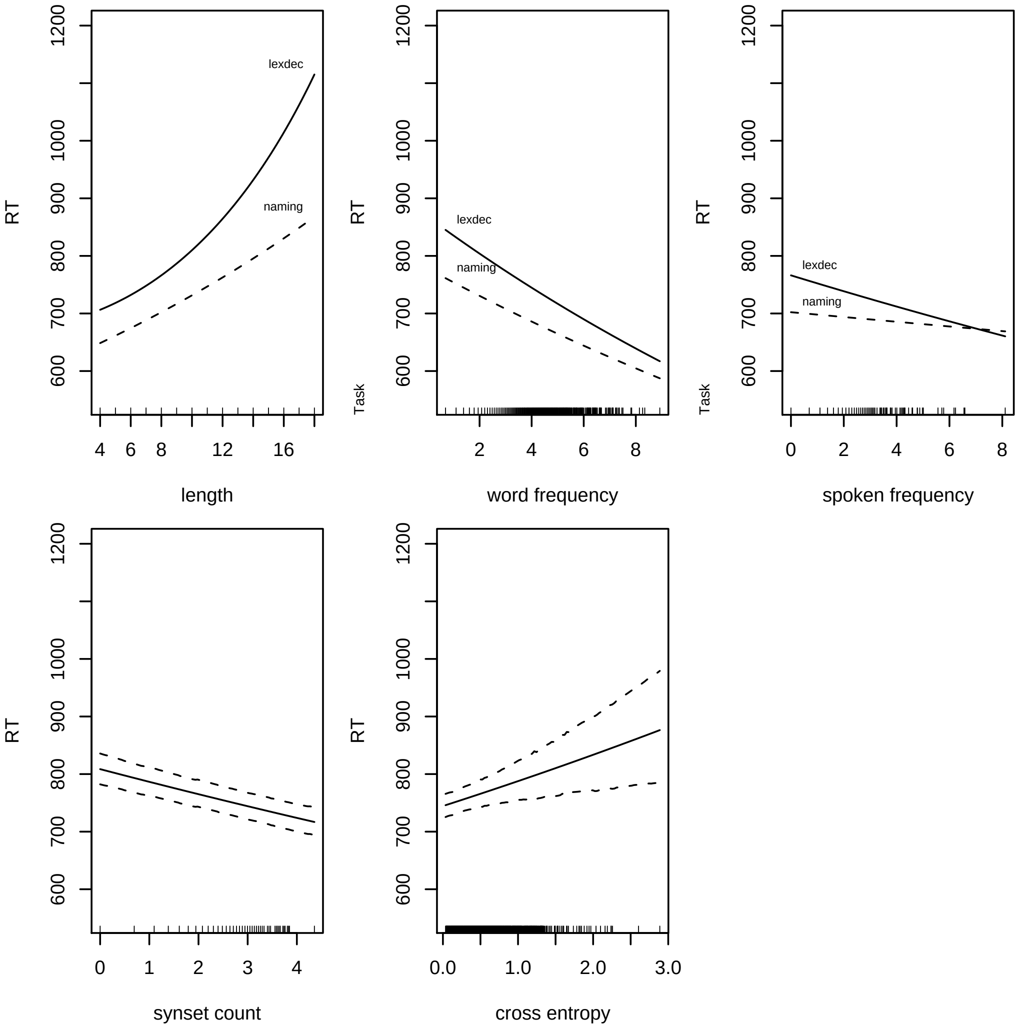

Figure 5: Partial effects of the predictors for word naming and visual lexical decision latencies for derived words. The lower panels are calibrated for visual lexical decision, and come with 95% highest posterior density confidence intervals.

	slope
-*est* (superlative)	0.097
-*ly* (adverbial)	0.090
-*ness* (complex base)	0.086
-*able*	0.068
-*er* (comparative)	0.054
-*er* (deverbal)	0.031
un-	0.021
-*ness* (simple base)	0.004

Table 8: Estimated slopes for derived words for the different mini-classes, positioned in decreasing order.

5.3 Base words

Because complex base words (e.g., *surprising*) come with predictors such as the frequency of the stem (*surprise*) that do not apply to the simple base words, we analyzed the simple and complex base words separately. We proceeded in the same way as for the derived words. We fitted a mixed-effects model to the data, observed that again the coefficients for entropy and relative entropy were very similar and statistically indistinguishable in magnitude and had the same sign, replaced the two measures by the cross entropy measure, refitted the model and removed overly influential outliers.

The coefficients of a mixed-effects model fitted to the lexical decision and naming latencies to the complex base words are listed in Table 9. The corresponding partial effects are graphed in Figure 6.

As for the preceding data sets, we find effects of word length (longer words elicit longer latencies, upper left panel) and word frequency (more frequent words elicit shorter latencies, upper center panel). Adding frequency of use in spoken English as a predictor again contributes significantly to the model over and above the written frequency measures (upper right panel). The frequency of the base word (lower left panel of Figure 6) also emerged as a significant predictor, but with a slope that is substantially shallower than that of the word frequency effect. The Synset Count of the embedded base word is predictive as well, and facilitatory just as observed for the derived words (lower center panel). Finally, the lower right panel shows that there is a small effect of cross entropy. But while for the derived words, the effect of cross entropy was inhibitory, it is facilitatory for the base words.

Before discussing this unexpected change in sign, we first inquire whether facilitation for cross entropy also characterizes the set of simple base words. Table 10 lists the partial effects of the predictors that were retained after stepwise

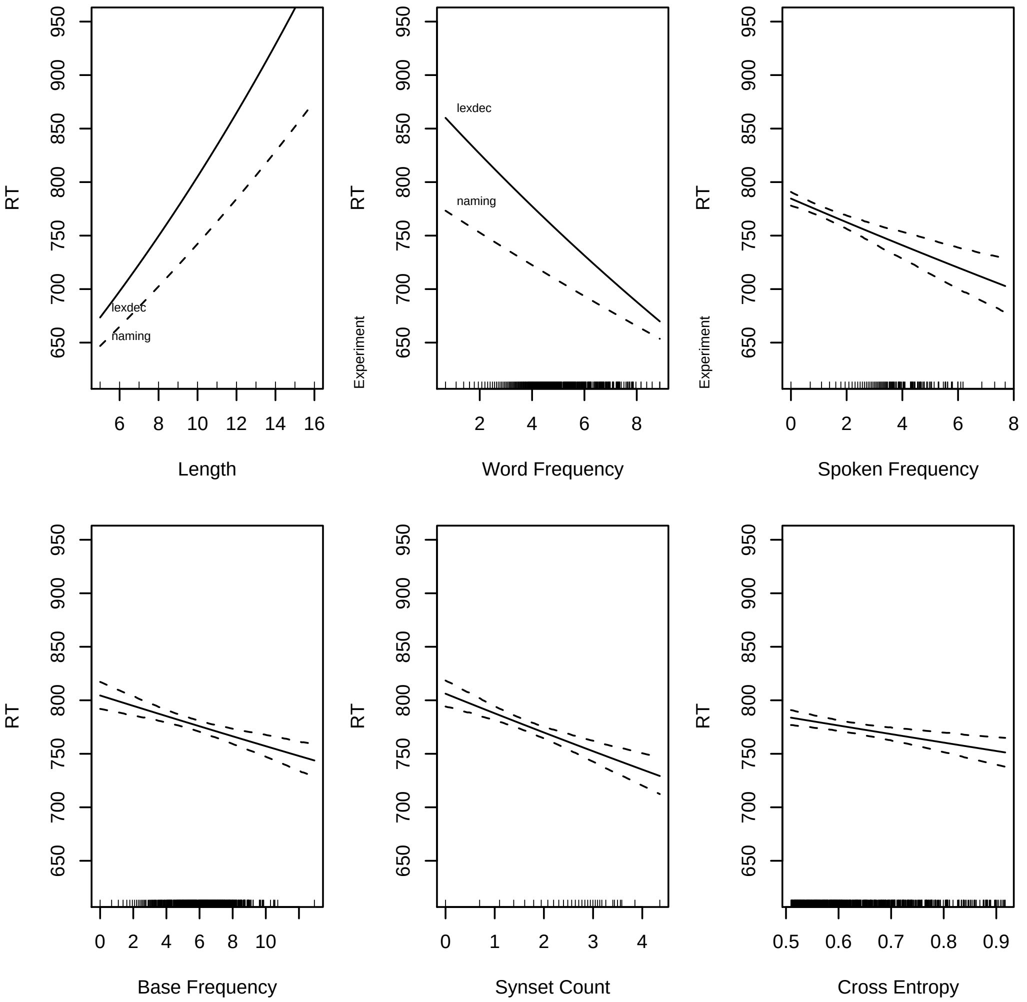

Figure 6: Partial effects of the predictors for word naming and visual lexical decision latencies for complex base words. Markov chain Monte Carlo based 95% confidence intervals are shown for those predictors that do not enter into interactions.

	Estimate	Lower	Upper	P
Intercept	6.6006	6.5428	6.6596	0.0001
Experiment=naming	-0.0397	-0.0750	-0.0031	0.0326
Length	0.0357	0.0325	0.0387	0.0001
Word Frequency	-0.0305	-0.0363	-0.0250	0.0001
Spoken Frequency	-0.0143	-0.0195	-0.0090	0.0001
Base Frequency	-0.0061	-0.0086	-0.0035	0.0001
Synset Count	-0.0230	-0.0311	-0.0147	0.0001
cross entropy	-0.1038	-0.1605	-0.0483	0.0002
Experiment=naming:Length	-0.0082	-0.0115	-0.0052	0.0001
Experiment=naming:Word Frequency	0.0100	0.0057	0.0141	0.0001

Table 9: Partial effects of the predictors for word naming and visual lexical decision latencies for complex base words. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

variable elimination. Figure 7 visualizes these partial effects. The upper left panel shows the effect of orthographic length, which shows a clear minimum near the median length (5 letters) for visual lexical decision but not for word naming. For the latter task, the shorter the word, the easier it is to articulate. For the former task, 5-letter words emerge as most easily read. The upper right panel shows that, as for the derived words, spoken frequency allows greater facilitation for visual lexical decision than for word naming.

	Estimate	Lower	Upper	P
Intercept	6.8433	6.7756	6.9097	0.0001
Experiment=naming	-0.2520	-0.3213	-0.1885	0.0001
Length (linear)	-0.0613	-0.0797	-0.0430	0.0001
Length (quadratic)	0.0067	0.0052	0.0080	0.0001
Spoken Frequency	-0.0251	-0.0286	-0.0216	0.0001
Family Size	0.0107	0.0021	0.0193	0.0158
Word Frequency	-0.0090	-0.0125	-0.0054	0.0001
cross entropy	-0.1316	-0.1823	-0.0869	0.0001
Synset Count	-0.0235	-0.0321	-0.0154	0.0001
Experiment=naming:Length (linear)	0.0507	0.0305	0.0722	0.0001
Experiment=naming:Length (quadratic)	-0.0034	-0.0050	-0.0018	0.0002
Experiment=naming:Spoken Frequency	0.0173	0.0141	0.0202	0.0001

Table 10: Partial effects of the predictors for word naming and visual lexical decision latencies for simple base words. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

The lower left panel presents the expected facilitatory effect of the Synset Count, and illustrates that words with more meanings elicit shorter latencies, for both word naming and lexical decision. Surprisingly, the lower central panel shows that the partial effect of Family Size is inhibitory, instead of facilitatory, as reported for previous experiments. We return to this finding below. The partial effect of cross entropy is presented in the lower right panel of Figure 7. As for the complex base words, the effect of cross entropy for simple base words is again facilitatory.

The analyses of the two sets of base words leave us with two questions. First, how should we understand the change in sign of the cross entropy effect between derived words and base words? Second, why do we have inhibition from the morphological family size for simple base words, and no effect of family size for complex base words?

With respect to the first question, we note that for base words there is bottom-up support for only the base word, and no such support for their derivatives. In the case of derived words, by contrast, there is bottom-up support for the derived word itself, its base word, and its affix. In other words, for derived words, three of the four elements in a proportional analogy such as

$$\underbrace{\textit{great} : \textit{greatest}}_{\text{mini paradigm}} = \underbrace{\text{A} : \textit{-est}}_{\text{mini class}} \tag{25}$$

are actually present in the signal. For derived words, we can therefore understand the effect of cross entropy as reflecting the cost of resolving the proportional analogy between mini-paradigm and mini-class. More specifically, the cross entropy reflects the average complexity of identifying the derived word in its mini-paradigm on the basis of the generalized probability distribution of the mini-class. Thus, the cross entropy can be understood as reflecting the cost of resolving the ambiguity in the visual input with the help of generalized knowledge in long-term memory about the corresponding mini-class. From this perspective, the inhibitory effect of cross entropy for derived words makes perfect sense: The higher the cross entropy, the more information has to be retrieved from memory to resolve the proportional analogy.

Let us now consider the facilitatory effect of cross entropy for simple base words. For simple base words, the visual input is unambiguous, with bottom-up support only for the word itself. There is no cost of a call on proportional analogy to resolve morphological ambiguity. In the absence of a morphological parsing problem, the cross entropy effect apparently reverses and emerges as a measure of the amount of support the base receives from related derived words co-activated by the base. Crucially, it is not simply the count of related derived words (we checked that this count is not predictive for the present data) but rather the

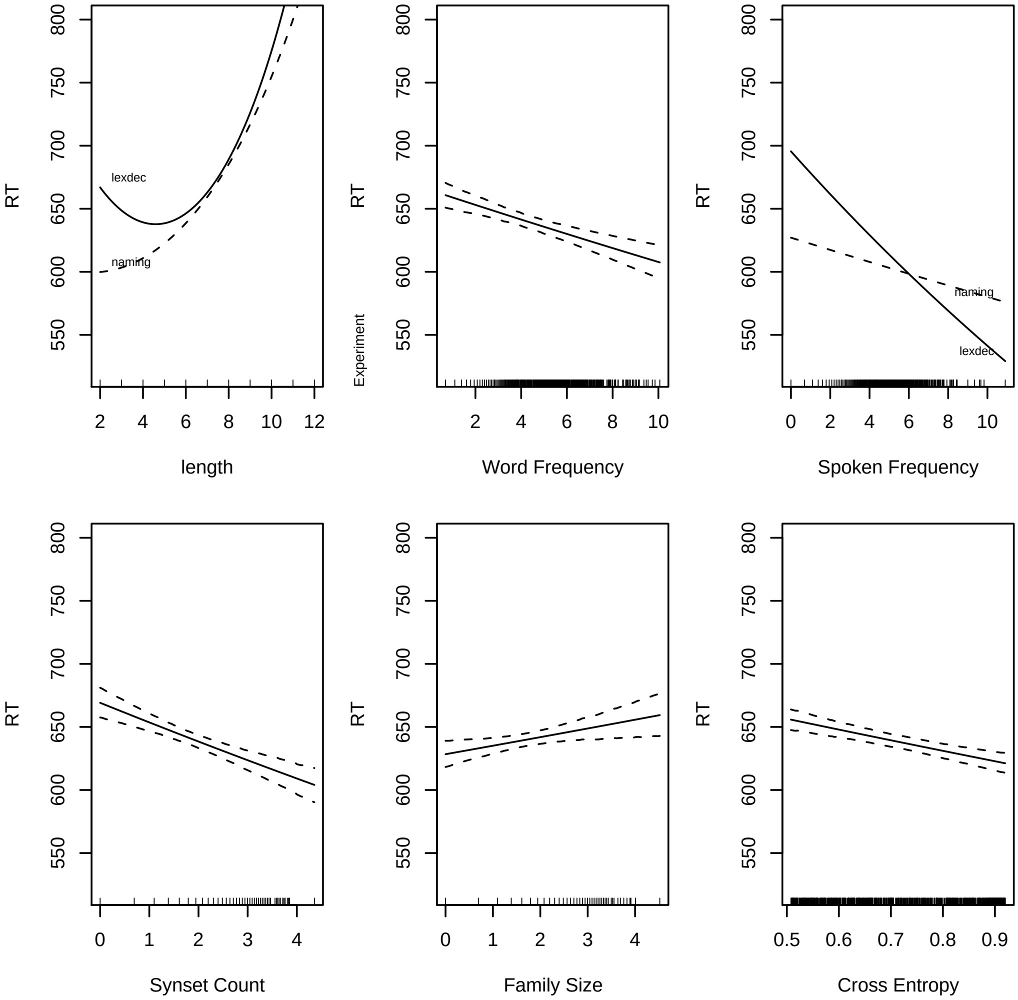

Figure 7: Partial effects of the predictors for word naming and visual lexical decision latencies for simple base words. Markov chain Monte Carlo based 95% confidence intervals are shown for those predictors that do not enter into interactions.

	Frequency	Family	Synset	cross	RT	RT
		Size	Count	entropy	lexdec	naming
Frequency	1.000	0.320	0.345	-0.527	-0.379	-0.266
Family Size	0.320	1.000	0.643	0.245	-0.473	-0.392
Synset Count	0.345	0.643	1.000	0.092	-0.552	-0.434
cross entropy	-0.527	0.245	0.092	1.000	-0.085	-0.101
RT lexical decision	-0.379	-0.473	-0.552	-0.085	1.000	0.648
RT naming	-0.266	-0.392	-0.434	-0.101	0.648	1.000

Table 11: Pairwise correlations between key predictors and lexical decision (lexdec) and naming latencies for the set of simple base words.

analogical support for the base given its derivative (defined in the mini-paradigm) and the general likelihood of a base word having derivatives (defined in the mini-class).

The second question to be considered is why we observe inhibition from the morphological family size for simple base words, and no effect of family size for complex base words. The unexpected inhibitory effect of family size is probably due to what is known in the statistical literature as suppression (see, e.g., Friedman and Wall, 2005): When predictor variables are correlated, and both are correlated with the dependent variable, then, depending on the strength of the former correlation, the beta coefficient of one of the predictors can become non-significant or even change sign. Table 11 presents the correlation matrix for key predictors, and reveals a large positive coefficient for the correlation of Family Size and the Synset Count, and the expected negative correlations for Family Size and response latencies in lexical decision and naming. This by itself is a warning that suppression might be at issue here.

We therefore inspected whether Family Size was significant in a model for the simple base words, excluding the Synset Count as predictor. It was not ($p > 0.8$). When cross entropy was also removed as predictor, the Family Size measure emerged as significant ($p < 0.01$), now with a negative slope, as expected given previous studies. For the complex base words, excluding only the Synset measure was sufficient to allow a facilitatory effect of Family Size to emerge. What this suggests is that the Family Size effect, which has always been understood as a semantic effect (see, e.g., Schreuder and Baayen, 1997; Moscoso del Prado Martín et al., 2004a), would be a composite effect that bundles effects of semantic similarity and effects of paradigmatic structure. Effects of similarity would then be better captured by means of the Synset Count, and effects of derivational paradigmatic structure would then be better captured by means of the cross entropy measure.

The question that arises at this point is whether the semantic aspect of the Family Size effect has no specifically morphological component whatsoever. To answer this question, we first partioned the Synset Count into two disjunct counts, a count for morphologically related synsets, and a count for morphologically unrelated synsets. A morphologically related synset is a synset in which at least one of the synset members is morphologically related to the target word (not counting the target word itself). A morphologically related synset, therefore, is a family size count that only includes semantically highly related family members.

In the model for the simple base words, we then replaced the Family Size measure and the Synset Count by the counts of morphologically related and unrelated synset counts. A mixed-effects analysis revealed that for visual lexical decision both counts were significant predictors with very similar coefficients (-0.018 and -0.015 respectively). For the naming latencies, however, only the synset count of morphologically unrelated synsets was significant. This interaction ($p = 0.0049$) shows that in a task such as word naming, which does not require deep semantic processing, semantic ambiguity that arises through morphological connectivity does not play a role. By contrast, the lexical decision task, which invites deeper semantic processing, allows the effect of morphologically related words that are also very similar in meaning to become visible. We therefore conclude that morphologically related words that are also semantically very similar have a special status compared to semantically similar but morphologically unrelated words (see also Moscoso del Prado Martín et al., 2004a).

6 Concluding remarks

In the preceding sections, we reviewed and presented studies each of which addressed a specific aspect of the complexities of paradigmatic structure in lexical processing. In order to obtain a model for the full complexity for an inflected variant w_e, we combine equations (10), (14), and (15) and add the effects of the entropy and relative entropy measures, leading to the following equation:

$$\begin{aligned} I \propto \beta_0 \quad &+ \quad \beta_1 \log_2 \mathrm{Pr}_N(w_e) + \beta_2 \log_2 \mathrm{Pr}_N(w) + \\ &+ \quad \beta_3 \log_2 \left(\frac{\mathrm{Pr}_\pi(e)/R_e}{\sum_e \mathrm{Pr}_\pi(e)/R_e} \right) + \\ &+ \quad \beta_4 \log_2 \left(\frac{\mathrm{Pr}_\pi(w_e)/R_e}{\sum_e \mathrm{Pr}_\pi(w_e)/R_e} \right) + \\ &+ \quad \beta_5 H_d + \\ &+ \quad \beta_6 H_i + \beta_7 RE. \end{aligned} \tag{26}$$

Large regression studies are called for that bring all these variables into play simultaneously. However, even though (26) is far from simple, it is only a first step towards quantifying the complexities of inflectional processing. We mention here only a few of the issues that should be considered for a more comprehensive model.

First, Kostić et al. (2003) calculated the number of functions and meanings R_e of exponent e conditionally on a lexeme's inflectional class. For instance, the number of functions and meanings listed for the exponent *a* for masculine nouns in Table 2, 109, is the sum of the numbers of functions and meanings for masculine genitive and the masculine accusative singular. This provides a lower bound for the actual ambiguity of the exponent, as the same exponent is found for nominative singulars and genitive plurals for regular feminine nouns. The justification for conditioning on inflectional class is that the stem to which an exponent attaches arguably provides information about its inflectional class. This reduces the uncertainty about the functions and meanings of an exponent to the uncertainty in its own class. Nevertheless, it seems likely that an exponent that is unique to one inflectional class (e.g., Serbian *ama* for regular feminine nouns) is easier to process than an exponent that occurs across all inflectional classes (e.g., *a, u*), especially when experimental items are not blocked by inflectional class. (Further complications that should be considered are the consequences of, for instance, masculine nouns (e.g., *sudija* "judge", *sluga* "servant") taking the same inflectional exponents as regular feminine nouns do, and of animate masculine nouns being associated with a pattern of exponents that differs from that associated with inanimate masculine nouns.)

Second, the standard organization of exponents by number and case has not played a role in the studies that we discussed. Thus far, preliminary analyses of the experimental data available to us have not revealed an independent predictive role for case, over and above the attested role of ambiguity with respect to numbers of functions and meanings. This is certainly an issue that requires further empirical investigation, as organization by case provides insight into the way that functions and meanings are bundled across inflectional classes.

Third, we have not considered generalizations across, for instance, irregular and regular feminine nouns in Serbian, along the lines of Clahsen et al. (2001). The extent to which inflected forms inherit higher-order generalizations about their phonological form provides further constraints on lexical processing.

Fourth, the size of inflectional paradigms has not been investigated systematically. Although the nominal inflectional classes of Serbian are an enormous step forward compared to the nominal paradigms of English or Dutch, the complexities of verbal paradigms can be orders of magnitude larger. From an information-theoretic perspective, the entropy of the complex verbal paradigms of Serbian must be much larger than the entropy of nominal paradigms, and one would expect this difference to be reflected in elongated processing latencies for inflected verbs. The study by Traficante and Burani (2003) provides evidence supporting this prediction. They observed that inflected verbs in Italian elicited longer processing latencies than inflected adjectives.

Fifth, all results reported here are based on visual comprehension tasks (lexical decision, word naming). Some of the present results are bound to change as this line of research is extended to other tasks and across modalities. For instance, the effect of inflectional entropy reported by Baayen et al. (2006) for visual lexical decision and word naming was facilitatory in nature. However, in a production study by Bien (2007), inflectional entropy was inhibitory. In lexical decision, a complex paradigm is an index of higher lexicality, and may therefore elicit shorter response latencies. In production, however, the paradigm has to be accessed, and a specific word form has to be extracted from the paradigm. This may explain why in production a greater paradigm complexity goes hand in hand with increasing processing costs. More in general, it will be important to establish paradigmatic effects for lexical processing in natural discourse using tasks that do not, or only minimally, impose their own constraints on processing.

Sixth, it will be equally important to obtain distributional lexical measures that are more sensitive to contextual variation than the abstract frequency counts and theoretical concepts of functions and meanings that have been used thus far. Interestingly, Moscoso del Prado Martín et al. (2008) and Filipović Đurđević (2007) report excellent predictivity for lexical processing of more complex information theoretic measures of morphological and semantic connectivity derived bottom-up from a corpus of Serbian.

To conclude, it is clear that the information theoretic measures that we have proposed and illustrated in this chapter capture only part of the multidimensional complexity of lexical processing. Each measure by itself presents, as it were, only one plane cross-cutting this multidimensional space. In spite of these limitations, the extent to which the present information-theoretic approach converges with Word and Paradigm morphology is striking. Across our experimental data sets we find evidence for exemplars, irrespective of whether the language under investigation is Dutch, English, or Serbian. At the same time, we observe the predictivity of entropy measures, which generalize across probability distributions tied to subsets of these exemplars, and evaluate the complexity of paradigms and the divergence between different levels of morphological organization. However, all the results discussed here pertain to the processing of familiar words. In order to properly gauge the processing complexity of new inflected and derived words, it will be necessary to combine Word and Paradigm morphology and the present information theoretic approach with memory-based computational models of language processing (Daelemans and Van den Bosch, 2005).

References

Anderson, S. R. (1992). *A-morphous morphology*. Cambridge University Press, Cambridge.

Aronoff, M. (1994). *Morphology by itself: stems and inflectional classes*. The MIT Press, Cambridge, Mass.

Baayen, R. (2003). Probabilistic approaches to morphology. In Bod, R., Hay, J., and Jannedy, S., editors, *Probability theory in linguistics*, pages 229–287. The MIT Press.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge University Press, Cambridge (in press).

Baayen, R. H., Davidson, D. J., and Bates, D. (2008a). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, page in press.

Baayen, R. H., Feldman, L., and Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 53:496–512.

Baayen, R. H., McQueen, J., Dijkstra, T., and Schreuder, R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In Baayen, R. H. and Schreuder, R., editors, *Morphological structure in language processing*, pages 355–390. Mouton de Gruyter, Berlin.

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

Baayen, R. H., Wurm, L. H., and Aycock, J. (2008b). Lexical dynamics for low-frequency complex words. a regression study across tasks and modalities. *The Mental Lexicon*, 2:419–463.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Bates, D. (2006). Linear mixed model implementation in lme4. Department of Statistics, University of Wisconsin-Madison.

Bates, D. M. (2005). Fitting linear mixed models in R. *R News*, 5:27–30.

Beard, R. (1995). *Lexeme-morpheme base morphology: a general theory of inflection and word formation*. State University of New York Press, Albany, NY.

Beckwith, R., Fellbaum, C., Gross, D., and Miller, G. (1991). WordNet: A lexical database organized on psycholinguistic principles. In Zernik, U., editor, *Lexical Acquisition. Exploiting On-Line Resources to Build a Lexicon*, pages 211–232. Lawrence Erlbaum Associates, Hillsdale, NJ.

Bien, H. (2007). *On the production of morphologically complex words with special attention to effects of frequency*. Max Planck Institute for Psycholinguistics, Nijmegen.

Blevins, J. P. (2003). Stems and paradigms. *Language*, 79:737–767.

Blevins, J. P. (2006). English inflection and derivation. In Aarts, B. and McMahon, A. M., editors, *Handbook of English Linguistics*, pages 507–536. Blackwell, London.

Burnard, L. (1995). *Users guide for the British National Corpus*. British National Corpus consortium, Oxford university computing service.

Clahsen, H., Hadler, M., Eisenbeiss, S., and Sonnenstuhl-Henning, I. (2001). Morphological paradigms in language processing and language disorders. *Transactions of the Philological Society*, 99(2):247–277.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons, New York.

Daelemans, W. and Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge University Press, Cambridge.

Filipović Đurđević, D. (2007). *The polysemy effect in processing of Serbian nouns*. PhD thesis, University of Belgrade, Serbia.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, 59:127–136.

Gagné, C. (2001). Relation and lexical priming during the interpretation of noun-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27:236–254.

Gagné, C. and Shoben, E. J. (1997). The influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:71–87.

Halle, M. and Marantz, A. (1993). Distributed morphology and the pieces of inflection. In Hale, K. and Keyser, S. J., editors, *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, volume 24 of *Current Studies in Linguistics*, pages 111–176. MIT Press, Cambridge, Mass.

Hockett, C. (1954). Two models of grammatical description. *Word*, 10:210–231.

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., editor, *Morphological Aspects of Language Processing*. Lawrence Erlbaum Inc. Publishers, New Jersey.

Kostić, A. (2008). The effect of the amount of information on language processing.

Kostić, A., Marković, T., and Baucal, A. (2003). Inflectional morphology and word meaning: Orthogonal or co-implicative domains? In Baayen, R. H. and Schreuder, R., editors, *Morphological Structure in Language Processing*, pages 1–44. Mouton de Gruyter, Berlin.

Kuperman, V., Bertram, R., and Baayen, R. H. (2008). Morphological dynamics in compound processing. *manuscript submitted for publication*, Radboud University Nijmegen:1–37.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177.

Luce, R. D. (1959). *Individual choice behavior*. Wiley, New York.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge.

Matthews, P. H. (1974). *Morphology. An Introduction to the Theory of Word Structure*. Cambridge University Press, London.

Milin, P., Filipović Đurđević, D., and Moscoso del Prado Martín, F. (2008). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Manuscript submitted for publication*.

Miller, G. A. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–312.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., and Baayen, R. H. (2004a). 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*, 30:1271–1278.

Moscoso del Prado Martín, F., Kostić, A., and Baayen, R. H. (2004b). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, 94:1–18.

Moscoso del Prado Martín, F., Kostić, A., and Filipović Đurđević, D. (2008). The missing link between morphemic assemblies and behavioral responses: a Bayesian Information-Theoretical model of lexical processing. *Manuscript submitted for publication.*

New, B., Brysbaert, M., Segui, F. L., and Rastle, K. (2004). The processing of singular and plural nouns in French and English. *Journal of Memory and Language*, 51:568–585.

Pinker, S. (1991). Rules of language. *Science*, 153:530–535.

Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. Weidenfeld and Nicolson, London.

Ross, S. M. (1988). *A First Cource in Probability*. Macmillan Publishing Company, New York.

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

Stemberger, J. P. and MacWhinney, B. (1986). Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition*, 14:17–26.

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

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes*, 9(3):271–294.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, 57A:745–765.

Traficante, D. and Burani, C. (2003). Visual processing of Italian verbs and adjectives: the role of the inflectional family size. In Baayen, R. H. and Schreuder, R., editors, *Morphological structure in language processing*, pages 45–64. Mouton de Gruyter, Berlin.