# An amorphous model for morphological processing in visual comprehension based on naive discriminative learning

R. Harald Baayen
University of Alberta

Petar Milin
University of Novi Sad
Laboratory for Experimental Psychology, University of Belgrade

Dusica Filipović Đurđević
University of Novi Sad
Laboratory for Experimental Psychology, University of Belgrade

Peter Hendrix
University of Alberta

Marco Marelli
University of Milano-Bicocca

**Abstract**

A two-layer symbolic network model based on the equilibrium equations of the Rescorla-Wagner model (Danks, 2003) is proposed. The study starts by presenting two experiments in Serbian, which reveal for sentential reading the inflectional paradigmatic effects previously observed by Milin, Filipović Đurđević, and Moscoso del Prado Martín (2009) for unprimed lexical decision. The empirical results are successfully modeled without having to assume separate representations for inflections or data structures such as inflectional paradigms. In the next step, the same naive discriminative learning approach is pitted against a wide range of effects documented in the morphological processing literature. Frequency effects for complex words as well as for phrases (Arnon & Snider, 2010) emerge in the model without the presence of whole-word or whole-phrase representations. Family size effects (Schreuder & Baayen, 1997; Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 2004) emerge in the simulations across simple words, derived words, and compounds, without derived words or compounds being represented as such. It is shown that for pseudo-derived words no special morpho-orthographic segmentation mechanism as posited by Rastle, Davis, and New (2004) is required. The model also replicates the finding of Plag and Baayen (2009), that, on average, words with more productive affixes elicit longer response latencies, while at the same time predicting that productive affixes afford faster response latencies for new words. English phrasal paradigmatic effects modulating isolated word reading are reported and modelled, showing that the paradigmatic effects characterizing Serbian case inflection have cross-linguistic scope.

**Keywords:** naive discriminative learning, morphological processing, reading, compound cue theory, Rescorla-Wagner equations, weighted relative entropy, a-morphous morphology.

In traditional views of morphology, just as simple words consist of phonemes, complex words are composed of discrete morphemes. In this view, morphemes are signs linking form to meaning. A word such as *goodness* is analysed as consisting of two signs, the free morpheme *good*, and the bound morpheme *-ness*. When reading *goodness*, the constituents *good* and *-ness* are parsed out, and subsequently the meaning of the whole word, "the quality of being good" (in any of the various senses of good) is computed from the meanings of the constituent morphemes.

The morphemic view has been very influential in psycholinguistic studies of morphological processing. Many studies have addressed the question of whether the parsing of a complex word into its constituents is an obligatory and automatic process (e.g., Taft & Forster, 1975; Taft, 2004; Rastle et al., 2004) and have investigated the consequences of such obligatory decomposition for words that are not morphologically complex (e.g., *corner* versus *walk-er*, *reindeer* (not *re-in-de-er*) versus *re-in-state*). Priming manipulations have been used extensively to show that morphological effects are stronger than would be expected from form or meaning overlap alone (e.g., Feldman, 2000). Other studies have addressed the consequences of the breakdown of compositionality, both for derived words *business* ('company', not 'the quality of being busy') and compounds (*hogwash*, 'nonsense') (see, e.g., Marslen-Wilson, Tyler, Waksler, & Older, 1994; Libben, Gibson, Yoon, & Sandra, 2003; Schreuder, Burani, & Baayen, 2003). Furthermore, frequency effects have often been used as diagnostics for the existence of representations, with whole-word frequency effects providing evidence for representations for complex words, and morphemic frequency effects pointing to morpheme-specific representations (e.g., Taft & Forster, 1976a; Taft, 1979, 1994; Baayen, Dijkstra, & Schreuder, 1997).

In current theoretical morphology, however, the morpheme does not play an important role. One reason is that, contrary to what one would expect for a linguistic sign, bound morphemes often express a range of very different meanings. In English, the formative *-er* is used for deverbal nouns (*walk-er*) but also for comparatives (*greater*). The suffix *-s* indicates plural on nouns (*legs*), singular on verbs (*walks*), and also the possessive (*John's legs*). In highly inflecting languages such as Serbian, the case ending *-i* indicates dative or locative singular for regular feminine nouns (*a* class), but nominative plural for masculine nouns.

A second reason is that formatives often pack together several meanings, often only semi-systematically. For instance, in Latin, the formatives for the present passive contain an *r* as part of their form, but this *r* can appear initially (*-r, -ris*, first and second person singular) or final (*-tur, -mur, -ntur*, third person singular, first and third person plural). The exception is the formative for the second person plural, which does not contain an *r* at all (*-mini*). Thus, the presence of an *r* in a verb ending is a good, although not perfect, indicator of passive meaning. To complicate matters even further, the very same passive formatives are used on selected verbs to express active instead of passive meaning, indicating that the interpretation of these formatives is highly context-dependent. This is not what

We are indebted for comments, discussion, and feedback to two reviewers, Jorn and Corine Baayen, Jim Blevins, Geert Booij, Paula Chesley, Victor Kuperman, Janet Pierrehumbert, Ingo Plag, Michael Ramscar, and audiences at presentations in Tübingen, Tucson, Freiburg, Pisa, Siegen, Groningen, York, and San Diego. This work was partially supported by the Ministry of Science and Environmental Protection of the Republic of Serbia (grant number: 149039D).

one would expect if these formatives were bona fide linguistic signs.

A third reason is that some languages shamelessly reuse inflected forms as input for further case inflections, as exemplified by Estonian non-nominative plural case endings attaching to the partitive singular (Erelt, 2003). For instance, *jalg* ('foot', nominative) has as singular case endings forms such as *jalga* (partitive), *jala* (genitive) and *jalast* (elative). The corresponding plural case endings are *jalad* (nominative), *jalgasid* (partitive), *jalgade* (genitive) and *jalgadest* (elative). Even though the form of the partitive singular is present in the plural non-nominative case endings, it does not make any semantic contribution to these plural forms (and therefore often analysed as a stem allomorph).

A fourth reason is that form-meaning relationships can be present without the need of morphemic decomposition. Phonaesthemes, such as *gl-* in *glow, glare, gloom, gleam, glimmer* and *glint*, provide one example, the initial *wh* of the question words of English (*who, why, which, whether, where, . . .* ) provides another (Bloomfield, 1933). Furthermore, blends (e.g., *brunch*, from *breakfast* and *lunch*) share aspects of compositionality without allowing a normal parse (see, e.g., Gries, 2004, 2006).

A fifth reason is that inflectional formatives often express several grammatical meanings simultaneously. For instance, the inflectional exponent *a* for Serbian regular feminine nouns expresses either nominative and singular, or genitive and plural. Similarly, normal signs such as *tree* may have various shades of meaning (such as 'any perennial woody plant of considerable size', 'a piece of timber', 'a cross', 'gallows'), but these different shades of meaning are usually not intended simultaneously in the way that nominative and singular (or genitive and plural) are expressed simultaneously by the *a* exponent.

A final reason is that in richly inflecting languages, the interpretation of an inflectional formative depends on the inflectional paradigm of the base word it attaches to. For instance, the abovementioned Serbian case ending *-a* can denote not only nominative singular or genitive plural for regular feminine nouns, but also genitive singular and plural for regular masculine nouns. Moreover, for a subclass of masculine animate nouns, accusative singular forms make use of the same exponent *-a*. The ambiguity of this case ending is resolved, however, if one knows dative/instrumental/locative plural endings for feminine and masculine nouns (*-ama* vs. *-ima*, respectively). In other words, resolving the ambiguity of a case ending depends not only on contextual information in the preceding or following discourse (syntagmatic information), but also on knowledge of the other inflected forms in which a word can appear (paradigmatic information).

Considerations such as these suggest that the metaphor of morphology as a formal calculus with morphemes as basic symbols, and morphological rules defining well-formed strings as well as providing a semantic interpretation, much as a pocket calculator interprets $2 + 3$ as 5, is inappropriate. Many studies of word formation have concluded that more insightful analyses can be obtained by taking the word as the basic unit of morphological analysis (for details, and more complex arguments against a beads-on-a-string model of morphology (also known as 'item-and-arrangement morphology'), see, e.g., Matthews, 1974; Hockett, 1987; S. Anderson, 1992; Aronoff, 1994; Beard, 1995; Blevins, 2003, 2006; Booij, 2010).

The following quote from Hocket (1987:84) is informative, especially as in early work Hockett himself had helped develop an 'item-and-arrangement' model of morphology that he later regarded as inadequate:

> In 1953 Floyd Lounsbury tried to tell us what we were doing with our clever morphophonemic techniques. We were providing alternations by devising an 'agglutinative analog' of the language and formulating rules that would convert expressions in that analog into the shapes in which they are actually uttered. Of course, even such an agglutinative analog , with its accompanying conversion rules, could be interpreted merely as a descriptive device. But it was not in general taken that way; instead, it was taken as a direct reflection of reality. We seemed to be convinced that, whatever might superficially appear to be the case, every language is 'really' agglutinative.

It is worth noting that in a regular agglutinating language such as Turkish, morphological formatives can be regarded as morphemes contributing their own meanings in a compositional calculus. However, in order to understand morphological processing across human languages, a general algorithmic theory is required that covers both the many non-agglutinative systems as well as more agglutinative-like systems.

If the trend in current linguistic morphology is moving in the right direction, the questions of whether and how a complex word is decomposed during reading into its constituent morphemes are not the optimal questions to pursue. A first relevant question in 'a-morphous' approaches to morphological processing is how a complex word activates the proper meanings, without necessarily assuming intermediate representations supposedly negotiating between the orthographic input and semantics. A second important question concerns the role of paradigmatic relations during lexical processing.

Of the many models proposed for morphological processing in the psycholinguistic literature, the insights of a-morphous morphology fit best with aspects of the the triangle model of Harm and Seidenberg (1999); Seidenberg and Gonnerman (2000); Plaut and Gonnerman (2000); Harm and Seidenberg (2004). This connectionist model maps orthographic input units onto semantic units without intervening morphological units. The triangle model also incorporates phonological knowledge, seeking to simulate reading aloud within one unified system highly sensitive to the distributional properties of the input, where other models posit two separate streams (orthography to meaning, and orthography to phonology, see, e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Borowsky et al., 2006).

In what follows, we propose a computational model, the "naive discriminative reader", which models morphological processing with an architecture directly mapping form onto meaning, without using specific representations for either bound morphemes or for complex words. The model follows the triangle model, but differs in various ways. First, it works with just two levels, orthography and meaning. In this study, we do not address reading aloud, focusing instead on developing a model that properly predicts morphological effects in comprehension. Second, there are no hidden layers mediating the mapping of form onto meaning. Third, the representations that we use for coding the orthographic input and semantic output are symbolic rather than subsymbolic. Fourth, our model makes use of a simple algorithm based on discriminative learning to efficiently estimate the weights on the connections from form to meaning, instead of backpropagation. The research strategy pursued in the present study is to formulate the simplest probabilistic architecture that is sufficiently powerful to predict the kind of morphological effects documented in the processing literature.

Of special interest to our modeling effort are two general classes of phenomena that suggest a form of 'entanglement' of words with morphologically related words during lexical processing. Schreuder and Baayen (1997) documented for simple words that the type count of morphologically related words co-determines processing latencies in visual lexical decision. This 'family size' effect has been replicated for complex words and emerges also in languages such as Hebrew and Finnish (De Jong, Schreuder, & Baayen, 2000; Moscoso del Prado Martín, Kostić, & Baayen, 2004; Moscoso del Prado Martín et al., 2005; Moscoso del Prado Martín et al., 2004; Baayen, 2010). One interpretation of the family size effect, formulated within the framework of the multiple read-out model of Grainger and Jacobs (1996), assumes that a word with a large family co-activates many family members, thereby creating more lexical activity and hence providing more evidence for a yes-response in lexical decision. Another explanation assumes that resonance within the network of family members boosts the activation of the input word (De Jong, Schreuder, & Baayen, 2003). In the present study, we pursue a third explanation, following Moscoso del Prado Martín (2003, chapter 10), according to which family size effects can emerge straightforwardly in networks mapping forms onto meanings.

The second class of phenomena of interest to us revolves around the processing of inflected words that enter into extensive, highly structured paradigmatic relations with other inflected words. Milin, Filipović Đurđević, and Moscoso del Prado Martín (2009) showed, for Serbian nouns inflected for case and number, that response latencies in the visual lexical decision task are co-determined by both the probabilities of a word's other case endings, and the probabilities of these case endings in that word's inflectional class. More precisely, the more a given word's probability distribution of case inflections differs from the corresponding distribution of its inflectional class, the longer response latencies are.

There are two main options for understanding these results. Under one interpretation, case-inflected variants are stored in memory, with computations over paradigmatically structured sets of exemplars giving rise to the observed effects. This explanation is extremely costly in the number of lexical representations that have to be assumed to be available in memory. We therefore pursue a different explanation, one that is extremely parsimonious in the number of representations required. We will show that these paradigmatic effects can arise in a simple discriminative network associating forms with meanings. Crucially, the network does not contain any representations for complex words — the network embodies a fully compositional probabilistic memory activating meanings given forms.

Although in generative grammar, morphology and syntax have been strictly separated (for an exception, see, e.g., Lieber, 1992), approaches within the general framework of construction grammar (Goldberg, 2006; Booij, 2005, 2009; Dabrowska, 2009; Booij, 2010) view the distinction between morphology and syntax as gradient. In this framework, the grammar is an inventory of constructions relating form to meaning. From a structural perspective, morphological constructions differ from phrasal or syntactic constructions only in lesser internal complexity. From a processing perspective, morphological constructions, being smaller, should be more likely to leave traces in memory than syntactic constructions. However, at the boundary, similar familiarity effects due to past experience are predicted to arise for both larger complex words and smaller word n-grams. Interestingly, frequency effects have been established not only for (regular) morphologically complex words (see, e.g., Baayen et al., 1997; Baayen, Wurm, & Aycock, 2007; Kuperman, Schreuder, Bertram, & Baayen, 2009), but recently for short sequences of words as well (Arnon & Snider, 2010; Bannard & Matthews, 2008; Shaoul, Westbury, & Baayen, 2009; Tremblay & Baayen, 2010).

If phrasal frequency effects are of the same kind as frequency effects for complex words, it becomes highly questionable that frequency effects should be interpreted as reflecting whole-word or whole-phrase representations, given the astronomical numbers of words and phrases that would have to be stored in memory. We will show that whole-word frequency effects as well as phrasal frequency effects can arise in the context of discriminative learning, without having to posit separate representations for words or phrases.

Finally, we will also document, as well as model, phrasal paradigmatic effects for English monomorphemic words that parallel the paradigmatic effects for Serbian number and case inflection.

In what follows, we first introduce two experiments that provide further evidence for inflectional paradigmatic effects for Serbian nouns first reported by Milin, Filipović Đurđević, and Moscoso del Prado Martín (2009). These experiments will shed further light on whether these effects persist in sentential reading, on whether they survive the presence of a prime, and whether they are modulated by sentential context. The remainder of the paper addresses the computational modeling of lexical processing. After presenting the naive discriminative reader model, we first show that this model provides a close fit to the Serbian experimental data. We then proceed with pitting the predictions of the naive discriminative reader against the observed visual lexical decision latencies available in the English Lexicon Project (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004). We discuss a range of data subsets for English: simple words, inflected words, derived words, pseudo-derived words, words with phonaesthemes, compounds, and finally phrasal effects on the reading of simple words. In the general discussion, we compare the present approach to other computational models, including a more detailed comparison with the Bayesian Reader of Norris (2006).

## Experiment 1

Inflectional paradigms in English are extremely simple compared to the paradigms for case inflection on nouns or the paradigms for verbal inflections found in languages such as Finnish, Italian, or Serbian. Whereas English nouns distinguish between singular and plural forms, nouns in Serbian are inflected for both number and case, distinguishing between six cases: nominative, genitive, dative, accusative, locative, and instrumental. (In classical Serbian, there is a seventh case, the vocative. This case is hardly functional in modern Serbian (Kostić, 1965), and will therefore not be considered in the present study.) In addition, Serbian nouns belong to one of the three genders, masculine, feminine, and neuter, and fall into four inflectional classes, each of which realize combinations of number and case in their own distinct way. As in Latin, inflectional endings (exponents) can be ambiguous. For instance, for regular feminine nouns, the nominative singular and the genitive plural are identical, and the same holds for the genitive singular and the nominative and accusative plural. Further such examples can be found in the example paradigms shown in Table 1.

Milin, Filipović Đurđević, and Moscoso del Prado Martín (2009) addressed the processing of Serbian case paradigms by focusing on the unique forms of a noun, while differentiating between inflectional classes. For each inflectional class, these authors calculated the

Table 1: Examples of inflectional paradigms for Serbian nouns: "žena" (*women*, feminine) and "prozor" (*window*, masculine). Frequencies taken from Kostić (1999).

| | | FEMININE | | | MASCULINE | | |
|---|---|---|---|---|---|---|---|
| Case | Number | Form | Frequency | Lemma | Form | Frequency | Lemma |
| nominative | singular | žena | 576 | žena | prozor | 91 | prozor |
| genitive | singular | žene | 229 | žena | prozora | 157 | prozor |
| dative | singular | ženi | 55 | žena | prozoru | 10 | prozor |
| accusative | singular | ženu | 167 | žena | prozor | 211 | prozor |
| instrumental | singular | ženom | 39 | žena | prozorom | 54 | prozor |
| locative | singular | ženi | 16 | žena | prozoru | 111 | prozor |
| nominative | plural | žene | 415 | žena | prozori | 81 | prozor |
| genitive | plural | žena | 336 | žena | prozora | 83 | prozor |
| dative | plural | ženama | 33 | žena | prozorima | 3 | prozor |
| accusative | plural | žene | 136 | žena | prozore | 211 | prozor |
| instrumental | plural | ženama | 24 | žena | prozorima | 33 | prozor |
| locative | plural | ženama | 4 | žena | prozorima | 48 | prozor |

relative entropy (henceforth RE) of a noun on the basis of the probabilities $p$ (relative frequencies) of a word's unique inflected variants (stem + case endings) and the corresponding probabilities $q$ (relative frequencies) of the exponents in the word's inflectional class (see Table 2):

$$\mathrm{RE} = \sum_{i=1}^{6} p_i \log_2(p_i/q_i). \tag{1}$$

The probability distributions of the exponents in an inflectional class can be viewed as the prototypical distribution of case endings for that class. The probability distribution of a given word's inflected variants can be viewed as the distribution of a specific exemplar. The relative entropy quantifies how different the exemplar is from the prototype. When the two distributions are identical, the log in (1) evaluates to zero, and hence the relative entropy is zero. Another way of looking at the relative entropy measure is that it quantifies how many extra bits are required to code the information carried by a given exemplar when the theoretical distribution of its class is used instead of its own distribution. Milin, Filipović Đurđević, and Moscoso del Prado Martín (2009) showed empirically that a greater relative entropy, i.e., a greater distance from the prototype, goes hand in hand with longer visual lexical decision latencies. We will return to a more detailed discussion of the interpretation of relative entropy as a measure of lexical processing costs once our computational model has been introduced.

Experiment 1 was designed to ascertain whether these paradigmatic effects extend to sentential reading, and are not artificially induced by the task requirements of the visual lexical decision paradigm. We therefore exchanged the visual lexical decision task used by Milin, Filipović Đurđević, and Moscoso del Prado Martín (2009) for self-paced reading. As we were also interested in ascertaining how a subliminal prime might modulate the effect of relative entropy, we combined self-paced reading with a priming manipulation.

The introduction of a priming manipulation raises the question of how the prime might affect the processing consequences of the divergence between the target's inflectional

Table 2: The two probability distributions determining the relative entropy of “planina” (*mountain*).

| Unique noun forms | Frequency | Probability $p$ | Exponent | Frequency | Probability $q$ |
|---|---|---|---|---|---|
| planin-*a* | 169 | 0.31 | *a* | 18715 | 0.26 |
| planin-*u* | 48 | 0.09 | *u* | 9918 | 0.14 |
| planin-*e* | 191 | 0.35 | *e* | 27803 | 0.39 |
| planin-*i* | 88 | 0.16 | *i* | 7072 | 0.10 |
| planin-*om* | 30 | 0.05 | *om* | 4265 | 0.06 |
| planin-*ama* | 26 | 0.05 | *ama* | 4409 | 0.06 |

paradigm and the prototypical paradigm of its inflectional class. With the introduction of a prime, three inflectional probability distributions are potentially involved instead of just two, and four plausible relative entropy measures could be introduced: one for the prime and the inflectional class, and one for the target and the inflectional class. Furthermore, prime and target could mask the probability distribution of the inflectional class and serve as each other’s reference distribution.

Instead of developing a series of different relative entropy measures, we have adopted a measure from information theory that allows us to evaluate three probability distributions with a single measure, a weighted relative entropy. The use of this weighted entropy measure, is grounded in two assumptions. First, the hypothesis is carried over from previous work that it is the divergence of the target’s probability distribution from that of its inflectional class that is at issue. Second, we assume that the presence of the prime affects the target’s probability estimates, interfering with the target’s paradigmatic relation to its inflectional class.

The weighted relative entropy measure that we have adopted is the one developed in (Belis & Guiasu, 1968; Taneja, 1989; Taneja, Pardo, Gil, & Gil, 1990). The distorting effect of the prime on the probabilities of the target’s inflectional variants is captured through weights on these probabilities:

$$D(P||Q;W) = \sum_i \frac{p_i w_i}{\sum_i p_i w_i} log_2 \frac{p_i}{q_i}. \tag{2}$$

In (2), the index $i$ ranges over inflectional variants. The $p_i$ denote the probabilities of the target’s own inflected variants (probability distribution $P$). The $q_i$ denote the corresponding probabilities of the exponents of the target’s inflectional class (probability distribution $Q$). The weights $w_i$ represent the odds ratio of the form frequency of the target’s $i$-th inflectional variant and the form frequency of the prime’s $i$-th inflectional variant:

$$w_i = \frac{f(target_i)}{f(prime_i)}, \tag{3}$$

with the condition that both frequencies are greater than zero. $W$ represents the vector of these weights. The denominator $\sum_i p_i w_i$ is the expectation for the distribution $p_i$ modulated by weights $w_i$ ($E(P;W)$).

Table 3 provides an example of how the weighted relative entropy is calculated for the feminine target noun “planina” (*mountain*) with the noun “struja” (*electric current*) as its prime. Both nouns belong to the same inflectional class. In the second and fifth column of the Table 3 we find the form frequency counts ($f(a_i)$ and $f(b_i)$) for each inflected form, of the target and the prime, respectively. By dividing these frequencies by the column totals ($f(a) = 552$ and $f(b) = 162$), we obtain estimates of the probabilities of these forms in their paradigms. These estimated probabilities (relative frequencies) are shown in the third and sixth columns ($p(a_i) = f(a_i)/f(a)$ and $p(b_i) = f(b_i)/f(b)$). The seventh column contains the vector of weights — the odds ratio of the form frequency of the target and the form frequency of the prime ($w_i = f(a_i)/f(b_i)$). In the eighth column we find the weighted probabilities ($p_i w_i$) of the inflected variants of the target. The expectation $E(P; W)$ is obtained by summing the values in this eighth column ($\sum p(a_i)w_i = 4.53$). The ninth column represents the frequencies of the inflectional exponents in the target's inflectional class ($f(e_i)$). The $f(e_i)$ are obtained by summation over the frequencies of all words in the inflectional class with the $i$-th inflectional ending. Finally, the tenth column lists the estimated probabilities of the exponents given their class, obtained by dividing each entry in the ninth column by their total ($f(e) = 72182$): $q(e_i) = f(e_i)/f(e)$).

In summary, the questions addressed by Experiment 1 are: first, whether paradigmatic entropy effects are present in sentential reading; and second, whether the effect of a prime on paradigmatic processing, if present, is adequately captured using a weighted relative entropy measure.

*Participants*

A total of 171 undergraduate students of psychology from the University of Novi Sad (150 females and 21 males) participated in the experiment for partial course credit. All participants were fluent speakers of Serbian, with normal or corrected-to-normal vision.

*Materials and predictors*

We retrieved the full set of nouns that appeared at least once in each combination of case and number in the *Frequency Dictionary of Contemporary Serbian Language* (Kostić, 1999). For each gender separately, nouns were randomly divided into two groups: a group of target nouns (henceforth targets), and a group of prime nouns (henceforth primes). Each noun from the list of targets was randomly assigned to a noun from the corresponding list of primes (belonging to the same gender). The final list consisted of 50 masculine, 54 feminine and 16 neuter pairs of targets and primes. For each prime and target word, we compiled information on word length (in letters), word (surface) frequency and stem (lemma) frequency.

We used a normalized Levenshtein distance (Levenshtein, 1966; Jurafsky & Martin, 2000) to assess the orthographic similarity of prime and target. The Levenshtein or edit distance of two strings is the number of deletions, additions, or substitutions required to transform one string into the other. The normalized Levenshtein distance is the Levenshtein distance rescaled to the interval $[0, 1]$. This rescaling is obtained by dividing the Levenshtein distance by the length of the longest sting:

$$\text{Normalized Levenshtein distance} = \frac{\text{Levenshtein distance}}{\text{max(string length)}}. \tag{4}$$

Table 3: The inflected variants of the feminine nouns “planina” (*mountain*) and “struja” (*electric current*). Columns present frequencies and relative frequencies of the respective inflectional paradigms and the inflectional class to which they belong.

| TARGET INFLECTED VARIANT | | | PRIME INFLECTED VARIANT | | | WEIGHT | | CLASS EXPONENT | |
|---|---|---|---|---|---|---|---|---|---|
| | frequency | relative freq. | | frequency | relative freq. | | | frequency | relative freq. |
| | $f(a_i)$ | $p(a_i) = f(a_i)/f(a)$ | | $f(b_i)$ | $p(b_i) = f(b_i)/f(b)$ | $w_i = f(a_i)/f(b_i)$ | $p_i w_i$ | $f(e_i)$ | $q(e_i) = f(e_i)/f(e)$ |
| planin-*a* | 169 | 0.31 | struj-*a* | 40 | 0.25 | 4.23 | 1.31 | 18715 | 0.26 |
| planin-*u* | 48 | 0.09 | struj-*u* | 23 | 0.14 | 2.09 | 0.19 | 9918 | 0.14 |
| planin-*e* | 191 | 0.35 | struj-*e* | 65 | 0.40 | 2.94 | 1.03 | 27803 | 0.39 |
| planin-*i* | 88 | 0.16 | struj-*i* | 8 | 0.05 | 11.0 | 1.76 | 7072 | 0.10 |
| planin-*om* | 30 | 0.05 | struj-*om* | 9 | 0.06 | 3.33 | 0.17 | 4265 | 0.06 |
| planin-*ama* | 26 | 0.05 | struj-*ama* | 17 | 0.10 | 1.53 | 0.08 | 4409 | 0.06 |
| | $f(a) =$ 552 | | | $f(b) =$ 162 | | | $E(P;W) = \sum p(a_i)w_i =$ 4.53 | $f(e) =$ 72182 | |

Following Lund and Burgess (1996a); Landauer and Dumais (1997); McDonald and Ramscar (2001); Moscoso del Prado Martín, Kostić, and Filipović Đurđević (2009) and Filipović Đurđević, Đurđević, and Kostić (2008), we used a cosine similarity measure to represent the semantic proximity of the target and the prime in the hyper-space of their realized textual contexts. This measure reflects the angle between two contextual vectors in hyper-dimensional semantic space:

$$\cos(v_1, v_2) = \frac{v_1 v_2}{|v_1||v_2|}, \tag{5}$$

where $v_1$ represents the context vector of the first, and $v_2$ the context vector of the second word. A context vector $v_i$ is defined by the co-occurrence frequencies of word $i$ with a predefined set of high-frequency context words. The more often two vectors occur with the same context words, the smaller the angle between their corresponding context vectors, and the larger the similarity between the two words, with $\cos \rightarrow 1.0$. To calculate the cosine similarity, we used the 1000 most frequent words of the Serbian language, as retrieved from the *Frequency Dictionary of Contemporary Serbian Language* (Kostić, 1999), as context words list. Co-occurrence of the prime and target with the context words was represented by 1000-dimensional vector, which was built using electronic database of journal articles of *Media Documentation Ebart* (`http://www.arhiv.rs`), containing approximately 70 million words.

For each of the 120 target nouns, three grammatical Serbian sentences were constructed such that each target noun appeared exactly once in nominative singular, once in accusative singular and once in dative/locative singular. Sentences consisted of five words. The position of the target word was counterbalanced: in 50% of the sentences it was the second word in the sentence, and in 50% of the sentences it was the third. In the full set of 360 sentences, each target therefore appeared three times, once in each of three cases. Primes were not considered during the construction of the sentences. The sentences contained various other nouns in addition to the targets. These additional nouns appeared only once across all experimental sentences, with 6 exceptions which appeared twice. They did not belong to the previously selected set of targets and primes.

### *Design and procedure*

Our experimental design included two fixed-effect factors. The first factor was *target case* with three levels: nominative singular, accusative singular and dative/locative singular. The second factor was *prime condition* with five levels: no prime (only hash marks presented), a different stem in a different case, a different stem in the same case, the same stem in a different case, and the same stem in the same case. Primes and targets always belonged to the same inflectional class. The same case and same stem condition implements the identity priming condition. This experimental design with $3 \times 5$ levels is summarized in Table 4.

A Latin-square design with 15 lists ensured that all target words appeared in all of the selected cases, and that each participant was presented with all of the target words only once. Each list consisted of eight sentences per each of the fifteen experimental conditions (three target cases by five priming conditions), totalling to 120 sentences. The presentation sequence was randomised within each list, and for each participant separately.

The presentation of non-target words in each sentence was preceded by a 53.2 ms (exactly four ticks, 13.3 ms each, adjusted with the monitor refresh rate) presentation of hash marks. The stimulus preceding the target word was also presented for 53.2 ms. However, depending on the priming condition, the target word was preceded either by hash marks, its random noun pair in the same case, its random noun pair in a different case, the same noun in a different case, or the same noun in the same case (identity priming).

Participants were instructed to read the words silently in order to understand the meaning of a sentence. The beginning of each sentence was announced on the screen, and initiated by a participant's button-press. Each word remained on the screen until the participant's response. The next word of the sentence was shown on the screen immediately after this response (preceded by hash marks or its prime). We measured reading latencies for the target words as the time elapsed from the onset of the target word to the participant's response.

The *stationary-window* variant of the self-paced sentence reading task was used as a compromise between a task such as lexical-decision and natural sentence reading. On the one hand, priming is much more engaged in lexical-decision experiments where isolated words are presented on the center of the screen, preceded (or sometimes succeeded) by the prime. On the other hand, the *moving-window* paradigm is a more natural variant of the self-paced sentence reading task, as it requires the eye to move through the sentence. Nevertheless, the *stationary-window* paradigm has been found to be a reasonable alternative (c.f., Just, Carpenter, & Woolley, 1982; and Juola, Ward, & McNamara, 1982 for their discussion of gains and losses in reading when eye movements are made unnecessary).

In order to prevent participants from pressing the button automatically, and to make sure that they read the sentences for meaning, 15% of the sentences were followed by a yes/no question querying for comprehension. Prior to the experiment, participants were presented with twelve practice trials.

The experiment was carried out using the *SuperLab Pro 2.0* experimental software (`http://www.cedrus.com`), running on a PC, with a 266 MHz Pentium II processor, and a standard video-card. The monitor was set to 75 Hz refresh rate and a resolution of 1024 x 768 pixels. The stimuli were presented in light-grey, 40 pt Yu Helvetica capital letters, on a black background.

### *Results and discussion*

Five participants were excluded due to large numbers of erroneous answers to the questions (error rates exceeding 30%). Analysis of reaction times (RTs) revealed a small number of extreme outliers (0.5% of the data) that were excluded from further analysis. Response latencies and word (surface) and stem frequencies for both targets and primes were log-transformed to approximate normality. In order to remove autocorrelational structure from the residual errors (Baayen & Milin, 2010), we included two control predictors, the trial number of an item (Trial) in a subject's experimental list (rescaled to $Z$-scores to bring its magnitude in line with that of other predictors), and the response latency at the preceding trial (Previous RT). We used linear mixed-effect modeling (Bates, 2005, 2006; Baayen, Davidson, & Bates, 2008) with participant and word as crossed random-effect factors.

We probed for non-linear effects of the covariates, and for a significant contribution

Table 4: Characteristics of the sentence stimuli. The target is presented in bold, and primes in small capitals.

| Target Case | Prime Condition | Example of sentence stimuli | Prime | Target | |
|---|---|---|---|---|---|
| | hash marks | | ##### | | |
| | different stem, different case | | KULOM | | |
| Nominative | different stem, same case | NJEGOVA | KULA | **PORODICA** | GA JE VOLELA. |
| | same stem, different case | | PORODICOM | | |
| | same stem, same case | | PORODICA | | |
| | | *His family loved him.* | | | |
| | hash marks | | ##### | | |
| | different stem, different case | | KULOM | | |
| Accusative | different stem, same case | OSRAMOTIO JE | KULU | **PORODICU** | SVOJIM PONAŠANJEM. |
| | same stem, different case | | PORODICOM | | |
| | same stem, same case | | PORODICU | | |
| | | *He embarrassed (his) family with his behaviour.* | | | |
| | hash marks | | ##### | | |
| | different stem, different case | | KULOM | | |
| Dative/Locative | different stem, same case | U NJENOJ | KULI | **PORODICI** | NEMA PLAVOOKIH. |
| | same stem, different case | | PORODICOM | | |
| | same stem, same case | | PORODICI | | |
| | | *In her family no one is blue-eyed.* | | | |

of by-word or by-participant random slopes. The latency at the previous target required by-participant random slopes. The order of a trial turned out to be characterized by a significant non-linearity and also required by-participant weights for the linear slope. After removal of potentially influential outliers with absolute standardized residuals exceeding 2.5, we refitted the model. Results are summarized in Table 5 and presented in Figure 1.

Table 5: Initial modelling of target word reading latencies: Partial effects for fixed-effect factors and covariates. The reference level for Prime condition was *no prime* (hash marks), and *nominative* for Target case. Lower, Upper: 95% highest posterior density credible intervals based on 10,000 samples from the posterior distribution of the parameters; P: Markov chain Monte Carlo p-value.

| | Estimate | Lower | Upper | P |
|---|---|---|---|---|
| Intercept | 5.5081 | 5.4019 | 5.6096 | 0.0001 |
| Previous RT | 0.1250 | 0.1086 | 0.1394 | 0.0001 |
| Target position (3rd) | -0.4261 | -0.5592 | -0.3538 | 0.0001 |
| Trial Order (linear) | -0.1146 | -0.1250 | -0.1045 | 0.0001 |
| Trial Order (quadratic) | 0.0213 | 0.0179 | 0.0252 | 0.0001 |
| Word Length | 0.0109 | 0.0070 | 0.0145 | 0.0001 |
| Prime Condition (diff. stem diff. suff.) | 0.1301 | 0.1200 | 0.1406 | 0.0001 |
| Prime Condition (diff. stem same suff.) | 0.0782 | 0.0678 | 0.0881 | 0.0001 |
| Prime Condition (same stem diff. suff.) | 0.0660 | 0.0555 | 0.0758 | 0.0001 |
| Prime Condition (same stem same suff.) | -0.0305 | -0.0408 | -0.0206 | 0.0001 |
| Target Case (accusative) | 0.0246 | 0.0150 | 0.0340 | 0.0001 |
| Target Case (dative/locative) | 0.0262 | 0.0141 | 0.0387 | 0.0002 |
| Target Lemma Frequency | -0.0119 | -0.0177 | -0.0058 | 0.0001 |
| Previous RT x Target Position (3rd) | 0.0703 | 0.0593 | 0.0912 | 0.0001 |

The first two panels of Figure 1 present the effects of control variables. The positive slope for the previous target latency as a predictor of the current targets' reading latency is indicative of consistency and/or inertia in the participants' behaviour across trials. The slope for the target in the third position in the sentence was greater than that for the slope of the target in the second position. The somewhat richer preceding syntactic context for targets in the third position may have afforded enhanced sentential integration, with a spillover effect from the difficulty of integration at the previous trial. The negatively decelerating effect of trial indicates that participants gained experience with the task as they progressed through the experiment. The positive slope for word length and the negative slope for target lemma frequency are as expected.

As to the fixed-effect factor Prime Condition: The identity condition (same stem, same suffix, SS) elicited the shortest latencies, the different stem, different suffix condition (DD) showed the longest latencies, with the different-stem same suffix (DS) and same-stem different-suffix (SD) conditions occupying intermediate positions. The condition in which only hash marks were shown elicited longer latencies than the identity condition, but shorter latencies than the other three priming conditions.

The advantage of the identity condition is as expected, given that the target word

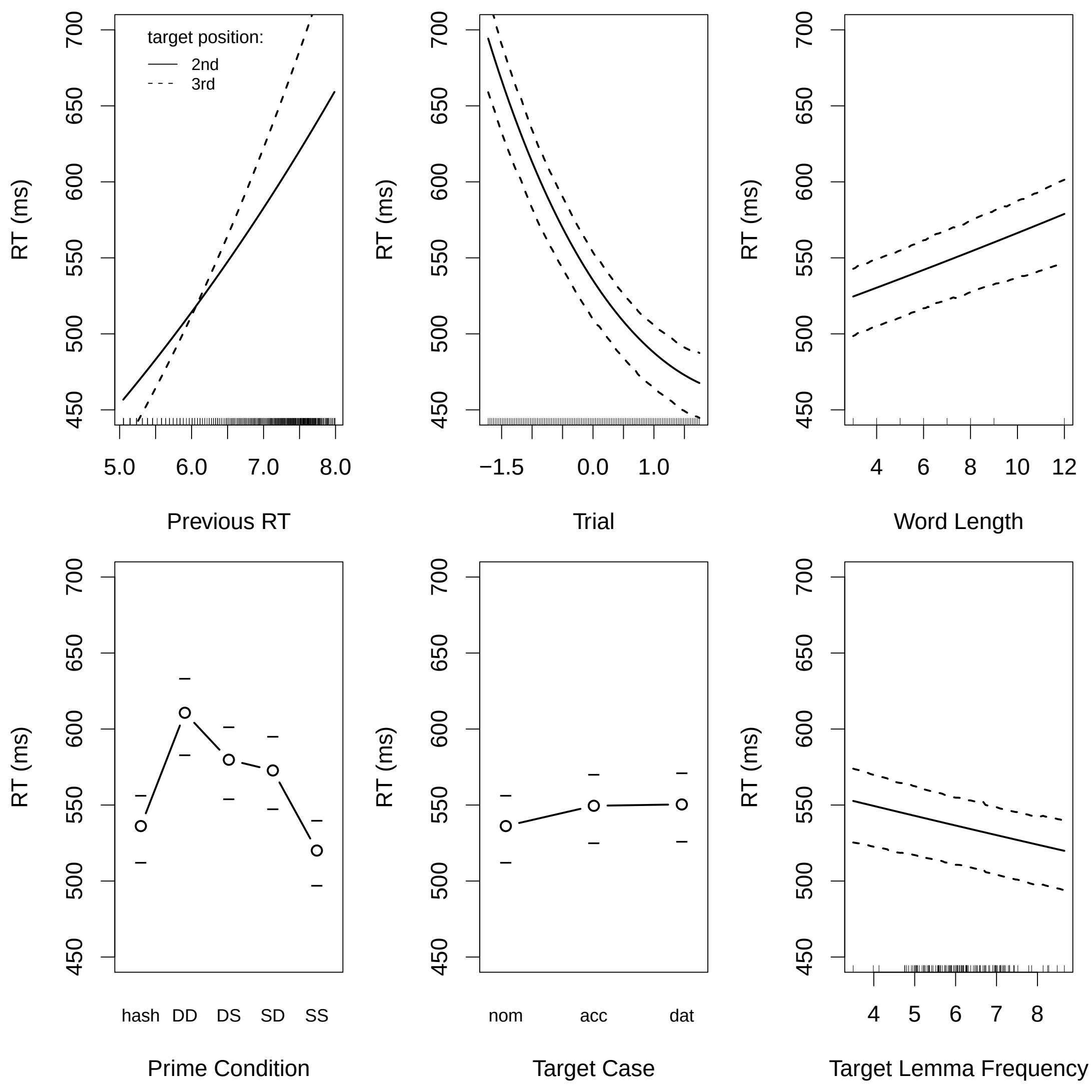


*Figure 1.* Initial modelling of target word reading latencies: Partial effects for fixed-effect factors and covariates. The reference level for Prime condition was *no prime* (hash marks), the other factor levels are labeled DD (different stem, different suffix), DS (different stem, same suffix), SD (same stem, different suffix), and SS (same stem, same suffix). The reference level for Target Case was *nominative*. Lower, Upper: 95% highest posterior density intervals based on 10,000 samples from the posterior distribution of the parameters; P: Markov chain Monte Carlo p-value.

is available 53.2 ms prior to the point in time at which it becomes (fully) available in the other prime conditions, and never disrupted by a mask or mismatching information. The fast average response to the no-prime condition (hash marks only) compared to the DD, DS and SD prime conditions is of special interest, as it indicates that the conflicting information provided by a different stem, a different suffix, or both, disrupt processing more than the presentation of linguistically neutral hash marks.

Turning to the effect of *Target Case*, we find that nouns with nominative case elicited shorter latencies, compared to the other two oblique cases (accusative and dative/locative), irrespective of gender. This is in line with previous findings on Serbian (cf. Lukatela et al., 1978; Lukatela, Gligorijević, Kostić, & Turvey, 1980; Kostić & Katz, 1987). One possible interpretation is that it mirrors the difference in number of syntactic functions and meanings of Serbian noun cases, where nominative has only three functions/meanings, as compared to a magnitude larger number for the other (oblique) cases used in this study (more about the role of syntactic functions and meanings in Serbian in Kostić, Marković, & Baucal, 2003; also, syntactic functions and meanings are further discussed in the framework of information theory by Milin, Kuperman, Kostić, & Baayen, 2009).

In what follows, we excluded the no-priming condition from the data set, as this makes it possible to include predictors bound to the prime. Although target words occurred in three cases (nominative, or accusative, or dative/locative), an initial survey of the data revealed that the relevant contrast was between nominative and non-nominative case. Hence, we used *Target Case* as a binary factor contrasting whether nominative case is TRUE or FALSE. As the prime's stem frequency and the target's word frequency were irrelevant as predictors, in contrast to the prime's word frequency and the target's stem frequency, only the latter two frequency measures will be considered further. Finally, as the two priming conditions in which exactly one constituent differed between prime and target revealed very similar mean latencies, we collapsed these two factor levels, resulting in a new factor for prime condition with three levels: DD (different stem and different inflection), DSSD (different stem and same inflection, or different inflection and same stem), and SS (identical stem and inflection).

The condition number $\kappa$ characterizing the collinearity of the predictors was too high (35.6) to proceed straightforwardly with the regression analysis. We reduced $\kappa$ to 21.7 as follows. First, we regressed the Cosine similarity measure on prime condition, weighted relative entropy, and Levenshtein distance. The residuals of this model constituted our orthogonalized Cosine measure. Second, we replaced prime frequency by the residuals of a model regressing prime frequency on target frequency. Both orthogonalized measures were significantly and positively correlated with the original measures ($r = 0.66$ and $r = 0.94$, respectively).

The same random slopes were required as in the preceding analysis. After removal of outliers and refitting, the model summarized in Table 6 was obtained. As can be seen in Figure 2, the frequency of the prime had a facilitatory effect (mid upper panel) that was smaller in magnitude than the effect of the lemma frequency of the target (left upper panel). The normalized Levenshtein distance (orthogonalized with respect to the prime condition) failed to reach significance (right upper panel). The cosine similarity measure revealed the expected facilitation (left lower panel). The more similar the prime and the target were in terms of their textual occurrences, the faster processing completed.

Finally, the weighted relative entropy measure revealed the predicted inhibitory main

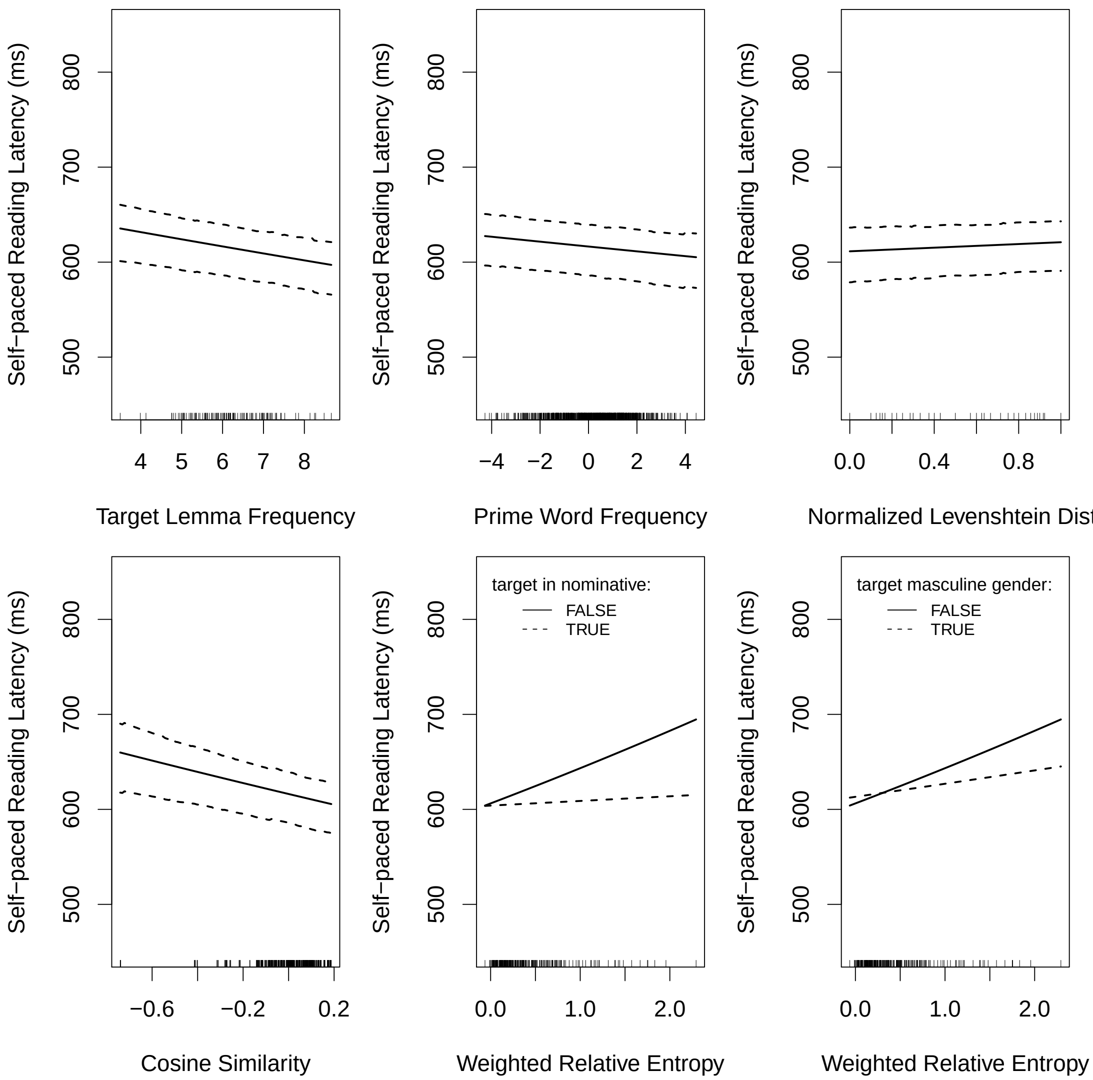


*Figure 2.* Partial effects of selected predictors in a mixed-effects model for the reading latencies in Experiment 1, excluding the no-prime condition. For simple main effects, dashed lines represent 95% highest posterior density credible intervals.

Table 6: Partial effects of the predictors in a mixed-effects model for the latencies in Experiment 1, excluding the no-prime condition. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

| | Estimate | Lower | Upper | P |
|---|---|---|---|---|
| Intercept | 5.6787 | 5.5598 | 5.7954 | 0.0001 |
| Previous RT | 0.1173 | 0.0979 | 0.1328 | 0.0001 |
| Target Position (3rd) | -0.4017 | -0.5593 | -0.3231 | 0.0001 |
| Trial Order (linear) | -0.1170 | -0.1280 | -0.1064 | 0.0001 |
| Trial Order (quadratic) | 0.0212 | 0.0172 | 0.0255 | 0.0001 |
| Length | 0.0099 | 0.0057 | 0.0139 | 0.0001 |
| Prime Condition DSSD | -0.0455 | -0.0575 | -0.0322 | 0.0001 |
| Prime Condition SS | -0.1321 | -0.1550 | -0.1056 | 0.0001 |
| Weighted Relative Entropy | 0.0594 | 0.0388 | 0.0795 | 0.0001 |
| Nominative Case | -0.0038 | -0.0175 | 0.0101 | 0.5832 |
| Masculine Gender | 0.0114 | -0.0061 | 0.0280 | 0.2092 |
| Normalized Levenshtein Distance | 0.0155 | -0.0061 | 0.0401 | 0.1668 |
| Cosine similarity | -0.0925 | -0.1379 | -0.0459 | 0.0002 |
| Target Lemma Frequency | -0.0121 | -0.0190 | -0.0057 | 0.0002 |
| Prime Word Frequency | -0.0041 | -0.0076 | -0.0011 | 0.0122 |
| Previous RT x Target Position (3rd) | 0.0664 | 0.0536 | 0.0905 | 0.0001 |
| Nominative Case x Weighted Relative Entropy | -0.0513 | -0.0740 | -0.0288 | 0.0001 |
| Masculine Gender x Weighted Relative Entropy | -0.0372 | -0.0607 | -0.0107 | 0.0026 |

effect (not shown). The more atypical the probability distribution of an exemplar's case inflections compared to the prototype (its inflectional class), the longer it takes to read that exemplar. Interestingly, the effect of weighted relative entropy was modulated by Case and Gender: Inhibition was present only for words in the oblique cases, of neuter or feminine gender. For masculine nouns, and for nouns in nominative case, the effect vanished (dashed lines in the mid and right lower panels).

The emergence of a significant effect of weighted relative entropy in sentential reading shows that the effects of inflectional paradigmatic structure are not restricted to isolated word reading, and indicate that paradigmatic entropy effects may have broader ecological validity. Furthermore, for oblique cases, the effect of the prime is properly captured by the weighted relative entropy measure. The greater the frequency odds between the target's inflected variants as compared to those of the prime, the greater the delay in processing time.

Are the interactions of Case and Gender with Weighted Relative Entropy contingent on nouns being presented in sentential context? To address this question, we carried out a second experiment in which the prime and target pairs of Experiment 1 were presented in isolation, using lexical-decision with masked priming.

## Experiment 2

### *Participants*

142 undergraduate students of psychology from the University of Novi Sad (125 females and 17 males), participated in experiment for partial course credit. None of them participated in Experiment 1.

### *Materials*

We used the same set of 50 masculine, 54 feminine and 16 neuter pairs of target and prime nouns as in Experiment 1.

### *Design and procedure*

We implemented the same $15 \times 15$ Latin-square design as in Experiment 1. To each list we added an equal number of matched Serbian pseudo-words (with legal Serbian ortho-phono-tactics), with the same inflected endings. In this way we obtained fifteen experimental lists, with 240 items each. Participants were randomly assigned to one of these experimental lists. Presentation sequence was randomised within each list, and for each participant. The experiment was preceded by 10 practice trials.

The target stimuli (words or pseudo-words) were presented for 1500 ms, preceded by a 53.2 ms prime. In the no-prime condition, the target was preceded by hash marks. In the other priming conditions, the target word immediately followed the prime word. We measured lexical decision latencies for the target words as the time elapsed from the onset of the target word to the participant's response. An experimental session lasted 10 minutes, approximately. Stimuli were presented with *SuperLab Pro 2.0*, using Serbian Latin letters (light-grey capital 40 pt Yu Helvetica on a black background).

### *Results and discussion*

Inspection of the data revealed 7.3% of word items that frequently produced erroneous answers. Typically less frequent words such as "brid" (*blade*, *edge*), "srez" (*district*), "mena" (*phase*), and "nota" (*note*), in combination with less frequent inflectional ending (like dative/locative), provoked error responses. Such error-prone words were removed from the data set. As for Experiment 1, we log-transformed response latencies, word (surface) frequencies, and stem frequencies. We used exactly the same predictors as in Experiment 1, decorrelated and transformed in the same way. Subject and item were included as random-effect factors.

Table 7 and Figure 3 summarize the mixed-effects model fitted to the lexical decision data. We tested for possible non-linearities and by-word or by-participant random slope effects in the model, removed outliers, and refitted the model to the data. The control predictors Previous RT and Trial were significant predictors, with inhibitory and facilitatory effects respectively. Trial was the only predictor for which by-participant random slopes (for the quadratic term of Trial only) were supported by a likelihood ratio test. Word Length was inhibitory, as expected. Response latencies were delayed by the presence of a prime, with the greatest disadvantage for primes composed of a different stem and a different inflectional ending, as expected.

Response latencies increased with Weighted Relative Entropy. Unlike in the sentence reading experiment, interactions with Case and Gender received no statistical support whatsoever, and were therefore removed from the model specification. The Normalized Levenshtein Distance reached full significance in Experiment 2 as an inhibitory predictor. For the lexical decision latencies, the target's form frequency was a slightly better predictor than the target's lemma frequency. As in sentence reading, there was a facilitatory effect of the frequency of the prime, and as before this effect was reduced compared to the frequency effect of the target. The (orthogonalized) Cosine Similarity measure was not significant.

The presence or absence of sentential context explains some important differences in the results of the lexical decision and self-paced reading experiments, which both used priming. In the primed lexical decision experiment, words appeared in isolation, without any context that would otherwise allow the participant to anticipate the upcoming word and its case. Without such contextual support, the cognitive system apparently falls back on the de-contextualized probability of the word's form, as indicated by the significance of the target's inflectional form (surface) frequency outperforming its lemma frequency, and the full significance of the Levenshtein measure of orthographic similarity. Furthermore, the presence of a prime in the absence of sentential context rendered the Cosine Similarity measure insignificant.

It is less clear why in sentential reading, but not in isolated word reading, the effect of Weighted Relative Entropy is restricted to oblique case forms of non-masculine gender. A processing advantage for nominative forms is in line with the results reported by Lukatela et al. (1978) and Lukatela et al. (1980); Kostić and Katz (1987). As argued above when discussing the base model (Table 5, and Figure 1), this processing advantage for forms in nominative case might be due to its syntactic simplicity, encompassing only three functions and meanings.

Since only a relatively small number of neuter nouns was included in the materials, the interaction of Gender with Weighted Relative Entropy basically contrasts masculine with feminine nouns. It turns out that the interaction of Weighted Relative Entropy by Gender is matched by an imbalance in average Relative Entropy in the Serbian lexicon. Leaving the primes in the present experiment aside, it turns out that the average Relative Entropy was was 0.17 for feminine nouns and 0.25 for masculine nouns, a difference of 0.08 that received ample statistical support ($p < 0.0001$). The greater Relative Entropy for masculine case forms indicates a more challenging learning problem for masculine nouns compared to feminine nouns, resulting in a weaker inflectional class prototype and reduced effects of dissimilarity to the prototype in the priming context. This empirical finding is in line with the fact that the masculine noun class is less regular then the feminine noun class: The masculine noun class exhibits exponent (affixal) differences between animate and inanimate nouns and various other inconsistencies which are not present in the feminine noun class (see, e.g., Stevanović, 1989; Stanojčić & Popović, 2005, etc.).

Since there is no difference between the case forms with respect to Relative Entropy, it is unlikely that the interaction of Weighted Relative Entropy by Case is driven by the distributional properties of the input.

Considering Experiments 1 and 2 jointly, we conclude that the present entropy-based measures are well-supported as probes for paradigmatic effects in lexical processing. This raises the question of how to interpret these paradigmatic entropy effects. One possibility

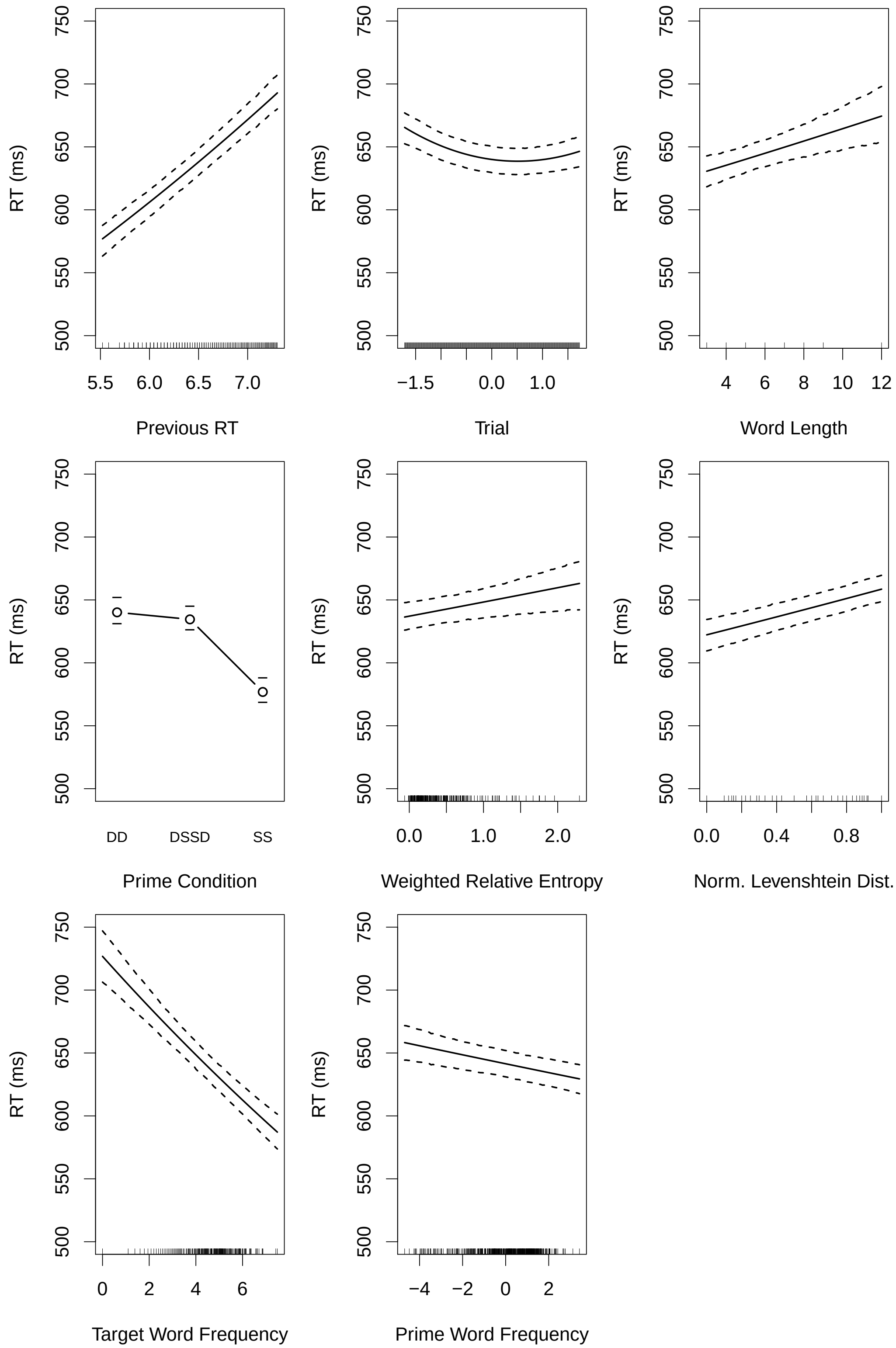


*Figure 3.* Partial effects in the mixed-model fitted to the lexical decision latencies (Experiment 2), excluding the no-prime condition. Dotted lines represent 95% highest posterior density credible intervals.

Table 7: Coefficients of the mixed-effects model fitted to the lexical decision latencies of Experiment 2: Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

| | Estimate | Lower | Upper | P |
|---|---|---|---|---|
| Intercept | 5.8485 | 5.7344 | 5.9298 | 0.0001 |
| Previous RT | 0.1028 | 0.0919 | 0.1191 | 0.0001 |
| Trial order (linear) | -0.0085 | -0.0125 | -0.0044 | 0.0001 |
| Trial order (quadratic) | 0.0083 | 0.0052 | 0.0113 | 0.0001 |
| Length | 0.0075 | 0.0032 | 0.0119 | 0.0006 |
| Prime Condition DSSD | -0.0088 | -0.0168 | -0.0005 | 0.0336 |
| Prime Condition SS | -0.1041 | -0.1191 | -0.0893 | 0.0001 |
| Weighted relative entropy | 0.0174 | 0.0031 | 0.0277 | 0.0160 |
| Normalized Levenshtein Distance | 0.0567 | 0.0408 | 0.0733 | 0.0001 |
| Target Word Frequency | -0.0285 | -0.0337 | -0.0232 | 0.0001 |
| Prime Word Frequency | -0.0055 | -0.0081 | -0.0029 | 0.0001 |

would be to assume that inflected variants are stored and organized into paradigmatic tables in long-term memory. In this line of reasoning, however, it remains unclear how entropy effects might actually arise during lexical access. We therefore explored a different possibility, namely, that paradigmatic entropy effects emerge straightforwardly as a consequence of discriminative learning. Specifically, we predict that an interaction of WRE by Gender, but not the interaction of WRE by Case, will be replicable in an input-driven associative learning approach.

## A model based on naive discriminative learning

Our interest in discriminative learning was sparked by the studies of Ramscar and Yarlett (2007); Ramscar, Yarlett, Dye, Denny, and Thorpe (2010). Ramscar and colleagues made use of the Rescorla-Wagner equations to simulate the time-course of lexical learning. However, there are other relevant psycholinguistic studies which made use of Rescorla-Wagner model, for example, Hsu, Chater, and Vitányi (2010) and Clair, Monaghan, and Ramscar (2009) on language acquisition, and Ellis (2006), who studied second language learning.

The Rescorla-Wagner model is deeply rooted in the cognitive psychology tradition (cf. Miller, Barnet, & Grahame, 1995; Siegel & Allan, 1996). Amazingly fruitful, it has been closely linked with several well-known and well-defined probabilistic algorithms, such as the connectionist delta-rule (cf. Gluck & Bower, 1988; J. R. Anderson, 2000), and the Kalman filter (cf. Dayan & Kakade, 2001). Recently, it has been discussed as an instance of a general probabilistic learning mechanism (see, e.g., Chater, Tenenbaum, & Yuille, 2006; Hsu et al., 2010, etc.).

Complementing the approach of Ramscar and colleagues (Ramscar & Yarlett, 2007; Ramscar et al., 2010), our modeling effort focuses on the end result of the lexical learning process, when the system is in a state of equilibrium. In this incarnation of the model of

Wagner and Rescorla (1972), cues are associated with an outcome. Both cues and outcomes can be either present or absent. For our purposes, cues are segment (letter) unigrams and bigrams (for a more complete orthographic coding scheme, see Whitney, 2001), and outcomes are meanings, ranging from the meanings of words (*house, table*), and inflectional meanings (e.g., case: *nominative, genitive, dative, accusative, instrumental, locative*; number: *singular*, *plural*) to affixal meanings (e.g., *-ness* or *un-*).

Let $\textsc{present}(X,t)$ denote the presence of cue or outcome $X$ at time $t$, and $\textsc{absent}(X,t)$ denote its absence at time $t$. The Rescorla-Wagner equations specify the association strength $V_i^{t+1}$ of cue $C_i$ with outcome $O$ at time $t+1$ as

$$V_i^{t+1} = V_i^t + \Delta V_i^t, \tag{6}$$

with the change in association strength $\Delta V_i^t$ defined as

$$\Delta V_i^t = \begin{cases} 0 & \text{if } \textsc{absent}(C_i, t) \\ \alpha_i\beta_1 \left(\lambda - \sum_{\textsc{present}(C_j,\,t)} V_j\right) & \text{if } \textsc{present}(C_j, t) \ \& \ \textsc{present}(O, t) \\ \alpha_i\beta_2 \left(0 - \sum_{\textsc{present}(C_j,\,t)} V_j\right) & \text{if } \textsc{present}(C_j, t) \ \& \ \textsc{absent}(O, t) \end{cases} \tag{7}$$

with the standard settings for the parameters: $\lambda = 1$, all $\alpha$'s equal, and $\beta_1 = \beta_2$. The association strength of a cue to an outcome is strengthened when cue and outcome co-occur. The association strength is decreased whenever the cue occurs without the outcome being present. The extent to which an association strength is adjusted depends on the number of other cues present. When there are more cues present simultaneously, positive adjustments are smaller while negative adjustments are larger, and vice versa. It is worth noting, as pointed out by Rescorla (1988), that this approach to learning differs fundamentally from the theories of Pavlovian learning which were dominating the field until the early sixties of the previous century. Current emphasis is on the context of learning and the learning of relations among events, allowing an organism to build a representation of its environment. In particular, the information that one event provides about another is crucial. Thus, Gallistel (2003) argues that only informative events can elicit conditioning (p. 93). More specifically, he claims that learning can occur if and only if there is a divergence between the observed entropy of a potentially informative event and the maximum entropy — that is, if the event has non-random property (see also Gallistel & Gibbon, 2002. For neurobiological results, see Schultz, 2002 and Daw & Shohamy, 2008).

As an illustration of how association strengths develop over time, consider Table 8 and Figure 4. Table 8 presents a small artificial lexicon with word forms, their frequencies of occurrence, and their meanings. For ease of exposition, we use examples from English. The letters (unigrams) of the word constitute the cues for the model, the meanings represent the outcomes. When the 419 tokens of the 10 words are presented 25 times in randomized order, association strengths develop over time, as illustrated in Figure 4. The upper left panel presents the association strength for *h* and HAND. The *h* occurs only in the words *hand* and *hands*. As it is a perfect cue for the meaning HAND, its association strength is increased whenever *hand* or *hands* is encountered. It is never decreased, as there are no words containing an *h* that do not map onto the meaning HAND.

The upper right panel shows the development of the association strength of the *s* with the PLURAL meaning. As the *s* occurs not only in plurals, but also in *sad, as* and *lass*, it is

Table 8: Example lexicon for naive discriminative learning.

| Word | Frequency | Lexical Meaning | Number |
|---|---|---|---|
| *hand* | 10 | HAND | |
| *hands* | 20 | HAND | PLURAL |
| *land* | 8 | LAND | |
| *lands* | 3 | LAND | PLURAL |
| *and* | 35 | AND | |
| *sad* | 18 | SAD | |
| *as* | 35 | AS | |
| *lad* | 102 | LAD | |
| *lads* | 54 | LAD | PLURAL |
| *lass* | 134 | LASS | |

not an unequivocal cue for plurality. Depending on the order in which plural and non-plural exemplars are encountered, its association strengths with the plural meaning increases or decreases. The general trend over time, for this small lexicon, is for this association strength to increase. The remaining two panels illustrate that for the short word *as* the *a* becomes strongly linked with its meaning, wherease the *s* becomes a negative cue.

In this simple example, the *s* becomes a marker of plurality, irrespective of its positions, which linguistically doesn't make sense. In our actual simulations, we included as cues not only letter unigrams, but also letter bigrams, with a word-initial *s* represented as *#s* and word-final *-s* represented as *s#*. In our model for English, discussed below, the association strength for the unigram *s* to plurality in the stable state is negative (-0.008), for *#s* it is positive but small (0.003), and for *s#* it is also positive but large (0.018). With a better coding scheme, and realistic language input, linguistically sensible results are obtained.

What is worth noting for the purpose of present study is that the Rescorla-Wagner algorithm performs Maximum-Likelihood estimation of the parameters for models of causal learning and/or causal inference (Yuille, 2005; Yuille, 2006), clarifying the often intricate probabilistic interrelationship between a system and its environment. The Rescorla-Wagner algorithm provides the Maximum-Likelihood estimates of the weights on the connections between letter unigrams and bigrams and word meanings.

Danks (2003) proposed an efficient way for obtaining these maximum likelihood estimates by examining the system when it is in a stable state. Danks calls attention to the fact that an asymptote for the Rescorla-Wagner model is in general not well-defined. However, one can think of the model settling down eventually into a state where the expected changes in the weights are zero ($V_i^{t+1} = V_i^t$, i.e., $V_i^{t+1} - V_i^t = 0$). Danks shows that in this equilibrium state the association strengths $V_i$ of the cues $C$ to a specific outcome $O$ can be obtained by solving the the system of equations (8), where $n + 1$ denotes the number of

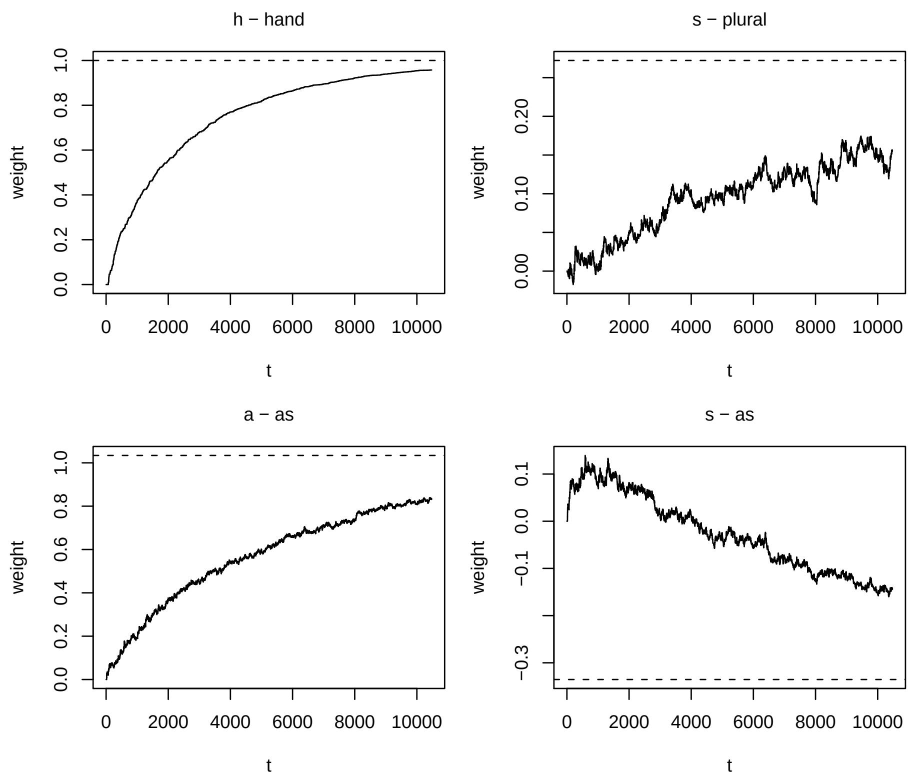


*Figure 4.* Development of association strengths of unigram cues to meanings when the Rescorla-Wagner model is exposed to the words in Table 8. Word tokens are presented in random order. To show the long-term development of the association strength, each token is presented 25 times (i.e., each word frequency is multiplied by 25). Dashed lines represent the stable-state association strength obtained with the equilibrium equations of Danks (2003).

different cues (input features) and where the indices $i$ and $j$ range over the different cues:

$$\begin{pmatrix} \Pr(C_0|C_0) & \Pr(C_1|C_0) & \dots & \Pr(C_n|C_0) \\ \Pr(C_0|C_1) & \Pr(C_1|C_1) & \dots & \Pr(C_n|C_1) \\ \dots & \dots & \dots & \dots \\ \Pr(C_0|C_n) & \Pr(C_1|C_n) & \dots & \Pr(C_n|C_n) \end{pmatrix} \begin{pmatrix} V_0 \\ V_1 \\ \dots \\ V_n \end{pmatrix} = \begin{pmatrix} \Pr(O|C_0) \\ \Pr(O|C_1) \\ \dots \\ \Pr(O|C_n) \end{pmatrix} \tag{8}$$

or, in short,

$$\Pr(O|C_i) - \sum_{j=0}^{n} \Pr(C_j|C_i)V_j = 0. \tag{9}$$

Here, $\Pr(C_j|C_i)$ represents the conditional probability of cue $C_j$ given cue $C_i$, and $\Pr(O|C_i)$ the conditional probability of outcome $O$ given cue $C_i$. Informally, we can think of the association strengths $V_j$ as optimizing the conditional outcomes given the conditional probabilities characterizing the input space. The estimation of the association strengths (or weights on the connections from cues to outcomes) with (8) is parameter-free, and totally determined by the training data. The appendix provides detailed information on the steps required to calculate the equilibrium association strengths for the example lexicon in Table 8. The stable-state association strengths (connection weights) for the examples in Figure 4 are represented by dashed lines.

We model the association strengths from the letter unigrams and bigrams (cues) to a given meaning (outcome) separately and independently of all other outcomes. In other words, for each meaning a different $O$ is substituted in (8), and a different set of equations has to be solved. The assumption of independence for the association strengths to the different meanings involves an obvious simplification. This simplifying assumption, which is similar to the independence assumption in naive Bayes classifiers, affords efficient computation while yielding adequate results. Our model therefore implements a form of associative learning that we will refer to as *naive discriminative learning*.

Let $i$ range over the outcomes (meanings), and $j$ over the cues (unigrams and bigrams), and define the association strength $V_{ji}$ to denote the equilibrium association strength $V_j$ as estimated for cue $C_j$ and outcome $O_i$. Given the set of input cues $\{C_k\}$, the activation $a_i$ of outcome $O_i$ is given by

$$a_i = \sum_{j \in \{C_k\}} V_{ji}. \tag{10}$$

The activation of the $i$-th meaning $a_i$ represents the total posterior evidence for this meaning given the unigrams and bigrams in the input. (In our experience, adding trigrams and higher-order n-grams leads to only a minute increase in goodness of fit.) Response latencies and self-paced reading times are assumed to be negatively correlated with this total posterior evidence. When the weights are estimated from small data sets, it is sufficient to model RTs simply as

$$\text{simulated RT}_i \propto -a_i. \tag{11}$$

For large training data, it is preferable to model response latencies as inversely proportional to the amount of activation $a_i$. Similar to the empirical distributions of lexical decision latencies, the distribution of activations $a$ tends to have a rightward skew. This skew is often largely eliminated by a log-transform,

$$\text{simulated RT}_i = \log(1/a_i). \tag{12}$$

In what follows, we use the transform that best approximates normality, just as for the observed latencies, in order to obtain maximally adequate statistical models (see Baayen & Milin, 2010, for further discussion of transformations of the response variable in statistical modeling). This completes the definition of our model, to which we will refer as the *naive discriminative reader*.

## Modeling the processing of case inflections in Serbian

In what follows, we restrict ourselves to a discussion of simulating the self-paced reading latencies of Experiment 1. The results obtained for the lexical decision latencies of Experiment 2 were similar quantitatively and qualitatively, and will not be discussed further. The results of Experiment 1 revealed a slightly more complex pattern of results, and therefore provides the more interesting data set to model and report. A first challenge to the naive discriminative reader is that its predictions should reflect the observed effect of weighted relative entropy. Since exponents can be multiply ambiguous, a second challenge for the discriminative learning approach is how effective these exponents can be as cues for case and number.

The model was trained on the set of 270 nouns and their case inflected forms (3240 wordforms in all), which appeared at least once in each combination of case and number in the *Frequency Dictionary of Contemporary Serbian Language* (Kostić, 1999). For this data set, training simply proceeded on the basis of individual words, without context.

For unprimed reading, the total activation predicting a word's processing time is defined as the sum of the activations of its lexical meaning and the grammatical meanings for number (singular and plural) and case (nominative, genitive, dative, accusative, locative, and instrumental). In other words, we assume that a word's inflectional paradigm comes into play at the level of (grammatical) meaning. It is important to distinguish this 'semantic paradigm' from a traditional form paradigm, comprising a word's different forms as in Table 1. The model has no representations for word forms, and there is no competition between word forms in the model, nor the merging of evidence for them. Our hypothesis is that a self-paced reading time, and a lexical decision latency, is proportional to the cumulative activation of the word' meaning and its semantic paradigm (as activated by its orthographic cues).

In order to simulate priming, we first calculated for prime and target separately, the activation of the meanings of the two nouns, as well as the activation of the meanings for singular and plural number, and those of nominative, genitive, dative, accusative, locative, and instrumental case. The two resulting sets of 10 activations were then used to estimate a primed decision time.

For the modeling of priming, we explored two alternatives, both of which turned out to yield good results. The first alternative builds on the way priming is modelled in the original Rescorla-Wagner framework, and the second alternative follows the retrieval theory of priming proposed by Ratcliff and McKoon (1988). To illustrate the two alternatives, let $\boldsymbol{a}_P$ be the 10-element vector of the meaning activations for the prime, and let $\boldsymbol{a}_T$ denote the corresponding vector of activations for the target, with each meaning activation as defined in (10).

In the original Rescorla-Wagner model, when input cues (stimuli) are presented in isolation, the total activation amounts to a simple sum of association strengths ($\sum_{i=1}^{10} a_i$). The maximum strength $\lambda = 1$ in (7) cancels out in the derivation of the equilibrium equations (8). However, in the case of 'compound cues' consisting of a prime and a target, simple learning becomes competitive learning, and the maximum strength ($\lambda$) must be shared between the competing cues (see J. R. Anderson, 2000; Brandon, Vogel, & Wagner, 2003; Vogel, Brandon, & Wagner, 2003) For $\lambda \neq 1$, this leads to a revised system of equations, also defined by Danks (2003):

$$\lambda \Pr(O|C_i) - \sum_{j=0}^{n} \Pr(C_j|C_i)V_j = 0. \tag{13}$$

In the case of priming, we have two sets of competing cues. The maximum activation is shared between them: $\lambda_P + \lambda_T = \lambda$. Setting $\lambda$ to 1, the compound activation ($a_{PT}$) follows straightforwardly:

$$\begin{aligned} a_{PT} &= \sum_{i=1}^{10} (\lambda_P a_{Pi} + \lambda_T a_{Ti}) \\ &= \sum_{i=1}^{10} (\lambda_P a_{Pi} + (1 - \lambda_P) a_{Ti}) \quad (0 \leq \lambda_P \leq 1). \end{aligned} \tag{14}$$

Competitive learning in the Rescorla-Wagner model predicts that the addition of a prime leads to decreased activation of the target's meaning, and hence to longer response latencies. For a prime totally unrelated to the meaning of the target, for instance, the weights on the links of the prime to the target's meaning will be small or even negative, leading to small or even negative $a_{Pi}$ in (14).

The retrieval theory of priming developed by Ratcliff and McKoon (1988) defines the familiarity of a compound cue $S$ as follows:

$$S = \sum_{i=1}^{10} (a_{Pi}^{w} \cdot a_{Ti}^{1-w}) \quad (0 \leq w \leq 0.5), \tag{15}$$

with $w$ a weight for capturing the relative importance of the prime compared to the target. Good fits were obtained with both (14) and with (15), for a wide range of values of $\lambda_P$ and $w$. The results for the compound cue theory were slightly better, hence, we restrict ourselves to reporting the results using (15), with $w = 0.2$.[1]

Using Compound Cue Strength as dependent variable, with $w = 0.4$, a distribution of simulated response latencies was obtained for which log or inverse transforms did not lead to improved approximation of normality. Therefore, simulated response latencies were defined as negative compound cue strength. The simulated latencies correlated well with the observed latencies: $r = 0.24$ ($t(1185) = 8.58, p < 0.0001$). We fitted the same regression model to the simulated latencies as fitted to the observed latencies, with as exception the experimental control predictors *Trial*, *Previous RT*, and the sentential predictor *Target Position* (2nd or 3rd position in the sentence). The Cosine Similarity measure, and the interaction of Weighted Relative Entropy by Case did not reach significance. We then refitted the models with these latter two predictors excluded.

There is one predictor for which the model makes the opposite prediction. Whereas Word Length is inhibitory for the observed latencies, it is facilitatory for the simulated latencies. In the model, a longer word provides more activation to its associated lexical meaning (as well as to its grammatical meanings). A longer word has more active orthographic cues, and hence more connection strengths are summed to obtain the activation of its lexical meaning (and its grammatical meanings). Due to greater activation of its lexical meaning (and its grammatical meanings), the response latency to a longer word is predicted to be shorter. The model is blind to the increasing likelihood of multiple fixations for longer words, and the associated increase in processing costs.

[1] Data and code are available in the `ndl` package for the R free software environment for statistical computing and graphics, available in the CRAN archives at `http://cran.r-project.org/`.

To bring the cost of additional fixations for longer words into the model, the simulated response latencies were redefined as follows, with $S_i$ the compound cue strength for the $i$th word, and $l_i$ the length (in letters) of that word:

$$\text{simulated RT}_i = S_i + \phi \mathrm{I}_{[l_i>5]}. \tag{16}$$

For words with more than 5 letters, the expression $\mathrm{I}_{[l_i>5]}$ evaluates to 1, and a fixation penalty $\phi$ is added to the simulated latency. Table 9 and Figure 5 summarize the resulting model.

Table 9: Coefficients estimated for the simulated self-paced reading latencies.

| | Estimate | Standard Error | t-value | p-value |
|---|---|---|---|---|
| Intercept | -12.084 | 0.104 | -115.854 | 0.0000 |
| Word Length | 0.058 | 0.007 | 8.131 | 0.0000 |
| Weighted Relative Entropy | 0.185 | 0.038 | 4.823 | 0.0000 |
| Masculine Gender = TRUE | 0.169 | 0.033 | 5.173 | 0.0000 |
| Normalized Levenshtein Distance | 1.201 | 0.062 | 19.303 | 0.0000 |
| Target Lemma Frequency | -0.135 | 0.011 | -11.909 | 0.0000 |
| Prime Form Frequency | -0.019 | 0.008 | -2.282 | 0.0227 |
| Prime Condition = DSSD | -0.028 | 0.035 | -0.797 | 0.4257 |
| Prime Condition = SS | 0.158 | 0.068 | 2.346 | 0.0192 |
| W.Rel.Entropy : Masculine Gender = TRUE | -0.252 | 0.053 | -4.783 | 0.0000 |

Here, and in the simulations following below, we accept as a valid insight the prediction of the model that longer words provide more evidence for a word's meaning than shorter words. This probabilistic advantage of a longer word for making contact with its meaning may help explain the U-shaped functional form of the effect of word length reported by Baayen (2005) and New, Ferrand, Pallier, and Brysbaert (2006) for English. For the shorter word lengths, a greater length combines with shorter response latencies. For the longer word lengths, facilitation reverses into inhibition. The facilitation for the shorter word lengths fits well with the prediction of the model that more bottom-up information provides more support for a word's lexical meaning (as well as its grammatical meanings). The increased processing costs for longer words are, in the present approach, the straightforward consequence of multiple fixations and saccades, a physiological factor unrelated to discriminative learning. Crucially, it is not the length in letters that is inhibitory in our approach, but whether more than one fixation is required. With $\phi = 0.3$, the by-item correlation of the observed and simulated latencies improved slightly from 0.24 to $r = 0.26$ ($t(1185) = 9.17, p < 0.0001$). Qualitatively, the effect of the other predictors in the model were not affected.

The second panel of Figure 5 shows the combined effect of Prime Condition and Normalized Levenshtein Distance. The Normalized Levenshtein Distance is zero for the identity primes, and nonzero for the other two prime conditions. We therefore plot their joint effect. Although the statistical models for the observed and simulated latencies assign different weights to the treatment contrasts of Prime Condition and to the slope of the Levenshtein

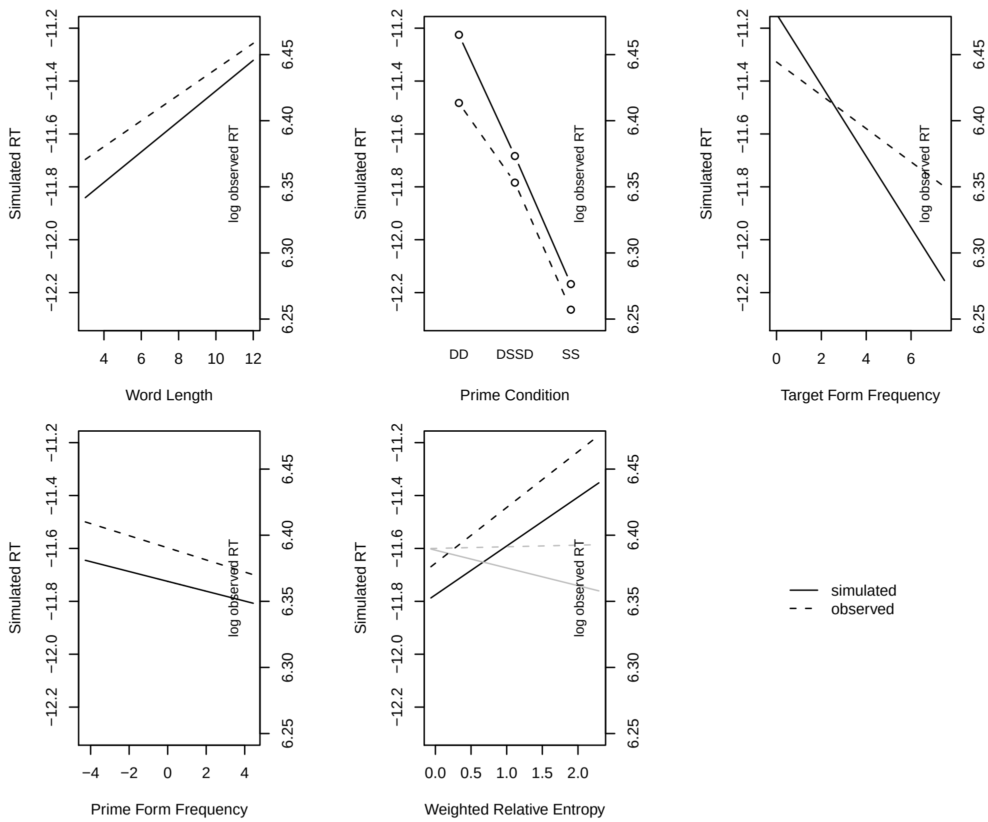


*Figure 5.* Partial effects of the significant predictors for the simulated self-paced reading latencies. The effect of Prime Condition in the second panel represents the combined effect of this factor and of the Normalized Levenshtein Distance. Solid lines represent simulated latencies, dashed lines the observed latencies. The grey lines in the last panel represent masculine nouns, the black lines represent neuter and feminine nouns.

Distance, the predictions of the two models are very similar: The simulated latencies faithfully reflect the priming effects. The main difference between the two regression models is that for the simulated latencies, the Levenshtein Distance is assigned greater weight, unsurprisingly, as the model has not been provided with any information on the discrete category mismatches of stems and case endings.

The third panel clarifies that the model adequately captures the facilitation of Target Lemma Frequency, although it underestimates the magnitude of the slope. The fourth panel shows it properly accounts for the form frequency of the prime. It is worth noting that the model's association strengths are estimated on the basis of absolute word frequencies, but that in the regression model the log-transformed frequency is used, exactly as for the observed reaction times. The effect of frequency of occurrence expresses itself linearly on a logarithmic scale in both observed and simulated latencies.

The final panel presents the interaction of Weighted Relative Entropy by Gender, which reaches significance for the simulated latencies just as for the observed latencies. The model even predicts a slight processing advantage for masculine nouns as entropy

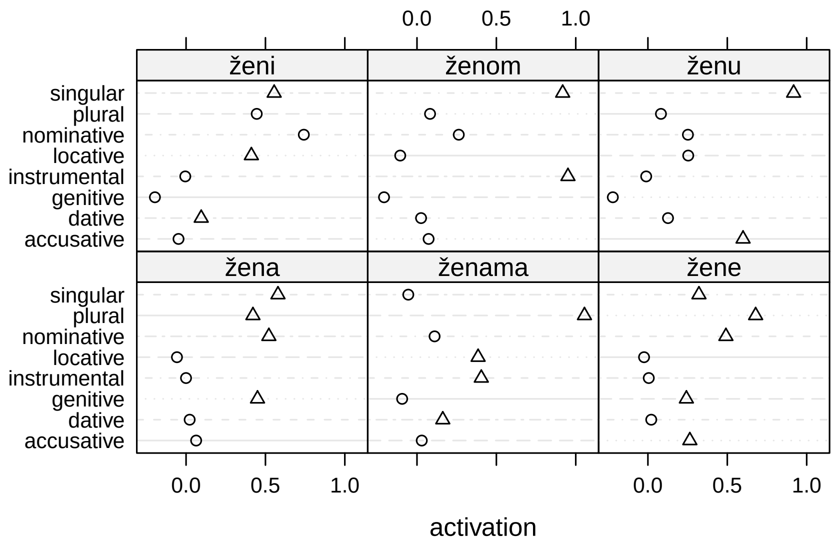


*Figure 6*. Activation of case and number meanings for the six inflected forms of *žena*. Triangles represent meanings that, in the absence of context, are possible and appropriate for a given form.

increases, which was not detectable for the observed latencies. The emergence of a significant interaction of WRE by Gender is exactly as predicted by the greater relative entropy that characterizes masculine nouns in Serbian.

As discussed above, the Cosine Similarity Measure reached significance only in the sentence reading task, and not in isolated word recognition. The absence of Cosine Similarity as a predictor for the simulated latencies is therefore as expected. The insignificant effect of Case (in interaction with Relative Entropy) in the model for the simulated latencies is expected, given that we have no evidence suggesting that the distributional properties of the nominative case forms are very different from those of the oblique case forms. Since the nominative case carries the lowest number of functions and meanings compared to the other case endings (Kostić et al., 2003), and since in our model the different functions and meanings of the cases are not specified, with all cases being treated equal, no special advantage for the nominative can emerge.

In summary, the present discriminative learning approach has succeeded in approximating well the effects of a series of lexical variables, including Weighted Relative Entropy. The model captures the effect of Weighted Relative Entropy without having to posit exemplars for individual inflected forms, and without having to specify explicitly prototypes for each inflectional class, and with only two free parameters, the compound cue weight $w$ for the prime in (15), and the fixation penalty parameter $\phi$.

The second challenge for the naive discriminative reader mentioned above is whether the model is sufficiently sensitive for handling the multiple ambiguity of the exponents expressing case and number. Some indication of the model's performance is provided by Figure 6, which presents, for each form of the feminine noun *žena*, the activations of the number and case meanings. For four forms, *žena, žene, ženu*, and *ženom*, the possible meanings (marked by triangles) have the highest activations. *Ženu*, for instance, is the accusative singular form, and these two meanings are appropriately activated most strongly in the model. For *ženama*, we see interference from the *-a* exponent, which is embedded in the *ama* exponent. For *ženi*, we have interference from the *-i* exponent expressing nominative plural in masculine nouns. In other words, all the model does is to make available the most likely readings of the exponents given the input. Further top-down processes will need to be brought into the model in order to account for the selection of the appropriate subsets of meanings given additional lexical and contextual information.

A final question awaiting clarification at this point is what exactly the relative entropy is capturing, both in the human data, as well as in the simulation. We address this question with the help of a simple constructed example. Table 10 presents a lexicon with four case-inflected forms for each of six lemmas. There are three different cases, nominative (nom), genitive (gen) and accusative (acc). The accusative is indexed by two different exponents, *a* and *u*. The *a* also occurs as a marker of the nominative, and hence is ambiguous as to which case it represents. In this example, exponents (in lower case) never occur as stem segments (in upper case). For each lemma, we calculated the relative entropy of the cases. For the first wordform, $p = \{10/100, 20/100, 70/100\}$, and since $q = \{0.246, 0.246, 0.501\}$, the relative entropy is $\sum(p \log 2(p/q)) = 0.134$. The $\{p\}$ distribution represents the exemplar, and $\{q\}$ the prototypical probability distribution.

To this data set, we fitted a logistic mixed-effects model, predicting Nominative versus other cases from Exponent (*a, i, u*) as fixed-effect factor and Lemma as random-effect factor. The random intercepts estimated for the lemmas are listed in Table 10 as Ranef Nom. The left panel of Figure 7 graphs the functional relation between relative entropy and the random intercepts. A greater (positive) random intercept implies that the lemma has a stronger preference for being used in the nominative. Conversely, large negative values indicate that the nominative is disfavored by a lemma. A random intercept equal to zero indicates no divergence from the average preference (log odds) for the nominative. The quadratic relation between entropy and the random intercepts is due to relative entropy being an unsigned measure. It captures the extent to which a lemma's preference for the nominative deviates from the population average, without indicating whether this preference is due to attraction or repulsion. What this example illustrates is that relative entropy provides a non-parametric, unsigned, alternative to the random intercepts of a logistic mixed-effects classifier.

We also fitted the naive discriminative reader model to this data set, using only unigrams as orthographic cues. Table 10 lists several statistics derived from the model's weight matrix. The support provided by the stem letters to the nominative, normed to the probability scale, is provided in the column listed as Stem Support Nom. As can be seen in the second panel of Figure 7, Relative Entropy and Stem Support Nom are again related through a quadratic polynomial, which is virtually identical to the one for Relative Entropy and the Random Intercepts. This does not come as a surprise, as the Random Intercepts and StemSupportNom are nearly perfectly linearly related, as shown in the third panel of Figure 7. In other words, the function of the random intercepts in the logistic mixed-effects model, the calibration of the lemmas' individual preferences for the nominative, is carried in the naive discriminative reader by the support from the cues comprising the stem.

Table 10: Simulated lexicon and associated statistics

| Word Form | Freq. | Case | Lemma | Relative Entropy | Ranef Nom | Ranef Acc | Stem Support Nom | Stem Support Gen | Stem Support Acc | Exponent Support |
|---|---|---|---|---|---|---|---|---|---|---|
| AQEa | 10 | nom | A | 0.134 | -1.121 | 1.121 | -0.014 | 0.260 | 0.533 | 0.353 |
| AQEi | 20 | gen | A | 0.134 | -1.121 | 1.121 | -0.014 | 0.260 | 0.533 | 0.740 |
| AQEu | 30 | acc | A | 0.134 | -1.121 | 1.121 | -0.014 | 0.260 | 0.533 | 0.595 |
| AQEa | 40 | acc | A | 0.134 | -1.121 | 1.121 | -0.014 | 0.260 | 0.533 | 0.127 |
| ABCa | 15 | nom | B | 0.053 | -0.676 | 0.676 | 0.037 | 0.260 | 0.482 | 0.353 |
| ABCi | 22 | gen | B | 0.053 | -0.676 | 0.676 | 0.037 | 0.260 | 0.482 | 0.740 |
| ABCu | 28 | acc | B | 0.053 | -0.676 | 0.676 | 0.037 | 0.260 | 0.482 | 0.595 |
| ABCa | 35 | acc | B | 0.053 | -0.676 | 0.676 | 0.037 | 0.260 | 0.482 | 0.127 |
| APQa | 20 | nom | C | 0.010 | -0.288 | 0.288 | 0.087 | 0.260 | 0.432 | 0.353 |
| APQi | 24 | gen | C | 0.010 | -0.288 | 0.288 | 0.087 | 0.260 | 0.432 | 0.740 |
| APQu | 26 | acc | C | 0.010 | -0.288 | 0.288 | 0.087 | 0.260 | 0.432 | 0.595 |
| APQa | 30 | acc | C | 0.010 | -0.288 | 0.288 | 0.087 | 0.260 | 0.432 | 0.127 |
| ZPEa | 30 | nom | D | 0.007 | 0.243 | -0.243 | 0.162 | 0.260 | 0.357 | 0.353 |
| ZPEi | 26 | gen | D | 0.007 | 0.243 | -0.243 | 0.162 | 0.260 | 0.357 | 0.740 |
| ZPEu | 24 | acc | D | 0.007 | 0.243 | -0.243 | 0.162 | 0.260 | 0.357 | 0.595 |
| ZPEa | 25 | acc | D | 0.007 | 0.243 | -0.243 | 0.162 | 0.260 | 0.357 | 0.127 |
| EPBa | 35 | nom | E | 0.039 | 0.583 | -0.583 | 0.210 | 0.260 | 0.309 | 0.353 |
| EPBi | 28 | gen | E | 0.039 | 0.583 | -0.583 | 0.210 | 0.260 | 0.309 | 0.740 |
| EPBu | 22 | acc | E | 0.039 | 0.583 | -0.583 | 0.210 | 0.260 | 0.309 | 0.595 |
| EPBa | 20 | acc | E | 0.039 | 0.583 | -0.583 | 0.210 | 0.260 | 0.309 | 0.127 |
| DPBa | 40 | nom | F | 0.139 | 1.269 | -1.269 | 0.289 | 0.260 | 0.230 | 0.353 |
| DPBi | 30 | gen | F | 0.139 | 1.269 | -1.269 | 0.289 | 0.260 | 0.230 | 0.740 |
| DPBu | 20 | acc | F | 0.139 | 1.269 | -1.269 | 0.289 | 0.260 | 0.230 | 0.595 |
| DPBa | 10 | acc | F | 0.139 | 1.269 | -1.269 | 0.289 | 0.260 | 0.230 | 0.127 |

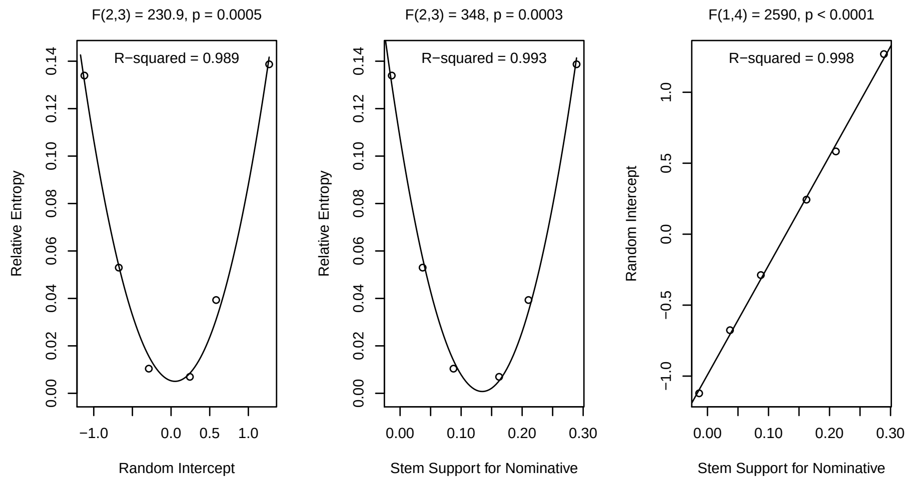


*Figure 7.* Scatterplots and Pearson and Spearman correlation coefficients for the probabilities of the ergative predicted by the naive discriminative reader, the by-lemma random intercepts in a logistic mixed-effects model, and the relative entropies of the lemmata, calculated for a constructed lexicon with a binary case distinction.

Table 10 also lists the support of the stem for genitive case (Stem Support Gen) and for accusative case (Stem Support Acc), and the support of the wordform's exponent for its corresponding case. As genitive case corresponds one-to-one with the presence of the *i* exponent, genitive case is well supported whenever the *i* is present (0.74), while there is no differentiation in the support provided by the cues provided by the stem (0.26 for all lemmas). This example illustrates that the naive discriminative learning algorithm, when presented with a truly agglutinative exponent, detects the one-to-one mapping of form to meaning. The special case of agglutination is captured naturally in our approach, without requiring any further mechanisms. By contrast, theories that assume processing is grounded in an item-and-arrangement (representation-plus-rules) architecture cannot be extended to account for the complexities of non-agglutinative morphological systems without many ad-hoc assumptions.

In this example, the *a* exponent is ambiguous between nominative and accusative. The random intercepts in a model predicting the accusative and the stem support for the accusative show the same correlational structure as shown in Figure 7 for the nominative. However, the stem support for the accusative is negatively correlated with the stem support for the nominative ($r = -1$), due to the way in which the form frequencies were assigned to the inflected variants of the lemmas, gradually changing from $\{10, 20, 30, 40\}$ to $\{40, 30, 20, 10\}$.

When case endings are ambiguous, as is the case for the *a* exponent in the present example, the weights from such an exponent to its case meanings cannot differentiate between the individual preferences of the lemmas. In other words, these weights do not enter into correlations with relative entropy. They represent the model's best guess, its probabilistic generalization, about the most likely meanings, optimizing across all lemmas.

We note here that the quadratic functions in Figure 7 are due to the gradual changes in the frequencies of the inflected forms, ranging from $\{10, 20, 30, 40\}$ for lemma A to $\{40, 30, 20, 10\}$ for lemma F. For distributions for which a majority of relative entropies reflect attraction (or all reflect repulsion), the relation between relative entropy and random intercepts (and stem support) can be roughly linear, with a slope that can be both positive and negative. For an example discussing the statistical analysis of an empirical data set, see Baayen (2011).

How exactly the effects of relative entropy work out for a given data set is highly dependent on the distributional characteristics of that data set. For the Serbian data, interpretation is complicated further by the presence of subliminal primes. Nevertheless, the prediction that follows from the above considerations for the Serbian data is that the effect of weighted relative entropy should reflect the support provided by the cues of the noun stems for the meanings of the case endings. We therefore calculated the stem support for each of the cases (nominative, genitive, etc.). We first inspected whether the weighted relative entropy can be predicted from the summed support for the cases in interaction with grammatical gender. This was indeed the case: For non-masculine nouns, the total support correlated negatively with weighted relative entropy, while for masculine nouns the correlation was positive ($p < 0.0001$ for the coefficients of both the two main effects and the interaction). We then defined a simplified simulated response latency as negative total support (as greater total support should contribute to a shorter response latency), and examined whether this simulated RT is predictable from weighted relative entropy in interaction with grammatical gender. We obtained results very similar to those listed in Table 9, with a positive slope for weighted relative entropy for non-masculine nouns, and a negative slope for masculine nouns (all $p < 0.0001$).

In summary, the (weighted) relative entropy measure, as applied to case paradigms, is a non-parametric and unsigned measure of the degree of attraction (or repulsion) for a given lemma to the average (population) probability distribution of case endings. The naive discriminative reader explains this attraction (or repulsion) in the empirical data as originating in the association of stem cues to case meanings as a result of discriminative learning. We note here that the association of stem cues to case meanings, and the relative entropy effects that bear witness to these associations, challenge linguistic theories that regard agglutination, in the sense of item-and-arrangement, as the underlying formal property of morphology.

## Modeling morphological processing in English: from simple words to prepositional paradigms

We have seen that a model based on naive discriminative learning correctly replicates the effects of a wide range of predictors, including weighted relative entropy, observed to co-determine the sentential reading of Serbian case-inflected nouns. The input to the model, however, is limited to just 3240 wordforms of those 270 nouns for which all case forms are attested in the *Frequency Dictionary of Contemporary Serbian Language* (Kostić, 1999). As a consequence, the results obtained might be due to overfitting.

To rule out this possibility, and to obtain further insight in the potential of naive discriminative learning for understanding morphological processing, we trained the model on a substantial part of the British National Corpus (henceforth BNC, Burnard, 1995), and pitted the predictions of the model against the by-item average lexical decision latencies available in the English Lexicon Project (henceforth ELP, Balota et al., 2004) as well as in previously published data sets.

In what follows, we first introduce the corpus data used to set the weights of the Rescorla-Wagner network that is the engine of the naive discriminative reader. We then discuss morphological effects across simple words, inflected words, derived words, pseudo-derived words, and compounding. Next, we consider whether phrasal frequency effects might also be captured within the same modeling framework. We conclude with showing that the paradigmatic exemplar-prototype effects characterizing the reading of Serbian nouns are also present in English, using as example the English equivalent of Serbian case paradigms: prepositional paradigms.

*The training data*

From the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), we extracted all monomorphemic nouns, verbs, and adjectives, as well as all compounds and derived words with a monomorphemic noun, verb or adjective as base word. For each of these words, forms inflected for number, person, tense and aspect were also extracted. This set of words was complemented with the word stimuli used in the studies of Rastle et al. (2004), Bergen (2004), and Christianson, Johnson, and Rayner (2005), resulting in a lexicon with 24710 different words (word types).

All instances of the words in our lexicon that occurred in the constructions listed in Table 11 were retrieved from the BNC, together with the preceding words in these constructions. Function words in the constructions were restricted to those occurring in a precompiled list of 103 determiners, prepositions, pronouns, and adverbs. Those words that did not appear in these constructions but that were used as stimuli in published experiments were extracted from the BNC, together with the preceding word (when not sentence-initial). Constructions with non-ASCII characters were discarded. The resulting phrasal lexicon comprised 1,496,103 different phrase types, 11,172,554 phrase tokens, to a total of 26,441,155 words (tokens), slightly more than a quarter of the total corpus size.

In summary, the input to the naive discriminative reader in the simulation studies below is a realistic sample of English words with simple morphological structure, in a wide range of locally restricted syntactic contexts as attested in the BNC. The connection weights of the Rescorla-Wagner network were calculated by solving the equilibrium equations (8). All following simulation are based on the resulting matrix of connection weights.

*Simple words*

Although simple words such as *shoe* or *think* have no internal syntagmatic morphemic structure, they enter into paradigmatic relations with inflected words (*shoes, thinks*, as well as with derived words and compounds *snowshoe, thinker*. The consequences for lexical processing of the entanglement of a simple word with its inflectional paradigm has been gauged in previous studies with Shannon's entropy (Shannon, 1948), a measure which estimates the amount of information carried by an inflectional paradigm (Baayen, Feldman, & Schreuder, 2006; Baayen et al., 2007; Baayen, Levelt, Schreuder, & Ernestus, 2008):

$$H_i = -\sum_k p_k \log_2(p_k). \tag{17}$$

Table 11: Constructions retrieved from the BNC. Words marked as $X$ were included even when not available in the initial 24710-word lexicon.

| | |
|---|---|
| Preposition + Article + Noun | *about a ballet* |
| Preposition + Possessive Pronoun + Noun | *about her actions* |
| Preposition + X + Noun | *about actual costs* |
| Preposition + Noun | *about achievements* |
| X's + Noun | *protege's abilities* |
| Article + Noun | *a box* |
| Article + X + Noun | *the abdominal appendages* |
| Possessive Pronoun + Noun | *their abbots* |
| Article + X's + Noun | *the accountant's bill* |
| Pronoun + Auxiliary + Verb | *they are arrested* |
| Pronoun + Verb | *he achieves* |
| Auxiliary + Verb | *is abandoning* |
| Article + Adjective | *the acute* |

In (17), $k$ ranges over a word's inflectional variants (for *shoe*, the singular *shoe* and the plural *shoes*, for *think* the verb forms *think, thinks, thinking*, and *thought*). The probability $p_k$ is the conditional probability of the $k$-th word in the paradigm:

$$p_k = \frac{f_k}{\sum_i f_i}, \tag{18}$$

where $f_i$ denotes the frequency of the $i$-th form in a word's inflectional paradigm. In visual lexical decision, inflectional entropy enters into a negative correlation with response latencies. For simple words, the kind of words under consideration here, Baayen et al. (2006) show that information-rich inflectional paradigms tend to afford shorter reaction times in the visual lexical decision task.

Simple words are entangled not only with their inflectional variants, but also with the derived words and compounds in which they occur. The type count of such words, its morphological family size, has also been observed to co-determine response latencies, such that words with greater morphological families are responded to more quickly (Schreuder & Baayen, 1997; Bertram, Baayen, & Schreuder, 2000; De Jong et al., 2000; Dijkstra, Moscoso del Prado Martín, Schulpen, Schreuder, & Baayen, 2005; Moscoso del Prado Martín et al., 2005; Moscoso del Prado Martín et al., 2004). Moscoso del Prado Martín et al. (2004) showed that the family size count is the upper bound of the entropy of the conditional probabilities of the family members given the family.

We pitted the predictions of the naive discriminative reader for the simple nouns studied by Baayen et al. (2006), using as predictors, in addition to family size and inflectional entropy, a word's mean bigram frequency, its length, its written frequency, its neighborhood density (using the N-count measure), its number of synonyms as listed in WordNet (Fellbaum, 1998), and the frequency ratio of the word's use as a noun or a verb. We also included a new predictor, prepositional relative entropy, which will be discussed in more detail below.

For the simulation, we selected from our lexicon the 1289 monomorphemic words that can be used as nouns for which lexical decision latencies are available in the ELP. The observed latencies were inverse-transformed ($-1000/RT$) to remove most of the right skew from the distribution. Table 12 lists the coefficients obtained with a regression model fitted to the empirical lexical decision latencies.

Shorter latencies were typical for more frequent words (Written Frequency), for words with large morphological families (Family Size), for words with more morphologically complex synonyms (Complex Synset Count), for words with more information-rich inflectional paradigms (Inflectional Entropy), and for words used more often as nouns than as verbs (Noun Verb Ratio). The effects of Word Length and Neighborhood Density (N-Count) did not reach significance. Words with greater Mean Bigram Frequency elicited longer latencies.

The question of interest is whether the processing costs predicted by the naive discriminative reader reflect the same set of predictors, with effect sizes of similar magnitude. A good fit can be obtained by defining the simulated RT simply as $\log(1/a_{\text{word}})$, in which case the model is completely parameter-free and driven entirely by the corpus-based input. The fit improves slightly by taking a word's strongest competitors into account. We first define the probability of identification of a word $i$ in its competitor set as

$$\text{Pid}_i = \frac{a_i}{a_i + \sum_{j=1}^{n} a_j}, \tag{19}$$

where $a_{\text{i}}$ is the activation of the $i$-th word, $a_j$ is the activation of a competitor and $n$ the number of highest-activated competitors taken into account. As Yarkoni, Balota, and Yap (2008) report that their Levenshtein-distance based neighborhood measure performs optimally when the 20 nearest neighbors are considered, we set $n$ to 20. Response latencies are taken to be proportional to the reciprocal of the probabilities of identification. To remove the rightward skew in the distribution of these reciprocals, simulated RTs were defined as

$$\text{RT}_i = \log(1/\text{Pid}_i). \tag{20}$$

The correlation for the observed and simulated response latencies was $r = 0.56$, ($t(1293) = 24.09, p = 0$). This correlation is comparable to the correlations reported by Moscoso del Prado Martín (2003) for the goodness of fit of his connectionist model to the lexical decision latencies in the English Lexicon Project. The correlations of simulated and observed response latencies reported by Norris (2006) for his Bayesian Reader model, for 4-letter words, were slightly higher, at 0.56 (for recognition threshold 0.95) and 0.61 (for recognition threshold 0.99).

However, as for the Serbian data, the model predicts facilitation from word length. We therefore adjusted (20) to bring into the model the costs of additional fixations for longer words.

$$\text{RT}_i = \log\left(\frac{1}{\text{Pid}_i} + \phi \text{I}_{[l_i > 5]}\right). \tag{21}$$

A regression model with exactly the same model specification that was used for the empirical latencies was fitted to the simulated latencies, with $\phi = 3.2$. The coefficients of this model are listed in Table 13. All coefficients in the simulated model have the appropriate sign, and the correlation of the coefficients for the regression models fitted to the observed and the simulated latencies reached significance r = 0.87, ($t(7) = 4.73$, $p = 0.0021$), see Figure 8, indicating that effect sizes are modeled reasonably well. It is only the N-count measure for which the model predicts a small but significant positive slope while the observed slope is effectively zero.

Table 12: Coefficients for the model fitted to the observed response latencies of monomorphemic English nouns.

| | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Intercept | -1.451 | 0.051 | -28.278 | 0.0000 |
| Mean Bigram Frequency | 0.023 | 0.008 | 2.713 | 0.0068 |
| Written Frequency | -0.048 | 0.003 | -16.408 | 0.0000 |
| Family Size | -0.015 | 0.006 | -2.545 | 0.0111 |
| Length | -0.003 | 0.008 | -0.382 | 0.7022 |
| Noun Verb Ratio | -0.004 | 0.002 | -2.109 | 0.0351 |
| Inflectional Entropy | -0.039 | 0.010 | -3.927 | 0.0001 |
| Complex Synset Count | -0.014 | 0.004 | -4.059 | 0.0001 |
| Prepositional Relative Entropy | 0.009 | 0.003 | 3.241 | 0.0012 |
| N-Count | -0.000 | 0.001 | -0.033 | 0.9739 |

Table 13: Coefficients for the model fitted to the simulated response latencies of monomorphemic English nouns.

| | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Intercept | 0.217 | 0.493 | 0.441 | 0.6594 |
| Mean Bigram Frequency | 1.113 | 0.081 | 13.714 | 0.0000 |
| Written Frequency | -0.540 | 0.028 | -19.246 | 0.0000 |
| Family Size | -0.404 | 0.057 | -7.108 | 0.0000 |
| Length | -0.117 | 0.080 | -1.465 | 0.1433 |
| Noun Verb Ratio | -0.114 | 0.019 | -6.073 | 0.0000 |
| Inflectional Entropy | -0.625 | 0.095 | -6.606 | 0.0000 |
| Complex Synset Count | -0.104 | 0.034 | -3.033 | 0.0025 |
| Prepositional Relative Entropy | 0.166 | 0.027 | 6.194 | 0.0000 |
| N-Count | 0.043 | 0.009 | 4.970 | 0.0000 |

It is noteworthy that an effect of morphological family size emerges in the simulated reaction times without the presence of any separate representations for complex words in the model. Similarly, we find an effect of inflectional entropy without the presence of separate representations for inflected words, and without any explicit paradigmatic organization imposed on such representations. These effects all fall out straightforwardly from naive discriminative learning.

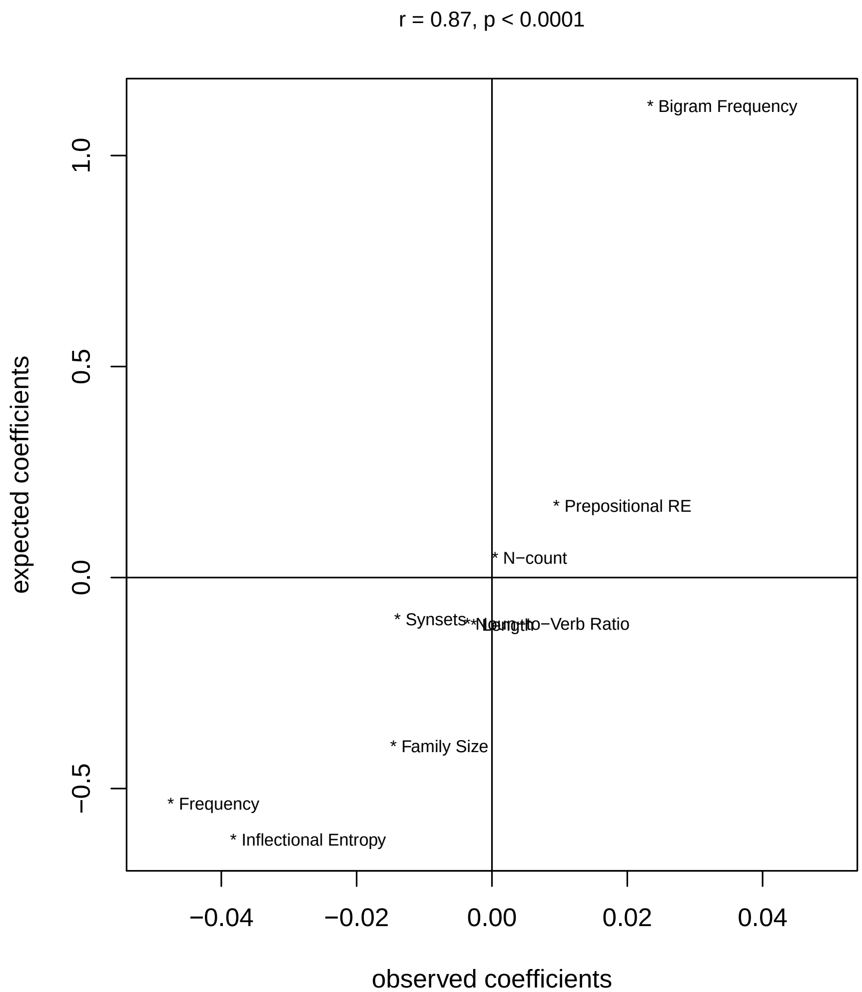


*Figure 8.* Observed and expected coefficients for the linear models for 1295 monomorphemic English nouns.

In addition to the effects described above, other orthographic frequency measures have been shown to play a role in word processing, at least in terms of eye-movement patterns during reading. One measure that is particularly relevant here is orthographic familiarity (White, 2008), defined as the sum of the token frequencies of the n-grams within a given word (e.g., in a four-letter word, the two trigrams, the three bigrams, and four unigrams). Orthographic familiarity has a significant (albeit small) facilitatory effect on several reading time measures, independently of word-frequency effects. Since n-grams are the very representation the present model adopts to implement orthographic information, it is no surprise that an effect of orthographic familiarity emerges in the naive discriminative reader in the form of a strong bigram frequency effect, with a positive slope as in the model for the observed latencies. (We have experimented with including higher-order n-grams as cues, but the increase in prediction accuracy was tiny compared to using letter pairs as cues in addition to letter unigrams.) The model correctly predicts inhibition for mean bigram frequency because high-frequency bigrams are shared between many different words, and hence have a low cue validity for their meanings. As the current implementation of our model is blind to how the eye extracts information from the visual input, the precise modeling of the early facilitatory effect of orthographic familiarity, which is believed to emerge at initial stages of fixation programming (White, 2008), is beyond the scope of the present study.

*Inflected words*

We begin our evaluation of the potential of naive discriminative learning for the comprehension of morphologically complex words with a study of present and past-tense inflection in English. Although the semantics of inflection tend to be straightforwardly regular, the formal expression of tense can be quite irregular, as is the case for the irregular verbs of English. Of specific interest to us is whether naive discriminative learning is sufficiently powerful to model the effects of (ir)regularity in visual comprehension of English verbs.

From the CELEX lexical database Baayen et al. (1995), we selected all verbs listed as monomorphemic. For these verbs, we took the (uninflected) present and past-tense plural forms (*walk, walked, come, came*) and extracted (where available) the corresponding lexical decision latencies from the English Lexicon Project (Balota et al., 2004), together with their frequency in the HAL corpus (Lund & Burgess, 1996b), their orthographic length, and their number of neighbors at Hamming distance 1 (the N-count measure). This resulted in a data set with 1326 different verb lemmas, of which 1209 were regular and 131 were irregular verbs. The total number of different verbal word forms was 2314. Response latencies were inverse transformed ($-1000/RT$), and HAL frequency was log-transformed (base $e$). For each verb, its log-transformed morphological family size and its inflectional entropy were included as additional covariates, together with two factors specifying whether a verb was regular or irregular (Regularity), and whether a verb form was in the past tense (PastTense).

A mixed-effects model fitted to the empirical lexical decision latencies with random intercepts for verb lemma (as a given verb contributes a present and a past-tense form) revealed the expected negative slopes for Frequency, Family Size, and Inflectional Entropy, and the expected positive slope for Length. The N-Count measure did not reach significance. Regularity and Tense interacted as shown in the upper left panel of Figure 9, with a difference in the group means for past and present tense forms for irregulars but not for

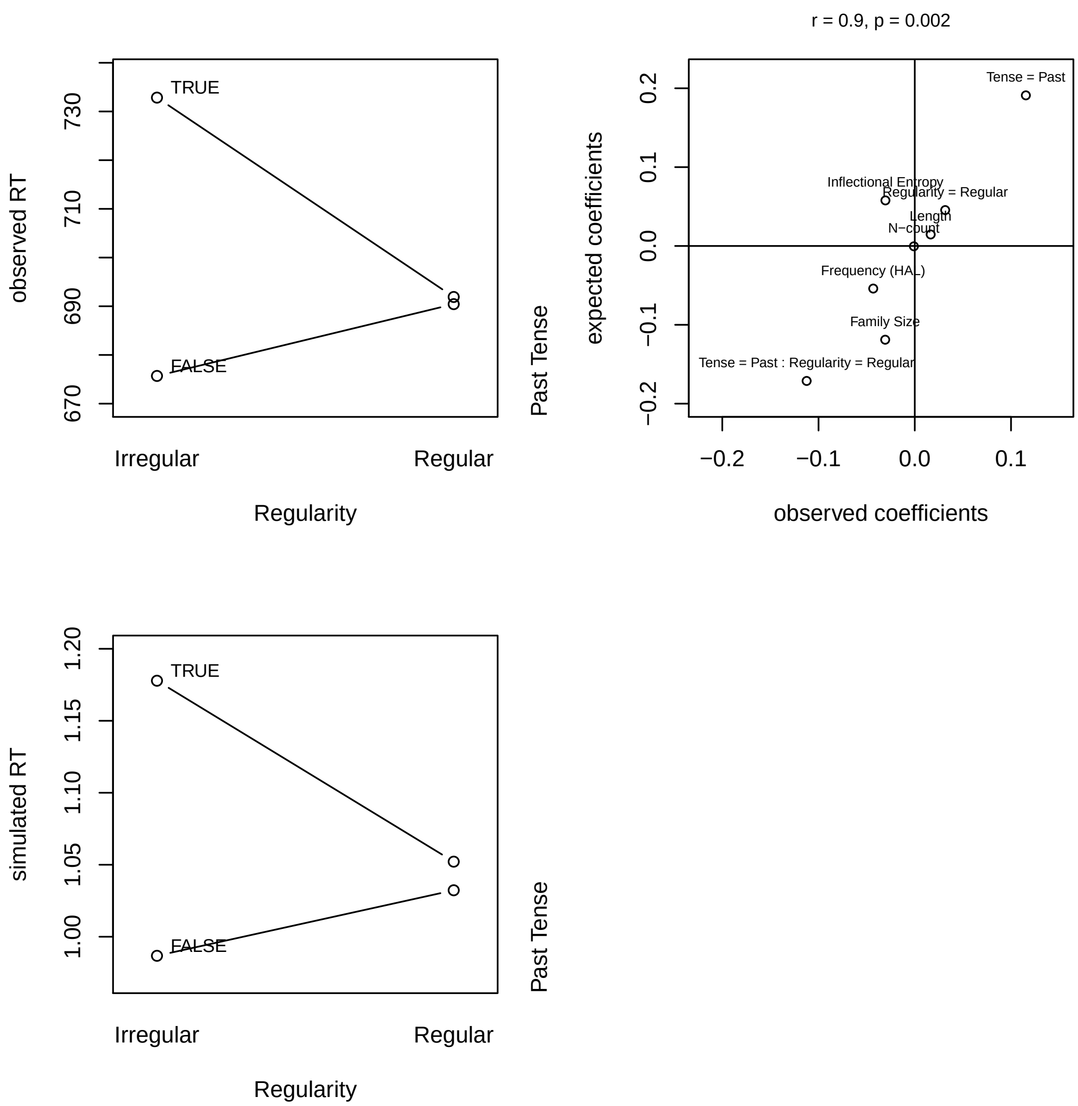


*Figure 9.* Observed (upper left) and simulated (lower left) interaction of Regularity by Past Tense, and a scatterplot for the observed and expected coefficients (upper right) for the models fitted to the observed and simulated latencies for English present and past-tense verbs.

regulars (see Table 14).

Table 14: Coefficients for the mixed-effects model fitted to the observed response latencies for inflected verbs. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

| | Estimate | Lower | Upper | P |
|---|---|---|---|---|
| Intercept | -1.1828 | -1.2491 | -1.1297 | 0.0001 |
| Frequency | -0.0433 | -0.0474 | -0.0414 | 0.0001 |
| Tense = Past | 0.1160 | 0.0789 | 0.1433 | 0.0001 |
| Family Size | -0.0306 | -0.0358 | -0.0189 | 0.0001 |
| N-count | -0.0009 | -0.0025 | 0.0008 | 0.3662 |
| Length | 0.0169 | 0.0127 | 0.0238 | 0.0001 |
| Inflectional Entropy | -0.0305 | -0.0474 | -0.0123 | 0.0006 |
| Regularity = Regular | 0.0314 | 0.0071 | 0.0558 | 0.0104 |
| Tense = Past : Regularity = Regular | -0.1136 | -0.1498 | -0.0802 | 0.0001 |

The modeling of tense inflection raises three implementational issues. A first issue is how to represent tense, as an equipollent opposition (with a past-tense semantic representation as well as a present-tense representation) or as a single graded representation representing the amount of evidence supporting the (marked) past-tense interpretation. We opted for the second, more parsimonious solution. The binary distinction between present and past tense was modeled with a single semantic representation capturing the amount of evidence supporting a past tense interpretation. We rescaled the activation $a_{\text{past}}$ of the past-tense meaning for a given verb into a probability. Defining $\boldsymbol{a}_{\text{past}}$ to denote the vector of activations of the past-tense meaning across all inflected words, the rescaled activation of a paste-tense meaning is given by

$$a'_{\text{past}} = \frac{a_{\text{past}} + \min(\boldsymbol{a}_{\text{past}})}{\max(\boldsymbol{a}_{\text{past}}) - \min(\boldsymbol{a}_{\text{past}})}. \tag{22}$$

For present-tense verbs, we take the complementary probability:

$$p_{\text{tense}} = \begin{cases} a'_{\text{past}} & \text{for past-tense verb forms} \\ 1 - a'_{\text{past}} & \text{for present-tense verb forms.} \end{cases} \tag{23}$$

A second issue concerns what semantic information is made available to the model during training. Many verb forms are ambiguous with respect to tense: *come* is either a finite present-tense form (*I come*) or the infinitive (*I want to come*), *walked* is either a past-tense form (*I walked*) or a past participle (*I have walked*), and *hit* can be either a present or past tense finite form (*I hit*) or a past participle (*I have hit*). The interpretation of these forms is context-dependent. The constructions extracted from the BNC providing information about verb inflections to our model contained a verb preceded by a pronoun, an auxiliary, or a pronoun and an auxiliary. A verb was coded as expressing past tense semantics if and only if it appeared in the context of an immediately preceding auxiliary that unambiguously signals paste-tense semantics (e.g., *had, was, were*). As a consequence, the model critically depends on contextual cues for learning the past-tense semantics of regular verbs with the *-ed* suffix. Irregular past-tense forms, by contrast, were associated with past-tense meaning independently of context (*had, was, were, came, went, saw, ...*).

A third issue concerns how to weight the contributions of the semantic representations for the verb and past tense. As the semantics of the verb itself is, in general, much richer than that of the more abstract semantics of present or past, we expect that a better fit is obtained when the weight for the past tense meaning is smaller than that of the verb meaning. We therefore introduce a weight $0 < w_{\text{tense}} < 1$.

As for the monomorphemic words, the simulated RT was defined as the log-transform of the reciprocal of the probability of identification of a word in the set of its most highly activated competitors:

$$\begin{aligned} \text{Pid} &= \frac{w_{\text{tense}} p_{\text{tense}} + a_{\text{verb}}}{w_{\text{tense}} p_{\text{tense}} + a_{\text{verb}} + w_c \sum_{i=1}^{n} a_i} \\ \text{simulated RT} &= \log(1/\text{Pid}), \end{aligned} \tag{24}$$

where $w_c$ is a weight for the summed activations of the $n$ strongest competitors. As for the monomorphemic words, $n$ was fixed at 20. A good fit was obtained for $w_{\text{tense}} = 0.15$ and $w_c = 0.1$. Finally, the effect of multiple fixations is brought into the model as before,

$$\text{simulated RT} = \log\left(\frac{1}{\text{Pid}} + \phi \mathrm{I}_{[l>5]}\right), \tag{25}$$

with $\phi = 0.2$.

For two words (*bade, whiz*), the simulated activation was less than zero. These verbs were removed from the data set. The correlation between the observed and simulated response latencies was r = 0.47, ($t(2312) = 25.39$, $p = 0$). Table 15 lists the coefficients of the mixed-effects model fitted to the simulated latencies. The correlation between the coefficients of the regression models for the observed and expected latencies was r = 0.9, ($t(6) = 5.15$, $p = 0.0021$). This correlation is illustrated in the right panel of Figure 9.

Frequency and Family Size were significantly facilitatory for both observed and simulated response latencies. The N-count measure was not predictive in both mixed-effects models. Longer words elicited significantly longer response latencies, which the model attributes to additional fixations being required for longer words.

Whereas Inflectional Entropy was facilitatory for the observed latencies, it emerged as inhibitory for the simulated latencies, indicating that the model fails to learn the inflectional meanings associated with verbal meanings properly. Again, there is a good reason for this. The model was trained on data that specified past tense wherever possible, but did not provide information on the aspectual meanings such as the present/past perfect (*she has/had walked*) or the continuous (*she is/was walking*). Hence, the empirical inflectional entropy (based on CELEX) does not match the model's learning experience. For the simple nouns studied above, the empirical inflectional entropies provided a much better characterization of the model's learning opportunities — number specification was available to the model through disambiguating pronouns and determiners in the context. As a consequence, inflectional entropy could emerge with the correct sign in the simulation of the simple nouns.

Table 15: Coefficients for the mixed-effects model fitted to the simulated response latencies for inflected verbs. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

| | Estimate | Lower | Upper | P |
|---|---|---|---|---|
| Intercept | 1.4467 | 1.5663 | 1.8048 | 0.0001 |
| Frequency | -0.0557 | -0.0916 | -0.0795 | 0.0001 |
| Tense = Past | 0.1939 | 0.0704 | 0.1991 | 0.0001 |
| Family Size | -0.1352 | -0.1123 | -0.0792 | 0.0001 |
| N-count | 0.0003 | -0.0018 | 0.0048 | 0.3580 |
| Length | -0.0041 | -0.0223 | -0.0008 | 0.0410 |
| Inflectional Entropy | 0.0660 | 0.0249 | 0.0955 | 0.0001 |
| Regularity = Regular | 0.0397 | -0.0274 | 0.0700 | 0.3884 |
| Tense = Past : Regularity = Regular | -0.1816 | -0.2082 | -0.0695 | 0.0004 |

The contrast coefficients for Tense, Regularity, and their interaction all agreed in sign and reached significance for both the observed and simulated latencies. The lower panel of Figure 9 visualizes the interaction of Regularity by Past Tense in the simulated latencies. The interaction in the simulated RTs mirrors well the interaction characterizing the observed latencies, with a large difference between present and past-tense irregulars forms, and similar latencies for regular present and past tense forms.

The interaction of regularity by tense is difficult to interpret in current models assuming parsing of regulars and storage for irregulars. Under such accounts, regular past-tense forms, requiring some decompositional processing, should elicit longer latencies than the corresponding present-tense forms, contrary to fact. Furthermore, the processing advantage for irregular present-tense forms compared to regular present-tense forms is not expected. Crucially, the interaction of regularity by tense occurs in a model in which frequency and other variables are included as covariates.

Our model suggests a very different interpretation. Recall that during training, the information available to the model for discriminating between present and past meanings is very different for regulars and irregulars. For irregular past-tense forms, the past-tense interpretation is made directly available during learning, independently of the context. For regulars, by contrast, the past-tense reading is available for learning only in the presence of past-tense auxiliaries.

The observed lexical decision latencies were elicited for words presented in isolation, without context, and our model likewise simulates reaction times for isolated word reading. Since the present/past distinction is not context-dependent for irregulars, a large difference emerges for the means of the simulated latencies of irregulars. By contrast, the low cue validity of *-ed* as a marker for the past tense causes regular past-tense forms to be highly context-dependent for their tense interpretation. Regular past-tense forms do become associated with the past-tense meaning to a greater extent than present-tense forms, but compared to irregular verbs, the association is weaker. With only a small weight for the Tense meaning ($w_1 = 0.14$), the group means for present and past-tense regulars collapse, potentially reflecting a list effect in the English Lexicon Project, in which many different words were presented and in which Tense was not a prominent feature.

In summary, this simulation study shows that a reasonable fit to the data can be obtained with the basic engine introduced for the simulation of simple nouns, combined with four free parameters: the weight for the tense meaning, two parameters defining the weight and size of the competitor set, and a fixation penalty. The model faithfully reflects the interaction of Regularity by Tense, an interaction that challenges classical, purely representation-based theories of morphological processing.

We suspect that the pattern of results observed for lexical decision will change when these forms are read in sentential context. In sentential context, information is available for disambiguation of the ambiguous *-ed* suffix. As a consequence, we expect that in context, the past tense meaning will be activated more strongly for regular verbs. It is important to note, however, that the naive discriminative reader models only the initial stage of visual comprehension, in which orthographic information contacts meanings. Subsequent processes of context-driven disambiguation and interpretation are not accounted for. Therefore, two assumptions are crucial to the present explanation. Firstly, if our explanation is on the right track, then the lexical decision task provides a window on this context-free initial activation of lexical meanings from orthographic forms. Secondly, it is assumed that discriminative learning of the mapping from form to meaning is informed only by meanings that are unambiguous and that do not need higher-level cognitive processes to resolve their ambiguity. In other words, our hypothesis is that the initial mapping from form to meaning is learned not on fully specified meanings that are the outcome of complex and late processes of sentential interpretation, but rather on the underspecified meanings that form the input to those processes.

### *Derived words*

Whereas the forms of inflected words typically tend to mark aspects of meaning that are relevant for syntactic coreferential processing (e.g., number and person agreement marking), derived words tend to express meanings that differ more substantially from those of their base words. While for many words, the semantics of the base are transparently visible in the semantics of the derived word (e.g., *true* in *truth*), some derived words can have meanings for which this is no longer true (e.g., *busy* in *business*). Derivation is therefore described as involving word formation, in the sense that it allows for the creation of labels for new concepts that have gained currency in the speech community.

Inflectional morphology tends to be quite regular (the irregular past tenses of English being exceptional), but derivational processes are characterized by degrees of productivity. Some suffixes are hardly ever used for the creation of new words (e.g., English *-th* in *warmth*), while others give rise to large numbers of new formations (e.g., *-ness* in English). The extent to which a derivational affix is available for the formation of new words is known as its degree of productivity (see, e.g., Baayen & Renouf, 1996; Bauer, 2001; Baayen, 2008).

In what follows, we first consider what lexical distributional properties predict the processing of derived words, following Baayen et al. (2007), and examine whether naive discriminative learning replicates the importance of these properties. We then consider whether the model also predicts that more productive affixes require longer processing latencies, as observed by Plag and Baayen (2009). Next, we consider whether it is necessary to postulate a special early morphographic parsing process, as claimed by Rastle et al. (2004). Finally, we examine whether the notion of the morpheme, a theoretical construct that many current theories of morphology consider to be obsolete (Beard, 1977; Aronoff, 1994; S. Anderson, 1992) can be dispensed with in the discriminative learning framework by considering the processing of phonaesthemes.

*Derived word processing.* We selected 3003 derived words (569 prefixed words and 2434 suffixed words) with 81 different affixes and 1891 different base words for analysis.

As predictors for the observed and simulated lexical decision latencies, we considered the frequency and length of the derived word, the frequency of its base, the family size of its base, the family size of the suffix, and the frequency of the letter bigram spanning the transition from base into suffix or prefix into base (the boundary bigram frequency). Effects of the frequency and family size of the base have often been interpreted as evidence of the orthographic input being parsed into stem and affix representations (see, e.g., Taft & Forster, 1976b; Bertram, Schreuder, & Baayen, 2000; Kuperman, Bertram, & Baayen, 2008), Whole-word frequency effects, by contrast, would indicate non-compositional, holistic processing. Furthermore, it has been argued that morphological effects are due to complex words typically having low-frequency boundary bigrams (Seidenberg, 1987; Seidenberg & McClelland, 1989, for discussion and counterevidence, see Rapp, 1992). Given these traditional diagnostic measures for morphological processing, the question we need to address is whether the present discriminative learning framework can properly reflect the importance of these predictors for lexical processing.

As these predictors are highly collinear with a condition number $\kappa = 33.6$ (Belsley, Kuh, & Welsch, 1980), we orthogonalized them as follows. Base frequency was residualized on word frequency. The residualized base frequency strongly correlated with the original count ($r$ =0.96). Base family size was residualized on word frequency ($r$ =0.98). Suffix family size was residualized on (residualized) family size and word frequency ($r$ =0.99). Finally, the boundary bigram frequency was residualized on all other predictors ($r$ =0.99). The condition number for the resulting set of predictors was substantially reduced to $\kappa =$ 12.9. At the same time, the high correlations of the new variables with their originals ensure that the new variables remain well interpretable.

We fitted a mixed-effects regression model to the observed latencies with random intercepts for base and affix. The estimated coefficients are listed in Table 16. Response latencies increased with word length. Words with a higher boundary bigram frequency elicited longer latencies as well. More frequent words, words with more frequent base words, and words with large base families or suffix families, elicited shorter response latencies.

The simulated response latencies were defined along the same lines as for inflected words:

$$\text{Pid} = \frac{w_{\text{affix}} a_{\text{affix}} + a_{\text{base}}}{w_{\text{affix}} a_{\text{affix}} + a_{\text{base}} + w_c \sum_{i=1}^{n} a_i} \qquad (26)$$

$$\text{simulated RT} = \log\left(\frac{1}{\text{Pid}} + \phi \mathrm{I}_{[l>5]}\right) \qquad (27)$$

The number of competitors $n$ was fixed at 20 as in the preceding simulations. A good fit was obtained for affix weight $w_{\text{affix}} = 0.25$, for competitor weight $w_c = 0.1$, and for

Table 16: Coefficients for the mixed-effects model fitted to the observed response latencies for derived words. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

| | Estimate | Lower | Upper | P |
|---|---|---|---|---|
| Intercept | -1.2956 | -1.3282 | -1.2541 | 0.0001 |
| Length | 0.0277 | 0.0230 | 0.0310 | 0.0001 |
| Word Frequency | -0.0664 | -0.0701 | -0.0636 | 0.0001 |
| Base Frequency | -0.0071 | -0.0107 | -0.0037 | 0.0001 |
| Base Family Size | -0.0119 | -0.0182 | -0.0039 | 0.0040 |
| Affix Family Size | -0.0151 | -0.0272 | -0.0050 | 0.0066 |
| Boundary Bigram Frequency | 0.0068 | 0.0047 | 0.0112 | 0.0001 |

$\phi = 0.2$. The correlation between the observed and simulated latencies was $r = 0.25$, $(t(3001) = 13.86, p < 0.0001)$. We fitted the same mixed-effects model to the simulated latencies. The coefficients of this model are reported in Table 17. Figure 10 visualizes the correlation between the coefficients of the model fitted to the observed and expected latencies.

Table 17: Coefficients for the mixed-effects model fitted to the simulated response latencies for derived words. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

| | Estimate | Lower | Upper | P |
|---|---|---|---|---|
| Intercept | 1.0359 | 1.0267 | 1.0598 | 0.0001 |
| Length | 0.0041 | 0.0021 | 0.0055 | 0.0001 |
| Word Frequency | -0.0110 | -0.0144 | -0.0118 | 0.0001 |
| Base Frequency | -0.0172 | -0.0192 | -0.0164 | 0.0001 |
| Base Family Size | -0.0064 | -0.0088 | -0.0024 | 0.0004 |
| Affix Family Size | -0.0081 | -0.0134 | -0.0014 | 0.0146 |
| Boundary Bigram Frequency | 0.0045 | 0.0046 | 0.0073 | 0.0001 |

Although the coefficients for the simulated reaction times have the right sign and reach significance, the correlation between the two sets of coefficients fails to reach significance, indicating that there is room for improvement. Most striking is the imbalance of word frequency and base frequency. For the observed latencies, the coefficient for word frequency is larger than that for base frequency. For the simulated latencies, the reverse holds. This is due to the model being a fully decompositional model that does not do justice to the loss of transparency of many derived words (e.g., the meaning of *business*, 'enterprise', is not straightforwardly related to the meaning of its base, *busy*). We expect more balanced results once opaque derived words are assigned separate meaning representations, distinct from those of their base words.

Whereas facilitatory effects of base frequency and family size are traditionally under-

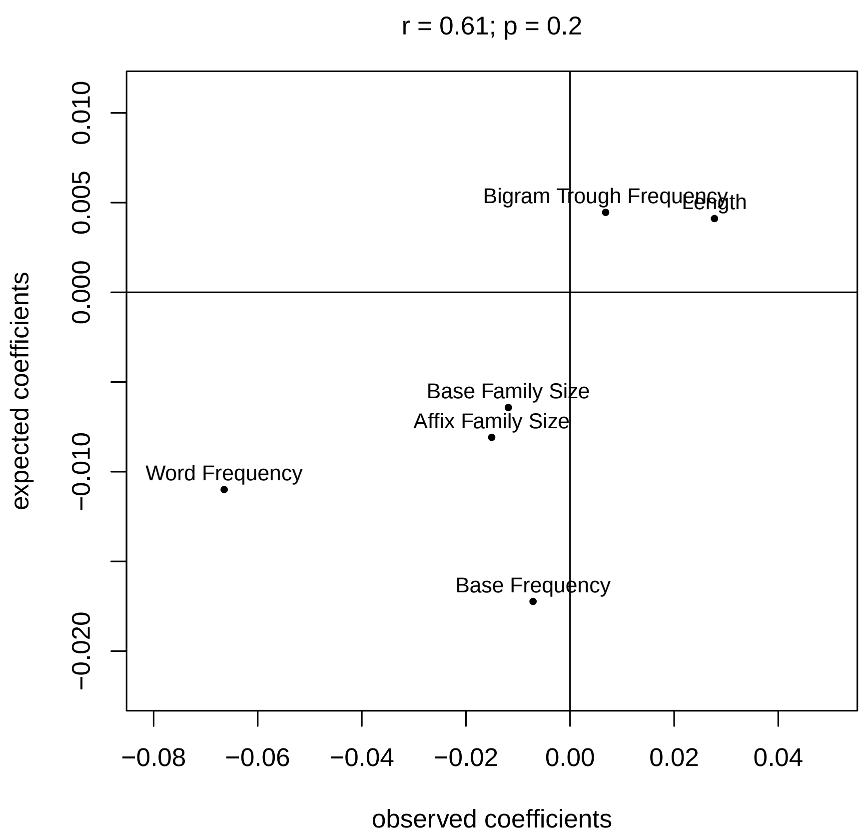


*Figure 10.* Coefficients for the observed and simulated response latencies for 3003 English derived words.

stood as the result of the input being parsed into its constituents, the present simulation shows that these facilitatory effects can arise without any explicit parsing process being involved. Furthermore, a whole-word frequency effect is present in the simulation in the absence of any whole-word representations.

As expected, words with higher boundary bigram frequencies emerge with both greater observed and greater simulated latencies. Conversely, processing is faster for lower-frequency boundary bigrams, or 'bigram troughs' (Seidenberg, 1987). This effect co-exists peacefully with stem-frequency and constituent family size effects, indicating that it is one of several processing diagnostics of morphological complexity. We note here that bigram trough effects are open to very different interpretations. The original hypothesis of Seidenberg (1978) was that low-level processing of letter-pairs is at issue. By contrast, Hay (2003) argues that affixes with deeper bigram troughs are easier to parse out, affording greater affix productivity.

The naive discriminative reader predicts that bigram troughs also should give rise to shorter response latencies, but not because morphological decomposition would proceed more effectively. The reason bigram troughs provide facilitation in our model is very different, although straightforward. High-frequency boundary bigrams are typically used word-internally across many words, and therefore have a low cue validity for the meanings of these words. Conversely, low-frequency boundary bigrams are much more typical for specific base+affix combinations, and hence are better discriminative cues, affording enhanced activation of meanings, and hence allowing faster processing.

*Affix productivity.* As mentioned above, derivational affixes differ in their degree of productivity. Affixal productivity can be gauged by considering the number of different word types with a given affix. In the present data set, unproductive *-th* is represented by 16 word types, and productive *-ness* by 177 word types. Although the number of types in which an affix occurs, referred to above as affix family size, provides a decent first approximation of affixal productivity, a more sensitive measure considers the likelihood of encountering new, previously unseen formations. The measure we examine here ($\mathcal{P}$) is the Good-Turing estimate of the probability mass of words present in the population but absent in a (corpus) sample (Good, 1953; Baayen, 1992):

$$\mathcal{P} = \frac{V_1}{N}, \tag{28}$$

where $V_1$ denotes the number of types with the affix appearing once only in the sample (corpus), and $N$ the total number of tokens of all words with the affix in the sample. An intuitive understanding of this measure can be obtained by considering a vase with marbles of different colors, with different colors occurring with varying frequencies (e.g., red 6, yellow 1, blue 15, purple 1, magenta 2, green 3, white 7, black 2, brown 3). When a marble is drawn from the vase without replacement, the likelihood that its color occurs once only is equal to the ratio of the number of colors with frequency 1 ($V_1$) to the total number of marbles ($N$), for the present example leading to the probability (2/40). Once sampled (without replacement), the color uniquely represented by the marble drawn from the vase will never be sampled again. By symmetry, the probability that the *last* marble sampled has a color that has not been seen previously equals $V_1/N$. In other words, $\mathcal{P}$ is the probability that, having seen $N-1$ tokens, an unseen type will be sampled at 'sampling time' $N$. For formal proofs, see, e.g., Baayen (2001). This productivity measure outperforms the affix family size in that it correctly predicts that an affix instantiated in a relatively small number of types can nevertheless be productive, see Baayen (1994) for experimental evidence.

Recently, Plag and Baayen (2009) observed for a selection of English derivational suffixes that the by-affix processing cost, estimated by averaging response latencies across all words with a given affix, entered into a positive correlation with degree of productivity $\mathcal{P}$. It is only for the most productive suffixes that this effect was slightly attenuated. The upper panel of Figure 11 replicates the main trend for a larger selection of affixes, including not only suffixes but also prefixes. As productivity increases, processing latencies increase (r = 0.34, ($t(73)$ = 3.08, $p$ = 0.0029)). The lower panel shows that the same pattern is present in the simulated latencies ($r = 0.37, (t(73) = 3.45, p = 0.0009)$).

The traditional psycholinguistic interpretation of the $\mathcal{P}$ measure is that (i) words with many low-frequency formations (such as formations which occur once only, contributing to $V_1$) are unlikely to have whole-word representations, and hence depend on rule-based parsing, and (ii) that the more high-frequency, lexicalised formations there are (contributing to a

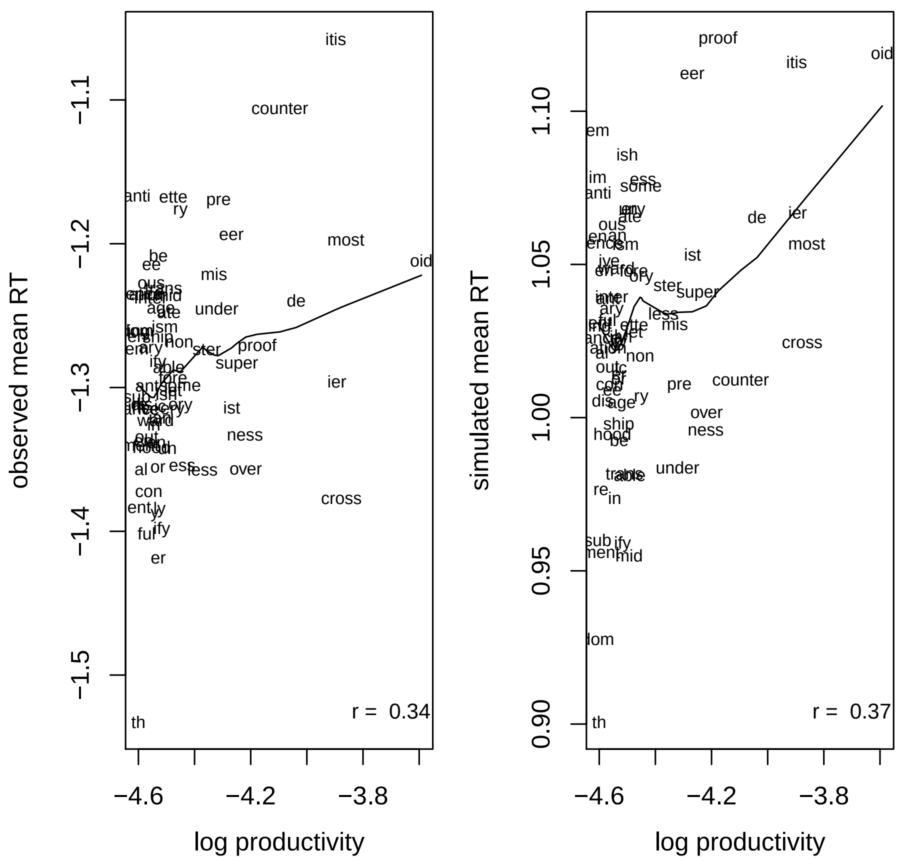


*Figure 11.* Log affix productivity ($\mathcal{P}$) as predictor of mean affix latency for observed (left panel) and simulated (right panel) data. Observed latencies are on the $-1000/RT$ scale. Regression lines are non-parametric lowess regression smoothers.

large $N$), the more rule-based processing will be superfluous. In other words, productive affixes have relatively few higher-frequency words, and many lower-frequency forms, providing a bias against whole-word based access and a bias in favor of decompositional processing (see, e.g., Baayen, 1992). Since our discriminative learning model does not incorporate whole-word representations for derived words, and yet faithfully reproduces the positive correlation of affix productivity and average processing latency, a different explanation is required.

What Figure 11 shows is that less productive suffixes, which tend to be suffixes with relatively fewer but higher-frequency formations, guarantee shorter latencies, on average. Their higher frequency of occurrence ensures better by-item learning. Furthermore, because a less productive suffix is more constrained in its morphological micro-context — it co-occurs with fewer stems — it should become a relatively good cue for these few stems. Conversely, low token frequencies and many types lead to reduced item-specific learning, with as flip side better generalisation to previously unseen words. This line of reasoning predicts that for a productive suffix such as *-ness* the activation of the suffix in new words should be greater than the activation of an unproductive suffix such as *-th*.

To test this prediction, we made a list of formations that were new for the model, by adding these suffixes to the 322 monomorphemic adjectives in the training set, and subsequently removing all forms to which the model had been exposed during training. For *-ness* (as in *goodness*, $\mathcal{P} = 0.0047$), there were 208 unseen derived words. There were 321 unseen formations for *-th* ($\mathcal{P} < 0.0001$). For each of these new words, we then calculated the predicted activation of the suffix meaning. A Wilcoxon rank sum test indicated that, as expected, the average activation of the suffix was greater ($W = 38171, p = 0.0053$) for productive *-ness* (0.382) than for unproductive *-th* (0.273) The model therefore predicts reduced processing costs for neologisms with productive suffixes. In other words, the processing advantage that less productive suffixes enjoy over more productive suffixes for existing words is reversed into a processing disadvantage for unseen words.

Interestingly, even though *-th* is typically described as no longer productive in English (see, e.g. Bauer, 2001), occasionally new words emerge (Baayen, 2009). Among the novel words predicted by the model to have a higher activation for *-th*, we find *strength, slowth, firmth* and *oldth*, all attested on Urban Dictionary (`http://www.urbandictionary.com`). Although these new words have the flavor of being quite unusual, it is not difficult to deduce their meanings (as in "Caution: installation of too many Firefox add-ons may induce slowth", Urban Dictionary s.v. *slowth*). This fits well with the relatively high activation levels predicted for *-th*: Even though smaller on average than those for *-ness*, they tend to be larger than zero. In fact, there is considerable overlap in the distributions of activations for the two suffixes. The model predicts that even for new words, the meaning of the unproductive suffix is activated, hence, neologisms such as *slowth* are correctly predicted to be comprehensible, even though most speakers would not spontaneously produce such words themselves.

We conclude that naive discriminative learning succeeds in capturing important aspects of morphological productivity, without having to posit separate representations for complex words or separate morphological rules associated with some kind of probability that would specify their likelihood of application.

*Pseudo-derived words.* Thus far, we have shown that morphological effects arise in our model in the absence of any specifically morphological processes or representations. However, a well known and widely discussed phenomenon in the recent psycholinguistic literature is a pattern of morphological priming effects emerging in masked priming experiments that would support the existence of an early morpho-orthographic parsing process. In what follows, we focus on the study of Rastle et al. (2004), who observed that the magnitude of the priming effect for target words preceded by a derived prime was comparable irrespective of whether the prime was a semantically related morphological relative (e.g., *dealer-deal*) or whether the prime-target relationship was semantically opaque (e.g., *corner-corn*). The priming effects obtained for these conditions were significantly larger than those obtained in a form condition in which no suffix is present (*brothel-broth*). This evidence has been interpreted as indicating that complex words are decomposed at an early morpho-orthographic level of processing, and that this decomposition process is triggered by apparent morphological structure. The hypothesis of an early purely form-driven morpho-orthographic decomposition process is not uncontested, and may depend on the kind of filler materials in the experimental list (Feldman, O'Connor, & Moscoso del Prado Martín, 2009). Since our model does not comprise a morpho-orthographic processing module, it is important to clarify whether or not the data of Rastle et al. (2004) can nevertheless be adequately simulated.

A first question that we need to address for modeling the data of Rastle et al. (2004) is how to represent the meanings of the pseudo-derived items in their study: words such as *early, fleeting, fruitless, archer*, and *cryptic* (examples taken from the appendix of Rastle et al., 2004). Linguistically, these pseudo-derived words are a very heterogeneous set. The stem of *early* is related historically to modern English *ere*, a link not many native speakers will be aware of, but the suffix *-ly* is still functional as an adverbial marker.

The adjective *fruitless* is opaque when considered in isolation: the meaning 'in vain', 'unprofitable' seems unrelated to the meaning of the base, *fruit*. Yet there are metaphors in English that build on this meaning, as found in expressions such as *the fruits of his labors*, and *fruitless labors* are then 'labors that did not bear fruit'. Moreover, one finds expressions such as *a fruitless tree*, in which the literal meaning, 'without fruit' is appropriate. For this example, it is debatable whether the meaning of the base is totally irrelevant for the meaning of the derived word. What is clear, however, is that the privative meaning of *-less*, 'without fruit', or 'without success', is still present in the complex word.

The etymological origin of *archer*, 'someone who wields a bow', is Latin *arcus* (bow, arc). It is similar in structure to a denominal formation such as *trucker*, 'someone who drives a truck'. Again, the suffix is transparently present in the complex word, marking it as an agent noun, even if the base is no longer synchronically that clearly visible.

For the adjective *cryptic*, Rastle et al. must have had in mind the free-standing base word *crypt*, 'a vault wholly or partly under ground'. And indeed, the meaning of the adjective *cryptic*, 'hidden, secret, incomprehensible' is unrelated to this meaning of the base. Leaving aside that the meaning of *crypt* goes back to a proto-Indo-European root meaning 'to hide', and that English does make use of a transparent bound root *crypt-* as in *cryptography*, it is clear that the suffix *-ic* is contributing to the meaning of the adjective just as it does in *rhythm-ic* or *Semit-ic*. For *fleeting*, the suffix *-ing* is contributing to the adjectival reading 'transient' just as it does in words such as *daring* or *humbling*.

It should be noted that functional suffixes in words with bases that do not contribute to the meaning of the derived word are sometimes active in the grammar. In Dutch, simple words take a prefix for their past participle ("zitten" - "gezeten", *sit*; "wandelen" - "gewandeld", *walk*). Complex verbs don't take this prefix ("behandelen" - "behandeld", *treat*). Although Dutch does not have a verb "ginnen", the derived word "beginnen" (*begin*) behaves as a complex word by not taking the prefix for its past participle, which is "begonnen" and not "gebegonnen".

Although these examples show that the degree of opacity of the pseudo-complex words is debatable for at least a subset of the items, we have chosen to assign these pseudo-complex words their own meanings, rather than the meanings of their base words. However, where a suffix is synchronically active, as in the examples discussed above, the word is also linked to the suffix meaning. For words such as *ample* and *trolley*, in which there is no synchronic suffix, no suffix meaning was assigned. This coding scheme is probably conservative, as the example of *fruitless* shows.

Table 18: Assignment of meanings to selected words in the opaque, transparent, and form conditions in the study of Rastle et al. (2004).

| Word | Type | Lexical Meaning | Suffix Meaning |
|---|---|---|---|
| *archer* | opaque | archer | er |
| *cryptic* | opaque | cryptic | ic |
| *fruitless* | opaque | fruitless | less |
| *trolley* | opaque | trolley | - |
| *employer* | transparent | employ | er |
| *alcoholic* | transparent | alcohol | ic |
| *cloudless* | transparent | cloud | less |
| *arsenal* | form | arsenal | - |
| *brothel* | form | brothel | - |
| *candidacy* | form | candidacy | - |

For both prime and target, we estimate probabilities analogous to (25),

$$\mathrm{Pid}_{\mathrm{word}} = \frac{w_{\mathrm{affix}} a_{\mathrm{affix}} + a_{\mathrm{word}}}{w_{\mathrm{affix}} a_{\mathrm{affix}} + a_{\mathrm{word}} + \sum_{i=1}^{n} a_i}, \tag{29}$$

where $w_{\mathrm{affix}}$ is a weight for the affixal meanings, and where $n$ represents the number of strongest competitors taken into account.

To model the masked priming results of Rastle et al. (2004), we again make use of the compound cue theory of Ratcliff and McKoon (1988). We allow prime and target to have different weights, by defining the compound cue strength as

$$\begin{aligned} S' &= \mathrm{Pid}_{\mathrm{P}}^{w}\, \mathrm{Pid}_{\mathrm{T}}^{1-w}, \\ \text{Simulated RT} &= \log(1/S'), \end{aligned} \tag{30}$$

with as prime weight $w = 0.05$. The correlation of the simulated and observed latencies was 0.51. Crucially, the magnitude of the priming effects matched those from the empirical study. The transparent and pseudo-derived words had empirical priming effects of 22 and 24 ms that were both significant and did not differ significantly between them. Similarly in the model, priming effects of 0.064 and 0.071 were obtained, that were both highly significant ($t = 13.92$ and 15.59 respectively), and that did not differ (for both treatment coefficients, the standard error was 0.0046).

It is noteworthy that a morpho-orthographic effect is replicated in a model without a morpho-orthographic parsing component. If our model is on the right track, the reason that the transparent and opaque conditions give rise to a similar priming effect is not that a semantically blind orthographic parser separates affix from stem, allowing the parsed-out stem to prime the target. Instead, due to discriminative learning, the orthographic representations for the suffix (unigrams, bigrams) have become associated with the suffix meaning. Crucially, these associations emerge because for the majority of opaque items, the suffix is fully functional in the meaning of the complex word.

Table 19: Coding of the meanings of the items in the simulation of the primed lexical decision experiment of Bergen (2004). Each word pair in the Meaning condition was assigned an arbitrary and unique semantic label.

| CONDITION | PRIME | | | TARGET | | |
|---|---|---|---|---|---|---|
| | input | meaning1 | meaning2 | input | meaning1 | meaning2 |
| Phonaestheme | *glimmer* | gl | glimmer | *gleam* | gl | gleam |
| Baseline | *dial* | - | dial | *ugly* | - | ugly |
| Meaning | *collar* | x1 | collar | *button* | x1 | button |
| Form | *druid* | - | druid | *drip* | - | drip |
| Pseudo-phon. | *bleach* | bl | bleach | *blank* | bl | blank |

Is independent evidence available that morphological units can be fully functional even when there is no obvious semantic contribution from the base? To answer this question, we consider the processing of phonaesthemes.

*Phonaesthemes.* Phonaesthemes are frequently recurring sound-meaning pairings in the absence of a stem. Classic examples from Bloomfield (1933) are word initial *gl* in *glow, glare, gloom, gleam, glimmer* and *glint.* Bergen (2004) observed that 38.7% of the types and 59.8% of all tokens in the Brown corpus beginning with *gl* have dictionary definitions that refer to light or vision. For *sn*, 28% of the word types and 19% of the word tokens have meaning related to 'nose' or 'mouth' (e.g., *sniff, snore, snort, snot, snout, sneeze*).

Bergen studied the processing of phonaesthemes using a primed visual lexical decision with a prime duration of 150 ms and a 300 ms interval between the presentation of prime and target. Stimuli fell into five categories. The set of Phonaesthemes shared a phonological onset and a meaning well supported across a large number of word types and tokens (e.g., *glitter, glow*). Then, in the Form condition words shared an onset, but no meaning (*druid, drip*). In the Meaning condition they shared meaning (*cord, rope*). The set of Pseudo-phonaesthemes comprised words sharing onset and meaning, but in this case the potential phonaestheme was not well-supported distributionally (*crony, crook*). Finally, the Baseline condition included words unrelated in form and meaning (*frill, cook*). Stimuli were matched for frequency, number of letters, number of phonemes, and number of syllables. The words in the Phonaestheme condition elicited significantly shorter latencies than the words in any of the other four conditions, indicating that distributionally well-supported phonaesthemes enjoy a processing advantage of nearly 60 ms compared to the baseline condition.

Using the lists of stimuli listed in Appendix B of Bergen (2004), we calculated the simulated response latencies for his materials. The meanings of the words were coded as shown in Table 19, with phonaesthemes and pseudo-phonaesthemes receiving a second meaning (in addition to the meaning of the whole word) represented simply by the phonaestheme. Words in the Meaning condition were also assigned a second meaning, which varied from pair to pair. Probabilities of identification were defined as

$$p_{\text{word}} = \frac{w_m a_{\text{shared meaning}} + a_{\text{word}}}{w_m a_{\text{shared meaning}} + a_{\text{word}} + \sum_{i=1}^{n} a_i}, \tag{31}$$

with $w_m$ the weight for the shared meaning (the equivalent of the weight for affix meanings

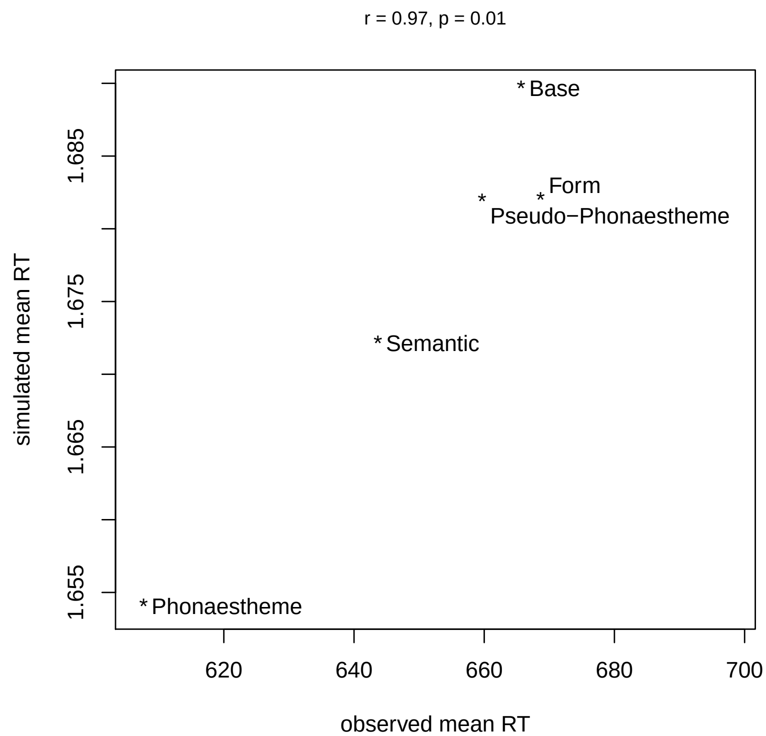


*Figure 12.* Observed and simulated group means for the baseline condition (Base), the form condition (Form), the semantic condition (Semantic), and the Pseudo-phonaestheme and Phonaestheme conditions, using the data of Bergen (2004).

for derived words), and $n$ the number of highest-activated competitors taken into consideration. A good fit requires approximately the parameter values $w_m = 0.01$ and $n = 40$. As for the pseudo-derived words, we made use of the compound cue theory, setting the prime weight to 0.2 (cf. equations 30).

As Bergen (2004) does not provide item means, we calculated the mean simulated latency for each condition. As illustrated in Figure 12, the model captures the main trend in the observed group means: r = 0.97, ($t(3) = 6.95$, $p = 0.0061$). An analysis of covariance of the simulated latencies with word frequency as covariate indicated that the group mean for the phonaesthemes contrasted significantly with the joint group mean of the other four groups ($\hat{\beta} = 0.020, t(48) = -2.339, p = 0.024$). With only 10 observations for each condition, the model was not accurate enough to support the significance of the contrasts of the Phonaesthemes with each of the other four conditions separately.

This simulation study suggests, albeit tentatively, that priming effects for phonaesthemes similar to those found for regular morphemes can emerge within the framework of naive discriminative learning. Morpheme-like effects can be present without an input string requiring a parse into a sequence of morphemes that jointly span the input. The model therefore dovetails well with theories in linguistic morphology which have challenged the explanatory value of the theoretical construct of the morpheme (see, e.g., Matthews, 1974; Aronoff, 1994; S. Anderson, 1992; Stump, 2001; Blevins, 2003, 2006; Booij, 2010).

Although we can label particular forms such as *gl* as phonaesthemes, setting them apart from 'accidental' series where aspect of form and meaning would coincide supposedly by chance, the phenomenon itself is in all likelihood a gradual one. We suspect that it is only for the strongest and semantically most consistent series that morpheme-like effects are detectable in on-line behavioral experiments. Yet a discriminative learning approach predicts that even small local consistencies in the fractionated chaos of local form-meaning correspondences will be reflected in the weights, and that they will codetermine lexical processing, however minute these contributions may be. Interestingly, even in this chaos there seems to be some order, as Shillcock, Kirby, McDonald, and Brew (2001) observed that in English, for the most frequent monosyllabic words, there is a small but significant correlation between the phonological distance and the semantic distance between each pair of words. Words with more similar meanings tend to have more similar forms. Morphology, as the study of the relation between form and meaning in words, can begin to account for these kinds of phenomena only by freeing itself from the chains of the morpheme.

In the next section, we consider compounding, the most productive word formation process in English, and the one that comes closest to syntactic phrase formation.

### *Compounds*

Studies on compound processing (e.g., Pollatsek, Hyönä, & Bertram, 2000; De Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2002; Juhasz, Starr, Inhoff, & Placke, 2003; Kuperman et al., 2008, 2009; Baayen, 2010) have documented a wide range of variables explaining processing latencies, including compound frequency, word length, and both the family size and frequency of the head and modifier constituents. In what follows, we consider 921 compounds for which lexical decision latencies are available in the English Lexicon Project.

Following Baayen (2010), we analyse the response latencies with a generalized additive model (Wood, 2006), with as well-established predictors the positional family size of the modifier (the number of compounds sharing the left constituent as modifier), the frequency of the modifier, the positional family size of the head, the length of the compound, and the frequency of the compound.

A factor specifying whether the head of the compound (e.g., *worm* in *silkworm*) is also used as a modifier in other compounds (e.g., *worm* in *wormwood*) was included, as well as a factor specifying whether a compound is part of the strongly connected component of the compound graph. The strongly connected component of a directed graph is that part of the graph in which any node can be reached from any other node by following the directed links between the nodes. For the present data, being part of the strongly connected component of the directed compound graph implies that it is possible, by following modifier-to-head links, to reach a compound's modifier by starting at its head, as in the cycle *silkworm wormwood woodcock cockhorse horsehair hairoil oilsilk*.

Recall that the head and modifier family sizes count the number of compounds sharing head or modifier. The count of the number of compounds that these compounds share a constituent with, henceforth the secondary family size (Baayen, 2010), was also included as a predictor. (The secondary family size count was orthogonalized with respect to the head and modifier family sizes by taking residuals from a model regressing secondary family size on the head and modifier family sizes.)

These three predictors (whether a constituent is used both as head and as modifier, secondary family size, and being part of the strongly connected component of the compound graph) are of special interest as the original motivation for exploring these measures came from a spreading-activation approach to lexical organization. If meanings are connected in a network, then the properties of that network can be summarized using concepts from graph theory, and the consequences of network organization should then be visible in processing costs as gauged by visual lexical decision and word naming latencies. This is indeed what Baayen (2010) found. If activation spreading in a lexical network is indeed the underlying process, then these measures should not be predictive for the simulated latencies generated by the naive discriminative reader, as it addresses only the mapping from orthographic cues to meanings, and not subsequent semantic processes. Therefore, these measures provide an excellent opportunity for falsifying the naive discriminative learning approach.

As a final predictor we included the amount of information carried by the compound as gauged by means of Shannon's entropy applied to the probability distribution of the compound's constituents:

$$H_{\text{compound}} = -\sum_{i=1}^{2} p_i \log_2 p_i, \tag{32}$$

with $p_i$ the probability of the $i$-th constituent given the compound:

$$p_i = \frac{f_i}{\sum_{j=1}^{2} f_j}. \tag{33}$$

A nonlinear interaction involving head family size, secondary family size, and being part of the strongly connected component was modeled with a tensor product, using generalized additive modeling (GAM). Table 20 lists the coefficients of the linear terms of the resulting GAM model fitted to the observed response latencies of the English Lexicon Project. The regression surfaces for the compounds outside and in the strongly connected component required 7.036 and 6.124 estimated degrees of freedom respectively, and reached significance (both $p < 0.0003$). These regression surfaces are shown in the upper panels of Figure 13. For compounds not in the strongly connected component, longer latencies are found for small head family sizes and large secondary productivity. For compounds in the strongly connected component, head family size is facilitatory, but mainly for secondary productivity values around zero, i.e., for secondary family sizes near the mean of the distribution. In other words, for less probable secondary family sizes, longer latencies are found.

Table 21 lists the linear coefficients obtained when the same generalized additive model specification is used for the simulated latencies, defined as

$$\text{simulated RT} = \log\left(\frac{1}{a_{\text{mod}} + w_h a_{\text{head}}} + \phi \mathrm{I}_{[l>8]}\right), \tag{34}$$

with the expectation that $w_h < 1$ because modifiers tend to be read before heads. A good fit to the data was obtained for $w_h = 0.5$ and $\phi = 3.5$. Since the lengths of compounds were longer than those of the simple, inflected, and derived words, ranging from 6 to 14, the cutoff point for multiple fixations is placed slightly further into the word, and $\phi$ is set at a larger

Table 20: Coefficients for the generalized additive model fitted to the observed response latencies of two-constituent English compounds.

| | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| Intercept | 6.776 | 0.036 | 187.177 | 0.0000 |
| Modifier Family Size | -0.016 | 0.006 | -2.595 | 0.0096 |
| Compound Frequency | -0.042 | 0.003 | -13.526 | 0.0000 |
| Modifier Frequency | -0.008 | 0.003 | -2.524 | 0.0118 |
| Head also used as Modifier | -0.021 | 0.012 | -1.758 | 0.0791 |
| Compound Entropy | -0.061 | 0.013 | -4.611 | 0.0000 |
| Compound Length | 0.017 | 0.003 | 5.196 | 0.0000 |

Table 21: Coefficients for the generalized additive model fitted to the simulated response latencies of two-constituent English compounds.

| | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| Intercept | 2.477 | 0.232 | 10.675 | 0.0000 |
| Modifier Family Size | -0.192 | 0.040 | -4.771 | 0.0000 |
| Compound Frequency | -0.111 | 0.020 | -5.651 | 0.0000 |
| Modifier Frequency | -0.148 | 0.020 | -7.304 | 0.0000 |
| Head also used as Modifier | -0.206 | 0.076 | -2.709 | 0.0069 |
| Compound Entropy | 0.086 | 0.085 | 1.006 | 0.3148 |
| Compound Length | 0.160 | 0.021 | 7.502 | 0.0000 |

value to reflect that more than one additional fixation may have been required. The by-item correlation of observed and simulated latencies was $r = 0.31$, ($t(919) = 9.71, p < 0.0001$). The two tensor products both reached significance (both $p < 0.0001$) for 8.07 and 7.78 estimated degrees of freedom.

A comparison of Table 20 and Table 21 shows that the model correctly predicts facilitatory effects of compound frequency, modifier family size, modifier frequency, and also mirrors the shorter latencies for compounds with heads that are also used as modifiers.

The empirical decision latencies are characterized by a facilitatory effect of compound entropy. The facilitatory effect of compound entropy is consistent with the facilitatory effect of inflectional entropy. When a word is characterized by a higher amount of information, carried by its inflectional paradigm or carried by its constituents as in the case of compounds, there is more information about the word available in long-term memory. Therefore, a higher entropy (more information in memory) predicts shorter response latencies in the lexical decision task.

The model, however, does not capture the facilitation from compound entropy, which suggests to us that the compound entropy effect in the observed latencies reflects a processing stage subsequent to the initial process of activating meanings from orthographic cues.

The magnitude of the effects of compound frequency and modifier frequency are out of balance in the model, which overestimates the effect size of modifier frequency and underestimates the effect size of compound frequency. As in the simulation of derived words, this is due to the model being withheld information about semantic opacity. Nevertheless, even though the model assumes full transparency, whole-word frequency effects do emerge, indicating that semantic opacity is not the only force underlying whole word frequency effects.

The regression surfaces estimated for the simulated latencies are shown in the bottom panels of Figure 13. It is clear that the model does not capture the full details of the interaction of head family size by secondary productivity by membership in the strongly connected component. Nevertheless, there are some encouraging similarities. For compounds outside the strongly connected component (left panels), the model captures the facilitation for large head families, and part of the inhibition for small head families and low secondary productivity. The model fails to capture that inhibition is strongest for small head family size and large secondary productivity. Turning to the compounds in the strongly connected component, the model faithfully replicates the trough that is visible for the observed latencies for zero secondary productivity (the mode of the distribution of secondary productivity values).

The ability of the model to approximate the effects of secondary productivity and membership in the strongly connected component came as a surprise to us. We thought that without further knowledge of the semantic relations between the semantic nodes, and without a mechanism of spreading activation in this network of semantic relations, these effects would not emerge in the naive discriminative reader. Since these effects are nevertheless present in the simulated latencies, it must be the case that the distributional information on the basis of which the weights are estimated is not uniformly distributed with respect to secondary productivity and membership in the strongly connected component. We therefore examined more closely how (dis)similar words are as a function of their membership of the strongly connected component and their secondary productivity.

The Levensthein distance of the modifier to the head in the compounds in the strongly connected component is significantly smaller than the corresponding distance for compounds that are not part of the strongly connected component ($t(663.2)$ = -2.34, $p$ = 0.0194). Furthermore, the mean of the average Levenshtein distances of constituents in the strongly connected component to any other constituent is significantly smaller than the mean of the average distances calculated for constituents outside the strongly connected component ($t(426.22)$ = -1.97, $p = 0.0496$).

Finally, while there is no correlation of this average Levenshtein distance for modifiers and secondary productivity, r = 0, ($t(919)$ = -0.07, $p = 0.9432$), the corresponding correlation for the head is markedly present, r = -0.21, ($t(919)$ = -6.51, $p = 0$), such that greater average Levenshtein distances predict reduced secondary productivity. In other words, the more similar a word is to other words, the greater its secondary productivity is. This pattern of results helps explain why the interaction displayed in Figure 13 pivots around the head and not around the modifier: It is only heads, and not modifiers, that are non-uniformly distributed with respect to their similarity to other constituents.

The non-uniform distribution of form similarity with respect to secondary productivity and membership in the strongly connected component implies a non-uniform distri-

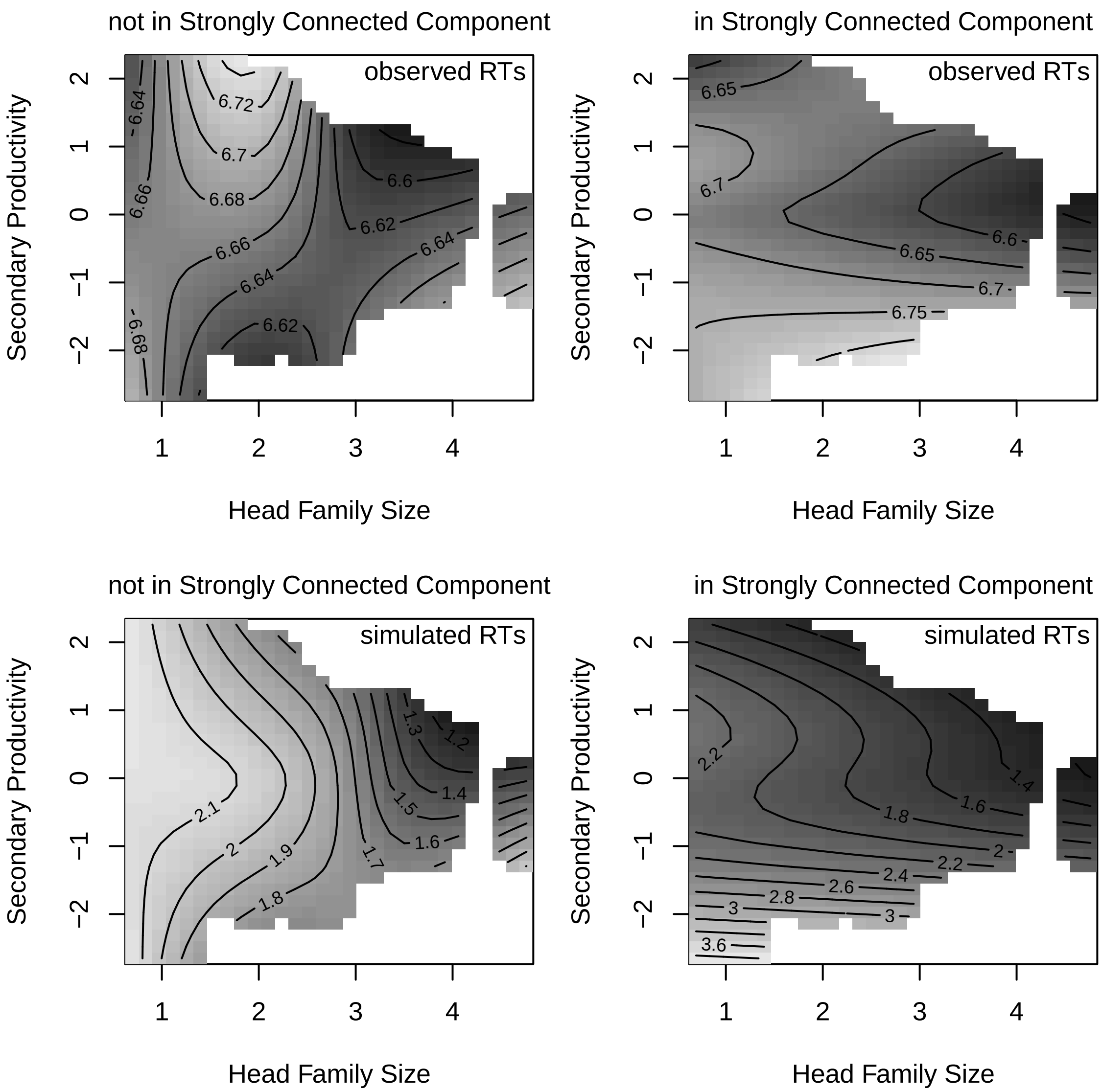


*Figure 13.* Partial regression surfaces (modeled with tensor products) for observed (upper panels) and simulated (lower panels) response latencies, for the interaction of head family size by secondary productivity by membership of the strongly connected component of the compound graph. Fitted observed latencies specified on the contour lines are on the log scale. Fitted simulated latencies are also on a log scale, as defined by (34).

bution of the difficulty of discriminative learning. Words with denser neighborhoods are more difficult to associate with their meanings. As a consequence, the effects of secondary productivity and membership in the strongly connected component may, at least in part, be effects of neighborhood similarity in disguise.

We conclude our discussion of compounds with a comment on the observation, coming from recent eye-tracking studies, that compound frequency effects can be present already at the first fixation (Kuperman et al., 2008, 2009). Following Bruza, Kitto, Nelson, and McEvoy (2009a, 2009b), one could attribute such an early effect as arising due to quantum entanglement.

Nelson, McEvoy, and Pointer (2003), in a study on cued recall, showed that connections among a target word's associates facilitate recall regardless of the number of connections returning from those associates to the target. They proposed an 'activation at a distance' equation that outperformed a spreading activation account. Bruza et al. (2009a) explore the possibility of accounting for such 'spooky distance effects' in terms of quantum entanglement. For compound processing, quantum entanglement might likewise account for the apparently simultaneous activation of the first constituent and the full compound.

However, within the context of naive discriminative learning, the early effect of compound frequency in the eye-movement record follows straightforwardly. To simulate first fixation durations, we assume that only the first constituent of the compound is visible, and that there is sufficient parafoveal information to clarify that the modifier is not followed by a space. Modeling the first fixation duration as proportional to $\log(1/a_{\text{mod}})$, we obtain significant facilitatory effects of modifier frequency, modifier family size, and also compound frequency ($\hat{\beta} = -0.08, t(916) = -3.44, p = 0.0006$). If we assume that the first character of the head is also available, the facilitatory effect of compound frequency increases and remains significant ($\hat{\beta} = -0.17, t(916) = -6.47, p < 0.0001$). In other words, naive discriminative learning obviates the need for an appeal to quantum entanglement. The activation at a distance phenomenon reported by Nelson et al. (2003) may likewise find an alternative explanation in discriminative learning.

### *Phrasal effects*

The frequency effects observed for inflected words, derived words, and compounds were replicated in our simulation studies without any assumptions about processes or representations that would be specifically morphological in nature. The naive discriminative reader therefore predicts frequency effects also for multi-word sequences that are not compounds or other morphological units: The model is trained on sequences of words (see Table 11) and not on isolated words.

A model that does not presuppose a strict distinction between lexicon and syntax fits well with recent linguistic theories rejecting a strict boundary between morphology and syntax (as typically assumed in mainstream generative grammar), and that instead situate morphology and syntax on a continuum with pairings of form and meaning (often referred to as constructions) exhibiting different degrees of complexity (Goldberg, 2006; Jackendoff, 2009; Booij, 2010).

In what follows, we explore two kinds of phrasal effects: a phrasal frequency effect (facilitating phrasal comprehension), and a phrasal exemplar-prototype effect affecting the processing of individual words that is structurally similar to the relative entropy effect discussed above for Serbian nouns.

*Phrasal frequency effects.* Multi-word frequency effects have recently been reported (Bannard & Matthews, 2008; Arnon & Snider, 2010; Tremblay & Baayen, 2010), even for n-grams that are fully transparent and fully compositional. Within the present framework of discriminative learning, conditional on learning not being restricted to words in isolation, such phrasal frequency effects should emerge.

To test this prediction, we selected from the 11,000 prepositional phrases in the model's lexical input, 558 phrasal pairs such as *in a flour* (low-frequency, 2 occurrences in the BNC) and *in the flour* (high-frequency, 37 occurrences in the BNC), comprising 133 different nouns and 39 different prepositions. For each pair, one phrase had a high frequency, and the other a low frequency. For each phrase, a comprehension latency was simulated on the basis of the (unweighed) activations of the three constituents:

$$\text{Simulated Latency} = \log\left(\frac{1}{a_{\text{noun}} + a_{\text{preposition}} + a_{\text{determiner}}}\right). \tag{35}$$

The pairwise differences in the simulated latencies and the corresponding pairwise differences in trigram frequencies were significantly correlated r = -0.17, ($t(556)$ = -4.07, $p = 1e-04$). The facilitation from frequency was confirmed by a mixed-effects regression model fitted to the simulated latencies with as predictors the log-transformed frequency of the trigram in the BNC, the identity of the determiner (*a* or *the* as fixed-effect factor, and preposition and noun as random-effect factors. In addition to random intercepts, random slopes for frequency were supported by likelihood ratio tests for both random effect factors (all $p < 0.0001$). A third potential random-effect factor, the combination of preposition and noun, did not explain any variance and was removed from the model specification. The identity of the determiner and trigram frequency emerged with independent main effects, with phrases with *the* and higher frequency phrases having shorter simulated latencies (frequency: $\hat{\beta} = -0.01, t = -3.169$).

Crucially, an effect of phrasal frequency is predicted by our model without there being explicit representations for prepositional phrases in a model that is fully compositional and extremely economical in the number of semantic representations that it admits. What this simulation study shows is that the benefits of experience for compositional phrases, as attested recently in the behavioral study of Arnon and Snider (2010) and the electrophysiological study of Tremblay and Baayen (2010), may be understood without postulating that phrases are somehow "stored", which would lead to a combinatorial explosion of "supra-lexical" representations. Further evidence for this possibility is reported in Baayen and Hendrix (2011), who successfully simulated a phrasal frequency effect for the four-word materials used in Experiment 1 of Arnon and Snider (2010).

*Phrasal paradigmatic effects on single-word lexical processing.* In the framework of discriminative learning, morphological family size and inflectional paradigmatic effects do not arise due to co-activation of morphologically related words. Instead, experience with morphologically related words makes it possible for their base words to be learned better, resulting in stronger connections from form to meaning. If a strict division between morphology and syntax is abandoned, then experience with words in phrasal rather than morphological contexts should also affect learning. More specifically, just as a Serbian noun

Table 22: Phrase frequency and probability, and prepositional frequency and probability, for 7 prepositions in indefinite prepositional phrases with *plant*. For phrasal probabilities, we backed off from zero by adding one to the phrasal frequencies. The relative entropy for this example is 0.143.

| phrase | phrasal frequency | phrasal probability | preposition | prepositional frequency | prepositional probability |
|---|---|---|---|---|---|
| *on a plant* | 28608 | 0.279 | *on* | 177908042 | 0.372 |
| *in a plant* | 52579 | 0.513 | *in* | 253850053 | 0.531 |
| *under a plant* | 7346 | 0.072 | *under* | 10746880 | 0.022 |
| *above a plant* | 0 | 0.000 | *above* | 2517797 | 0.005 |
| *through a plant* | 0 | 0.000 | *through* | 3632886 | 0.008 |
| *behind a plant* | 760 | 0.007 | *behind* | 3979162 | 0.008 |
| *into a plant* | 13289 | 0.130 | *into* | 25279478 | 0.053 |

incurs a processing cost if its use of case inflections is different from the prototypical use of case inflections in its inflectional class, we may expect that nouns in English will occur a processing cost if their phrasal use is atypical.

To test this prediction, we considered simple prepositional phrases in English, consisting of a preposition, a determiner, and a noun. The prepositions were taken from the set *above, across, against, along, amid, amidst, among, amongst, around, at, atop, before, behind, below, beneath, beside, besides, between, beyond, following, from, in, inside, into, near, next, off, on, onto, outside, over, past, round, through, to, toward, towards, under, underneath, up, upon, with, within*, and *without*, the determiners were *a* and *the*, and the noun was selected from the 1452 simple words that have a nominal reading (and possibly a verbal reading) and for which response latencies are available in the English Lexicon Project.

Using the Google 1T n-gram data (Brants & Franz, 2006), we compiled a data set of 38577 trigrams with the definite article, and a data set of 14851 trigrams with the indefinite article. For both data sets, we extracted the Google 1T 3-gram frequency, from which we also calculated the frequencies of the prepositions summed across all the trigrams in which they occurred. As illustrated for a sample of 3-grams in Table 22, the n-gram frequencies and prepositional frequencies were transformed into probabilities, which served as input for the calculation of relative entropies to which we will refer as prepositional relative entropies.

The prepositional relative entropies for indefinite and definite phrases were highly correlated ($r = 0.659$), and both were predictive for the response latencies to simple nouns. In the models for the observed and simulated latencies of simple nouns presented above (see Tables 12 and 13), we used the indefinite prepositional relative entropy, which seemed slightly more robust, possibly because lexical decisions for words presented in isolation are elicited for words in an indefinite context.

The predictivity of prepositional relative entropy for isolated word reading in English provides further support for our hypothesis that the processing of words is co-determined not only by the morphological contexts in which that word occurs, but also by its syntactic contexts. Crucially, it is not simply the frequency of such syntactic contexts, but also how such syntactic contexts are structured paradigmatically. The unconditional probabilities with which prepositions are used represent a noun's prototypical prepositional usage. The nouns' own probabilities of occurrence with these prepositions represent exemplar profiles. The prepositional relative entropy captures the distance between an exemplar and its prototype.

It is possible that exemplars have their own representation, and that in lexical processing the distance of that representation to the prototypical representation is somehow taken into account. However, explanations positing exemplar representations place high demands on memory capacity. The naive discriminative learning framework, in which relative entropy effects emerge naturally, by contrast, imposes very limited demands on memory, and also does not require a separate process evaluating an exemplar's distance to the prototype.

It is important to realize that prepositional paradigms capture only one paradigmatic aspect of phrasal syntax. For instance, our theory predicts that the prepositions used in verbal adjuncts constitute a second paradigmatic domain for which a relative entropy can be defined. This relative entropy should correlate positively with response latencies to verbs. Phenomena typically explored with collostructional analysis (Stefanowitsch & Gries, 2003; Gries & Stefanowitsch, 2004) may similarly constitute dimensions of paradigmatic variation affecting lexical processing.

## Discussion

The experimental data on the reading of Serbian case-inflected nouns reported in the present study, combined with the data previously obtained by Milin, Filipović Đurđević, and Moscoso del Prado Martín (2009), indicate that the processing of a word form is co-determined by the probabilities of all inflectional variants of this particular word, and the probabilities of the exponents of inflectional class to which a given word belongs. Similarly, for English, we have shown that the processing latencies of simple nouns are co-determined by the probabilities with which these nouns co-occur with prepositions vis-a-vis the unconditional probabilities of these prepositions. These experimental results fit well with previous data documenting the importance of paradigmatic structure for lexical processing, as witnessed by the effects of inflectional entropy and morphological family size.

We have shown that a naive discriminative learning architecture suffices to capture these paradigmatic effects for both morphological and phrasal processing. Although the good fit to the Serbian data initially obtained with the naive discriminative reader could have been due to the restricted data set on which the model was trained, the subsequent good fits obtained for the English data, based on a broad and general instance base extracted from the BNC, indicates that overfitting is not at issue. We have also shown that the naive discriminative reader is able to account for a wide range of phenomena, from morphological effects to pseudo-prefixed words, and from phonaesthemes to phrasal frequency effects.

The success of the naive discriminative reader raises the question of whether other models might be equally succesful. In what follows, we therefore compare the naive discriminative reader in some detail with the Bayesian Reader of Norris (2006), and briefly discuss other models of word recognition.

In the Bayesian Reader model for word recognition the probability of identifying a word $w_i$ given input $I$ is defined as

$$\Pr(w_i|I) = \frac{\Pr(w_i)\Pr(I|w_i)}{\sum_{i=j}^{m}\Pr(w_j)\Pr(I|w_j)}. \tag{36}$$

The likelihood function $\Pr(I|w_i)$ is defined stochastically as a function of time and the Euclidian distance of $w_i$ to the input as available at a given point in time. We skip the details of the stochastic modeling of the time course of lexical activation, which renders the model computationally extremely demanding. Instead, we discuss a simplified version that we have implemented, henceforth the Easy Bayesian Reader.

Recall that for the naive discriminative reader, a word's orthographic input was coded as the set of its unigrams and bigrams. For the Easy Bayesian reader, we encoded a word's form as a binary vector indicating which of the $27 + 27^2 = 756$ unigrams and bigrams was instantiated for that word. For a prime-target pair, following Norris and Kinoshita (2008) that in masked priming prime and target are blurred into one percept, the input was encoded as binary vector representing both the prime and the target simultaneously. the likelihood $\Pr(I|w_i)$ was assessed as the Euclidean distance of the binary orthographic vectors of the visual input $I$ and word $w_i$, normed to the interval $[0, 1]$. The probability $\Pr(w_i)$ was estimated by its relative frequency in our corpus. Finally, the response latency for $w_i$ was defined as $1/\Pr(w_i|I)$, and log-transformed to obtain an approximately normal response variable.

Table 23: Coefficients estimated for the simulated self-paced reading latencies using the Easy Bayesian Reader model.

| | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Length | 0.046 | 0.017 | 2.761 | 0.0058 |
| Weighted Relative Entropy | 0.224 | 0.066 | 3.379 | 0.0008 |
| Target Gender = masculine | -0.227 | 0.053 | -4.300 | 0.0000 |
| Normalized Levenshtein Distance | 0.236 | 0.145 | 1.629 | 0.1036 |
| Target Lemma Frequency | -0.999 | 0.026 | -37.928 | 0.0000 |
| Target Case = Nominative | -0.854 | 0.054 | -15.746 | 0.0000 |
| Prime Word Frequency | -0.079 | 0.019 | -4.134 | 0.0000 |
| Prime Condition = DSSD | -0.076 | 0.082 | -0.932 | 0.3515 |
| Priming Condition = SS | -0.270 | 0.157 | -1.715 | 0.0865 |

We investigated how well the predictions of the Easy Bayesian Reader fitted the primed self-paced reading latencies of our Serbian case-inflected nouns (Experiment 1). The simulated latencies entered into a significant correlation with the observed by-item mean self-paced reading latencies, r = 0.15, ($t(1185) = 5.16$, $p = 0$), albeit to a lesser extent than the latencies simulated using naive discriminative learning ($r = 0.23$). The model does not capture the interaction of Weighted Relative Entropy by Case nor the interaction of Weighted Relative Entropy by Gender. After removal of these interactions from the model specification, the model summarized in Table 23 was obtained. The model correctly predicts an inhibitory effect of Weighted Relative Entropy. It also correctly predicts inhibition from Word Length, and shorter latencies for nouns in nominative case. The effect size for nominative case, however, is four times as large as the effect of the identity priming condition, instead of being an order of magnitude smaller (compare Table 5 and Figure 1). Apparently, the Easy Bayesian Reader captures important aspects of the data, but with reduced precision. Since the way we coded the orthographic input differs from the implementation in the original Bayesian Reader model, it is possible that the original full model will provide enhanced results.

Assuming that this is indeed the case, three important differences between the two approaches should be noted. First, the Bayesian Reader compares the orthographic input with orthographic representations in memory. In our implementation of the Easy Bayesian Reader, it is assumed that Serbian words inflected for case and number have such orthographic representations. In other words, the Easy Bayesian Reader is a full-form based model. By contrast, the naive discriminative reader is a full-decomposition model in which inflected forms do not have their own representations in memory, and in which orthographic form information is directly mapped onto meaning representations.

Second, the growth-rate of the Easy Bayesian Reader is quadratic in the number of entries $N$ in the lexicon. Simulating a response latency for each of the 1776 distinct Serbian case-inflected wordforms requires the calculation of $1776^2 = 3154176$ distances. For the discriminative learning model, we have $27 + 27^2 = 756$ orthographic representations for unigrams and bigrams (including the space character), and 278 semantic representations (270 noun lemmas, 6 cases, and 2 numbers), in all $M = 278 + 756 = 1034$ representations. The number of weights in the model is $756 * 278 = 210168$. Since each additional meaning node requires only 756 additional weights, the growth-rate of the discriminative learning model is linear. Even for the small data set of Serbian nouns, the number of distances the Easy Bayesian Reader has to compute is already 15 times the number of weights that need to be set in the discriminative learning model.

Third, the Bayesian Reader simulates the time course of lexical activation, the Easy Bayesian Reader and our discriminative learning model do not. A time course for the activation of a word $w_i$ can in principle be generated by using the probability $\Pr(w_i|I)$ to estimate the word's drift rate in a lexical diffusion model (Ratcliff, Gomez, & McKoon, 2004).

This example comparison illustrates the dilemma facing computational models of simple word recognition that build on a lexicon of representations for simple words, not only the Bayesian Reader, but also models such as ACT-R (see, e.g., Van Rijn & Anderson, 2003) and DRC (Coltheart et al., 2001). For the modeling of morphological and phrasal effects, this family of models has two options.

A first option is to add complex words to the lexicon, as if they were simple words. For the small data set of Serbian case-inflected nouns, the results obtained with the Easy Bayesian Reader suggest this may work in principle. For realistic lexicons, the price of a lexicon with huge numbers of entries may become prohibitive. For instance, the number of n-gram types on which our model was trained, 1,496,103, represents only a fraction of the number of n-grams occurring in the BNC alone. Yet no fewer than 2,238,324,000,000 distances would have to be evaluated to estimate the posterior probabilities of just these phrases in the Bayesian Reader approach.

A second option is to restrict the lexicon to monomorphemic words, and to supplement current models with a probabilistic parser. However, for such a parser to work, a morpheme-based theory of morphology would have to be assumed, which, apart from being linguistically unattractive, makes the wrong predictions for phonaesthemes and pseudo-derived words. Furthermore, it is unclear to us how and why such a parser would give rise to paradigmatic entropy effects in lexical processing.

The naive discriminative learning approach that we have pursued is similar to the triangle model (Harm & Seidenberg, 1999; Seidenberg & Gonnerman, 2000; Harm & Seidenberg, 2004) in that the orthographic input is mapped onto meaning without intervening lexical representations and without requiring explicit rules for parsing. It differs from the triangle model in several ways, however. First, we have not made any attempt to model phonology. Hence, our model is more limited and does not provide accurate predictions for word naming and reading aloud. Given the neurophysiological evidence for two cortical streams in reading (a ventral, occipital-temporal, stream used when accessing familiar words encoded in lexical memory, and a dorsal, occipital-parietal-frontal, stream used when mapping sublexical spelling onto sounds, see, e.g., Borowsky et al., 2006), we believe it is worth exploring whether our model could function as part of the lexical (ventral) route in, for instance, the DRC architecture (see Hendrix & Baayen, 2010, for a proposal).

A second difference with the triangle model is that we have substantially simplified the computational engine, which does not incorporate hidden layers and does not use backpropagation for estimating connection weights. All we need for modeling morphological effects is a (symbolic) layer of orthographic nodes (unigrams and bigrams) and a (symbolic) layer of meanings. This offers the advantages of simplicity and interpretability: the activation of a meaning is the model's discriminative learning estimate of the posterior probability of that meaning given its unigrams and bigrams and the co-occurrence probabilities of these unigrams, bigrams, and meanings.

A disadvantage of the present model is that it is blind to the semantic relations between words. The connectionist model presented in chapter 10 of Moscoso del Prado Martín (2003), in which orthographic input units map, via a hidden layer, onto independently established corpus-based semantic vector representations of word meanings, offers the advantage of better modeling the role of semantic similarity in word processing. Thus, the effect of the cosine distance in semantic space between prime and target, that reached significance as predictor for the self-paced reading latencies of Serbian case-inflected words in Experiment 1, is not captured by our model.

Finally, we note that the naive discriminative reader is compatible with theories assigning hierarchical structures to complex words. For instance, for *rethinking*, a structure such as [[REPEAT[THINK+CONTINUOUS]]] specifies scope relations that are part of the meaning of this word. All that the naive discriminative reader does is assign probabilities to the meanings REPEAT, THINK, and CONTINUOUS. Therefore, the current implementation is consistent with the possibility that semantic rules build such hierarchical structures on the basis of these meanings. Crucially, the present simulation results indicate that for explaining the consequences of morphological structure as gauged by the lexical decision task, it is not necessary to duplicate such hierarchical structure at a morphemic level with structures such as [[*re*[*think*+*ing*]]]. Given discriminative learning, such morphemic structures are completely redundant.

This also absolves the modeler from thorny implementational problems such as how to represent allomorphic variants. By way of example, consider the Dutch diminutive suffix, which appears in five different forms: *je* (*muis-je*, 'small mouse'), *pje* (*bloem-pje*, 'small flower'), *etje* (*wang-etje*, 'small cheeck'), *kje* (*woning-kje*, 'small house'), *tje* (*bever-tje*, 'small beaver'). Models with morpho-orthographic morphemic representations have to posit five different orthographic morphemes for the diminutive, they need some competition mechanism between these (highly similar) allomorphs, as well as a disambiguation mechanism distinguishing the allomorph *je* from the personal pronoun *je* ('you'). These kinds of complications do not arise for the naive discriminative reader.

## Concluding remarks

We have shown that basic principles of discriminative learning applied to the mapping of form to meaning suffice to explain a wide range of phenomena documented for the processing of complex words and n-grams in reading. The naive discriminative reader model is parsimonious in its parameters. The basic engine estimating the connection weights of the Rescorla-Wagner network is parameter-free. We introduced one parameter that allows the weight of syntactic adjustments (affixal meanings) to be less than the weight of lexical meanings. We also made use of two further parameters for modeling the influence of the highest-activated competitors. Longer words often require more than one fixation. As the current implementation of the naive discriminative reader is blind to how the eye moves through longer words, a parameter was invested in accounting for the costs of planning and executing additional saccades. Finally, for the modeling of priming, we needed one additional parameter, the weight for the relative importance of the prime using the compound cue theory of Ratcliff and McKoon (1988). The naive discriminative reader is also sparse in the number of representations required: at the orthographic level, letter unigrams and bigrams, and at the semantic level, meaning representations for simple words, inflectional meanings such as case and number, and the meanings of derivational affixes. As a consequence, the number of connections required is a linear function of meaning representations.

The model contrasts with the many unimplemented verbal models proposed for morphological processing. According to the supralexical model of Giraudo and Grainger (2001), whole-word representations would mediate access to constituents. According to the obligatory decomposition model of Taft (2004), constituents would mediate access to whole-word representations. The parallel dual route models of Frauenfelder and Schreuder (1992); Schreuder and Baayen (1995); Baayen et al. (1997) allow whole-word and constituent access representations to race for word recognition. Computational implementations correctly replicating paradigmatic effects, as gauged by family size and entropy measures, are not available. We doubt that insightful computational implementations of such models can ever be made to work, given the subtlety of, e.g., the prepositional entropy effect in English, which is only one of the many paradigmatic dimensions that we suspect co-determine single-word reading.

Although our model can be viewed as a simplified connectionist model, it can also be viewed as a symbolic Bayesian model specifying, for a given orthographic input, a distribution of probabilities over the meaning representations. In other words, the naive discriminative reader is as a statistical classifier grounded in basic principles of human learning. Baayen (2011) shows, for a binary classification task, that the naive discriminative reader performs with a classification accuracy comparable to state-of-the-art classifiers such as generalized linear mixed models and support vector machines.

We note here that the naive discriminative reader is compatible with the results of Bowers, Davis, and Hanley (2005) (for a replication in visual lexical decision, see Baayen et al., 2007), which suggest that the meanings of partially matching words become accessible irrespective of whether they are legitimate morphological constituents.

Although the naive discriminative reader does not incorporate explicit parsing rules, it is sensitive to the different degrees of productivity of derivational suffixes, and therefore fits well with 'a-morphous' theories of morphology (S. Anderson, 1992). The Rescorla-Wagner engine of the model can be viewed as a formal, computational implementation of the notion of analogy in word and paradigm morphology (Matthews, 1974; Blevins, 2003), a tantalizing notion in linguistics that remains frustratingly vague without computational implementation.

The inductive modeling approach that we have pursued in this work contrasts with deductive styles of modeling, in which processes (literal brain processes or metaphorical cognitive processes) are posited, from which processing consequences are derived and then tested against observed data. The advantage of the deductive style is that observed effects can be related to and understood in terms of the processes originally posited and implemented in the model. The present inductive approach shares with the deductive approach that at the start a cognitive process is posited, in our case, discriminative learning as formalized in the Rescorla-Wagner equations. However, we find it extremely difficult to derive predictions for the consequences of discriminative learning for the adult system, as formalized by the equilibrium equations of Danks (2003), when the weights are set on the basis of realistic language input. This is why we have adopted an inductive approach in which simulated processing costs generated from the combination of real data and a cognitive learning principle are pitched against an array of empirical results. The advantage is precision and model simplicity, the disadvantage is 'explanatory disappointment' — results now follow from the data and a simple learning principle, rather than from more intuitively accessible higher-order explanatory principles. Nevertheless, we think it is worth considering that the simpler explanation may be on the right track.

In a recent review article, Evans and Levinson (2009) argued that there are no language universals, and that we are the only species with a communication system that is fundamentally variable at all levels of structure, across time, and across space. One of the central questions for the cognition of language that they put forward is whether the very different language systems of the world can be acquired by the same general learning strategies (p. 447). It is our hope that naive discriminative learning provides a step forwards as a powerful, flexible, computationally implementable, and computationally efficient learning algorithm.

Of course, many questions and challenges remain to be addressed. For instance, staying within the domain of morphology, it is currently unknown whether naive discriminative learning can predict the specific processing effects documented for the non-concatenative morphological systems of Arabic and Hebrew (Deutsch, Frost, & Forster, 1998; Boudelaa & Marslen-Wilson, 2001). For languages with reduplication, we anticipate that higher-order n-gram orthographic representations will be essential, as well as more sophisticated positional encoding. Another open issue is whether the present approach generalizes to different writing systems such as Chinese and Japanese. Furthermore, the current level of simplicity achieved for English lexical decision cannot be maintained for reading aloud, for which a dual route extension based on the same principles of discriminative learning is required (and sufficient) to obtain accurate predictions for word naming latencies (Hendrix & Baayen, 2010). For multiple-fixation reading, as well as for auditory comprehension, even more complex architectures will be required. Furthermore, we expect a complete comprehension model to require a hierarchy of discriminative learning systems. Finally, even for responses in visual lexical decision, the naive discriminative reader provides a high-level characterization of contextual learning that at the level of cortical learning may be more adequately modeled by hierarchical temporal memory systems (Hawkins & Blakeslee, 2004; Numenta, 2010).

However, for understanding single word reading as gauged by the lexical decision task, the naive discriminative reader provides a computational model that is as simple and economical as possible, while providing good fits to the empirical data. When dealing with the intricacies of language as a complex dynamic system, and when probing the possible role of context-sensitive, discriminative learning, there is no harm in starting small.

## References


Anderson, J. R. (2000). *Learning and memory: An integrated approach.* New York: Wiley.

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

Arnon, I., & Snider, N. (2010). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Journal of Memory and Language*, *62*, 67–82.

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

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. E. Booij & J. van Marle (Eds.), *Yearbook of morphology 1991* (pp. 109–149). Dordrecht: Kluwer Academic Publishers.

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

Baayen, R. H. (2001). *Word frequency distributions.* Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (2005). Data mining at the intersection of psychology and linguistics. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 69–83). Hillsdale, New Jersey: Erlbaum.

Baayen, R. H. (2008). Morphological productivity. In M. Kytö & A. Lüdeling (Eds.), *Handbook of corpus linguistics* (pp. 899–919). Berlin: Mouton de Gruyter.

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Luedeling & M. Kyto (Eds.), *Corpus Linguistics. An International Handbook* (pp. 900–919). Berlin: Mouton De Gruyter.

Baayen, R. H. (2010). The directed compound graph of English. an exploration of lexical connectivity and its processing consequences. In S. Olsen (Ed.), *New impulses in word-formation (Linguistische Berichte Sonderheft 17)* (pp. 383–402). Hamburg: Buske.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, submitted.

Baayen, R. H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, *36*, 94–117.

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

Baayen, R. H., & Hendrix, P. (2011). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *Empirically examining parsimony and redundancy in usage-based models, LSA workshop, January 2011*.

Baayen, R. H., Levelt, W., Schreuder, R., & Ernestus, M. (2008). Paradigmatic structure in speech production. *Proceedings Chicago Linguistics Society 43*, *1*, 1–29.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (cd-rom).* University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Baayen, R. H., & Renouf, A. (1996). Chronicling The Times: Productive Lexical Innovations in an English Newspaper. *Language*, *72*, 69-96.

Baayen, R. H., Wurm, L., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words. a regression study across tasks and modalities. *The Mental Lexicon*, *2*, 419–463.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology:General*, *133*, 283–316.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.

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

Bates, D. (2006). *Linear mixed model implementation in lme4.* Available from `http://spider.stat.umn.edu/R/library/lme4/doc/Implementation.pdf` (Department of Statistics, University of Wisconsin-Madison)

Bauer, L. (2001). *Morphological productivity.* Cambridge: Cambridge University Press.

Beard, R. (1977). On the extent and nature of irregularity in the lexicon. *Lingua*, *42*, 305-341.

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

Belis, M., & Guiasu, S. (1968). A quantitative-qualitative measure of information in cybernetics system. *IEEE Transactions on Information Theory*, *14*, 593–594.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics. Identifying influential data and sources of collinearity.* New York: Wiley.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, *80*, 290–311.

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, *42*, 390-405.

Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 419-511.

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

Blevins, J. P. (2006). English inflection and derivation. In B. Aarts & A. M. McMahon (Eds.), *Handbook of english linguistics* (p. 507-536). London: Blackwell.

Bloomfield, L. (1933). *Language.* London: Allen and Unwin.

Booij, G. E. (2005). Compounding and derivation: evidence for construction morphology. In W. U. Dressler, D. Kastovsky, O. E. Pfeiffer, & F. Rainer (Eds.), *Morphology and its demarcations* (pp. 109–132). Amsterdam: Benjamins.

Booij, G. E. (2009). Compounding and construction morphology. In R. Lieber & P. Štekauer (Eds.), *The oxford handbook of compounding* (pp. 201–216). Oxford: Oxford University Press.

Booij, G. E. (2010). *Construction Morphology.* Oxford: Oxford University Press.

Borowsky, R., Cummine, J., Owen, W., Friesen, C., Shih, F., & Sarty, G. (2006). FMRI of ventral and dorsal processing streams in basic reading processes: insular sensitivity to phonology. *Brain topography*, *18*(4), 233–239.

Boudelaa, S., & Marslen-Wilson, W. D. (2001). Morphological units in the Arabic mental lexicon. *Cognition*, *81*(1), 65-92.

Bowers, J., Davis, C., & Hanley, D. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory and Language*, *52*, 131–143.

Brandon, S. E., Vogel, E. H., & Wagner, A. R. (2003). Stimulus representation in sop: I: Theoretical rationalization and some implications. *Behavioural Processes*, *62*(1), 5–25.

Brants, T., & Franz, A. (2006). *Web 1t 5-gram version 1.* Philadelphia: Linguistic Data Consortium.

Bruza, P., Kitto, K., Nelson, D., & McEvoy, C. (2009a). Extracting Spooky-activation-at-a-distance from Considerations of Entanglement. *Quantum Interaction*, 71–83.

Bruza, P., Kitto, K., Nelson, D., & McEvoy, C. (2009b). Is there something quantum-like about the human mental lexicon? *Journal of Mathematical Psychology*, *53*, 362–377.

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

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Science*, *10*(7), 287–291.

Christianson, K., Johnson, R., & Rayner, K. (2005). Letter transpositions within and across morphemes. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(6), 1327-1339.

Clair, M. C. S., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, *33*(7), 1317–1329.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–258.

Dabrowska, E. (2009). Words as constructions. *New Directions in Cognitive Linguistics*, 201–224.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121.

Daw, N., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, *26*(5), 593–620.

Dayan, P., & Kakade, S. (2001). Explaining away in weight space. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 451–457). Cambridge, MA: MIT Press.

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.

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2003). Morphological resonance in the mental lexicon. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 65–88). Berlin: Mouton de Gruyter.

Deutsch, A., Frost, R., & Forster, K. I. (1998). Verbs and nouns are organized and accessed differently in the mental lexicon: Evidence from Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 1238-1255.

Dijkstra, T., Moscoso del Prado Martín, F., Schulpen, B., Schreuder, R., & Baayen, R. H. (2005). A roommate in cream: Morphological family size effects on interlingual homograph recognition. *Language and Cognitive Processes*, *20*, 7–41.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24.

Erelt, M. (Ed.). (2003). *Estonian language.* Tallinn: Estonian academy publishers.

Evans, N., & Levinson, S. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448.

Feldman, L. B. (2000). Are morphological effects distinguishable from the effects of shared meaning and shared form? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1431-1444.

Feldman, L. B., O'Connor, P. A., & Moscoso del Prado Martín, F. (2009). Early morphological processing is morpho-semantic and not simply morpho-orthographic: evidence from the masked priming paradigm. *Psychonomic Bulletin & Review*, *16*(4), 684–691.

Fellbaum, C. e. (1998). *WordNet: An electronic database*. Cambridge, MA: The MIT Press.

Filipović Đurđević, D., Đurđević, Đ., & Kostić, A. (2008). Vector based semantic analysis reveals absence of competition among related senses. *Psihologija*, *42*(1), 95–106.

Frauenfelder, U. H., & Schreuder, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In G. E. Booij & J. v. Marle (Eds.), *Yearbook of morphology 1991* (p. 165-183). Dordrecht: Kluwer Academic Publishers.

Gallistel, C. (2003). Conditioning from an information perspective. *Behavioural Processes*, *62*, 89–101.

Gallistel, C., & Gibbon, J. (2002). *The Symboloc Foundations of Conditioned Behavior*. Mahwah, New Yersey: Lawrence Erlbaum Associates.

Giraudo, H., & Grainger, J. (2001). Priming complex words: Evidence for supralexical representation of morphology. *Psychonomic Bulletin and Review*, *8*, 127–131.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology:General*, *117*(3), 227–247.

Goldberg, A. (2006). *Constructions at work. the nature of generalization in language*. Oxford: Oxford University Press.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*, 237-264.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*, 518-565.

Gries, S. (2004). Shouldn't it be breakfunch? A quantitative analysis of blend structure in English. *Linguistics*, *42*(3), 639–667.

Gries, S. (2006). Cognitive determinants of subtractive word-formation processes: a corpus-based perspective. *Cognitive Linguistics*, *17*(4), 535–558.

Gries, S., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective onalternations. *International Journal of Corpus Linguistics*, *9*(1), 97–129.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491-528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.

Hawkins, J., & Blakeslee, S. (2004). *On intelligence*. New York: Henry Holt and Company.

Hay, J. B. (2003). *Causes and Consequences of Word Structure*. New York and London: Routledge.

Hendrix, P., & Baayen, R. H. (2010). The Naive Discriminative Reader: a dual route model of reading aloud using naive discriminative learning. *Manuscript, University of Alberta*.

Hockett, C. (1987). *Refurbishing our foundations: Elementary linguistics from an advanced point of view*. J. Benjamins.

Hsu, A. S., Chater, N., & Vitányi, P. (2010). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Manuscript submitted for publication*.

Jackendoff, R. (2009). Compounding in the parallel architecture and conceptual semantics. In P. Lieber R. Stekauer (Ed.), *The Oxford handbook of compounding* (p. 105-128). Oxford: Oxford university press.

Juhasz, B., Starr, M., Inhoff, A., & Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from lexical decision, naming, and eye fixations. *British Journal of Psychology*, *94*, 223–244.

Juola, J., Ward, N., & McNamara, T. (1982). Visual search and reading rapid serial presentations of letter strings, words, and text. *Journal of Experimental Psychology: General*, *111*, 208–227.

Jurafsky, D., & Martin, J. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, NJ: Prentica Hall.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228-238.

Kostić, Đ. (1965). Sintaktičke funkcije padežih oblika u srpskohrvatskom jeziku *(Syntactic functions of cases in Serbo-Croatian).* Institute for Experimental Phonetics and Speech Pathology, Belgrade, Serbia.

Kostić, Đ. (1999). Frekvencijski rečnik savremenog srpskog jezika *(Frequency Dictionary of Contemporary Serbian Language).* Institute for Experimental Phonetics and Speech Pathology & Laboratory of Experimental Psychology, University of Belgrade, Serbia <http://www.serbian-corpus.edu.yu/>.

Kostić, A., & Katz, L. (1987). Processing differences between nouns, adjectives, and verbs. *Psychological Research*, *49*, 229–236.

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

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, *23*, 1089–1132.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP*, *35*, 876–895.

Landauer, T., & 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.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and Control Theory*, *10*(8), 707–710.

Libben, G., Gibson, M., Yoon, Y., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, *84*, 50–64.

Lieber, R. (1992). *Deconstructing morphology: Word formation in syntactic theory.* Chicago: University of Chicago Press.

Lukatela, G., Gligorijević, B., Kostić, A., & Turvey, M. T. (1980). Representation of inflected nouns in the internal lexicon. *Memory and Cognition*, *8*, 415-423.

Lukatela, G., Mandić, Z., Gligorijević, B., Kostić, A., Savić, M., & Turvey, M. T. (1978). Lexical decision for inflected nouns. *Language and Speech*, *21*, 166–173.

Lund, K., & Burgess, C. (1996a). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments, and Computers*, *28*(2), 203-208.

Lund, K., & Burgess, C. (1996b). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments and Computers*, *28*(2), 203–208.

Marslen-Wilson, W. D., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, *101*, 3-33.

Matthews, P. H. (1974). *Morphology. an introduction to the theory of word structure.* London: Cambridge University Press.

Moscoso del Prado Martín, F. (2003). *Paradigmatic effects in morphological processing: Computational and cross-linguistic experimental studies.* Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics.

Moscoso del Prado Martín, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H., & Baayen, R. H. (2005). Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch. *Journal of Memory and Language*, *53*, 496–512.

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

McDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context judgements of semantic similarity. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 611–616). Edinburgh, Scotland: Cognitive Science Society.

Milin, P., Filipović Đurđević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Journal of Memory and Language*, 50–64.

Milin, P., Kuperman, V., Kostić, A., & Baayen, R. H. (2009). Paradigms bit by bit: an information-theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: form and acquisition* (pp. 214–252). Oxford: Oxford University Press.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the rescorla-wagner model. *Psychological Bulletin*, *117*(3), 363–386.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R. H. (2004). 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., & Filipović Đurđević, D. (2009). The missing link between morphemic assemblies and behavioral responses: a Bayesian Information-Theoretical model of lexical processing. *Manuscript submitted for publication.*

Nelson, D., McEvoy, C., & Pointer, L. (2003). Spreading Activation or Spooky Action at a Distance? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 42–52.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Re-examining word length effects in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Norris, D., & Kinoshita, S. (2008). Perception as evidence accumulation and bayesian inference: Insights from masked priming. *Journal of Experimental Psychology*, *137*(3), 434–455.

Numenta. (2010). *Hierarchical temporal memory including HTM cortical learning algorithms.* Available from `http://www.numenta.com/htm-overview/education.php` (Version 0.1.1, November 23, 2010)

Plag, I., & Baayen, R. H. (2009). Suffix ordering and morphological processing. *Language*, *85*, 106–149.

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.

Pollatsek, A., Hyönä, J., & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human, Perception and Performance*, *26*, 820–833.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*(6), 927–960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957.

Rapp, B. (1992). The nature of sublexical orthographic organization: The bigram trough hypothesis examined. *Journal of Memory and Language*, *31*(1), 33–53.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*, 1090–1098.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159–182.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*(3), 385–408.

Rescorla, R. A. (1988). Pavlovian conditioning. it's not what you think it is. *American Psychologist*, *43*(3), 151–160.

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (p. 131-154). Hillsdale, New Jersey: Lawrence Erlbaum.

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

Schreuder, R., Burani, C., & Baayen, R. H. (2003). Parsing and semantic opacity. In E. Assink & D. Sandra (Eds.), *Reading complex words. cross-language studies* (pp. 159–189). Dordrecht: Kluwer.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2), 241–263.

Seidenberg, M. S. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy. In M. Coltheart (Ed.), *Attention and Performance XII* (pp. 245–264). Hove: Lawrence Erlbaum Associates.

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

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psycholgical Review*, *96*, 523–568.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379-423.

Shaoul, C., Westbury, C., & Baayen, R. H. (2009). *Agreeing with Google: We are sensitive to the relative orthographic frequency of phrases.* (Poster presented at Psychonomics 2009)

Shillcock, R., Kirby, S., McDonald, S., & Brew, C. (2001). Filled pauses and their status in the mental lexicon. In *Proceedings of the 2001 conference of disfluency in spontaneous speech* (pp. 53–56). Edinburgh: International Speech Communication Association.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the rescorla-wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314–321.

Stanojčić, v., & Popović, L. (2005). Gramatika srpskog jezika *(Serbian Language Grammar)*. Beogra: Zavod za udžbenike i nastavna sredstva.

Stefanowitsch, A., & Gries, S. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243.

Stevanović, M. (1989). Savremeni srpskohrvatski jezik *(Contemporary Serbian Language)*. Beograd: Naučna Knjiga.

Stump, G. (2001). *Inflectional Morphology: A Theory of Paradigm Structure.* Cambridge: Cambridge University Press.

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.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*, 638-647.

Taft, M., & Forster, K. I. (1976a). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, *15*, 607–620.

Taft, M., & Forster, K. I. (1976b). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, *15*, 607-620.

Taneja, I. (1989). On generalized information measures and their applications. *Advances in Electronic and Electron Physics*, *76*, 327–413.

Taneja, I., Pardo, L., Gil, P., & Gil, M. (1990). Generalized shannon-gibbs-type weighted inequalities. In *Proceedings of the 3rd International conference on Information Processing and Management of Uncertainty in Knowledge-based Systems* (pp. 398–399). London, UK: Springer-Verlag.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and erp study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 151–173). London: The Continuum International Publishing Group.

Van Rijn, H., & Anderson, J. R. (2003). Modeling lexical decision as ordinary retrieval. In *Proceedings of the fifth international conference on cognitive modeling* (pp. 207–212). Bamberg, Germany: Universitats-Verlag Bamberg.

Venables, W. N., & Ripley, B. D. (2003). *Modern applied statistics with S-Plus* (4th ed.). New York: Springer.

Vogel, E. H., Brandon, S. E., & Wagner, A. R. (2003). Stimulus representation in sop: II. an application to inhibition of delay. *Behavioural Processes*, *62*(1-3), 27–48.

Wagner, A., & Rescorla, R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii* (pp. 64–99). New York: Appleton-Century-Crofts.

White, S. J. (2008). Eye movement control during reading: effects of word frequency and orthographic familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 205-223.

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin & Review*, *8*(2), 221–243.

Wood, S. N. (2006). *Generalized additive models.* New York: Chapman & Hall/CRC.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979.

Yuille, A. (2005). The Rescorla-Wagner algorithm and maximum likelihood estimation of causal parameters. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17* (p. 1585-1592). Cambridge, MA: MIT Press.

Yuille, A. (2006). Augmented Rescorla-Wagner and maximum likelihood estimation. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 1561–1568). Cambridge, MA: MIT Press.

## Appendix

Given the example lexicon shown in Table 8, and using as cues the letter unigrams *a, d, h, l, n, s*, we first calculate the matrix of cooccurrence frequencies $\boldsymbol{C}$, which has as its elements the frequencies $f(i, j)$ with which unigrams $i$ and $j$ co-occur:

$$\boldsymbol{C} = \left(\begin{array}{crrrrrr} & a & d & h & l & n & s \\ a & 419 & 250 & 30 & 301 & 76 & 210 \\ d & 250 & 250 & 30 & 167 & 76 & 41 \\ h & 30 & 30 & 30 & 0 & 30 & 20 \\ l & 301 & 167 & 0 & 301 & 11 & 137 \\ n & 76 & 76 & 30 & 11 & 76 & 23 \\ s & 210 & 41 & 20 & 137 & 23 & 210 \end{array}\right) \tag{37}$$

The main diagonal of $\boldsymbol{C}$ contains the unigram frequencies, the off-diagonal elements the co-occurrence frequencies. In words such as *lass*, the *s* is counted once. In models with not only unigram but also bigram cues, geminates are accounted for by bigrams (e.g., *ss*).

The co-occurrence matrix is transformed into a conditional probability matrix $\boldsymbol{C}'$ the elements of which specify the conditional probability of unigram $j$ given unigram $i$.

$$p(j|i) = p(j,i)/p(i) = p(j,i)/\sum_j p(j,i) = f(j,i)/\sum_j f(j,i). \tag{38}$$

For the example lexicon, we have

$$\boldsymbol{C}' = \begin{array}{c} \\ a \\ d \\ h \\ l \\ n \\ s \end{array} \begin{array}{c} \begin{array}{cccccc} a & d & h & l & n & s \end{array} \\ \begin{pmatrix} 0.33 & 0.19 & 0.02 & 0.23 & 0.06 & 0.16 \\ 0.31 & 0.31 & 0.04 & 0.21 & 0.09 & 0.05 \\ 0.21 & 0.21 & 0.21 & 0.00 & 0.21 & 0.14 \\ 0.33 & 0.18 & 0.00 & 0.33 & 0.01 & 0.15 \\ 0.26 & 0.26 & 0.10 & 0.04 & 0.26 & 0.08 \\ 0.33 & 0.06 & 0.03 & 0.21 & 0.04 & 0.33 \end{pmatrix} \end{array} \tag{39}$$

For instance, $p(a|d) = 250/(250 + 250 + 30 + 167 + 76 + 41) = 0.31$. The rows of $\boldsymbol{C}'$ add up to unity ($\sum_j p(j|i) = 1$).

The outcome matrix $\boldsymbol{O}$ specifies for each outcome (meaning) $j$ and each cue (unigram) $i$ the frequency with which they co-occur:

$$\boldsymbol{O} = \begin{array}{c} \\ a \\ d \\ h \\ l \\ n \\ s \end{array} \begin{array}{c} \begin{array}{cccccccc} \textit{and} & \textit{lass} & \textit{sad} & \textit{as} & \textit{land} & \textit{plural} & \textit{lad} & \textit{hand} \end{array} \\ \begin{pmatrix} 35 & 134 & 18 & 35 & 11 & 77 & 156 & 30 \\ 35 & 0 & 18 & 0 & 11 & 77 & 156 & 30 \\ 0 & 0 & 0 & 0 & 0 & 20 & 0 & 30 \\ 0 & 134 & 0 & 0 & 11 & 57 & 156 & 0 \\ 35 & 0 & 0 & 0 & 11 & 23 & 0 & 30 \\ 0 & 134 & 18 & 35 & 3 & 23 & 0 & 20 \end{pmatrix} \end{array} \tag{40}$$

This matrix is transformed into a matrix of conditional probabilities $p(o|i)$ specifying the probability of an outcome $o$ given cue $i$:

$$p(o|i) = p(o,i)/p(i) = \boldsymbol{O}_{i,o}/\sum_j \boldsymbol{C}_{j,i}. \tag{41}$$

For instance,

$$p(hand|h) = \frac{30}{30 + 30 + 30 + 0 + 30 + 20} = 0.21.$$

The conditional outcome matrix for the example lexicon is

$$\boldsymbol{O}' = \begin{array}{c} \\ a \\ d \\ h \\ l \\ n \\ s \end{array} \begin{array}{c} \begin{array}{cccccccc} \textit{and} & \textit{lass} & \textit{sad} & \textit{as} & \textit{land} & \textit{plural} & \textit{lad} & \textit{hand} \end{array} \\ \begin{pmatrix} 0.03 & 0.10 & 0.01 & 0.03 & 0.01 & 0.06 & 0.12 & 0.02 \\ 0.04 & 0.00 & 0.02 & 0.00 & 0.01 & 0.09 & 0.19 & 0.04 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.14 & 0.00 & 0.21 \\ 0.00 & 0.15 & 0.00 & 0.00 & 0.01 & 0.06 & 0.17 & 0.00 \\ 0.12 & 0.00 & 0.00 & 0.00 & 0.04 & 0.08 & 0.00 & 0.10 \\ 0.00 & 0.21 & 0.03 & 0.05 & 0.00 & 0.04 & 0.00 & 0.03 \end{pmatrix} \end{array}. \tag{42}$$

Let $\boldsymbol{v}_j$ denote the $j$-th column of $\boldsymbol{O}'$. The vector $\boldsymbol{w}_j$ of weights on the connections from the cues to the $j$-th meaning is obtained by solving

$$\boldsymbol{C}'\boldsymbol{w}_j = \boldsymbol{v}_j. \tag{43}$$

The weight matrix $\boldsymbol{W}$ ensues when (43) is applied once to each of the columns of $\boldsymbol{O}'$, binding the resulting vectors column-wise, i.e.,

$$\boldsymbol{W} = \begin{array}{c} \\ a \\ d \\ h \\ l \\ n \\ s \end{array} \left( \begin{array}{rrrrrrrr} \textit{and} & \textit{lass} & \textit{sad} & \textit{as} & \textit{land} & \textit{plural} & \textit{lad} & \textit{hand} \\ 0.38 & -0.03 & -0.41 & 1.03 & -0.38 & -0.45 & 0.41 & 0 \\ -0.16 & -0.56 & 0.61 & -0.44 & 0.16 & 0.53 & 0.39 & 0 \\ -0.69 & -0.05 & -0.36 & 0.05 & -0.31 & 0.49 & 0.36 & 1 \\ -0.21 & 0.62 & -0.17 & -0.62 & 0.21 & 0.22 & 0.17 & 0 \\ 0.61 & 0.42 & -0.19 & -0.42 & 0.39 & -0.09 & -0.81 & 0 \\ -0.21 & 0.34 & 0.54 & -0.34 & 0.21 & 0.27 & -0.54 & 0 \end{array} \right) \tag{44}$$

The simplifying assumption that the estimation of the weights for a given meaning can proceed independently of the weights for the other meanings, is what makes the model a *naive* discriminative learning model.

Let $\boldsymbol{u}_j$ denote the vector specifying which unigrams are present in the input for meaning $j$. For *hand*,

$$\boldsymbol{u}_8 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}. \tag{45}$$

The activation of meaning $j$ is given by

$$a_j = \sum_i \boldsymbol{W}_{ij} = \boldsymbol{W}^T \boldsymbol{u}_j. \tag{46}$$

In this example, the activation of the meaning of *hand* is 1. As the unigram $h$ occurs only in *hand* and *hands*, its carries the full burden of activating this meaning.

The conditional cooccurrence matrix can be singular. For instance, when the words *as* and *lass* are removed from the example lexicon,

$$\boldsymbol{C}' = \begin{array}{c} \\ a \\ d \\ h \\ l \\ n \\ s \end{array} \left( \begin{array}{cccccc} a & d & h & l & n & s \\ 0.31 & 0.31 & 0.04 & 0.21 & 0.09 & 0.05 \\ 0.31 & 0.31 & 0.04 & 0.21 & 0.09 & 0.05 \\ 0.21 & 0.21 & 0.21 & 0.00 & 0.21 & 0.14 \\ 0.32 & 0.32 & 0.00 & 0.32 & 0.02 & 0.01 \\ 0.26 & 0.26 & 0.10 & 0.04 & 0.26 & 0.08 \\ 0.24 & 0.24 & 0.12 & 0.02 & 0.14 & 0.24 \end{array} \right)$$

is exactly singular, since the probabilities in the first two rows and those in the first two columns are identical. We therefore use the Moore-Penrose pseudoinverse of the matrix, implemented in R as `ginv` in the `MASS` package of Venables and Ripley (2003). The pseudoinverse of a matrix provides a unique solution that is optimal in the least squares sense. Let $\boldsymbol{C}^+$ denote the pseudoinverse of the conditional co-occurrence matrix $\boldsymbol{C}'$. Calculating the weight matrix amounts to solving a series of systems of equations

$$\boldsymbol{C}'\boldsymbol{W} = \boldsymbol{O}, \tag{47}$$

achieved with the pseudoinverse as follows:

$$\boldsymbol{W} = \boldsymbol{C}^+\boldsymbol{O}. \tag{48}$$