Breaking the tyranny of learning: A broad-coverage distributed connectionist model of visual word recognition

Fermín Moscoso del Prado Martín[a] and R. Harald Baayen[b,c]

[a] MRC Cognition and Brain Sciences Unit, Cambridge, U.K.
[b] Interfaculty Research Unit for Language and Speech, University of Nijmegen, The Netherlands
[c] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Address all correspondence to:
F. Moscoso del Prado Martín
MRC Cognition and Brain Sciences Unit
15 Chaucer Road
CB2 2EF Cambridge
United Kingdom
e-mail: fermin.moscoso-del-prado-martin@mrc-cbu.cam.ac.uk
tel: +44 1223 355 294 X294
fax: +44 1223 359 062

Abstract

In this study we describe a distributed connectionist model of morphological processing, covering a realistically sized sample of the English language. The purpose of this model is to explore how effects of discrete, hierarchically structured morphological paradigms, can arise as a result of the statistical sub-regularities in the mapping between word forms and word meanings. We present a model that learns to produce at its output a realistic semantic representation of a word, on presentation of a distributed representation of its orthography. After training, in three experiments, we compare the outputs of the model with the lexical decision latencies for large sets of English nouns and verbs. We show that the model has developed detailed representations of morphological structure, giving rise to effects analogous to those observed in visual lexical decision experiments. In addition, we show how the association between word form and word meaning also give rise to recently reported differences between regular and irregular verbs, even in their completely regular present-tense forms. We interpret these results as underlining the key importance for lexical processing of the statistical regularities in the mappings between form and meaning.

Introduction

Morphological paradigms

Words in human languages often exhibit some degree of internal structure, which is characterized by morphological constituents. Morphological constituents – morphemes – are traditionally considered as the minimal meaningful units in human languages. For instance, the English word *unbreakable* can be decomposed into the morphemes *un-*, *break*, and *-able*, each of them having a partial contribution to the meaning and to the orthographical and phonetic forms of the word. In this respect, the vocabulary of any human language can be viewed as being structured into hierarchically organized kindreds of words that share part of their constituent morphemes. We refer to these hierarchically organized 'bins' as *morphological paradigms*.[1] A great amount of research using behavioral and neurophysiological techniques has shown that these morphological paradigms have implications for the ways in which words are stored and processed in the mental lexicon.

Effects of morphological paradigms on lexical processing

The 'size' of the morphological paradigms in which a word is embedded influences the amount of effort involved in processing that word. For instance, the summed frequency of all inflectional variants of a word – its lemma frequency – has been shown to correlate negatively with response latencies in visual lexical decision experiments (Baayen, Dijkstra, & Schreduer, 1997; Taft, 1979). Similarly, Baayen and Schreuder (1997) showed that the number of existing morphological derivations of a Dutch word – its morphological family size – is positively correlated with reaction times and error rates in visual lexical decision. This finding has been replicated in a range of languages, including English (Ford, Marslen-Wilson, & Davis, 2003), German (Lüdeling & De Jong, 2002), Hebrew (Moscoso del Prado Martín, Deutsch, Frost, De Jong, Schreuder, & Baayen, in press), Finnish (Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 2005), and both languages spoken by Dutch-English bilinguals (Dijkstra, Moscoso del Prado Martín, Schulpen, Schreuder, & Baayen, 2005). In addition, the size of morphological paradigms has been shown to affect reaction times also in priming (Feldman & Pastizzo, 2003), and Pylkkänen, Feintuch, Hopkins, and Marantz (2004) reported neurophysiological correlates of this effect using magnetoencephalography (MEG).

The degree of semantic relatedness between a word and the members of its morphological paradigms also has consequences for its processing. Schreuder & Baayen (1997) and Bertram, Schreuder, & Baayen (2000) noted that the exclusion of semantically unrelated items from morphological family size counts improved the correlation between morphological family size and response latencies. For instance, not counting *casualty* in the morphological family size of *casual*, improves the correlation between the family size of *casual* and the times it takes to recognize it. In the same direction, Moscoso del Prado Martín et al. (2005; in press) showed that, in extreme cases, highly unrelated paradigm members can even to a complete reversal of the family size effect, with the number of unrelated members of the paradigm correlating **positively** with response latencies in Hebrew visual lexical decision. Kostić, Marković, and Baucal (2003) reported similar effects for Serbian inflectional paradigms, indicating that semantic context in which a word appears can restrict the effective contribution of each of the members of an inflectional paradigm. Ford (2004) showed that the density of the clustering of the Latent Semantic Analysis representations (Landauer & Dumais, 1997) of the members of a morphological paradigm correlates with the lexical decision RTs to those words, with words belonging to 'denser' (i.e., more homogenuous) paradigms being recognized faster than words from more heterogeneous paradigms.

Moscoso del Prado Martín, Kostić, and Baayen (2004) showed that the effects of the size of a morphological paradigm, both in inflectional and derivational morphology, are best captured by a probabilistic measure of the informational complexity of the nested structures of the morphological paradigms. Based on the evidence for the different counts used to describe the size of a morphological paradigm, they introduced a single unified measure of the support that morphological paradimgs provide to the recognition of its members – the paradigmatic entropy. This measure is calculated over a predefined tree-like structure in which inflected forms are linked to their base forms, which, if morphologically complex, are themselves linked to the simpler words from which they are derived.

Models of morphological processing

The predictive value for lexical decision latencies of measures calculated from discrete morphological lattice structures would arise naturally in decompositional models of morphological processing. In these models words are processed through activation of discrete units corresponding to their constituent morphemes. Such models would be able to account for the effects described above, as long as they allow probabilistic factors to affect the decomposition process (e.g., Burani & Laudanna, 1992; Chialant & Caramazza, 1995; De Jong, Schreuder, & Baayen, 2003; Frauenfelder & Schreuder, 1991; Marslen-Wilson, Tyler, Waksler, & Older, 1994). For these same reasons, this measure might seem problematic for non-decompositional models of morphological processing, such as distributed connectionist models (DCMs; e.g., (Gaskell & Marslen-Wilson, 1997; Devlin, Gonnerman, Andersen & Seidenberg, 1997; Plaut & Booth, 2000; Plaut & Gonnerman, 2000; Rueckl, Mikolinski & Raveh, 1997). DCMs do not make use of discrete representations of the morphological structure of complex words, so it is not clear how would the hierarchical structure of a morphological paradigm be captured by these models. Although DCMs do not assume any explicit discrete representation of morphemes as such, it is claimed that the information about the morphological structure of words is encoded in the patterns of activation that these words elicit in the networks (Seidenberg & Gonnerman, 2000). However, most DCMs to the date have modelled small, unrealistic, and mostly artificial samples or analogs of human languages, avoiding the higher degrees of morphological complexity that are the cornerstone of human languages. It thus remains unclear how the highly complex structures needed for encoding the relations within a large, hierarchically organized morphological paradigm can arise just from the regularities existing between word forms and meanings.

On the other hand, non-decompositional DCMs have the advantage of being able to capture various graded effects arising from systematic pairings between form and meaning that do not constitute morphological relations in the traditional sense. For instance, Bergen (2004) reports that groups of words that are systematically related in form and meaning, but not morphologically related (e.g., the cluster of English words all relating to LIGHT and starting with the letter sequence 'gl' such as *glitter*, *glow*, *glimmer*, *glisten*, ...) prime each other in a way similar to the priming effects that have been observed between morphologically related words. Bergen shows that this effect is not solely due to orthographic or semantic similarity between the prime-target pairs, but that it depends on the consistency of the form-meaning relations across the lexicon. Boudelaa and Marslen-Wilson (2001) report a similar priming effect for Arabic words that share groups of two consonants. These groups of two consonants by themselves do not constitute a full morphological unit in the Arabic lexicon, at least in the same sense that the three-consonantal roots do (Bentin & Frost, 2001). These findings suggest that the mental lexicon is sensitive to systematic correspondences between form and meaning, even when these do not come in the form of clearly decomposable units that could be clearly classified as morphemes. Non-decompositional theories of lexical processing such as DCMs or Bybee's discrete network model (Bybee, 1985; 1988) are better suited to account for these graded effects, as they do not depend on the explicit decomposition of a complex word into discrete morphemes.

The main goal of this paper is to investigate whether the hierarchical patterns of morphological structure that are employed by symbolic models of morphological processing, are a direct consequence of the statistical patterns that are found in the mappings from orthography to meaning in *real* language. The consequence of whis would be that any model that captures these sub-regularities, will exhibit effects that appear to arise from discrete structures. We address the question by constructing a broad-coverage distributed connectionist (BC-DC) model of lexical recognition, trained on a sample of human language that is as realistic as possible, both in terms of the nature of the representations employed and with respect to the size of the sample in comparison to human vocabularies. As a consequence, we will show that the adequacy of discrete structures to describe the effects of morphological structures that are reported in the literature should not be taken as evidence for a direct isomorphism between these descriptive linguistic structures and the underlying mental processes and structures. In addition, we will also explore the predictions that a model based on plain form-meaning associations makes for another debate that has generated considerable discussion in the literature: that of processing of English regular and irregular past-tenses.

English past-tense

The processing of English past-tense forms has given rise to a heated debate in the psycholinguistic literature over the las two decades. On the one hand, a large group of authors have argued for a dual-route system in which irregulars would be stored in an associative memory system in a non-decompositional fashion, while regulars would be processed by application of a symbolic rule (e.g., Clahsen, 1999; Pinker, 1997, 1999). The proponents of the dual-route processing model base their arguments on differences found in the processing of regular and irregular past-tense forms in behavioural studies (e.g., Clahsen, 1999), neuro-psychological double dissociations (e.g., Miozzo, 2003), and brain-imaging studies (e.g., Ullman, Bergida, & O'Craven, 1997; Indefrey, Brown, Hagoort, Sach, & Seitz, 1997). On the other hand, another group of authors have proposed a 'single-route' approach, by which both regulars and irregular verbs would be processed by the same basic mechanism (e.g., MacWhinney & Leinbach, 1991; Moscoso del Prado Martín, Ernestus, & Baayen, 2004; Plunkett & Marchman, 1993; Plunkett & Juola, 1999; Rumelhart & McClelland, 1986).

A factor that has not been given sufficient consideration in the past-tense debate (although already suggested to play a role by MacWhinney and Leinbach, 1991) is to what extent semantic information might interact with verb regularity. In this respect, Ramscar (2002) provided experimental evidence that the semantic context in which a pseudo-verb is presented influences people's choices of the past-tense form for that pseudo-verb. Ramscar noted that this fact is problematic for dual route theories. More recently, Baayen and Moscoso del Prado Martín (in press) have added a new dimension to the debate by showing that distinction between regular and irregular and irregular verbs correlates with subtle differences in the semantic properties of both types of verbs. The authors argue that these differences might partially explain the differences observed between these two types of verbs in behavioural and neuroimaging experiment. More recently, Tabak, Schreuder, and Baayen (2005) report that the differences in the visual lexical decision RTs that are observed between regular and irregular past-tense forms appear also when the targets are fully regular present-tense forms of those same verbs. This constitutes a challenge for the dual route mechanism in that, according to that theory, there should be no such differences, as in this case both regulars and irregulars should be processed using the same mechanism. These results raise the question of whether a single-route connectionist model of lexical processing might mirror the experimental processing differences observed for the uninflected stems of regular and irregular verbs. These results are consistent with the view that morphology arises as the convergence between representations of form and meaning (Seidenberg & Gonnerman, 2000). Therefore, a model that is trained on performing the mapping from orthography to word meaning should also be sensitive to such differences in the formal and semantic properties of regular and irregular verbs. This would provide additional support for the presence of such differences, that would also be difficult to explain within a 'dual-route' framework, thus casting doubts on the architectural conclussions on the cognitive system that have been drawn on the basis of such differences.

Overview of the paper

In what follows, we begin by discussing technical problems and general methodological issues on computational modelling that need to be considered in order to build a BC-DC model of lexical processing. We continue by describing a BC-DC model that was trained to produce the semantic representation of a word from a representation of its orthography. In order to simulate, in as much as possible the statistical structure of the English language, we trained the model using a large sample of the English vocabulary, including realistic representations of orthography and word meaning that were directly derived from a corpus. After the technical specifications of the model, we report the results of four experiments. These experiments investigate the degree of correspondence between the models error scores and lexical decision RTs, and to which extent has the model developed morphological representations. In all three experiments, we compare the error scores produce by the model with the lexical decision latencies to those same words, taken from Balota, Cortese, and Pilotti (1999)'s lexical decision database. Experiment 1 assesses the general correspondence between our model and human lexical processing, we investigate whether the errors of the model to a large set of words from the Balota et al. dataset correlate with the pattern of visual lexical decision RTs. After establishing the correspondence between model errors and human RTs, Experiment 2 investigates the key question of this study: Has the model developed a representation of morphological paradigms? We examine whether the participants in the Balota et al. (1999) database show the paradigmatic entropy effects observed for Dutch by Moscoso del Prado Martín, Kostić, and Baayen (2004), and whether our model shows equivalent effects of morphological paradigms. Here we also address the possible confounds that may arise between paradigmatic entropy measures, and other purely formal or semantic factors. In Experiment 3 we test whether our model captures differences in the processing of present-tense forms of regular and irregular verbs similar to those reported by Tabak et al. (2005). As above, the differences in processing regular and irregular verb are addressed in comparison to the differences found in the response latencies of the participants in the Balota et al. study to those same stimuli. Experiment 4 is aimed at obtaining a better understanding of which aspects of lexical processing can be consequences by the plain associations between orthography and meaning, and which aspects would require additional information. For this purpose, we employed the large lexical decision dataset described by

Baayen, Feldman, and Schreuder (submitted). In this study, the authors provide a description of the effects of a very large set of variables on lexical decision latenicies from the Balota database. Using the same materials, we investigate which of these effects arise also in our network, and which ones do not appear. Finally, we conclude by outlining the implications of our model for current research on visual lexical processing.

Modelling Lexical Processing: General Considerations

Large vocabularies

Modeling the effects of morphological paradigms on lexical processing requires using much larger vocabularies than those used in current computational models. The reason for this is that, in real human language, the morphological paradigm of a word generally includes many very low frequency items. According to Zipf's second law (Zipf, 1949; see also Baayen, 2001), very low frequency words make up the majority of the lexicon of human languages. Currently, mostly due to technological limitations, connectionist models are generally trained on very limited vocabularies rising, in the best of cases, up to some 10,000 words. These models usually consider only the highest part of the word frequency spectrum, thus ignoring the most of the members of morphological paradigms. In addition, in these models the differences in word frequency are generally smoothed by the usage of the log tranform. The underlying assumption is that the behavior of a system with a small vocabulary should reflect the properties of real human lexicons, with vocabulary sizes that are several orders of magnitude greater. Unfortunately, this assumption is over-optimistic. Human languages are characterized by Large Number of Rare Events probability distributions (LNRE; Khmaladze, 1987; see Baayen, 2001 for an introduction). In LNRE distributions, most of the possible events (i.e., words) occur with extremely low probabilities in comparison to a few very frequent events. These distributions show qualitatively different properties to the more normal-like distribution used in the smaller-models. In this respect, small-scale connectionist models are exposed to an environment that is not only quantitatively, but also qualitatively different from that of realistic language processing. This makes it dangerous to extrapolate from one case to the other.

Representation of word meaning

The key importance that semantic factors have on the effects of morphological paradigms, indicates that a connectionist model of morphological processing should include some realistic representation of word meaning. However, it is not feasible to hand-code a realistic representation of the meaning of all words, that is at the same time well-suited to be used in a connectionist model. For this reason most studies have been restricted to using randomly generated vectors as representations of the meaning of words (e.g., Plaut & Booth, 2000). A key problem with this approach is that, by definition, random vectors overlook the finely structured semantic properties of real morphological paradigms. An interesting alternative that has not been sufficiently explored in connectionist models of lexical processing, is the use of semantic vectors derived from co-occurrence statistics over large corpora (e.g., Schütze, 1992). These vectors provide detailed encodings of the meanings of all the words of a language, and can be automatically extracted from unlabelled corpora. In this study we will represent the meanings of words using a variant of the word co-occurrence technique that captures the co-occurrence between words as the activation patterns from a simple recurrent network (for more details, see Moscoso del Prado Martín & Sahlgren, 2002).

Representation of word form

An analogous problem concerns the representation of orthograhic or phonological form in conectionist models. Most current models encode orthographic and phonological forms of words using templates. These templates consist of slots representing the different sequential positions in which the letters or sounds that make up a word can occur (e.g., first letter, second letter, . . .). Such a representational scheme presents many problems for a large-scale model. First, and most obviously, it assumes a predetermined maximum length for the words that can be fitted into the template. Second, and of key importance for modelling morphological processing, these templates require taking crucial decisions on how words should be aligned in the pattern. Normally, this is done in an ad-hoc manner, depending of the particular task that is going to be performed by the network. For instance, networks that concentrate on the processing of suffixes, such as the English regular past tense *-ed*, tend to align words at their right, making sure that the *-ed* suffix appears on the same position for each word. This comes at the cost of loosing information on possible overlaps that might be occurring at the begining or centre of the word. Formal overlap between morphologically related words can occur simultaneously at different positions. Consider once more the word *unbreakable*. A decision to align words on their right, would loose information about other words sharing the *un-* prefix, such as *unstoppable*, and words sharing the *break* stem as could be *unbreakability* or *breaking*. And third, these ad-hoc decisions on word alignment have been criticized for implicitely hardwiring morphological decomposition rules into the system by means of making a particular morphological pattern more salient (Pinker & Prince, 1988). For instance, models of English past-tense formation that manually align the inflectional *-ed* implicitely underline the importance of this suffix for the task. However the same model would be not able to deal with tense inflection in languages such as Hebrew or Arabic, that distribute the tense information across the whole word, rather than concentrating it on a word-final suffix. An alternative approach that has also been used consist in using sets of 'wickelphones' (Wickelgren, 1969) to enconde the forms of words (e.g., Seidenberg & McClelland, 1989). However Prince and Pinker (1988) showed that wilckelphones are ambiguous if one attempts to represent all the words in a language. A more promising approach is the direct usage of the sequential structure in language, by means of introducing recurrence in the networks (Elman, 1990; 1993; Norris, 1990). Some models have managed to employ this technique for building large scale models of lexical processing (Moscoso del Prado Martín, Ernestus, & Baayen, 2004). However, introducing the recurrence directly in the model tends to increase network unstability and training times. As an alternative Moscoso del Prado Martín, Schreuder, & Baayen (2004), proposed the Accumulation of Expectations (AoE) technique. AoE uses the same simple recurrent networks employed by Elman (1990; 1993) and Norris (1990) to derive 'flat' representations of word orthography and phonology. The technique is based on summing the activation values on the hidden layer of an SRN, and later using the obtained vector as a representation of a word's form. Moscoso del Prado Martín, Ernestus, & Baayen (2004) showed that these AoE vectors are adequate for building large-scale connectionist models of morphological processing.

The tyranny of learning

Part of the bottleneck on building large-scale connectionist systems can be attributed to what we could call 'the tyranny of learning'. Before their outputs are taken into consideration, connectionist models are expected to achieve a high degree of accuracy in reproducing the target values on which they have been trained. We believe that this limitation is unjustified. Connectionist networks can also be viewed as statistical tools that assess the complexity of a certain task, given the information on which the task is assumed to be performed. For instance, a neural network can be used to investigate if some verbs are more difficult to learn or process than others, given only their frequencies and orthographical forms. In such an approach, one would train the model to a reasonable level and then perform statistical analyses on the errors produced by the model. The fact that the errors are non-uniformly distributed enables us to draw conclusions on which sorts of information are most salient on the data, and on how these correlate with different experimental variables. From this perspective the models constitute descriptions of the statistical regularities that are present in the data, that could be exploited by the human cognitive system. It is clear that, by relaxing the requirement on general accuracy of the outputs, we discard the possibility of the model being considered a direct description of the biological mechanism that performs the task. However, we do not consider this a drawback for a DCM. Even in the case when a connectionist system does achieve an acceptable performance in a task, it is still controversial to accept the model as a low-level description of the corresponding biological systems. Few connectionist models of complex cognitive tasks use architectures and learning mechanisms that approximate those employed in the human brain. Therefore the inferences that can be drawn from the models concern only the statistical regularities that exist in the environment, and the sources of information that are relevant, or sufficient, for the task. As we will show below, models of this kind can still provide useful insights on the possible causes of behavioural effects. The observation of a behavioural effect often leads the researchers to posit the presence of additional mechanism to account for it. However, in some cases, the models would allow us to see that the effects could also be direct consequences of the complex statistical regularities present in the environment, thus making the additional mechanisms superfluous (on the lack of any additional support for them). DCMs are particularly well suited to detect this type of covert – sometimes counterintuitive – patterns. In summary, the relaxation of the accuracy requirement makes it possible to train connectionist models on very large datasets, without it being a problem that the network may not be very succesful at retrieving particular instances on which it was trained.

Technical Specifications of the Model

Network Architecture

We built a three-layered backpropagation network (Rumelhart, Hinton, & Williams, 1986) whose general architecture is shown in Figure 1. The network consisted of 40 orthographic input units, 120 “hidden” units, and 150 semantic output units. The units in the input layer had all-to-all connections to the units in the hidden layer, which themselves had all-to-all connections to the units in the output layer.

INSERT FIGURE 1

Training Data

The training set consisted of 48,260 English words. These corresponded to those English words that appear with a frequency higher than 10 in the first 20 million words (following the file order on a Linux directory) of the British National corpus and were also listed in the English part of the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995). For each of these words, we employed the English orthographic AoE vectors described by Moscoso del Prado Martín, Schreuder, and Baayen (2004), and we associated them with the co-occurrence vectors for English words described by Moscoso del Prado Martín and Sahlgren (2002).

Training Procedure

The network was presented with a word's orthographic vector at its input layer, and was trained to produce the corresponding semantic vector at its output layer. We trained the network with $64 \cdot 10^6$ words that were chosen randomly from the $48,260$ words in the example set, each word being chosen a number of times directly proportional to its frequency of occurrence in the corpus from which the semantic vectors were built. The network was trained using the backpropagation, using the modified momentum descent algorithm (Rohde, 1999) with the cosine distance as the error measure. We used a momentum of 0.9 and an initial learning rate of 0.1. Five times during training (each $12.8 \cdot 10^6$ words), the learning rate was divided by two. After training, the network showed an average cosine error of 0.0370 on the words present in the training set.

Experiment 1: Network errors and RTs

As a first step, we assessed to which degree are the errors in our network related to the response latencies in visual lexical decision. For this purpose, we selected a large set of words and their visual lexical decision latencies from the Balota et al. (1999) dataset, and we compared them to the errors produced for those same words by the model.

Method

We selected all the monomorphemic nouns and verbs from the Balota et al. (1999)'s young participants visual lexical decision dataset that were also present in our training corpus. This set consisted of 2,090 words (1,295 verbs and 795 nouns), for which we also obtained the visual lexical decision RTs.

For each of these words, we calculated the cosine error produced by the network, that is, one minus the cosine between the vector output by the network on presentation of the form of the word, and the semantic vector corresponding to the word.

Results and Discussion

From the original dataset, we excluded nine words that elicited RTs or network cosine errors above or below three standard deviations from the mean. The analyses were performed on the remaining 2,081 words.

Figure 2 plots the relationship between the network cosine errors (horizontal axis) and the visual lexical decision RTs from the young participants in Balota et al. (1999) dataset (vertical axis).

INSERT FIGURE 2 APPROXIMATELY HERE

As it can be appreciated in Figure 2, there is a quasi-linear relationship between the network errors and the participants' response latencies. Indeed, the network's cosine errors by themselves account for approximately 30% of the variance in the RTs ($r = 0.55, p < 0.0001$). However, there is a possible confound in this correlation: During training, words were presented to the network a number of times that was directly proportional to their frequency. Just by the effect of this training regime, high frequency words will elicit lower error than low frequency words. It is widely known that word frequency has a strong inverse correlation with visual lexical decision latencies. Therefore, the correlation between network errors and RTs could just reflect the effect of frequency that is common to both measures.

To address this issue, we need to ensure that the RTs correlate with the nework errors *after* the contribution of frequency to the errors is partialled out. To assess this, we performed a sequential analysis of variance on a least-squares regression, with the network's cosine error as dependent variable, and the word's frequency and RT as independent predictors. In our regressions, we assesed the non-linearity of the effects by including restricted cubic splines (Baayen, Feldman, & Schreuder, 2005) for each predictor in the regression. This analysis revealed significant linear and non-linear effects of word frequency (linear: $t(2077) = -34.576, p < 0.0001$; non-linear: $t(2077) = 13.356, p < 0.0001$), and a an additional significant linear effect of the RTs ($t(2077) = 3.235, p = 0.0012$). The effect of RTs did not have any significant non-linear component. We tested the stability of the regression using a fast backwards elimination of factors (Lawless & Singhal, 1978). This method enables us to decide which factors should be removed from the regression (i.e., they do not have significant independent effects once the effects of the other predictors are partialled out). All values reported in the analyses are calculated after removing all non-significant components (factors or non-linear components). The fast backward elimination on factors did not delete any of the factors in the regression. The adjusted R^2 for the regression was 0.71. Figure 3 shows the linear relation between RTs and network errors (right panel), after the contribution of frequency has been partialled out (left panel).

INSERT FIGURE 3 APPROXIMATELY HERE

These results show that the statistical regularities in the associations between forms and meaning are a significant factor influencing the RTs in visual lexical decision. It is still necesary to investigate whether this parallelism between RTs and network errors is due to morphological relationships, or it is just reflecting the orthographic or semantic properties of the words. These issues are addressed in the following experiment.

Experiment 2: Morphological structures

Having ascertained that there is a significant correspondence between our network's errors and RTs, we now turn to investigate whether the network has created representations of the discrete tree-like structures that are characteristic of English morphology. In order to explore whether these lattice structures have been captured by the network, we test whether the network cosine error to a word correlates with the paradigmatic entropy of the morphological paradigms to which the word belongs.

Moscoso del Prado Martín, Kostić, and Baayen (2004) introduced a single measure of the complexity of the morphological paradigms in which a word is embedded – the entropy of its morphological paradigms or *paradigmatic entropy*.

Given a morphological paradigm $\mathcal{P}(s) = \{w_1, w_2, \ldots, w_n\}$, were the w_i are the different words or stems that can be directly derived – i.e., by adding a single morpheme – from the stem s, the entropy H of the morphological paradigm of s is defined as:

$$H(s) = -\sum_{w_i \in \mathcal{P}(s)} P(w_i|s) \log_2 P(w_i|s). \tag{1}$$

In this equation, $P(w_i|s)$ represents the probability of the word being w_i given that the stem is known to be s. If $F(w_i)$ is the frequency with which w_i occurs in a corpus, and $F(s) = \sum_{j=1}^{n} F(w_j)$ is the frequency of the stem s (the sum of the frequencies of all the words or stems that are directly derived from it), then the probability of the word w_i occurring in the paradigm of s is given by the expression $P(w_i|s) \simeq F(w_i)/F(s)$.

If a word is simulataneusly belongs to a series of nested morphological paradimgs $(s_1, \ldots, s_n)$, its total paradigmatic entropy $(H(s_1, \ldots, s_n))$ is defined as the sum of the entropies of the different paradimgs:

$$H(s_1, \ldots, s_n) = \sum_{s=1}^{n} H(s_i), \tag{2}$$

where the $H(s_i)$ are computed using (1). For instance, in order to compute the paradigmatic entropy of the word *likelihoods*, we would sum the paradigmatic entropies of the stems that are contained in it, that is, the paradigmatic entropy of *like*, that of *likely*, and the entropy of the inflectional paradigm of *likelihood*.

The total paradigmatic entropy of the morphological paradigms to which a word belongs is negatively correlated with response latencies in visual lexical decision (Moscoso del Prado Martín, Kostić, & Baayen, 2004). All other factors being equal, words that are embedded in 'heavy' paradigms (high entropy), are recognised faster than words that belong to 'lighter' paradigms (low entropy). By analogy, if the structure of morphological paradigms has been captured and is being exploited by our network, we would expect that words from high entropy paradigms elicit less error than word from low entropy paradigms.

In this experiment, we compare the effects of paradigmatic entropy on the cosine errors of our network, with the effect of the same measure on visual lexical decision latencies. In order to assess that any possible effect of of the paradigmatic entropy is not solely due to the orthographic similarity between the word and its morphological relatives, we also need to take orthographic neighborhood size (Coltheart, Davelaar, Jonasson, & Besner, 1977) into account in our analyses.

Method

From the dataset employed in Experiment 1, we selected all the words that had additional members of their morphological paradigms present on the network's training corpus. In this way we obtained 2,009 words. As in Experiment 1, we removed from this dataset thirteen outliers that were above or below three standard deviations from the mean reaction time or network cosine error.

For each of the 1,996 remaining words we calculated the paradigmatic entropy using (1) and (2). For each word, we also computed the orthographic neighborhood size using the definition provided by Coltheart et al. (1977): For each word, we counted the number of other existing words in the Celex database that could be obtained by changing a single letter. The calculation of the paradigmatic entropy size was performed considering only those words that were present in the network's training set.

We performed two parallel least-squares regressions. In the first one, included (log) word frequency, paradigmatic entropy, and orthographic neighborhood size as independent predictors, and the logarithm of the Balota et al. (1997)'s visual lexical decision RTs as dependent variable. On the second one we used the (log) network cosine errors as dependent variable, and (log) word frequency, paradigmatic entropy, and orthographic neighborhood size as independent variables. Possible non-linearities in the effects were investigated by including three restricted cubic spline term for each of the predictors in the regressions. The significance of the non-linear components was assessed using a sequential analysis of variance. For simplicity, we will only report the values of those non-linear components that reached significance in the analyses. We tested the decided which factors to keep in the regressions using fast backwards elimination of factors. All values reported in the analyses are calculated after removing all non-significant components (factors or non-linear components).

Results and Discussion

The regression of the RTs revealed significant main effects of word frequency (linear: $t(1989) = -18.387, p < 0.0001$; non-linear: $t(1989) = 6.668, p < 0.0001$), orthographic neighborhood size (linear: $t(1989) = -2.220, p = 0.0265$; non-linear: $t(1989) = 2.133, p = 0.0331$), and paradigmatic entropy (linear: $t(1989) = -7.406, p < 0.0001$; non-linear: $t(1989) = 4.701, p < 0.0001$). The fast backward elimination of factors did not delete any of the predictors. The adjusted R^2 for this regression was 0.40.

The regression of the network errors showed significant main effects of word frequency (linear: $t(1991) = -37.535, p < 0.0001$; non-linear: $t(1991) = 13.115, p < 0.0001$), a linear effect orthographic neighborhood size ($t(1991) = 4.242, p < 0.0001$), and a linear effect of paradigmatic entropy ($t(1991) = -5.511, p < 0.0001$). Neither the effect of orthographic neighborhood size, nor that of paradigmatic entropy showed significant non-linear components. As above, a fast backward elimination of factors did not delete any of the predictors. The adjusted R^2 for this regression was 0.71.

The regressions revealed that orthographic neighborhood size and paradigmatic entropies have significant independent contributions both to the RTs and to the network cosine errors. Figure 4 provides a summary of each of the effects in both regrassions. The left panels illustrate the effects found in the regression on the RTs, while the right panels illustrate the effects of the three predictors on the network cosine errors.

INSERT FIGURE 4 APPROXIMATELY HERE

The top two panels of figure 4 plots the effects of frequency on the RTs and on the network errors. In both cases, frequency has a strong facilitatory effect; high frequency words are responded to faster and elicit less network error than low frequency words. Notice the similarity between the slight non-linear smoothings of the frequency effect in both regressions (RTs and network errors).

The middle two panels plot the effects of orthographic neighborhood size in both regressions. The figures show a non-linear U-shaped effect of neighborhood size on the response latencies, contrasting with a small linear inhibitory effect of neighborhood size on the network errors. This dissimilarity is possibly due to our network not capturing the sequential properties of visual word recognition. Our network receives the whole word form of the word at once, whereas in reality, the form of the word would be obtained through groups of several letters in one or more eye-fixations. Orthographic representations using the AoE paradigm give rise to reversed word-length effects, by which longer words will generally produce less error than shorter words (Moscoso del Prado Martín, Ernestus, & Baayen, 2004). In any case, the inhibitory effect of orthographic neighborhood size found in the network errors is in line with the inhibitory effects of neighborhood size that are usually found in visual lexical decision (e.g., Coltheart et al., 1977). At least for the upper neighborhood size half of the dataset, the network behaviour seems to be replicating that observed in the RTs for the Balota et al. (1999) dataset.

Finally, the two panels at the bottom of Figure 4 show the effects of paradigmatic entropy, after having partialled out word frequency and orthographic neighborhood size. Both in the case of the network errors and in the reaction times, the effect of paradigmatic entropy has a significant linear component that is facilitatory (RTs: $\hat{\beta} = -0.0381$; Network errors: $\hat{\beta} = -0.0849$), high entropy words are recognized faster and elicit less network error than low entropy words. However, in the case of the RTs there is a significant non-linear smoothing of the effect on the high-entropy words. This non-linearity is not present in the network errors. There is no straightforward explanation for this difference, but we believe it might be a side-effect of the nature of our paradigmatic entropy counts: On the one hand, the paradigmatic entropy counts are exact in the case of the network, i.e., they were calculated only considering the words that are present in the training set, and using the precise frequency counts with which the network was trained. On the other hand, in the case of the RTs, these counts represent only a rough estimation, since many other words that are members of the paradigm have not been considered, and their frequencies are also an estimation.

Crucially, on the network errors, neighborhood size and paradigmatic entropy have effects of opposite direction. While the paradigmatic entropy has a facilitatory effect, which is in line with the effect reported on Dutch lexical decision latencies by Moscoso del Prado Martín, Kostić, & Baayen (2004), the effect of neighborhood size is inhibitory. This rules out the possibility of the effect of paradigmatic entropy arising from form similarity alone, as it shows that effects of pure, non-morphological form similarity have the opposite effect on the network. Moreover, as we mentioned above, the effect of orthographic neighborhood size in visual lexical decision is reported to be inhibitory (e.g., Coltheart et al., 1977). This contrasts with the *facilitatory* effects of orthographic neighborhood size that are reported in tasks with a lesser semantic component, such as word naming (e.g., Andrews, 1989; 1992).

We have therefore established that there is an effect of paradigmatic entropy in our network's error scores, and that this effect does not arise from orthographic similarity alone. However, we need to consider a possible confound before we can conclude that the network is forming a representation of morphology: The effect that we observed could just be a consequence of the semantic similarity between the members of the paradigm. Indeed, Moscoso del Prado Martín and Sahlgren (2002), report that the same co-occurrence vectors that we have used in the network output, are by themselves – without any orthographic information – capable of reliably detecting inflectional relations between the words. They also report that these vectors alone do not suffice to capture derivational relations reliably. If the effect is solely due to the semantic similarity within the members of the inflectional paradigm, one would expect only the entropy of the inflectional paradigm to have an effect on the network errors. The paradigmatic entropy measure that we have used is computed as the sum of the entropy of an inflectional paradigm plus the entropy of a derivational paradigm. Therefore, it could be the case that it is only the inflectional entropy that is giving rise to the effect, without any significant contribution of the entropy of the derivational paradigm.

We addressed this possible confound by decomposing the total paradigmatic entropy into two measures – the entropy of the inflectional paradigm, and the entropy of the derivational paradigm – and entering these measures separately into the regression instead of the total paradigmatic entropy. We performed a new regression with the (log) network cosine errors as the dependent variable, and (log) word frequency, neighborhood size, entropy of the inflectional paradigm, and entropy of the derivational paradigm as independent predictors. The regression revealed the effects of frequency and neighborhood size (as described above), and significant linear effects of the inflectional entropy ($t(1990) = -5.167, p < 0.0001$), and the derivational entropy ($t(1990) = -2.572, p = 0.0102$). As expected, both sub-effects were facilitatory (inflectional: $\hat{\beta} = -0.10$; derivational: $\hat{\beta} = -0.06$; see Figure 5). Once more, none of the predictors was deleted by a fast backward elimination of factors. The adjusted R^2 of this regression was also 0.71.

The effect of paradigmatic entropy is therefore not only a result of the inflectional paradigm, but a result of the full structure of morphological paradigms, inflectional and derivational. Since, according to Moscoso del Prado Martín and Sahlgren (2002), our semantic vectors do not contain sufficient information to capture derivational information, we conclude that the observed effect of paradigmatic entropy cannot be solely attributed to the semantic similarity between the members of the morphological paradigms.

INSERT FIGURE 5 APPROXIMATELY HERE

In sum, our analyses have shown that the complexity of the morphological paradigms to which a word belongs is a significant predictor for both human RTs, and the amount of error produced by our model. As we have also shown, these effect cannot be attributed neither to form similarity nor to semantic similarity. Therefore the paradigmatic structure of the paradigms is arising from the regularities in the mapping from orthography to meaning.

Experiment 3: Regular and Irregular Verbs

We now turn to investigate whether the differences between regular and irregular verbs are reflected in the errors produced by the BC-DC model. Baayen and Moscoso del Prado Martín (in press) found that English regular and irregular verbs show significant differences with respect to several semantic factors, including the entropy of their inflectional paradigms. Tabak et al. (2005) showed that these differences in the properties of regular and irregular verbs are also reflected in lexical decision latencies to Dutch verbs, with irregular verbs being recognized slower than regular verbs. A crucial point in that study is that the difference in RTs to regular and irregular verbs arise even for the (completely regular) present-tense forms of these verbs. This finding is in contradiction with dual route models of inflectional processing. These models would predict differences in RTs to past-tense forms only, as it is only for past-tense forms for which there are two possible routes (decomposition into stem plus affix or rote memory). If the difference in RTs between present-tense forms of regular and irregular verbs arises as a consequences of differences in the regularity of the form-to-meaning mappings, our model should show a similar effect on the magnitude of its errors. To test this hypothesis, we simulated an additional experiment contrasting present-tense forms of regular and irregular verbs.

Method

From the dataset of Experiment 1, we selected 250 monomorphemic present-tense verbal forms. Of these, 125 corresponded to verbs with an irregular past-tense, and 125 were of completely regular verbs. The two subsets (regular and irregular verbs) were matched for word length in the mean (regulars: 4.34 ± 0.80; irregulars: 4.32 ± 0.89) and median (4 in both subsets).

We performed to least squares-regressions, one on the Balota et al. (1999) RTs, and another on the BC-DC cosine errors. We included as an independent the verb's regularity (regular vs. irregular). In addition, we also included log word frequency and inflectional entropy as covariates, since both of these factors tend to be significantly different between regular and irregular verbs (cf., Baayen & Moscoso del Prado Martín, in press). As in the previous experiments, the independent contribution of the factors was assessed using fast backwards elimination of factors.

Results and Discussion

The regression on the RTs revealed only a main effect of word frequency (linear: $t(247) = -10.516, p < 0.0001$; non-linear: $t(247) = 5.763, p < 0.0001$). The fast backwards elimination of factors recommended removing inflectional entropy and verb regularity as predictors from this regression. The R^2 for the regression (after having removed the non-significant predictors) was 0.45.

The regression on the BC-DC errors revealed main significance main effects of word frequency (linear: $t(245) = -12.563, p < 0.0001$; non-linear: $t(245) = 4.219, p < 0.0001$), inflectional entropy (linear: $t(245) = -2.707, p = 0.0073$), and a small effect of verb regularity ($t(245) = -1.824, p = 0.0693$), by which regular verbs elicited less errors than irregular verbs ($\hat{\beta} = 0.14 \pm 0.07$ for irregular verbs). Although the effect of verb regularity appeared to be only marginally significant, the fast backwards elimination of factors did not recommend deleting it from the regression. The R^2 for this regression was 0.70.

These results of the regression on the model confirms our hypothesis that, even in their present-tense forms, the mapping between orthography and word meaning tends to be more consistent for regular than for irregular verbs. As shown by the regression, these differences cannot be fully attributed to the differences found in the frequencies or inflectional entropies of these verbs. Although the analysis of the RTs did not replicate this effect of verb regularity, the results of the regression on the network errors are fully in line with the results reported by Tabak and colleages: irregular verbs are more difficult to process than regular verbs, even in their present-tense forms.

Experiment 4: XXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

Method

We added the network cosine error to the words in lexical decision dataset of Baayen et al. (2005) that were also present in our training set. In this way we obtained a list of 2178 words (1317 nouns and 861 verbs), together with their lexical decision RTs (from the Balota database) and the cosine errors produced by the model for those words. In addition, for each word we also had the measures of phonological consistency (the PC1 measure from Baayen et al., 2005), voicing of the first phoneme, log surface frequency from CELEX, written-to-spoken frequency ratio, gramatical category (Noun vs. Verb), nominal-to-verbal frequency ratio, derivational entropy, inflectional entropy, and log number of complex synsets from the WordNet database (Miller, 1990).

In order to investigate to what extent the effects reported by Baayen et al. (2005) are reproduced in our network's errors, we performed the same regression analysis that they report on the Balota lexical decision RTs, using the network's cosine errors as dependent variable. Since we have excluded some items from the original dataset (those that were not present in the network's training set) we also replicated their original regression analysis on the RTs. As in the previous experiments, the independent contribution of the factors in both regressions was assessed using fast backwards elimination of factors. The significance of non-linear contributions to the effects was assesed by means of sequential analyses of variance on the restricted cubic splines.

Results and Discussion

All factors mentioned above had significant independent contributions on the regression on the lexical decision RTs, thus replicating the results reported by Baayen et al. (2005) for the slightly larger dataset. The sequential analysis of variance on the network error scores did not reveal any significant non-linear components of the effect of Derivational entropy, thus this component was removed from the analyses. The fast backwards elimination of factors on the network error's regresion suggested deleting from the regression the factors phonological consistency, voicing of the first phoneme, and gramatical category. After removing the non-linearity of the derivational entropy effect, and the gramatical category, phonological consistency, and voicing of the first phoneme, the adjusted R^2 of the regression on the network errors was of 65%. All other factors had significant main effects on the regression. The lack of effect of the two phonological variables (phonological consistency and voicing of the first phoneme) is not surprising, given that our model lacked any representation of the words' phonological properties. In addition, our model failed to show the effect of gramatical category reported by Baayen and colleagues for the lexical decision RTs.

As it can be observed in Figure 6, the effects of word frequency (top panels), inflectional entropy (middle panels), and derivational entropy (bottom panels). This confirms the robustness of the independent effects on the network errors of word frequency (linear: $t(2169) = -26.505, p < 0.0001$; non-linear: $t(2169) = 6.382, p < 0.0001$), inflectional entropy (linear: $t(2169) = -6.194, p < 0.0001$), and derivational entropy (linear: $t(2169) = -2.409, p = 0.0167$) even when a different frequency count is used, and the effects of additional variables are partialled out. Note that once again, the effect of derivational entropy on the network errors is slightly different than that observed on the RTs.

While the latter is a linear inhibitory effect (see bottom-right panel of Figure 6), that suffers a non-linear threshold effect on the high end of the derivational entropy range, the former is just linear across the whole range of derivational entropy values.[2]

INSERT FIGURE 6 APPROXIMATELY HERE

Figure 7 plots the effects of the additional variables from the Baayen et al. (2005) study that had significant effects on the network error scores: the written-to-spoken frequency ratio (Top panels. Linear: $t(2169) = 2.914, p = 0.0036$; non-linear: $t(2169) = -2.560, p = 0.0105$), the noun-to-verb frequency ratio (Mid panels. Linear: $t(2169) = -5.039, p < 0.0001$), and the number of complex WordNet syntets (Bottom panels. Linear: $t(2169) = -7.519, p < 0.0001$). The general pattern that these effects show on the network (right panels) is very reminiscent of the pattern observed on the RTs (left panels), except for the thresholding of the effect of the written-to-spoken frequency ratio observed on the network errors (top-left panel), that is unparalleled on the RTs (top-right panel).

On the network, the effect of the number of WordNet synsets is a reflection of the centrality of the word in the semantic (co-occurrence) space. Word in more central areas of this space are easier to process by the network because of its general tendency to output something closer to the 'average' meaning. The number of WordNet synsets is a measure of the number of senses/meanings with which the word can be used. Words that can be used with many different meanings tend to refer less specific meanings/contexts than words with very restricted usages. This is reflected in both the network and reaction times. Likewise, the effect of the noun-to-verb frequency on the network, indicating that words that are used predominantly as nouns produce less error (and shorter RTs) than words that are used predominantly as verbs, can be associated with the centrality of its about their meaning, which appears to be greater for nominal-dominant words than for verbal-dominant ones. However, whereas an additional independent effect of gramatical category was observed in the lexical decision latencies, this effect was not replicated by our network.

Finally, we believe that the effect of spoken-to-written frequency on the network errors and partially on the reaction times is a reflection of the greater abundance of simple, prototypical concepts, in spoken language, whereas very specific complex concepts might be more common in written language. However, as we mentioned above, the effect is clearly more marked on the RTs than on the network, where it dissappears in the higher (more abundant in spoken language) range. This thresholding might reflect the presence of additional factors that arise in people as a result of the diffenrent properties of the written and spoken modalities of language, and the differences in frequency distributions across both modalities at are present in the human environment but not in the network's training regime.

INSERT FIGURE 7 APPROXIMATELY HERE

General Discussion

In this study, we have presented the BC-DC model of visual word recognition. The model was trained to map distributed representation of word orthography onto distributed representations of a word's meaning (co-occurrence patterns). After training, we compared the model's cosine distances with the response latencies of participants performing visual lexical decision for large sets of English monomorphemic nouns and verbs. We found that, in both cases, the model produced output patterns that were remarkably similar to the pattern of responses of actual participants.

The model that we have introduced constitutes a considerable departure from previously implemented distributed connectionist models of lexical processing in that it has a much broader coverage and in that it avoids the traditional restrictions on word length and morphological complexity. We trained the model on a vocabulary of 48,260 different word forms. This represents a realistic sample of the lexicon, containing the full range of morphological phenomena present in English. In principle, a model of these characteristics could be exposed to even larger vocabularies, approximating the number of different words to which an average adult is exposed.

We have used a neural network to assess the importance of the statistical correspondence between form and meaning. Note here that we are not claiming that our network reproduces the mechanism employed by the brain to capture these regularities. In this sense, the neural network that we have used is nothing more than a sofisticated form of regression analysis (see McKay, 2003 for more details on this usage of neural networks), that enables us to explore the complex statistical correspondences in a very high-dimensional space.

The key to this broad coverage lies in the use of truly distributed representations for word forms, as provided by the AoE representational paradigm (Moscoso del Prado Martín, Schreuder, & Baayen, 2004), and the realistic semantic vectors based on co-occurrence statistics. A corpus-based co-occurrence approach to semantic representation relues only on the realistic assumption that word co-occurrence is one of the main sources of information employed by humans to determine the meaning of a word (Boroditsky & Ramscar, 2003; McDonald & Ramscar, 2001). The combination of these techniques, together with the relaxation of the requisite on accuracy of the system's output, overcome the bottleneck for building models of lexical processing on a realistic scale. In addition, the coding scheme used for word forms presents the advantage of obviating the need for slot-based templates. The use of these templates for the coding of word forms, together hand-crafted representations to code meaning has been criticized for assuming a great amount of hard-wired symbolic information about orthographic and semantic structure (e.g., Pinker & Ullman, 2002b).

The response patterns produced by the BC-DC model account for approximately 30% of the variance in the RTs produced by the human participants. This is remarkable given that many factors that are known to affect visual lexical processing are not taken into account by our system. In particular, as we observed inon Experiment 2 when discussing the effects of orthographic neighborhood size, our model does not mirror the sequential properties of visual word recognition. However, we must note here that, in reality, this model of word recognition should be viewed as a two-stage process. In the first stage, the orthographic representations of the words are built by accumulating the activations of an SRN's hidden layer that receives the letters of a word in sequential order. In the second stage, illustrated by the network presented here, these representations are mapped into word's meaning. If one included in the analyses the results of the first phase, we would expect to replicate also the aspects of visual word-form recognition that are overlooked by the system. We have not integrated these data on our analyses because this would require taking a decission on whether these two stages are sequential or cascaded, and to what degree there is feedback from the second to the first stage. This constitutes a independent debate in itself. Addressing it at the same time would obscure the central question that we are addressing in this study: whether complex morphological structures naturally arise as a by-product of the associations between word form and word meaning. With the variables that we have used in our study, we are able to account for more than two thirds of the variance present in the model's errors. This suggests that further research is necessary to understand the source of the remaining one third of the model's variance. In particular, we expect this additional variance to be related to additional factors that we have not explored, such as purely semantic issues and the integration with the processes involved in creating the distributed represetation of word form. We leave those for further research.

Crucially, although the model did not receive any explicit symbolic representation of the morphological relations between the words in its training set, it developed sensitivity to complex, hierarchical morphological structure, as indicated by the effect of paradigmatic entropy that we observed. In particular, the effects of derivational entropy, and the analyses including neighborhood size, showed that the model is sensitive to effects that cannot be attributed to just form or meaning similarity on their own. Instead, the effects emerge from the systematic form-meaning associations shared by the morphological variants of a word. We have shown that the effects of paradigmatic entropy computed over discrte lattice-like structures (Moscoso del Prado Martín, Kostić, & Baayen, 2004) do not constitute a problem for distributed connectionist models of lexical processing. In fact, we believe that such effects are a fundamental property of neural processing systems (e.g., Deco & Obradović, 1996).

Our model also illustrates how the differences found in the processing of the present-tense forms of regular and irregular verbs arise naturally in a single-route model of lexical processing. The fact that this model was never actually trained on the past-tense formation task confirms the results of Baayen and Moscoso del Prado Martín (in press) and Tabak et al. (2005), in that there are significant differences between both the orthographic and semantic properties of regular and irregular verbs. It is not unlikely that these differences underlie many of the double dissociations and processing differences that have been found between these two kinds of verbs. Additionally, it is not clear how the dual-route models could account for the differences in processing the present tense, especially since their proponents explicitly deny any possible influence of verbal semantics in the selection of a verb's past-tense form (e.g., Pinker & Ullman, 2002a).

In conclusion, we have shown that the statistical correspondences between word form and word meaning might be sufficient for the representations of morphological paradigms to arise in a fully distributed system. This confirms the hypothesis put forward by Seidenberg and Gonnerman (2000), and fits well into the general sensitivity of the morphological processing system to statistical factors (cf., Hay & Baayen, in press). These findings also indicate that the contradiction between symbolic linguistic structures and distributed, statistical systems, is mostly artificial. Symbolic structures are the natural consequence of a system that is sensitive to the statistical properties of language, even when these are represented using distributed patterns of activation. Although they might fail to account for some less clear cut phenomena, the usage of symbolic structures greatly clarifies the relationships between words, which are somewhat obscured when talking in terms of patterns of activation.

Acknowledgements

The authors are indebted to Mirjam Ernestus, William Marslen-Wilson, Robert Schreuder, and Maarten van Casteren for comments in earlier versions of this manuscript. This research was made possible by a PIONIER grant of the Dutch Science Foundation (NWO) to the second author. During the last stages of preparation of this manuscript, the first author has been supported by the European Commission through a Marie Curie Individual Fellowship

References

Andrews, S. (1989), 'Frequency and neighborhood size effects on lexical access: Activation or search?', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15**, 802–814.

Andrews, S. (1992), 'Frequency and neighborhood size effects on lexical access: Similarity or orthographic redundancy?', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**(2), 234–254.

Baayen, R. H. (2001), *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht.

Baayen, R. H. & Moscoso del Prado Martín, F. (in press), 'Semantic density and past-tense formation in three Germanic languages', *Language*.

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* **37**, 94–117.

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

Balota, D. A., Cortese, M. J. & Pilotti (1999), Item-level analyses of lexical decission performance: Results from a mega-study, *in* 'Abstracts of the 40th Annual Meeting of the Psychonomics Society', Los Angeles, CA, p. 44.

Bentin, S. & Frost, R. (2001), 'Linguistic theory and psychological reality: a reply to Boudelaa and Marslen-Wilson', *Cognition* **81**(1), 113–118.

Bergen, B. K. (2004), 'The psychological reality of phonaesthemes', *Manuscript submitted for publication, University of Hawai'i at Manoa*.

Boroditsky, L. & Ramscar, M. (2003), 'Guilt by association: Gleaning meaning from contextual co- occurrence', *Manuscript, Massachusetts Institute of Technology*.

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

Burani, C. & Laudanna, A. (1992), Units of representation for derived words in the lexicon, *in* R. Frost & L. Katz, eds, 'Orthography, Phonology, Morphology, and Meaning', Elsevier, Amsterdam, pp. 361–376.

Bybee, J. L. (1985), *Morphology: A study of the relation between meaning and form*, Benjamins, Amsterdam.

Bybee, J. L. (1988), Morphology as lexical organization, *in* M. Hammond & M. Noonan, eds, 'Theoretical Morphology: Approaches in Modern Linguistics', Academic Press, London, pp. 119–141.

Chialant, D. & Caramazza, A. (1995), Where is morphology and how is it processed? the case of written word recognition, *in* L. B. Feldman, ed., 'Morphological Aspects of Language Processing', Lawrence Erlbaum Associates, Hillsdale, N. J., pp. 55–78.

Clahsen, H. (1999), 'Lexical entries and rules of language: a multi-disciplinary study of German inflection', *Behavioral and Brain Sciences* **22**, 991–1060.

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

Coltheart, M., Davelaar, E., Jonasson, J. T. & Besner, D. (1977), Access to the internal lexicon, *in* S. Dornick, ed., 'Attention and performance', Vol. VI, Erlbaum, Hillsdale, New Jersey, pp. 535–556.

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', Mouton de Gruyter, Berlin, pp. 65–88.

Deco, G. & Obradović, D. (1996), *An Information-Theoretic Approach to Neural Computing*, Springer Verlag, New York.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S. & Seidenberg, M. S. (1997), 'Category-specific semantic deficits in focal and widespread brain damage: A computational account', *Journal of Cognitive Neuroscience* **10**(1), 77–94.

Dijkstra, T., Moscoso del Prado Martín, F., Schulpen, B., Schreuder, R. & Baayen, R. (2005), 'A roommate in cream: morphological family size effects on interlingual homograph recognition', *Language and Cognitive Processes* **20**(1), 7–42.

Elman, J. L. (1990), 'Finding structure in time', *Cognitive Science* **14**, 179–211.

Elman, J. L. (1993), 'Learning and development in neural networks: The importance of starting small', *Cognition* **48**, 71–99.

Feldman, L. B. & Pastizzo, M. J. (2003), Morphological facilitation: the role of semantic transparency and family size, *in* R. H. Baayen & R. Schreuder, eds, 'Morphological structure in language processing', Mouton de Gruyter, Berlin, pp. 233–258.

Ford, M. A. (2004), Morphology in the mental lexicon: Frequency, productivity and derivation, PhD thesis, University of Cambridge, UK.

Ford, M. A., Marslen-Wilson, W. D. & Davis, M. H. (2003), Morphology and frequency: Contrasting methodologies, *in* R. H. Baayen & R. Schreuder, eds, 'Morphological structure in language processing', Mouton de Gruyter, Berlin, pp. 89–124.

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', Kluwer Academic Publishers, Dordrecht, pp. 165–183.

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

Hay, J. B. & Baayen, R. H. (in press), 'Shifting paradigms: gradient structure in morphology', *Trends in the Cognitive Sciences*.

Indefrey, P., Brown, M. C., Hagoort, P., Sach, S. & Seitz, R. J. (1997), 'A PET study of cerebral activation patterns induced by verb inflection', *Neuroimage* **5**, 548.

Khmaladze, E. V. (1987), The statistical analysis of large number of rare events, Technical Report Report MS-R8804, Dept. of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

Kostić, A., Marković, T. & Baucal, A. (in press), Inflectional morphology and word meaning: orthogonal or co-implicative domains?, *in* R. H. Baayen & R. Schreuder, eds, 'Morphological structure in language processing', Mouton de Gruyter, Berlin.

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.

Lawless, J. F. & Singhal, K. (1978), 'Efficient screening of nonnormal regression models', *Biometrics* **34**, 318–327.

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

MacWhinney, B. & Leinbach, J. (1991), 'Implementations are not conceptualizations: revising the verb learning model', *Cognition* **40**, 121–157.

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.

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'.

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

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

Miozzo, M. (2003), 'On the processing of regular and irregular forms of verbs and nouns: evidence from neuropsychology', *Cognition* **87**, 101–117.

Moscoso del Prado Martín, F. & Sahlgren, M. (2002), An integration of vector-based semantic analysis and simple recurrent networks for the automatic acquisition of lexical representations from unlabeled corpora, *in* A. Lenci, S. Montemagni & V. Pirrelli, eds, 'Proceedings of the LREC'2002 Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data', European Linguistic Resources Association, Paris.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R. & Baayen, R. H. (2005), '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., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H. & Baayen, R. H. (in press), 'Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch', *Journal of Memory and Language*.

Moscoso del Prado Martín, F., Ernestus, M. & Baayen, R. H. (2004), 'Do type and token effects reflect different mechanisms: Connectionist modelling of Dutch past-tense formation and final devoicing', *Brain and Language* **90**, 287–298.

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.

Moscoso del Prado Martín, F., Schreuder, R. & Baayen, R. H. (2004), Using the structure found in time: Building real-scale orthographic and phonetic representations by Accumulation of Expectations, *in* H. Bowman & C. Labiouse, eds, 'Models of Cognition, Perception and Emotion. Proceedings of the VIII Neural Computation and Psychology Workshop', World Scientific, Singapore, pp. 264–272.

Norris, D. G. (1990), A dynamic-net model of human speech recognition, *in* G. Altmann, ed., 'Cognitive Models of Speech Processing: Psycholinguistic and cognitive perspectives', MIT Press, Cambridge, MA.

Pinker, S. (1997), 'Words and rules in the human brain', *Nature* **387**, 547–548.

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

Pinker, S. & Prince, A. (1988), 'On language and connectionism', *Cognition* **28**, 73–193.

Pinker, S. & Ullman, M. (2002*a*), 'Combination and structure, not gradedness, is the issue: Reply to McClelland and Patterson', *Trends in the Cognitive Sciences* **6**(11), 472–474.

Pinker, S. & Ullman, M. (2002*b*), 'The past and future of the past tense', *Trends in the Cognitive Sciences* **6**(11), 456–462.

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

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

Plunkett, K. & Juola, P. (1999), 'A connectionist model of English past tense and plural morphology', *Cognitive Science* **23**(4), 463–490.

Plunkett, K. & Marchman, V. (1993), 'From rote learning to system building: acquiring verb morphology in children and connectionist nets', *Cognition* **48**, 21–69.

Prince, A. & Pinker, S. (1988), 'Wickelphone ambiguity', *Cognition* **30**, 189–190.

Pylkkänen, L., Feintuch, S., Hopkins, E. & Marantz, A. (2004), 'Neural correlates of the effects of morphological family frequency and family size: an MEG study', *Cognition* **91**, B35–B45.

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

Rueckl, J. G., Mikolinski, M., Raveh, M., Miner, C. S. & Mars, F. (1997), 'Morphological priming, fragment completion, and connectionist networks', *Journal of Memory and Language* **36**(3), 382–405.

Rumelhart, D. E. & McClelland, J. L. (1986), On learning the past tenses of English verbs, *in* J. L. McClelland & D. E. Rumelhart, eds, 'Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models', The MIT Press, Cambridge, Mass., pp. 216–271.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986), Learning internal representations by error propagation, *in* D. E. Rumelhart & J. L. McClelland, eds, 'Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations', The MIT Press, Cambridge, Mass., pp. 318–364.

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

Schütze, H. (1992), Dimensions of meaning, *in* 'Proceedings of Supercomputing '92', pp. 787–796.

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', *Psychological Review* **96**, 523–568.

Tabak, W., Schreuder, R. & Baayen, R. H. (2005), Lexical statistics and lexical processing: semantic density, information complexity, sex, and irregularity in dutch, *in* S. Kepser & M. Reis, eds, 'Linguistic evidence – Empirical, Theoretical and Computational Perspectives', Mouton de Gruyter, Berlin.

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

Ullman, M., Bergida, R. & O'Craven, K. M. (1997), 'Distinct fMRI activation patterns for regular and irrgular past tense', *NeuroImage* **5**, 549.

Wickelgren, W. A. (1969), 'Context-sensitive coding, associative memory, and serial order in (speech) behavior', *Psychological Review* **76**, 1–15.

Zipf, G. K. (1949), *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology*, Hafner, New York.

Notes

[1] Throughout this paper, we will use the term morphological paradigm to denote the set of words or stems that share a particular morpheme, be it an inflectional suffix (e.g., the *-ed* in *loaded*), a derivational prefix (e.g., *un-* in *unload*), a derivational suffix (e.g., *-er* in *loader*), or a stem (e.g., *-load-* in *workload*, *unload*, or *loader*). Our usage of this term is a generalization of the more traditional 'inflectional paradigm'. The later being restricted to denote the set of different inflectional variations of a given word. According to our loose definition of morphological paradigm, any given word in a language will simultaneusly belong to at least as many different paradigms as different morphemes it contains. See Moscoso del Prado Martin, Kostić and Baayen (2004) for more details on this usage of the term.

[2] The non-linear component of the effect of derivational entropy on the RTs appears to be facilitatory in the high derivational entropy. Notice however, that the standard errors of the effect on this high ranges would allow the effect to go in either direction, thus what we are observing is just a thresholding of the effect rather than a reversal in its direction.

List of Figure Captions

Figure 1: General architecture of the model. The lines represent trainable all-to-all connections between the units in two layers.

Figure 2: Comparison (in bi-logarithmic scale) between the young participants' average reaction time to English monomorphemic nouns and verbs (vertical axis) provided by Balota et al. (1999), with the model's cosine distance for those same nouns (horizontal axis). The line represents a non-parametric regression (Cleveland, 1979).

Figure 3: *Left panel:* Plot (in bi-logarithmic scale) of the non-linear effect of word frequency on the network's cosine errors in Experiment 1. As estimated by a least-squares regression. *Right panel:* Plot (in bi-logarithmic scale) of the relationship between the RTs and the network's cosine errors in Experiment 1. As estimated by a least-squares regression, after partialling out the effect of word frequency. The discontinuous lines represent the standard deviations of the effect estimates.

Figure 4: Plots comparing the effects observed in Experiment 2 for word frequency (*top panels*), orthographic neighborhood size (*middle panels*), and paradigmatic entropy (*bottom panels*) on the Balota et al. (1999) lexical decision RTs (*left panels*) and on the error scores of the network (*right panels*). All effects were estimated using least-squares regressions, including non-linear components when necesary. Each effect is plotted at the mean of the other variables. The discontinuous lines represent the standard deviations of the effect estimates.

Figure 5: Decomposition of the paradigmatic entropy on the network errors of Experiment 2 into its inflectional (*left panel*) and derivational (*right panel*) components. The effects were estimated using a least-squares regression (where no non-linearities were found for these two variables), in which the effects of word frequency (non-linear) and othographic neighborhood size (linear) were first partialled out. Both effects are plotted on the mean of the remaining variables.

Figure 6: Plots comparing the effects observed in Experiment 2 for word frequency (*top panels*), inflectional entropy (*middle panels*), and derivational entropy (*bottom panels*) on the Baayen et al. (2005) lexical decision RTs (*left panels*) and on the error scores of the network (*right panels*). All effects were estimated using least-squares regressions, including non-linear components when necesary. Each effect is plotted at the mean of the other variables. The discontinuous lines represent the standard deviations of the effect estimates.

Figure 7: Plots comparing the effects observed in Experiment 2 for the written-to-spoken frequency ratio (*top panels*), noun-to-verb frequency ratio (*middle panels*), and number of WordNet synsets (*bottom panels*) on the Baayen et al. (2005) lexical decision RTs (*left panels*) and on the error scores of the network (*right panels*). All effects were estimated using least-squares regressions, including non-linear components when necesary. Each effect is plotted at the mean of the other variables. The discontinuous lines represent the standard deviations of the effect estimates.

Figure 1:

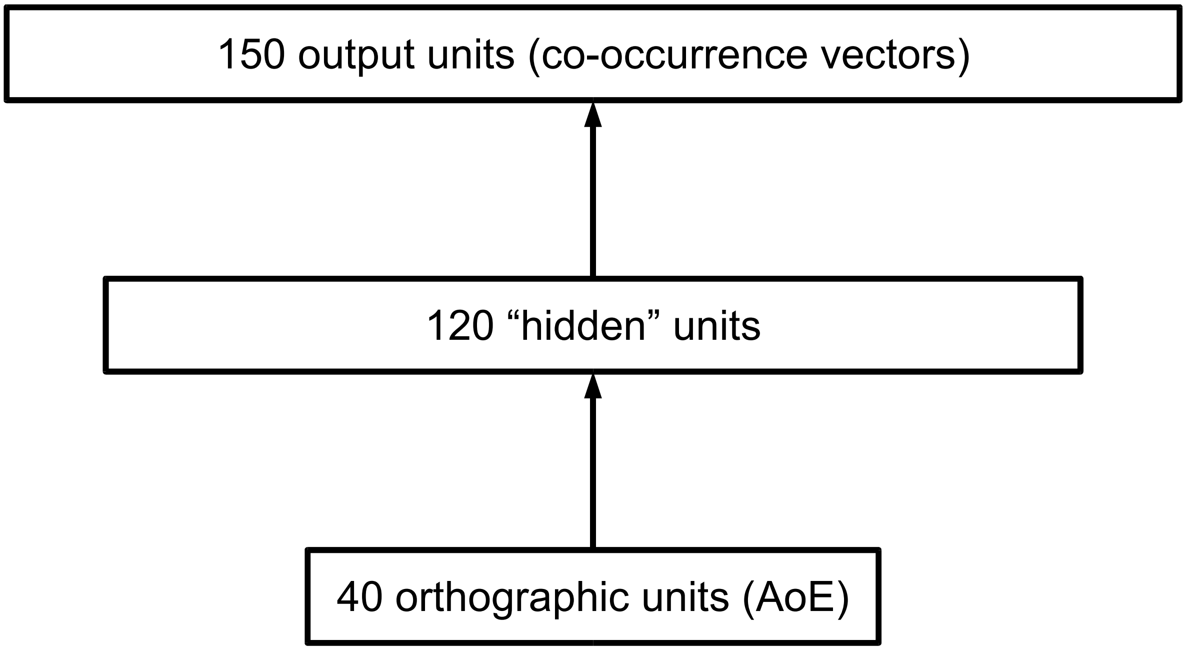

150 output units (co-occurrence vectors)
120 “hidden” units
40 orthographic units (AoE)

Figure 2:

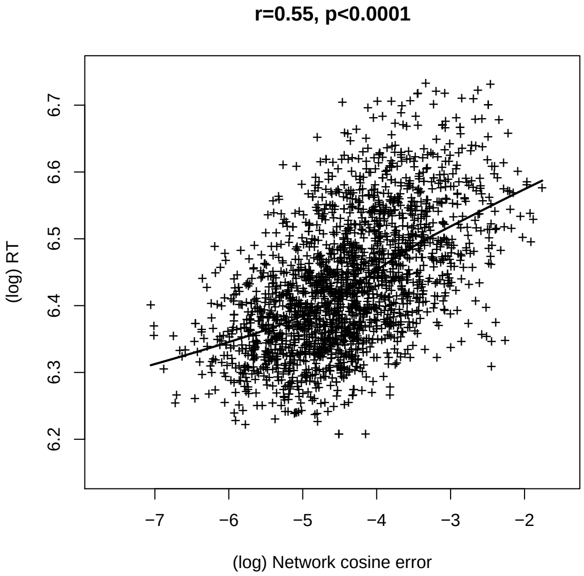

r=0.55, p<0.0001
(log) RT
6.2
6.3
6.4
6.5
6.6
6.7
−7
−6
−5
−4
−3
−2
(log) Network cosine error

Figure 3:

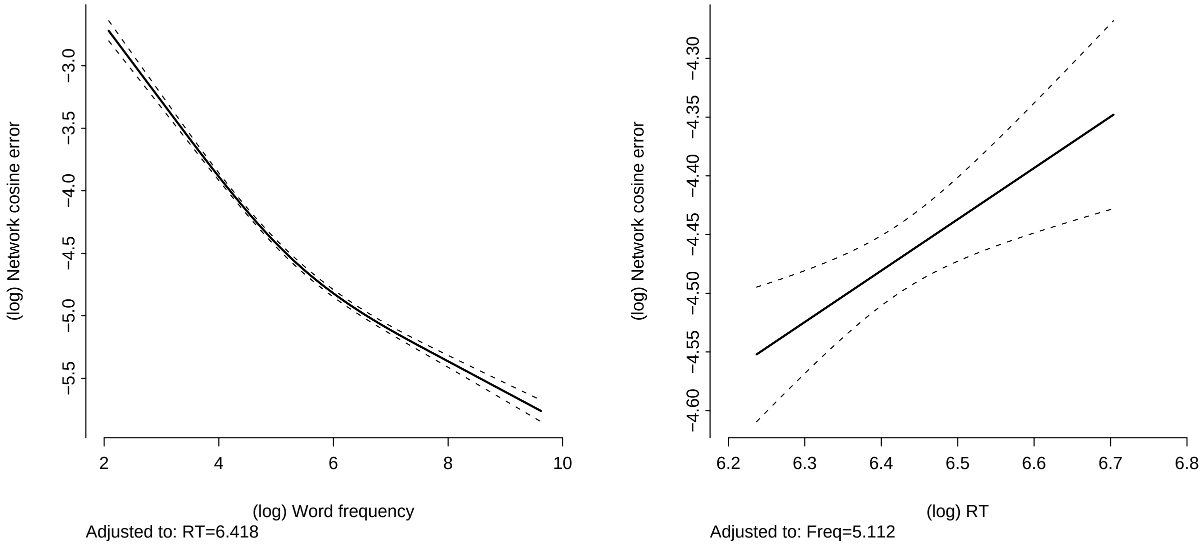
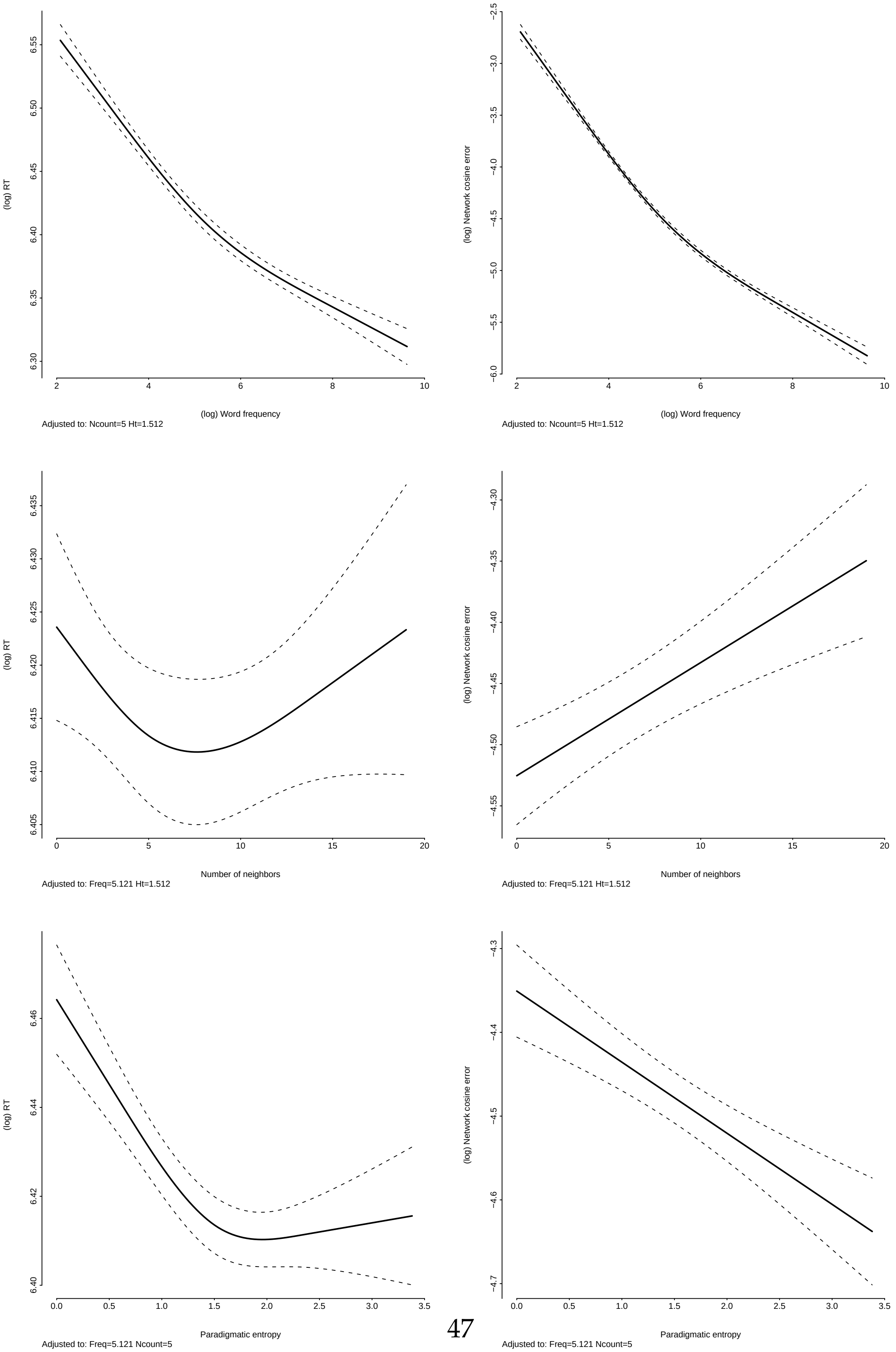

(log) Network cosine error
−3.0
−3.5
−4.0
−4.5
−5.0
−5.5
2
4
6
8
10
(log) Word frequency
Adjusted to: RT=6.418
(log) Network cosine error
−4.30
−4.35
−4.40
−4.45
−4.50
−4.55
−4.60
6.2
6.3
6.4
6.5
6.6
6.7
6.8
(log) RT
Adjusted to: Freq=5.112

Figure 4:



(log) RT
6.30
6.35
6.40
6.45
6.50
6.55
2
4
6
8
10
(log) Word frequency
Adjusted to: Ncount=5 Ht=1.512
(log) Network cosine error
−6.0
−5.5
−5.0
−4.5
−4.0
−3.5
−3.0
−2.5
2
4
6
8
10
(log) Word frequency
Adjusted to: Ncount=5 Ht=1.512
(log) RT
6.405
6.410
6.415
6.420
6.425
6.430
6.435
0
5
10
15
20
Number of neighbors
Adjusted to: Freq=5.121 Ht=1.512
(log) Network cosine error
−4.55
−4.50
−4.45
−4.40
−4.35
−4.30
0
5
10
15
20
Number of neighbors
Adjusted to: Freq=5.121 Ht=1.512
(log) RT
6.40
6.42
6.44
6.46
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Paradigmatic entropy
Adjusted to: Freq=5.121 Ncount=5
(log) Network cosine error
−4.7
−4.6
−4.5
−4.4
−4.3
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Paradigmatic entropy
Adjusted to: Freq=5.121 Ncount=5

Figure 5:

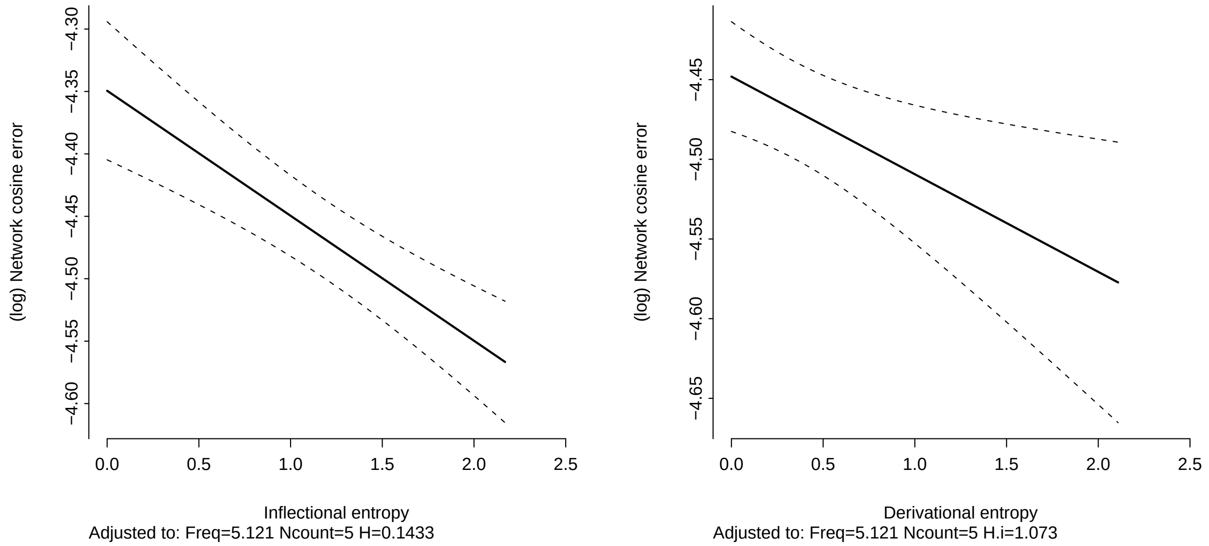
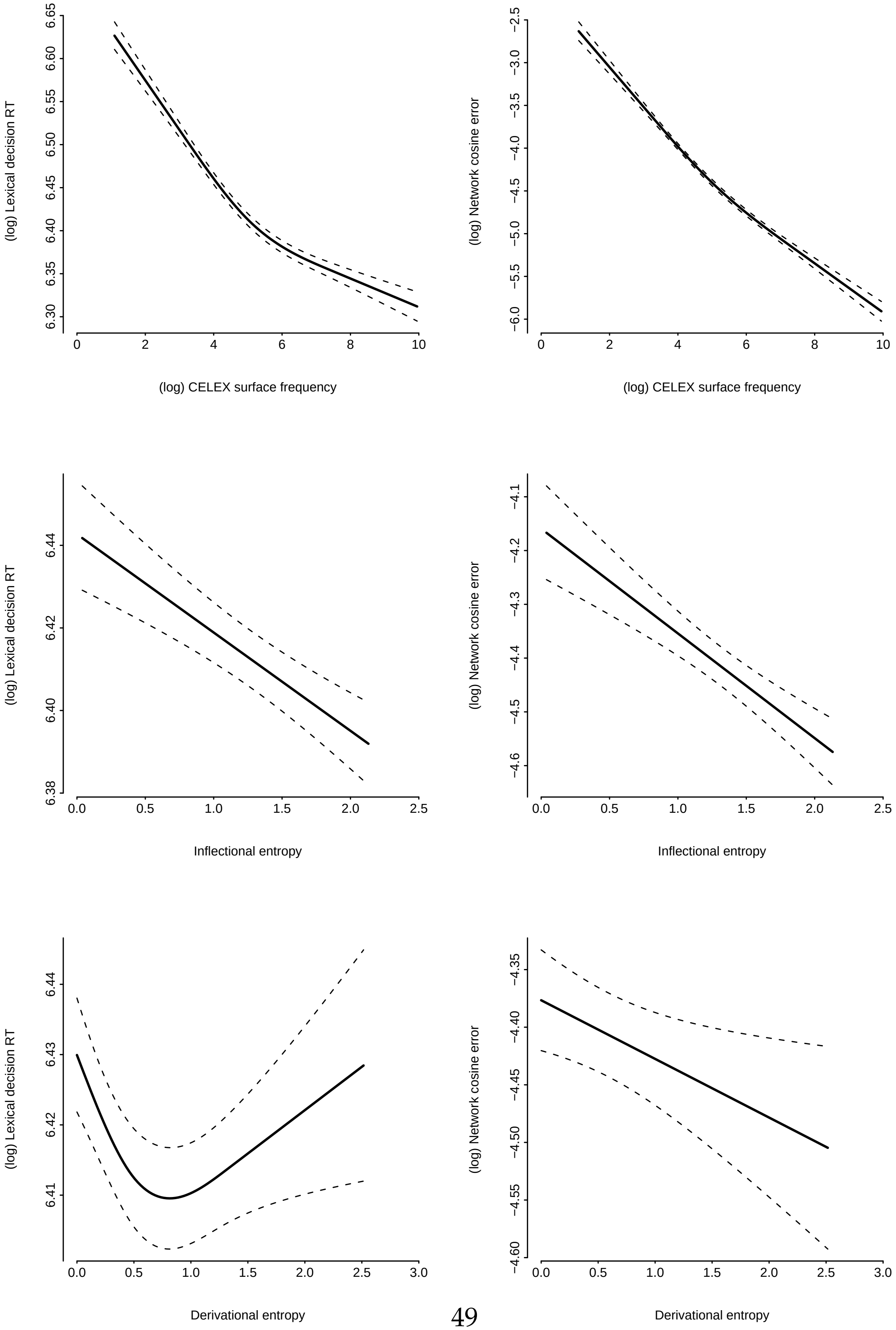
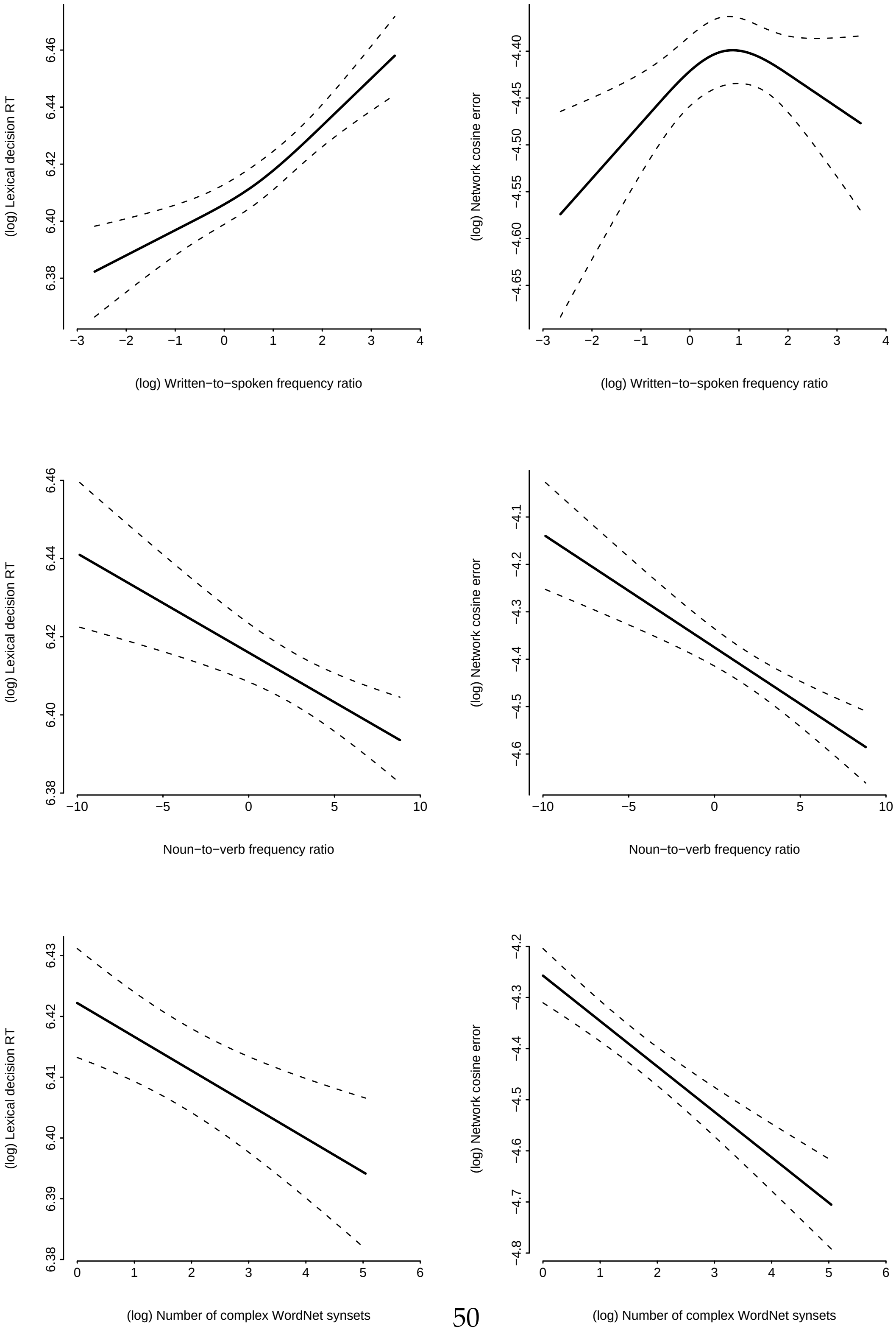

(log) Network cosine error
−4.30
−4.35
−4.40
−4.45
−4.50
−4.55
−4.60
0.0
0.5
1.0
1.5
2.0
2.5
Inflectional entropy
Adjusted to: Freq=5.121 Ncount=5 H=0.1433
(log) Network cosine error
−4.45
−4.50
−4.55
−4.60
−4.65
0.0
0.5
1.0
1.5
2.0
2.5
Derivational entropy
Adjusted to: Freq=5.121 Ncount=5 H.i=1.073

Figure 6:



(log) Lexical decision RT
6.30
6.35
6.40
6.45
6.50
6.55
6.60
6.65
0
2
4
6
8
10
(log) CELEX surface frequency
(log) Network cosine error
−6.0
−5.5
−5.0
−4.5
−4.0
−3.5
−3.0
−2.5
0
2
4
6
8
10
(log) CELEX surface frequency
(log) Lexical decision RT
6.38
6.40
6.42
6.44
0.0
0.5
1.0
1.5
2.0
2.5
Inflectional entropy
(log) Network cosine error
−4.6
−4.5
−4.4
−4.3
−4.2
−4.1
0.0
0.5
1.0
1.5
2.0
2.5
Inflectional entropy
(log) Lexical decision RT
6.41
6.42
6.43
6.44
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Derivational entropy
(log) Network cosine error
−4.60
−4.55
−4.50
−4.45
−4.40
−4.35
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Derivational entropy

Figure 7:


(log) Lexical decision RT
6.38
6.40
6.42
6.44
6.46
−3
−2
−1
0
1
2
3
4
(log) Written−to−spoken frequency ratio
(log) Network cosine error
−4.65
−4.60
−4.55
−4.50
−4.45
−4.40
−3
−2
−1
0
1
2
3
4
(log) Written−to−spoken frequency ratio
(log) Lexical decision RT
6.38
6.40
6.42
6.44
6.46
−10
−5
0
5
10
Noun−to−verb frequency ratio
(log) Network cosine error
−4.6
−4.5
−4.4
−4.3
−4.2
−4.1
−10
−5
0
5
10
Noun−to−verb frequency ratio
(log) Lexical decision RT
6.38
6.39
6.40
6.41
6.42
6.43
0
1
2
3
4
5
6
(log) Number of complex WordNet synsets
(log) Network cosine error
−4.8
−4.7
−4.6
−4.5
−4.4
−4.3
−4.2
0
1
2
3
4
5
6
(log) Number of complex WordNet synsets