Modelling Paradigmatic Effects in Visual Word Recognition

CHAPTER 10

An extended version of this chapter will be submitted as Fermín Moscoso del Prado Martín and R. Harald Baayen: BC–DC: A broad-coverage distributed connectionist model of visual word recognition.

Abstract

In this study we describe a distributed connectionist model of visual word recognition. The purpose of this model is to explore how the paradigmatic entropy effects described by Moscoso del Prado Martín, Kostić, and Baayen (2003) can arise in a probabilistic model of lexical processing. We present a model that learns to produce at its output the vectorized semantic representation of a word, the vectorized orthographic representation of which is presented at the input of the model. After training, we compare the outputs of the model with the lexical decision latencies for large sets of English monomorphemic nouns and verbs. Finally, we show that a network with these characteristics exhibits paradigmatic entropy effects similar to those observed by participants in visual lexical decision.

Introduction

Moscoso del Prado Martín, Kostić, and Baayen (2003) introduced a quantitative measure of the support that the recognition of a word receives from the inflectional paradigms to which that word belongs, its inflectional entropy. They define the inflectional entropy of an inflectional paradigm $\mathcal{P}(b) = \{w_1, w_2, \ldots, w_n\}$, were the w_i are the different inflectional variants of the base form b, as:

$$H_i(b) = -\sum_{w_i \in \mathcal{P}(b)} P(w_i|b) \log_2 P(w_i|b). \tag{10.1}$$

In this equation, $P(w_i|b)$ represents the probability of the surface (inflected) form being w_i given that the base form is known to be b. If $F(w_i)$ is the frequency of the inflected form w_i, and $F(b) = \sum_{w_i \in \mathcal{P}(b)} F(w_i)$ is the frequency of the base form b, the sum of the frequencies of all its inflectional variants, then the probability of the surface form w_i occurring, given that the base form is b, is $P(w_i|b) = F(w_i)/F(b)$.

Moscoso del Prado Martín and colleagues showed that the inflectional entropy of a word is a co-predictor of response latencies in visual lexical decision in Dutch. They presented evidence that the information residual of word, defined in terms of the inflectional entropy of a word, its surface frequency, and its derivational entropy (a measure akin to inflectional entropy calculated over derivational paradigms) predicts response latencies better than any combination of surface frequency, base frequency, cumulative root frequency, and morphological family size.

The measures introduced by Moscoso del Prado Martín and colleagues provide a simple way to quantify the support that a word receives from its morphological paradigms. These measures are calculated over a 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. The predictive value for RTs of measures calculated from such tree structures would arise naturally in decompositional models of morphological processing in which a word is processed through probabilistic activation of its constituent stems and affixes.

At first sight, the predictivity of inflectional entropy calculated on the basis of morphological trees might seem problematic for models of morphological processing that do not make use of discrete representations of the stems or base forms of complex words. In distributed connectionist models of lexical processing (Gaskell & Marslen-Wilson, 1997; Plaut & Booth, 2000, Plaut & Gonnerman, 2000; Seidenberg & Gonnerman, 2000), systematic correspondences between similarities in form and similarities in meaning lead to morphologically related words generating similar patterns of activation, that capture their morphological relations without explicit activation of a shared ‘stem’ unit.

These models have the advantage of being able to capture various graded effects arising from systematic pairings between form and meaning. Bergen (2003) 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 way a similar to the priming effects that have been observed for morphologically related words. Bergen shows that this effect is not solely due to orthographic or semantic similarity between the prime-target pairs. Boudelaa and Marslen-Wilson (2001) report a similar effect for Arabic words that share groups of two consonants. However, these groups of two consonants by themselves do not constitute a full morphological unit in the Arabic lexicon in the same sense that the three-consonantal stems 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 decomposable units. Non-decompositional theories of lexical processing such as distributed connectionist models and Bybee’s network model (Bybee, 1985) are better suited to account for these effects, as they do not depend on the decomposition of a complex word into discrete morphemes.

The question addressed in this study is whether non-decompositional distributed connectionist models can account for the observed predictivity of paradigmatic entropy, a measure that is calculated on the basis of the probability distributions of discrete morphological forms in a hierarchical tree representing the decompositional dependencies between the members of the paradigm.

Another issue that has caused considerable discussion in the psycholinguistic literature is that of past-tense formation. 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) and neuro-psychological double dissociations (e.g., Miozzo, in press), 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; Plunkett & Marchman, 1993; Plunkett & Juola, 1999; Rumelhart & McClelland, 1986).

A factor that has not been given enough 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 regularity. 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 (2003) have added a new dimension to the debate by showing that regulars and irregulars tend to be clustered in meaning, and crucially, that some of the processing differences between regular and irregular verbs that were found for past-tense forms appear also in the processing of the (completely 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. Baayen and Moscoso del Prado Martín also report that one of the main differences between regular and irregular verbs is that, in general, irregular verbs have a higher inflectional entropy than regular verbs.

This raises the question of whether a single-route model of lexical processing might mirror the experimental processing differences observed for the uninflected stems of regular and irregular verbs, as a function of the differences in their meanings.

In what follows, we first describe a distributed connectionist model that was trained to produce the semantic representation of a word from a representation of its orthography. As orthographic representations, we used the Accumulation of Expectations (AoE) vectors for English orthographic forms described by Moscoso del Prado Martín, Schreuder, and Baayen (2003). As a representation of a word's meaning, we used the semantic vectors developed by Moscoso del Prado Martín and Sahlgren (2002), which provide semantic representations for a large set of the inflected forms of English words.

Next, we investigate whether the responses of the model to a large set of words from the Balota, Cortese, and Pilotti (1999) database reflect the pattern of response latencies shown by the participants in that study. We then examine whether the participants in the Balota et al. database show the inflectional and derivational entropy effects observed for Dutch by Moscoso del Prado Martín, Kostić, and Baayen (2003), and whether the network shows similar paradigmatic effects. We then proceed to investigate whether our model captures the differences in the processing of present-tense forms of regular and irregular verbs. Finally, we address the possible confounds that may arise between paradigmatic entropy measures, and other purely formal measures such as neighborhood size. We conclude by outlining the implications of our model for current theories of visual lexical processing.

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 10.1. 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.

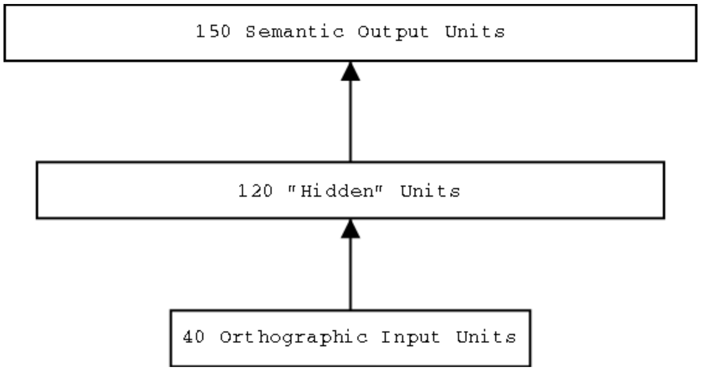

Figure 10.1: General architecture of the model. The lines represent trainable all-to-all connections between the units in two layers.

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 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 constructed orthographic vectors using the AoE technique described by Moscoso del Prado Martín, Schreuder, and Baayen (2003), and we associated them with the semantic vectors for English words described by Moscoso del Prado Martín and Sahlgren (2002).

Network Training

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 proportional to its frequency of occurrence in the corpus from which the semantic vectors were built. Training was done by 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.

Results

Once the network had been trained, we evaluated its performance on recognizing a large set of words from the Balota et al. (1999) study. In English, nouns and verbs differ with respect to the number of inflectional variants that they may have. While most English nouns have only two inflected forms (singular and plural), most English verbs have at least three inflectional variants (present-tense, past-tense/past-participle, and gerund), with a maximum of five different forms. This causes the distribution of inflectional entropies of nouns to be different from the distribution of inflectional entropies for verbs, with the latter having a higher inflectional entropy on average. This difference in the entropy distributions might also give rise to differences in the effects of inflectional entropy for nouns and verbs. We take this into consideration by analyzing separately the 1,295 monomorphemic English nouns and 795 monomorphemic English verbs in our dataset.

Nouns

We selected from the Balota et al. dataset those monomorphemic nouns that were also present in our training set. As in many other connectionist studies (e.g., Shillcock, Ellison, & Monaghan, 2000), we use the distance between the model's output for a word and its correct value as an analog of the reaction time measures for human participants. In other words, we view RTs as reflecting the processing load of mapping a word token in the input stream onto its associated semantic representation, and we are interested in ascertaining whether the cosine distances of our model, which quantify the complexity of this mapping, correlate with the RTs. Both RTs and cosine distances are lognormally distributed as revealed by quantile-quantile plots, and we therefore used their logarithm transform in all analyses. The Pearson correlation between the logarithm of the average reaction time produced by the group of young participants in the Balota et al. database with the logarithm of the cosine distance for the monomorphemic nouns was $0.55 (p < 0.0001)$. This correlation is illustrated in Figure 10.2. Note that the non-parametric regression line indicates that, in general, an increase in the average log reaction time of the young participants, corresponds linearly to the linear increase in the network's log cosine distance.

In order to ascertain that the participants in the English lexical decision study were showing inflectional entropy effects similar to those reported for Dutch by Moscoso del Prado Martín, Kostić, and Baayen (2003), we fitted a linear regression model with the logarithmic average RT for young participants as the dependent variable, and the logarithm of a word's surface frequency and inflectional entropy (calculated according to Equation 10.1) as independent variables. A sequential analysis of variance revealed significant main effects of surface frequency ($F(1, 1293) = 755.396, p < 0.0001$) and inflectional entropy ($F(1, 1293) = 34.78, p < 0.0001$, after having partialled out the effect of surface frequency), with no significant interaction ($F < 1$). The coefficients for the effects of both independent variables were negative ($\beta = -0.0365, t = -26.85, p < 0.0001$ for surface frequency, and $\beta = -0.0262, t = -5.90, p < 0.0001$ for inflectional entropy), indicating that both of these effects were facilitatory: Words with a high frequency or a high inflectional entropy were recognized faster. Introducing base frequency (i.e., the summed frequency of all the inflectional variants of a word) as an additional predictor after partialling out the effects of surface frequency and inflectional entropy resulted in a marginally significant main effect ($F(1, 1288) = 2.80, p < 0.0947$).

We examined whether the network was showing the surface frequency and inflectional entropy effects similar to those observed for the participants by means of a linear model fitting the model's logarithmic cosine distance of a word as a function of the logarithm of that word's frequency of occurrence during training

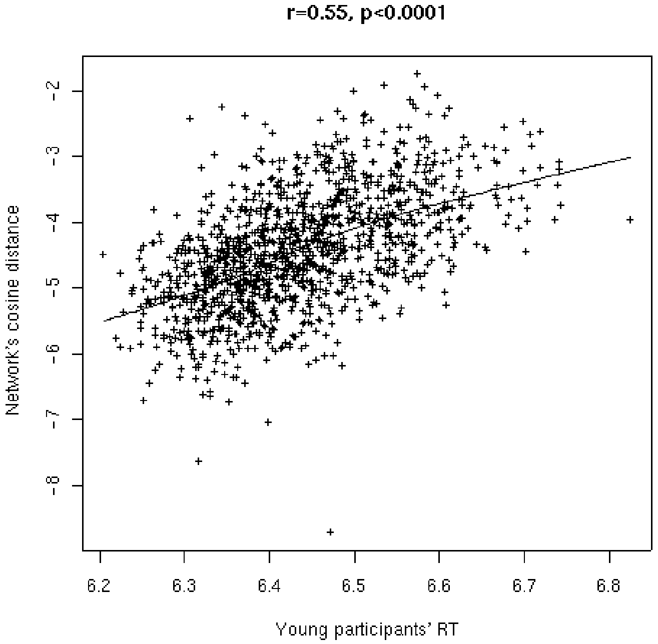

Figure 10.2: Comparison (in bilogarithmic scale) between the young participants' average reaction time to English monomorphemic nouns (horizontal axis) provided by Balota et al. (1999), with the model's cosine distance for those same nouns (vertical axis). The line represents a non-parametric regression (Cleveland, 1979).

(its surface frequency) and its inflectional entropy (calculated according to Equation 10.1 on the basis of only those inflectional variants of a word that appeared in the training set). This revealed significant main effects of surface frequency ($F(1,1293) = 2958.69, p < 0.0001$) and inflectional entropy ($F(1,1293) = 54.92, p < 0.0001$, after partialling out the effect of frequency) without any significant interaction ($F < 1$). As in the case of the participants, both of these effects had negative coefficients ($\beta = -0.4390, t = -53.52, p < 0.0001$ for word frequency, and $\beta = -0.1982, t = -7.41, p < 0.0001$ for inflectional entropy), indicating that words with a high frequency or a high inflectional entropy produce smaller cosine distances. Adding base frequency (considering only those inflectional variants of a word that appeared in the training set) into the regression after having partialled out the effects of surface frequency and inflectional entropy did not result in a significant main effect ($F < 1$).

Verbs

We selected from the Balota et al. (1999) dataset those monomorphemic verbs in first person singular form that were also present in our training set. The Pearson correlation between the logarithm of the average RT of the young participants and the network's log cosine distance for those verbs was $r = 0.54$ ($p < 0.0001$). Figure 10.3 illustrates this correlation. Again, we observe a roughly linear relation between the participants' log RTs for verbs and the network's log cosine distance for those verbs.

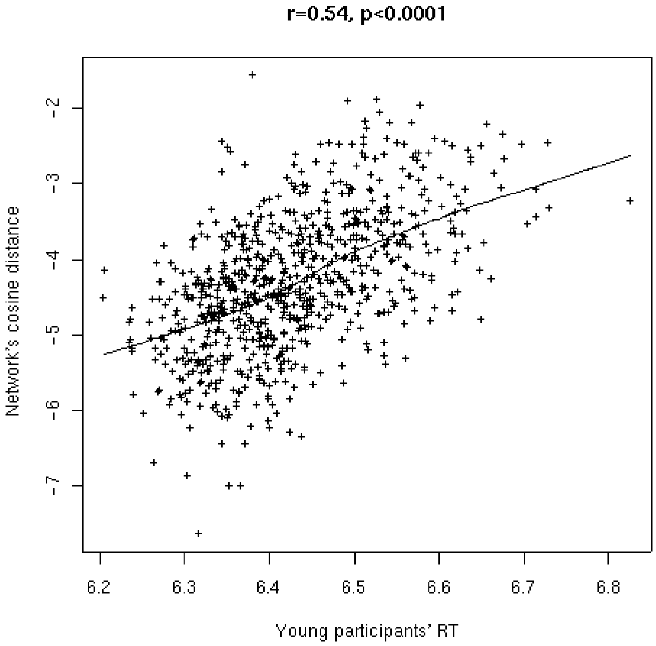

Figure 10.3: Comparison (in bilogarithmic scale) between the Balota et al. (1999) young participants' average reaction time to English monomorphemic verbs (horizontal axis), with the model's cosine distance for those same verbs (vertical axis). The line represents a non-parametric regression (Cleveland, 1979).

A linear regression model fitted to the young participants' average RT for the monomorphemic verbs, with surface frequency and inflectional entropy as the independent variables, revealed significant main effects of surface frequency ($F(1,793) = 453.73, p < 0.0001$), and inflectional entropy ($F(1,793) = 47.08, p < 0.0001$ after partialling out the effect of surface frequency), without any significant interaction ($F < 1$). Both main effects had facilitatory coefficients (surface frequency: $\beta = -0.0309, t = -20.09, p < 0.0001$; inflectional entropy: $\beta = -0.0385, t = -6.86, p < 0.0001$). As before, there was no significant main effect of base frequency ($F < 1$)

after partialling out the effects of surface frequency and inflectional entropy, and inflectional entropy still showed a significant main effect when we introduced it in the regression after base frequency ($F(1, 792) = 31.67, p < 0.0001$). Again, this confirms for English verbs the effect of inflectional entropy reported by Moscoso del Prado Martín and colleagues for Dutch words.

We fitted a linear regression model with the same independent variables to the logarithm of the network's cosine distances. The same effects emerged: a main effect of surface frequency ($F(1, 793) = 1609.57, p < 0.0001$), a main effect of inflectional entropy ($F(1, 793) = 29.64, p < 0.0001$ after partialling out the effect of surface frequency), and only a marginally significant effect of base frequency ($F(1, 792) = 2.83, p = 0.0929$) after partialling out surface frequency and inflectional entropy. The marginal effect of base frequency disappears ($F < 1$) when base frequency is included into the regression before inflectional entropy. Inflectional entropy remains significant when entered into the model after base frequency ($F(1, 792) = 29.94, p < 0.0001$). As in the case of the reaction times, the effects of surface frequency ($\beta = -0.4082, t = -38.91, p < 0.0001$) and inflectional entropy ($\beta = -0.2084, t = -5.45, p < 0.0001$) were both facilitatory.

Regular and Irregular Verbs

To investigate the issue of the differential processing of the present-tense forms of regular and irregular verbs, we included as additional independent variable the verb's regularity (i.e., *regular* vs. *irregular*).

The analysis of the network's cosine distance revealed a marginally significant main effect of regularity ($F(1, 791) = 3.62, p = 0.0574$, after partialling out the effects of surface frequency and inflectional entropy) and a significant frequency by regularity interaction ($F(1, 791) = 8.62, p = 0.0034$, after partialling out the main effects). The interaction had a negative coefficient for the regular verbs ($\beta = -0.9, t = -2.93, p = 0.0034$), indicating that, after partialling out the other variables, regular verbs are more sensitive to the frequency effect. A similar pattern emerged in the analysis of the RTs. We did not find a main effect of regularity ($F < 1$, after partialling out frequency and inflectional entropy), and we observed the same frequency by regularity interaction ($F(1, 791) = 6.03, p = 0.0143$, after partialling out the main effects) with a negative coefficient ($\beta = -0.01, t = -2.46, p = 0.0143$).

These results confirm those of Baayen and Moscoso del Prado Martín (2003) in that there is indeed and effect of regularity on the response latencies for the present-tense forms of verbs. However, neither in the participants nor in the network can this effect be completely attributed to the difference in inflectional entropy between regulars and irregulars, as shown by the observed frequency by regularity interactions. Recall here that the surface frequency and inflectional entropy counts that were used for the network were exact, in the sense they were calculated on the actual frequency distributions to which the network was exposed. Therefore, the crucial interaction in the network cannot be attributed to regularity being a mere correction to the other two counts.

Neighborhood Size

At this point, it is necessary to consider a possible confound in our data. Although we are arguing that the inflectional entropy effect that we are observing arises due to the network creating morphological generalizations over form-meaning regularities in its training set, it could be argued that we are only observing an effect of form similarity. Inflectionally related forms, independently of their relationship in meaning, tend to be (by definition) very similar in form. This could lead the model to be affected by the raw number of orthographically similar words in its training set. Such effects have been widely reported in the literature on visual word recognition. Coltheart, Davelaar, Jonasson, and Besner (1977) reported that words with large orthographic neighborhoods are recognized slower in a visual lexical decision task, where lexical neighborhood is defined as the number of words in the lexicon that differ in only one letter from the target. In contrast, the effects of inflectional entropy that we observed both in the network and in the experimental data were facilitatory. Words with large inflectional entropies were recognized faster by the participants and elicited lower distances in the network. However, a number of studies, using word naming tasks, have reported facilitatory effects of orthographic neighborhoods (e.g., Andrews, 1989; 1992). As our network is never actually performing a pure visual lexical decision task, it might be argued that we are observing effects more similar to those found in the word naming paradigm, with our inflectional entropy effect being unrelated to the true effects of inflectional entropy observed for the participants in visual lexical decision. This would imply that our inflectional entropy effect found in the network would be more related to the facilitatory neighborhood size effect in naming, and thus reflect only properties of orthographic form processing.

In order to address this potential confound, we fitted a new linear regression model with the log cosine distance to all nouns and verbs in the previous analyses as the dependent variable, and log surface frequency, inflectional entropy and log neighborhood size (as calculated from the CELEX database) as independent variables. A sequential analysis of variance revealed significant main effects of surface frequency ($F(1, 2001) = 4309.65, p < 0.0001$), inflectional entropy ($F(1, 2001) = 38.20, p < 0.0001$ after partialling out the effect of frequency), and neighborhood size ($F(1, 2001) = 12.83, p = 0.0005$, after partialling out the effect of frequency and inflectional entropy), and a significant interaction between frequency and neighborhood size ($F(1, 2001) = 6.21, p = 0.0128$, after partialling out the main effects). As in the previous analyses, we found negative, facilitatory, coefficients for both frequency ($\beta = -0.47, t = -30.27, p < 0.0001$) and inflectional entropy ($\beta = -0.13, t = -6.22, p < 0.0001$). The main effect of neighborhood size did not have a significant coefficient ($\beta = -0.06, t = -1.25, p = 0.2100$) and the frequency by neighborhood size interaction had a positive, inhibitory coefficient ($\beta = 0.02, t = 2.49, p = 0.0128$).

A similar analysis with the log RT as dependent variable and the same independent variables as above, revealed significant main effects of frequency ($F(1, 2001) = 1167.79, p < 0.0001$) and inflectional entropy ($F(1, 2001) = 88.88, p < 0.0001$), with the main effect of neighborhood size not reaching significance ($F(1, 2001) = 3.05, p = 0.0809$) and a significant interaction between surface frequency and neighborhood size ($F(1, 2001) = 13.92, p < 0.0002$, after partialling out all the main effects). Of the significant effects, the coefficients of the main effects of frequency ($\beta = -0.04, t = -17.56, p < 0.0001$) and inflectional entropy ($\beta = -0.03, t = -9.28, p < 0.0001$) were facilitatory, while the frequency by neighborhood size interaction was inhibitory ($\beta = 0.005, t = 3.73, p = 0.0002$).

These analyses indicate that inflectional entropy is facilitatory in nature both for the response latencies and for the model's cosine distances. By contrast, neighborhood size emerges as an inhibitory interaction with frequency, again for both the distances and the RTs, in line with the inhibitory neighborhood effects of neighborhood size in visual lexical decision reported by Coltheart et al. (1977). We therefore conclude that our model captures essential aspects of the participants' sensitivity to frequency, form similarity, and inflectional similarity in visual lexical decision. We also conclude that inflectional entropy and neighborhood size effects in our network reflect different aspects of processing, with inflectional entropy capturing the form-meaning correlations present in the inflectional paradigms of the training set, and with neighborhood size reflecting pure form similarities.

Derivational Entropy

After having ascertained that both our model and the participants showed comparable effects of inflectional entropy, we turn to investigate the effects of derivational entropy. Once more, we do this by means of two linear regression models, one with the RTs and the other with the cosine distances as the dependent variable. We calculated the derivational entropy according to the definition provided by Moscoso del Prado Martín, Kostić, and Baayen (2003), using only those words that appeared in our networks' training corpus.

A linear regression fit to the logarithm of the RT, with log surface frequency, inflectional entropy, and derivational entropy as independent variables, revealed main effects of frequency ($F(1, 2000) = 1165.75, p < 0.0001$), inflectional entropy ($F(1, 2000) = 88.72, p < 0.0001$, after partialling out the effect of frequency), and significant interactions of frequency by derivational entropy ($F(1, 2000) = 6.88, p = 0.0088$, after partialling out the main effects), and of inflectional by derivational entropy ($F(1, 2000) = 5.22, p = 0.0225$). The effect of derivational entropy did not reach significance after partialling out the effects of word frequency and inflectional entropy ($F(1, 2000) = 2.35, p = 0.1255$). The marginal coefficients of the main effects of frequency ($\beta = -0.04, t = -27.54, p < 0.0001$), inflectional entropy ($\beta = -0.03, t = -8.88, p < 0.0001$), and derivational entropy ($\beta = -0.06, t = -3.69, p = 0.0002$), were all negative, while the coefficients of the frequency by derivational entropy interaction ($\beta = 0.01, t = 2.52, p = 0.0119$) and the inflectional by derivational entropy ($\beta = 0.02, t = 2.28, p = 0.0225$) were both positive.

A similar linear regression model with log cosine distance as the dependent variable, and the same independent variables as before, revealed main effects of frequency ($F(1, 2000) = 1083.91, p < 0.0001$), inflectional entropy ($F(1, 2000) = 38.18, p < 0.0001$, after partialling out the effect of frequency), and derivational entropy ($F(1, 2000) = 6.16, p = 0.0132$, after controlling for frequency and inflectional entropy, and significant interactions of frequency by derivational entropy ($F(1, 2000) = 5.49, p = 0.0192$, after partialling out the main effects), and of inflectional by derivational entropy ($F(1, 2000) = 6.89, p = 0.0087$). The marginal coefficients of the main effects of frequency ($\beta = -0.44, t = -52.55, p < 0.0001$) and inflectional entropy ($\beta = -0.09, t = -3.42, p = 0.0007$) were again negative, with the coefficient of the interaction between frequency and inflectional entropy being positive ($\beta = 0.04, t = 4.46, p = 0.0140$), and the interaction between both entropies negative ($\beta = -0.12, t = -2.65, p = 0.0087$). The marginal coefficient for the effect of derivational entropy was negative, but not significant ($\beta = -0.13, t = -1.35, p = 0.1768$).

In both regressions, we found significant interactions of derivational entropy by frequency, and derivational by inflectional entropy. Both for the participants and for the network, the marginal coefficient of the effect of derivational entropy was negative, but not significant in the case of the network. In the case of the participants, this effect does not reach significance in a sequential analysis of variance, after having partialled out the effects of surface frequency and inflectional entropy. Recall here that the derivational entropy was calculated from the distribution of words in which the model was trained. Therefore, for the model, the derivational entropy is an exact measure, while it is only an approximation of its value for the participants. We think that using a more accurate measure of derivational entropy for the participants should also reveal a significant main effect, similar to that reported for Dutch by Moscoso del Prado Martín and colleagues. An additional problem is the difference in signs between marginal coefficients for the inflectional by derivational entropy interactions in the models of RTs and cosine distances. However, visual examination of the scatterplots for both models revealed that there is a lot of non-linearity in these interactions, making the coefficients for the linear effect unreliable. The emergence of this derivational entropy effect is a clear indication of the form-meaning interactions present in the model. Moscoso del Prado Martín and Sahlgren (2002), indicated that the semantic vectors that we have used in this simulation contained a detailed representation of inflectional relations between words, thus it could be argued that the effect inflectional entropy arises solely due to the semantic neighborhoods. However, Moscoso del Prado Martín and Sahlgren also reported that, using these semantic vectors only, one could not detect many derivational relations. This is a clear sign that the model must be exploiting the correspondences between form and meaning in order for the effect of derivational entropy to arise. However, further research is needed on these issues to clarify the consequences of using semantic vectors that are more sensitive to derivational relations, and more accurate calculations of derivational entropy for the participants.

Age of Acquisition

Another issue that has caused a considerable amount of debate in the literature is the effect of Age of Acquisition (AoA; Carroll & White, 1973). Words that are acquired early in development are recognized faster than words that are acquired later in life, independently of their frequency. Morrison and Ellis (1995) argued that connectionist networks would not be able to show effects of AoA given that they suffer from ‘catastrophic forgetting’, by which patterns that are acquired later in training 'overwrite' the representation of patterns that appeared earlier during training. In this respect, Ellis and Lambon Ralph (2000) proved that, when a network is trained on a set of early and a set of late patterns, it does show AoA effects. Smith, Cottrell, and Anderson (2001) reported effects of AoA on a network's error. Interestingly, instead of manipulating the moment during training at which a pattern was presented to the network for the first time, they measured the moment in training at which a given pattern is learned, without any manipulations on order of pattern presentation. This finding suggests that the AoA effect might arise from the inherent difficulty of learning (and processing) a particular pattern, independently of any developmental considerations. In simulations using artificial datasets, Anderson and Cottrell (2001) tested the hypothesis that the AoA effect reflects the patterns of similarity between the items in the dataset, that is, words which are similar (in form, meaning, or both) to many others, are easier to learn and faster to process. If the hypothesis put forward by Anderson and Cottrell is true, given that our model is trained on a realistic sample of the English language, it should show AoA effects in a similar way to participants, even though the order of presentation of the words during training is completely arbitrary.

In order to compare the effects of AoA in our network with those shown by human participants, we obtained AoA ratings for 521 words from the MRC Psycholinguistic Database (Coltheart, 1981) and we combined them with the reaction times for young participants to those same items from the Balota et al. (1999) dataset, and the cosine distances obtained by our network for those words.

We fitted a linear regression model to the young participants' average RT, with surface frequency, inflectional entropy, derivational entropy, and AoA as the independent variables. A sequential analysis of variance revealed significant main effects of surface frequency ($F(1, 517) = 241.67, p < 0.0001$), inflectional entropy ($F(1, 517) = 42.85, p < 0.0001$ after partialling out the effect of surface frequency), and AoA ($F(1, 517) = 80.94, p < 0.0001$ after partialling out the effects of surface frequency, and inflectional entropy). Additionally, we observed a significant interaction between frequency and AoA ($F(1, 517) = 7.79, p = 0.0055$). After partialling out the remaining main effects, we did not observe any additional effect of derivational entropy in this dataset ($F < 1$). The main effect of AoA had an inhibitory coefficient ($\beta = 6.128 \cdot 10^{-4}, t = 5.36, p < 0.0001$) while the interaction had a facilitatory coefficient ($\beta = -5.416 \cdot 10^{-5}, t = -2.79, p = 0.0055$).

A linear regression model fitted with the same independent variables to the logarithm of the network's cosine distances revealed the same effects as above: a main effect of surface frequency ($F(1,517) = 438.12, p < 0.0001$), a main effect of inflectional entropy ($F(1,517) = 10.96, p = 0.0010$ after partialling out the effect of surface frequency), and a main effect of AoA ($F(1,517) = 6.88, p = 0.0090$) after partialling out surface frequency and inflectional entropy. Once more, there was a significant frequency by AoA interaction ($F(1,517) = 56.99, p < 0.0001$) and no significant effect of derivational entropy ($F < 1$) after partialling out the remaining variables. As in the case of the reaction time analyses, the main effect of AoA had a positive coefficient in the regression ($\beta = 1.326 \cdot 10^{-4}, t = 7.99, p < 0.0001$) and the frequency by AoA interaction had a negative coefficient in the regression ($\beta = -2.128 \cdot 10^{-5}, t = -7.55, p < 0.0001$).

These results support the hypothesis advanced by Anderson and Cottrell (2001) that the AoA effect reflects, at least in part, the position of a word in morpho-semantic space, independently of development of neural plasticity.

General Discussion

In this study, we have presented a broad coverage distributed connectionist model of visual word recognition. The model was trained to map distributed orthographic representations onto distributed semantic representations. 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. In this study, we have used a vocabulary of 48,260 different words to train our model. 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.

The key to this broad coverage lies in the use of truly distributed representations of word forms and meanings, as provided by the AoE representational paradigm (Moscoso del Prado Martín, Schreuder, & Baayen, 2003), and the realistic context-based semantic vectors of Moscoso del Prado Martín and Sahlgren (2002). A corpus-based co-occurrence approach to semantic representation is based on realistic assumption of co-occurrence being one of the sources of information that humans use to determine the meaning of a word (e.g., Boroditsky & Ramscar, 2003; McDonald & Ramscar, 2001). Additionally, it overcomes the bottleneck for realistic models caused by having to rely on hand-crafted semantic representations.

The coding scheme used for word forms has the advantage that it obviates the need for slot-based templates that require manual preprocessing of the words form. The use of slot-based structures for the coding of word forms, and 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). Moreover, slot-based representations require arbitrary decisions on alignment for coding the similarities and dissimilarities of onsets, word-centers, and codas, as in the onsets of the Dutch words *sap*, *stap*, and *tap*.

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 (2003) in that there are indeed important 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).

The response patterns produced by the 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 mentioned above, our contextual semantic representations do not fully capture the type of semantic relations that are present in derivational morphology, and therefore our model does not completely mirror the effects caused by derivational paradigms. Additionally, other variables that are known to correlate with visual lexical decision latencies, such as concreteness or imageability, are absent. Such effects can only be captured by a model that also includes sensory-motor information in its semantic representations (cf., Pulvermüller, 2002).

Additionally, our model also mirrored participants' behavior with respect to the Age of Acquisition effect. Our model produced significantly lower error scores for words that are acquired early by people, according to Age of Acquisition norms. Crucially, our network's training regime did not follow any developmental considerations. This supports the proposal of Anderson and Cottrell (2001) that Age of Acquisition reflects the similarity structure of words in the lexicon, instead of decreases in neural plasticity during development.

With surface frequency, inflectional entropy, derivational entropy, and neighborhood size, we are able to account for approximately two thirds of the variance present in the model's cosine distances. This suggests that further research is necessary to understand the source of the remaining one third of the model's variance. We think that there are more psychologically relevant factors that are captured by the model. In particular, different types of effects that are claimed to arise at form recognition levels such as word length or bigram frequency need to be investigated. We leave these 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 morphological structure, as indicated by the effects of inflectional and derivational 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 just meaning similarity on its own. Instead, the effects emerge from the systematic form-meaning associations shared by the morphological variants of a word. In conclusion, we have shown that the paradigmatic entropy effects described by Moscoso del Prado Martín, Kostić, and Baayen (2003) 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).