Running head: The Bayesian Information-Theoretical Model of Lexical Processing

The Missing Link between Morphemic Assemblies and Behavioral Responses: a Bayesian Information-Theoretical model of lexical processing

Fermín Moscoso del Prado Martín[a], Aleksandar Kostić[b], and Dušica Filipović-Đurđević[b]

[a] MRC Cognition and Brain Sciences Unit, Cambridge, U.K.

[b] Laboratory for Experimental Psychology, University of Belgrade, Serbia and Montenegro

Address all correspondence to:
Dr. 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 X275
fax: +44 1223 359 062

Abstract

We present the Bayesian Information-Theoretical (BIT) model of lexical processing: A mathematical model illustrating a novel approach to the modelling of language processes. The model shows how a neurophysiological theory of lexical processing relying on Hebbian association and neural assemblies can directly account for a variety of effects previously observed in behavioral experiments. We develop two information-theoretical measures of the distribution of usages of a word or morpheme. These measures are calculated through unsupervised means from corpora. We show that our measures succesfully predict responses in three visual lexical decision datasets investigating the processing of inflectional morphology in Serbian and English languages, and the effects of polysemy and homonymy in English. We discuss how our model provides a neurophysiological grounding for the facilitatory and inhibitory effects of different types of lexical neighborhoods. In addition, our results show how, under a model based on neural assemblies, distributed patterns of activation naturally result in the arisal of discrete symbol-like structures. Therefore, the BIT model offers a point of reconciliation in the debate between distributed connectionist and discrete localist models. Finally, we argue that the modelling framework exemplified by the BIT model, is a powerful tool for integrating the different levels of the description of the human language processing system.

Introduction

Research in psycholinguistics during the last fifty years has provided us with a wealth of data on the detailed properties of lexical processing in the human mind. More recently, neuroimaging techniques have begun complementing this picture with detailed specifications of the spatio-temporal patterns of cortical activation that accompany language processing. Simultaneously, some theories detailing how language processing can take place in detailed neurobiological terms are currently becoming available, and receiving support from neuroimaging studies. However, there still seems to be a dissociation between the results obtained in behavioral studies, and the detailed neurobiological theories of language processing. In this study we argue that we are currently in the position to link both levels of explanation: behavioral and neurobiological. We demonstrate this by showing how a neurophysiological theory of lexical processing (Pulvermüller, 1999) can provide a direct explanation of several previously reported behavioral effects. For this purpose we develop a set of statistical information-theoretical tools that enable us to make quantitative predictions on behavioral responses based on an underlying neurophysiological theory, without the need for direct computational simulation.

Measures of lexical competition and facilitation

A large amount of psycholinguistic research has shown that the size of the phonological, orthographic, semantic and morphological 'neighborhoods' of words influence the time it takes for them to be recognized in lexical recognition tasks. Words with many phonological neighbors are reported to be recognized slower than words with few neighbors (Luce & Pisoni, 1998; Vitevitch & Luce ,1999). In contrast, in the orthographic domain, words with many orthographic neighbors are recognized faster than words with few neighbors in visual lexical decision (Andrews, 1989; 1992; 1997), while they appear to be responded to more slowly in visual identification tasks (Grainger & Seguí, 1990;). However, recent large-scale studies have shown that the effects of of orthographic neighborhood in lexical decision are more complex than previously thought (Baayen, 2005; Baayen, Feldman & Schreuder, 2005). These studies describe a non-linear u-shaped effect of neighborhood

size on lexical decision latencies (i.e., small neighborhoods produce facilitation, while large neighborhoods produce inhibition). The effect of orthographic neighborhood has also been found to correlate with the magnitude of the N400 component of the ERP signal (Holcomb, Grainger, & O'Rourke, 2002).

Similar effects have been observed in the domain of word meaning. Jastrzembski (1981) reported that, in visual lexical decision, semantically ambiguous words are responded to faster than semantically unambiguous words. Many other authors have replicated this result (e.g., Azuma & Van Orden, 1997; Borowsky & Masson, 1996; Kellas, Ferraro, & Simpson, 1988). An additional refinement to this ambiguity advantage was introduced by Rodd, Gaskell, and Marslen-Wilson (2002) who pointed out the need to distinguish between words having many unrelated meanings (homonymic) and words having many related senses (polysemous). They showed that, while polysemous words exhibit the previously described ambiguity advantage, homonymic words are in fact recognized slower. Parallel to what was found in the domain of word form (orthographic and phonological), the semantic neighborhood of a word can also have effects in opposite directions. This distinction has been confirmed in two recent neuromagnetic studies, that have also shown that both effects are reflected in different cortical sources of the M350 effect (Beretta, Fiorentino, & Poeppel, 2005; Pylkkänen, Llinás, & Murphy, in press).

Finally, in the domain of morphology, it is known that the summed frequency of all the words that share a morpheme is negatively related to the response latencies to those words in visual lexical decision (Colé, Beauvillain & Seguí, 1989; Taft, 1979). Similarly, the number of words that share a derivational morpheme – its morphological family size – also correlates negatively with visual lexical decision latencies (Schreuder & Baayen, 1997). Interestingly, as was the case for the effects observed in phonology, orthography, and semantics, it appears that the effects of morphological neighborhoods can also be modulated and even reversed in direction when one manipulates the degree of semantic relatedness between the morphologically related words (Moscoso del Prado Martín, Bertram, Häikiö, Schreuder & Baayen, 2005; Moscoso del Prado Martín, Deutsch, Frost, De Jong, Schreuder, & Baayen, 2005), or considers words that can have morphological relatives in both languages spoken by a bilingual (Dijkstra, Moscoso del Prado Martín, Schulpen, Schreuder, & Baayen, 2005). As with the

effects of semantic ambiguity, the morphological family size effect is also reflected in the M350 component of MEG experiments (Pylkkänen, Feintuch, Hopkins, & Marantz, 2004).

In summary, the neighborhood of a word, whether orthographic, phonological, morphological or semantic, influences the time it takes for that word to be recognized. However, in all domains mentioned above, it appears that, by itself, the size of a word's neighborhood can either facilitate the recognition of a word or, on the opposite, make it more difficult. All these effects appear to be reflected in the M350 and N400 components in magneto- and electro-encephalographic studies.

Information-theoretical measures and lexical recognition

Different lines of research on phonological and morphological neighborhoods are currently converging on the use of information-theoretical measures to describe the amount of support or competition that a word receives from its neighborhood. Kostić proposed an information-theoretical account of inflectional processing that was successful in explaining large proportions of the variance in lexical decision experiments to Serbian inflected words (Kostić, 1991; 1995; 2005; Kostić, Marković & Baucal, 2003). He considered the joint influence on response latencies of the distribution of frequencies of Serbian inflectional affixes, and their degree of syntactic and semantic heterogeneity. In the same direction, Moscoso del Prado Martín, Kostić, and Baayen (2004) showed that this account can be extended to provide a detailed description of the effects of Dutch morphological paradigms: The amount of support that a word receives from the morphological paradigms to which it belongs is best described by the entropy of the frequency distribution of the words that belong to that paradigm (i.e., the words that share an inflectional or derivational morpheme with it). Moscoso del Prado Martín and colleagues also pointed at the effects of semantic heterogeneity of morphological paradigms being directly accommodated in these information theoretical measures. More recently, Baayen and Moscoso del Prado Martín (2005) have shown that these measures also bear on issues like noun and verb regularity, and have implications for neuroimaging studies.

Interestingly, the success of information-theoretical measures in describing the effect of morphological paradigms on lexical processing is paralleled by information-theoretical mea-

sures characterizing the influence of phonological neighborhoods in spoken word recognition. Vitevitch and Luce (1999) showed that the amount of competition between words in the same phonological neighborhood is well described by the summed log frequency of the words in a particular neighborhood. This magnitude is in fact the same measure that Kostić (2005) calculated to describe the facilitation produced by morphological paradigms and, as shown by Moscoso del Prado Martín et al. (2004), it constitutes an upper bound estimate for the entropy measures. Indeed, Luce and Large (2001) showed that a similar entropy measure also plays a role in describing the effects of phonological neighborhoods.

Neural assemblies and lexical processing

Pulvermüller (1996; 1999; 2001) introduced a theory of lexical processing in the brain. It relies on the existence of neural assemblies (Hebb, 1949) distributed over broad cortical areas. These assemblies are tightly-coupled ensembles of neurons that automatically fire on presentation of a word. The assemblies would recruit neurons from left perisylvian areas (inferior frontal and superior temporal – including the traditional Broca's and Wernicke's language areas) relating to the phonological and orthographical forms of words, and from non-lateralized, widely distributed cortical areas relating to the meanings and grammatical properties of the words. A large amount of neurophysiological evidence has been provided in support of this theory (cf., Pulvermüller, 2003). These neural assemblies, although commonly termed 'lexical' or 'word' assemblies, can also correspond to sub-lexical morphemic units such as inflectional affixes (Shtyrov & Pulvermüller, 2002).

The lexical/morphemic assemblies are formed by Hebbian correlational learning: If the activation of neurons responding to the orthographic or phonological form of a word or morpheme consistently co-occurs in time with the firing of neurons responding to the meaning of that word, both sets of neurons will develop strong connections to each other via long-term potentiation of the corresponding synapses. Similarly, when either the neurons representing a particular meaning or word form fire independently of each other, long-term depression processes weaken the connections existing between them. This ensures that the connections will remain strong only among those pairs of word forms and word meanings that co-occur

together above chance level. When the connections have become sufficiently strong, the stimulation of one part of the network (e.g., the neurons responding to the orthographic properties) will result in the automatic firing and reverberation of the full network (including all properties of the word) within a short period of time.

A crucial aspect of the theory is the presence of inhibitory mechanisms that avoid the simultaneous activation of several word assemblies. Consider for instance the case of a polysemous or homonymic word: In Pulvermüller's theory, such words would be represented by multiple assemblies, each corresponding to one of the distinct meanings of the words. These assemblies would have different cortical topographies in relation to their meanings, but would overlap in their perisylvian areas representing their ortho-phonological properties, which are common for all their meanings. Therefore, in order to select one of the possible meanings of a word, some form of competition must take between the candidate assemblies for that particular form. Different mechanisms have been proposed to implement this mechanism. Pulvermüller (1999) suggested that direct inhibitory connections between different assemblies might be implemented by means of inhibitory striatal connections between the neurons in the assemblies (Miller & Wickens, 1991). In addition to this lateral inhibition, Pulvermüller (2003) argues for the presence of a more general regulatory mechanism that would deliver inhibition to all active assemblies when the overall level of cortical activation reaches a certain threshold. Such central regulation could be implemented through the thalamo-cortical loop. Indeed, neurophysiological evidence for thalamic modulation of cortical activity during semantic processing has been reported by Slotnik, Moo, Kraut, Lesser, and Hart (2002).

The missing link

As we have discussed above, a large amount of behavioral research has described in detail the effects of lexical neighborhoods on lexical recognition. At the same time, Pulvermüller's model offers – for the first time – a detailed, neurobiologically plausible theory of lexical processing supported by a large amount of neurophysiological and neuroimaging evidence. However, both lines of research seem to be somehow disconnected: On the one hand, the

behavioral results are currently presented without a low level anchor describing the neural processes and representations that give rise to these effects. Although MEG and EEG experiments have succeeded in showing that these effects have neurophysiological correlates (mainly in the M350 and N400 effects), no research has documented why and how do these particular differences arise in terms of the underlying neural structures. On the other hand, up to the moment, Pulvermüller's detailed neurophysiological theory has not attempted to make clear predictions at the behavioral level.

The effects of orthographic neighborhood size have been explained using a variety of computational models including the MROM model (Grainger & Jacobs, 1996), the DRC model (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) and the Bayesian Reader (Norris, in press). Vitevitch and Luce (1999) interpret their own results on phonological neighborhood within the framework of the adaptative resonance theory of speech perception (Grossberg, Boardman & Cohen, 1997), while Gaskell and Marslen-Wilson (1997) used a distributed connectionist network to show that the effect of phonological neighborhood can also arise due to competition between distributed representations of meaning. The effects of facilitation and competition caused by the semantic neighborhood have also been interpreted assuming both localist representations (Beretta et al., 2005; Pylkkänen et al., in press) or in terms of a distributed connectionist model (Rodd, Gaskell, & Marslen-Wilson, 2004). A similar contrast has been observed for the effects of morphological paradigms, which have been modeled using both localist, interactive-activation models (Baayen, Dijkstra, & Schreuder, 1997; De Jong, Schreuder, & Baayen, 2003; Taft, 1994) and distributed connectionist models (Moscoso del Prado Martín & Baayen, 2005).

All of these approaches are successful in replicating their targeted effects, and they provide plausible conceptual accounts of how the corresponding interactions take place. However neither of the models provides an account of how the words on which the measures are calculated, and the relationships existing between them, are actually represented in the brain. In addition, although the information-theoretical measures are providing a quite accurate description of the effects, both in the participants and in the computational models, no explanation is available of why these particular probabilistic measures best reflect the consequences of the underlying neural mechanisms. A traditional escape to these questions

comes from Marr's famous division of the levels of description of a computational system (Marr, 1982). It is argued that all the models above provide descriptions within Marr's computational level, which deals primarily with mental representations (see Norris, 2005 for an in-depth discussion of this issue). On the other hand, Pulvermüller's neurophysiological theory of language processing would lie between Marr's implementational and algorithmic levels, which are respectively concerned with neural representations and brain processes. Although Marr's division of labor is indeed useful for the study of the cognitive system, it must be kept in mind that Marr's explicit goal in positing this division was to obtain a unified theory (in his case of visual processing in the human brain). In order to achieve such a theory, the isolated investigation of each of the levels needs to be complemented with research aiming to link the results from the three levels. Some authors are pessimistic on the possibility of achieving an understanding of this link for higher cognitive processes in the near future (e.g., Norris, 2005). However, in other areas of cognitive processing, it has already been possible to approach this linkage. Specifically, the field of vision – Marr's own area of investigation – has recently come close such an integration (Rolls & Deco, 2001). The large set of results on human language processing at the behavioral, computational, and neurophysiological levels, suggests that we are beginning to be in a position to address such problems also for human language. In this direction, Edelman (in press) suggests three "general-purpose computational building blocks of biological information processing" that can be used to address the linkage of the different levels in the case of language: function approximation, density estimation, and dimensionality reduction. Edelman argues that these building blocks are implemented across multiple cognitive domains. In addition, Edelman describes how a combination of distributed representations with those building blocks is most likely to be successful in approaching the integrations of the different levels for the case of language.

In the present study, we show how a theory such as Pulvermüller's, can indeed be used to achieve detailed predictions on the behavioral level. We will show that, using the tools proposed by Edelman (in press), one can make predictions on behavioral measures of lexical processing following from the underlying neurophysiological theory. This provides us with a direct link between Pulvermüller's theory and the reported effects of lexical neighborhoods.

Crucially, our predictions also explain why the information-theoretical measures are proving the most suitable to describe these effects. In addition, the model that we present offers an insight on the debate between distributed and localist models to account for these effects. We will show that, in order to account for the effects, one needs to make use of distributed representations of the same type used in distributed connectionist models. However, a crucial component of our model is that it requires the explicit assumption that the distributed representations will give rise to a discrete number of localistic representations by effect of plain statistics. In such way, our model offers a meeting point for localist and distributed models of lexical processing.

In what follows we will begin by describing how Pulvermüller's neural assembly model can be used to make predictions on the behavioral level that would match the observed effects of lexical neighborhoods. We will continue by describing the forms of representations that we have used in the model, and how information theory enables us to extract measures of such representations that should link directly with the behavioral results. Next, we will use three lexical decision datasets to illustrate how the predictions made by the model with respect to morphological and semantic neighborhoods are indeed verified on the lexical decision latencies. Finally, we will discuss the implications of these results for current theories of lexical processing, and how our method offers a way to integrate the results from the different descriptive levels.

From Neural Assemblies to Behavioral Responses

We can represent the firing patterns of neurons in different cortical areas by means of multidimensional vectors. These vectors would define a space in which each point would correspond to a particular state of firing of the neurons in the system. The overlap between different patterns of firing can then be represented by a distance measure between their corresponding vectors, i.e., two patterns that involve the firing of many common neurons would be represented by two vectors whose distance in space is smaller than that between two vectors corresponding to patterns of firing with a lesser degree of overlap between them. In this multidimensional representational scheme, the firing probability of a particular combination

of neurons could be described by a multidimensional probability density function (PDF) over the vector space. From the Hebbian approach we can then predict that neural assemblies would become sensitive to the areas of the representational space in which the PDF has a higher value, indicating that points in that region space are relatively probable with respect to other parts of space. Moreover, given that neurons are known to have Gaussian-shaped receptive fields, one would assume that the probability of firing of a neural assembly would be determined by a multidimensional Gaussian distribution over the activation of the neurons that form it. This would imply that the PDF that the neurons are able to capture should correspond to a mixture of Gaussians (Dempster, Laird & Rubin, 1977), which each of the Gaussian components corresponding to one assembly. Therefore, in our vectorial representation, the corresponding PDF would also be a mixture of multidimensional Gaussians.

As reviewed above, Pulvermüller (1999) argues that lexical and morphemic assemblies recruit neurons in left perisylvian areas, related the the orthographic and phonological forms of words and their grammatical properties, and neurons in more broadly distributed cortical areas, which respond to the meanings of such words. Provided that we have adequate vectorial representations of the neural patterns responding to the forms and the meaning of words, and reasonable estimates of the frequency with which these are encountered, we could in principle predict the formation of neural assemblies as areas in the space which the value of the PDF is sufficiently high to develop an independent component in the Gaussian mixture. For illustration purposes, consider a hypothetical representation that enabled us to encode the patterns of activation related to word forms as a single real number, and the patterns of activation related to word meaning as another one. The scatterplot in Figure 1 represents a hypothetical sample of the distribution of firing patterns in such a space. The horizontal axis represents the neural activation pattern in the form-related areas, and the vertical axis represents the corresponding pattern in meaning-related areas. Each point in the graph would correspond to a particular occurrence of a combination of form and meaning neurons being active.

[INSERT FIGURE 1 AROUND HERE]

According to the Hebbian hypothesis, neural assemblies corresponding to words or mor-

phemes would be developed associating the formal and meaning neurons corresponding to the more densely crowded areas in the space, that is, the clusters of points in Figure 1. If the form and meaning of a particular instance of a word are represented by vectors $\mathbf{f}$ and $\mathbf{m}$ respectively, the scatter in Figure 1 would correspond to a PDF $p(\mathbf{f}, \mathbf{m})$ defined over the space of word forms and word meanings. Our assumption of a PDF composed of a discrete mixture of Gaussian would imply that each cluster would develop into a Gaussian distribution. The PDF resulting from fitting a mixture of Gaussians to Figure 1 is illustrated by Figure 2. The seven 'bumps' in Figure 2 describe how the receptive fields of the neurons in seven neural assemblies would map onto the areas of the space that could make those assemblies fire.

[INSERT FIGURE 2 AROUND HERE]

Following Pulvermüller's theory, during language comprehension, neural assemblies would automatically fire on presentation of a particular word form that falls within their influence. In many situations this could lead to the activation of multiple assemblies on the same stimulus. Consider the case in which a particular word form $\mathbf{f}$ has been encountered. The assemblies that are associated with it would all simultaneously receive activation. In our graphical scheme the probability of activation the the meaning-related neurons given a particular word form $p(\mathbf{m}|\mathbf{f})$ would be represented by a 'slice' of the PDF shown in Figure 2. Figure 3 illustrates how such a slice would look like: The selection of the particular word form $\mathbf{f}$, would correspond to constraining the overall joint probability distribution on the word-form margin with a a very sharp normal distribution centered on the particular word form. Such a spike would represent the activation in cortical areas responding to word-form resulting from the presentation of a particular visual or auditory stimulus.

[INSERT FIGURE 3 AROUND HERE]

Note that in Figure 3 the activation in the word-form neurons could correspond to at least three 'bumps' in the distribution. Accordingly, this would entail the simultaneous activation of the three corresponding assemblies. As these assemblies could in principle correspond to contradicting meanings of a word or morpheme, some mechanism must be at work to select

a single one being active and inhibit the firing of the other ones. In summary – as argued by Pulvermüller – some form of competition must take place to select a single assembly.[1] In this view, the competition would result in all active assemblies receiving inhibition. After a certain period, the assembly that receives the strongest activation – the one covering the greatest volume in Figure 3 – would become fully activated, while the activation of all other competing assemblies would die out as a result of the continued inhibition. The time it would take an assembly to become fully activated should therefore be related to two factors: (a) the initial degree of activation received by the assemblies and, (b) the degree of competition between the assemblies. Factor (a) would depend on the strength of activation delivered by the neurons representing the formal aspects of the word. This would depend on many factors: The frequency with which that particular combination of word form and word meaning is encountered, and different orthographic and phonological neighborhood effects. In addition, if one takes into account that there is deemed to be a certain amount of random activity in the system at any given moment, one could expect that when this random activity falls within the area of influence of a particular set of assemblies, it should add to their overall likelihood of being activated. This would entail that those groups of assemblies that cover a larger area of the representational space should have a certain advantage, as they would receive a larger amount of random activation.

Note that we have oversimplified the process of form identification as being an instantaneous process that renders a single form being active. Neither of this assumptions is true, the process is not instantaneous, and in principle can lead to multiple spikes for a particular stimulus (see Norris, in press for a detailed discussion and mathematical characterization of these issues). In this paper however, we will limit ourselves to the study of the interactions that happen once the form information has been reduced to a single spike. By this we are taking the simplifying assumption that we can sequentially separate the form identification processes from the activations at the level of meaning. In reality, these two processes are most likely to be cascaded. However, we believe the concussions we will draw from the inter-

[1] Whether this mechanism is implemented through lateral striatal connections (Pulvermüller, 1999) or a thalamo-cortical regulatory mechanism (Pulvermüller, 2003) would not make substantially different predictions for the purposes of this study.

actions at the level of meaning would also be true in a cascaded system, only with additional interactions with the word-form effects. These we leave for further research.

In turn, the amount of competition between neural assemblies that could correspond to a particular word form – factor (b) from the previous paragraph – should be influenced by:

i Number of components of the Gaussian mixture: As we described above, each component in the Gaussian mixture would correspond to an assembly that could fire in response to a particular word form. The amount of inhibition that all assemblies receive (either from the regulatory mechanism or through lateral connections) should then be related to the number of assemblies that are active, with more candidate assemblies resulting in more competition.

ii Relative probabilities of each of the components: The amount of competition between the assemblies should also be related to how unequal the activation of the candidate assemblies is. If one assembly receives much more activation than the remaining candidates, it is likely to resolve the competition faster than in a case where the level of activity of many of the candidates is roughly similar.

iii Degree of overlap between competing components: Neurons that could belong to more than one competing assembly will receive support from the activation of all of them, thus making their effective level of activation higher than would be expected according to a single assembly. Therefore, assemblies whose neurons receive additional support from other assemblies will be faster in reaching their ignition threshold. This entails that, for the overlapping parts of competing assemblies, the competition is reduced.

Measuring assembly coverage and competition between assemblies

A measure that would successfully index these three aspects would be the differential entropy of the probability distribution (Shannon, 1948; see Appendix A). This measure would grow with the number of components in the Gaussian mixture, i.e., a mixture with more components would have a higher level of uncertainty than a mixture with a single component. In a similar way, the degree of uniformness in the probabilities of the components of the measures

would also increase the differential entropy. Finally a mixture that contains two components that are very separated would imply a higher degree of uncertainty than a measure whose components that partially overlap.

Unfortunately, the differential entropy of the distribution would also be very influenced by the general width of the Gaussian mixture: PDFs with a large variance would increase the differential entropy in proportion to the log of the determiner of their covariance matrix. However, the width of the receptive fields of the different assemblies that are candidates for ignition should not influence the degree of competition between them. On the opposite, having a wide receptive field would be an advantage for the activation of a neural assembly, since it would increase the probability of random or noisy activation igniting one of the assemblies corresponding to a word or morpheme, thus reducing the average time it would take for the assemblies to be activated, and reducing the probability of the assemblies not being activated at all.

A more appropriate measure for our purposes is the negentropy of the PDF (Brillouin, 1956; see Appendix A). Negentropy is commonly used in techniques such as Independent Component Analysis (Comon, 1994) to assess the amount of useful information present in a combination of variables, that is, the degree of non-normality in their joint distribution. As in the case of differential entropy, negentropy is also sensitive to factors i, ii, and iii. Importantly, in contrast with differential entropy, this measure is mostly independent of the actual width of the distribution.

[INSERT FIGURE 4 AROUND HERE]

Figure 4 summarizes the variables of interest that we have highlighted here (for simplicity in a unidimensional space). The black curve is the PDF of the sample of points generated from a mixture of five Gaussian with different probabilities. The differential entropy of this PDF will comprise information about the number of Gaussians, their relative probabilities, the degree of separation between them, and the general spread of the distribution. Of these, the three first factors would have effects on the degree of competition between the corresponding assemblies, while the fourth one – the overall spread of the distribution – would not affect the competition at all, but would increase the probability of the assemblies

being activated. To separate this factor from the other three, we can subtract the entropy of the Gaussian mixture from the entropy of a single Gaussian distribution with equal mean and covariance (grey curve). The entropy of this single Gaussian is not sensitive to any factor related to the peaks (it has a single peak in any case), but is sensitive to the overall spread. Therefore, by this subtraction, we can separate the two variables of interest: the degree of inter-assembly competition is reflected by the negentropy, and the likelihood of the assemblies being activated which is indexed by the Equivalent Gaussian Entropy (EGE).

These two measures enable us to make predictions on behavioral responses based on Pulvermüller's neural assembly model: On the one hand, the negentropy of the distribution of meanings should correlate positively with response latencies to comprehension tasks that require the activation of an assembly, as it reflects the amount of competition that the winning assembly will have to overcome. On the other hand, the EGE measure should correlate negatively with both response latencies and error counts, since it reflects the general ease of activating a set of assemblies.

Probability Distributions on a High-dimensional Space

In the previous section we have outlined how information-theory enables us to make predictions at the behavioral level starting from Pulvermüller's neurophysiological theory. In order to test this idea we require a suitable vectorial representation of the meaning and grammatical function of each occurrence of a word or morpheme, and a technique to estimate the corresponding mixture of multidimensional Gaussians and the associated information-theoretical measures.

First and second order co-occurrence vectors

Schütze (1992, 1994) introduced a technique for building high-dimensional vectors representing the meaning of words, using the information derived from their usages in a large corpus. This technique consists in passing a small window through the corpus counting the number of times that words co-occur within that window. The result is a large square matrix with

as many rows and columns as different word types appeared in the corpus. The cells in the matrix correspond to the number of times with which the word corresponding to a column appeared within a small window centered on the word corresponding to the row. The rows (or the columns) in such a matrix provide a representation of the contexts in which that word is normally used. In turn, the contexts in which a word is used provide crucial information about the meaning and morpho-syntactic properties of the word itself (Wittgenstein's "meaning is use"). Indeed, Schütze observed that the distances between the vectors corresponding to the words provide useful information about their similarity in meaning.

A large amount of research has developed this idea of word co-occurrence vectors, and with different variations on the technique employed for collecting the vectors, transforming the frequencies, and reducing the dimensionality of the resulting matrix, has given rise to a family of related techniques such as Hyperspace Analog to Language (Lund & Burgess, 1996), Latent Semantic Aanalysis (Landauer & Dumais, 1997) or Random Indexing (Kanerva, Kristofersson & Holst, 2000). In addition, a large body of research has indicated that the distances between co-occurrence vectors correlate with human responses in different behavioral tasks (e.g., Landauer & Dumais, 1997; Landauer, Laham, Rehder, Schreiner, 1997; Lowe & McDonald, 2000; Lund, Burgess & Atchley, 1995; McDonald & Shillcock, 2001). In addition to capturing semantic properties of words, co-occurrence vectors have also been shown to capture the morpho-syntactic properties of words (Schütze, 1995) and inflectional affixes (Schone & Jurafsky, 2001).

As described in the previous paragraphs, co-occurrence vectors provide suitable representations of the average meaning and morpho-syntactic properties of words and morphemes. However, in order to employ such vectors for estimating the distributions of meanings, we require different vectors representing each individual usage of the words and morphemes. Schütze and Pedersen (1997) introduced a variation of the above techniques to deal with word sense disambiguation. Their second order co-occurrence vectors provide different representations for each occurrence of a word. The second order vectors are built in a two-stage process. First, using the techniques described above, a matrix of first-order vectors is constructed to represent the average meaning of each word type (the types can be further broken down in order to consider different meanings of a homonym as different vectors). Once the

first order vectors are constructed, each occurrence of the word of interest is represented by the sum of the first order vectors of the words occurring around it (also within a small window). In this way one obtains a set of different vectors for each word of interest, each vector corresponding to one instance of the word. Schütze and Pedersen obtained promising results on word sense disambiguation by using the distances between the second order vectors of an ambiguous noun in context, and the first order vectors of the different meanings of that noun.

We propose using these second order co-occurrence to represent the different instances of a word or morpheme in multidimensional space. This will enable us to obtain an estimate of its distribution of usages. A crucial point is that this technique enables us to build the vectors on the minimum possible assumption, that is, a corpus of language without any linguistic labeling.

It is clear that the co-occurrence vectors will not contain all the information that is relevant for the semantic and morpho-syntactic properties of words or morphemes, and that they are bound to be noisy. However, we believe that they will contain sufficient information as to provide a reasonable estimate of a word's or morpheme's variation in meaning and morpho-syntactic properties. Furthermore, a great deal of morpho-syntactic and semantic information has to be acquired through linguistic experience, rather than through direct exposure to the concept or referential meaning of the words. Therefore the information contained in these vectors could be closely related to some of the semantic and morpho-syntactic information about them that is actually captured by the cognitive system. Indeed, this hypothesis is supported by behavioral research (Boroditsky & Ramscar, 2003; McDonald & Ramscar, 2001). Furthermore, Pulvermüller (2002) argues that word co-occurrence information would also be exploited by a neural assembly model of language processing. In his view, the initial form-meaning associations would be built by direct co-occurrence between words and sensory-motor experience of their referents. Once some of these associations have developed, the sequential activation of different word assemblies in a short time window would lead to associations developing between the co-occurring word assemblies. This would result in a process of bootstrapping, by which sensory-motor information associated with one word could – through exclusively linguistic experience – end up being also associated to

other words that are used in similar contexts.

Estimation of the underlying distribution and informational measures

Second order co-occurrence techniques provide us with a method for estimating a sample of high-dimensional vectors describing the contexts in which a word or morpheme is used. In order to estimate the negentropy and EGE of the underlying multidimensional distribution, we could make use of direct estimation methods (Kraskov, Stögbauer, & Grassberger, 2004; Van Hulle, 2005a; 2005b). However, these methods are mathematically quite complex and make strong assumptions on the underlying distributions that are not justified in our case. The methods proposed by Kraskov et al. (2004) and Van Hulle (2005a) do not make use of any information on the underlying distribution. In our case, by the assumption of the neural assemblies, we have hypothesized that that distribution must be a multidimensional Gaussian mixture with an unknown number of components, thus our EGE and negentropy approximations should take this information into account. Van Hulle (2005b) introduces a method to estimate the differential entropy of a multidimensional mixture of Gaussians, but it is valid only when it can be assumed that the mixture components are "far enough apart". However, in our case, many of the mixture components are deemed to overlap. Instead, we can estimate our information-theoretical measures in two stages: First we estimate the underlying PDF as a Gaussian mixture, and then we estimate its negentropy and EGE.

The Expectation-Maximization (EM) algorithm (Dempster et al., 1997) has traditionally been used to estimate the PDF of multidimensional Gaussian mixtures where the number of components is known a priori. In our problem we need to estimate the PDF from a sample of points taken from the distribution. However, in contrast with the EM algorithm, we do not have any knowledge of the number of Gaussian components in the mixture. Instead we can use a infinite mixture model (Neal, 1991; 1998). Infinite mixture models assume that the underlying distribution is a mixture of an unknown, possibly very large (but finite), number of Gaussian components. Using Markov chain MonteCarlo methods one can sample from the space of possible distributions of this kind, and use Bayesian inference to find which one has a higher posterior probability of being the underlying distribution given the sample

of points. Note that in practice, after estimation, an infinite mixture model reduces to a normal Gaussian mixture with a finite number of components. Neal (2004) provides a set of software tools to estimate distributions of this type. This family of models corresponds well to our prior knowledge: We are assuming that the points in our sample have been generated by a Gaussian mixture with an unknown number of components.

The problem of estimating the negentropy is now simplified by having estimation of underlying PDF. According to the definition (see Appendix A), the negentropy of a distribution $p(\mathbf{x})$ is defined as the difference between the differential entropy of a Gaussian distribution $p_{\mathcal{N}}(\mathbf{x})$ of equal covariance matrix to $p(\mathbf{x})$ (EGE) and the differential entropy of $p(\mathbf{x})$ itself:

$$J(p) = h(p_{\mathcal{N}}) - h(p) \tag{1}$$

Provided we know the covariance matrix $\mathbf{K}$, which can be directly estimated from the sample of points, the entropy of the normal distribution $p_{\mathcal{N}}(\mathbf{x})$ can be calculated analytically as:

$$h(p_{\mathcal{N}}) = \frac{n}{2}\log_2(2\pi e) + \frac{1}{2}\log_2|\mathbf{K}| \tag{2}$$

There is no simple analytical way of calculating the differential entropy of a mixture of Gaussians. Instead, we can estimate it numerically using MonteCarlo integration: If $p(\mathbf{x})$ is a probability density function over a n-dimensional space S, and $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset S$ is a sufficiently large sample of points sampled according to $p(\mathbf{x})$, then the entropy of $p(\mathbf{x})$ can be approximated by:

$$h(p) \simeq -\frac{1}{N}\sum_{i=1}^{N}\log_2 p(\mathbf{x}_i) \tag{3}$$

Therefore we can calculate the EGE of our sample of points using (2), then calculate the differential entropy of our fitted Gaussian mixture using (3), and finally estimate the negentropy using (1).

Analysis 1: Negentropy and inter-assembly competition

In this section we will test the hypothesized relationship between the distribution of usages of an affix, and the competition that would take place between the different assemblies that could correspond to that affix. For this purpose we will reanalyze the results of the

experiments reported by Kostić et al. (2003) on the processing of Serbian nominal inflectional affixes.

Kostić et al. (2003) found that most of the variance in the average lexical decision RTs to Serbian inflected nouns is explained by the logarithmic ratio between the frequency of a particular suffix, and the number of syntactic functions and meanings that that affix might take (calculated through a detailed linguistic study described in Kostić, 1965). In a brief summary, Kostić's results show that the time it takes to recognize a Serbian suffix is directly related to the number of syntactic functions and meanings that it could have in a particular context (i.e., masculine nouns or feminine nouns).

If our hypothesis is correct, Kostić's number of syntactic functions and meanings should be related to the amount of competition between assemblies, and thus should also be correlated with the negentropy measure we have proposed. More importantly, the negentropy measure should play a similar role to that of Kostić's count in predicting lexical decision latencies.

Method

We obtained the frequency counts and counts of number of syntactic functions and meanings (defined according to Kostić, 1965) from Kostić et al. (2003)'s experiments on Serbian masculine and feminine nouns. From the same dataset, we obtained the average visual lexical decision RTs to Serbian masculine and feminine nouns in each of their nominal inflectional variants (uninflected, or suffixed with -a, -e, -i, -u, -om and -ima, for masculine nouns, and suffixed with -a, -e, -i, -u, -om and -ama for feminine nouns; see Kostić et al. (2003) for details on the Serbian nominal declension system).

From the Corpus of Serbian Language (CSL; Kostić, 2001), we sampled 500 random occurrences of masculine and feminine nouns in each of their inflectional variants. To ensure that the sample would be representative of the variation in usages of a particular suffix, and not biased to the usages of particularly frequent nouns, we constrained the sampling procedure to avoid selecting more than one instance of any particular noun. Each occurrence included the words located within a centered context window of seven words from the target

(three words to each side). To keep the level of linguistic information to a minimum, the words in the corpus were chosen to be masculine whenever their lemma in the CSL ended in a consonant, and feminine when it ended in vowel 'a'. This is the most basic rule to attribute gender to Serbian nouns but many Serbian nouns deviate from that rule (e.g., gazda – "boss" – is a neuter despite ending in 'a')'. However, using this simple rule enables us to ensure that most of the selected examples were of the targeted gender (this is reinforced by the constraint of selecting at most a single instance of any noun so that any exceptions would be selected at most once), while at the same time ensuring that the properties of the nouns would be detectable with a minimum requirement of linguistic annotation – i.e., just the lemmatization. In fact, these properties are shared with the nouns used in Kostić et al.'s experiments, and thus they provide a good reflection of the conditioning that the experimental situation induced on the participants. By the above method, we obtained a sample of 6,500 contexts, 500 for each of the possible inflectional variants in each gender (7 masculine and 6 feminine).

We constructed first-order co-occurrence vectors for all words that occurred in the CSL with a frequency equal or higher than one per two millions. We selected the 1,000 most frequent word types in the corpus as context words, without removing function words or very high frequency words. This was done for two reasons: First, Lowe and McDonald (2001) showed that function and high-frequency words tend to be most informative when constructing semantic co-occurrence vectors. Second, in this study we are especially interested in the variation in morpho-syntactic properties of inflectional affixes, and this information is most clearly reflected in the function words around them. The vectors we constructed were 'untransformed' in the sense that they plainly consisted of the raw counts of the number of times that a word would co-occur with each of the 1,000 context words within the seven word window. Although normalizing the vectors for different frequency counts or applying transformations such as the log-odds ratio appears to improve the quality of the semantic representations, keeping these transformations to a minimum enhances the biological plausibility of the model: Whichever transformations are adequate, should be detectable in an un-supervised manner from the distributional properties of the data.

We used these first order vectors to compute second-order co-occurrence vectors for each

of the 6,500 contexts. The second-order vectors for the contexts were computed as the average of the first-order vectors of the words in the window (excluding the word itself). The resulting second-order context vectors were subjected to a principal components analysis (PCA) (after centreing to zero, and scaling their components to unit variance). The first six principal components accounted for 92% of the variance. We selected the first six principal components of each of the vectors. This dimensionality reduction simplifies the estimation of the underlying distribution without affecting the underlying PDF or uncertainty of data points except for a factor of scale in the EGE measure. At the same time the neurophysiological plausibility of this transformation in a Hebbian system is ensured. Indeed, it is long known that neurons do perform operations which are equivalent to PCA (Oja, 1982). Finally, to ensure that the similarity space between the resulting vectors is defined by the Euclidean distance (in the untransformed vectors the distance would be defined by the angle formed between the vectors.), we normalized the vectors to having unit length. By this procedure we obtained a six-dimensional vector describing each of the 6,500 contexts.

For each suffix, using the software for flexible Bayesian modeling (FBM; Neal, 2004) we fitted an infinite mixture of Gaussians to the set of 500 six-dimensional vectors obtained above.[2] After estimating the most probable mixture of Gaussians for the distribution of context vectors, we used the samples of 500 points on which the density estimation was performed as a suitable sample of the distribution. Using the FBM tools we computed the probability $p(\mathbf{x}_i)$ of each of the points in our sample, according to the corresponding mixture, and then estimated the differential entropy $h(p)$ to be the negated average of the log probabilities using (3). We used these same samples to estimate the covariance matrix $\mathbf{K}$ for each suffix, and calculate the entropy $h(p_{\mathcal{N}})$ of the corresponding normal distribution (EGE) according to (2). Once both these entropies had been estimated, the value of the negentropy was computed using the definition in (1).

[2]The parameters used for the estimation of the Gaussian mixture were identical to those provided in Example 2 of the FBM documentation on mixture models ('A bivariate density estimation problem'), the only changes being that we set the number of dimensions to six, and 9 for the Dirichlet prior concentration parameter to account for the possibly large number of meanings that an affix might have.

Results and discussion

Figure 5 illustrates the correlation between Kostić's number of syntactic functions and meanings (vertical axis) and the negentropy of the contexts in which each Serbian suffix is used (horizontal axis). The correlation seems to be high ($r = .92, p < .0001$) but note that most of this correlation could be driven by the two points at bottom left of the figure. A non-parametric Spearman rank correlation confirmed that the correlation is not fully driven by those two outliers ($r_s = .64, p < .0215$).

[INSERT FIGURE 5 AROUND HERE]

The correlation between the number of syntactic functions and meanings of a Serbian suffix and the estimated negentropy of its distribution of usages (having assumed a Gaussian mixture) provides support for our hypothesis that the effect of number of meanings reported by Kostić is a consequence of the competition between neural assemblies, especially so if we consider that it becomes apparent on such a small set of points, and that our negentropy measure was derived using several levels of approximation (i.e., of the representation of the contexts, of the distribution, and of the actual measure). However, the crucial point is to ensure that negentropy has an effect on lexical decision RTs similar to that of number of functions and meanings.

As mentioned above, the RTs in Kostić et al. (2003) are explained by the logarithmic ratio between the frequency of the suffixes to their number of function and meanings. In order to directly compare the contribution to RTs of negentropy with that of number of functions and meanings, we need to consider separately their contributions to the RTs. For this purpose, we fitted a multilevel regression model with log average RT to a suffix as dependent variable, log suffix frequency and log number of meanings as fixed effects and experiment (masculine vs. feminine) as a random effect (to account for the fact that the RTs to both genders were collected in different experiments, each including nouns from a single gender). The analysis revealed significant effects of frequency ($F(1, 9) = 23.46, p = .0009$) and number of functions and meanings ($F(1, 9) = 16.15, p = .0030$; after partialing out the contribution of frequency). A similar analysis including negentropy instead of number of

functions and meanings revealed significant effects of frequency ($F(1,9) = 15.76, p = .0033$) and negentropy ($F(1,9) = 7.84, p = .0188$; after partialing out the contribution of suffix frequency). These analyses indicate that both number of functions and meanings have similar effects on the reaction times. Unfortunately, given the high correlation between both counts it is not advisable to include both of them as predictors in a single regression, as doing so would introduce a strong collinearity that would make it impossible to assess the independent contribution of the effects (Belsley, 1991). Instead, we considered their contributions to explaining the variance on the RTs.

[INSERT FIGURE 6 AROUND HERE]

Figure 6 shows how well the reaction times are predicted only on the basis of suffix frequency (and gender included as a random effect). As it can be observed, most of the variance (76%, uncorrected) in the data is already accounted for by frequency alone. Figure 7 illustrates the relationship between the residuals of the regression using only frequency as a fixed effect, and the number of syntactic functions and meanings (right panel) or negentropy (left panel).

[INSERT FIGURE 7 AROUND HERE]

Both of the panels in Figure 7 show very similar patterns of predictivity. Both counts are directly related to the residuals, and even the pattern of outliers is similar across both plots. The improvement of adding the number of syntactic functions and meanings into the model is shown by Figure 8. Note that, although the margin for improvement over the 76% of variance that is accounted for just by frequency is rather small, there is still a clear increase in the predictivity of the model (approximately 15% additional explained variance).

[INSERT FIGURE 8 AROUND HERE]

Figure 9 shows the effect of substituting the number of functions and meanings with the negentropy as a predictor in the model. The additional explained variance (approximately 10%) is less than in the regression using functions and meanings, but it is still a significant improvement over frequency, accounting for a large part of the improvement that the original count gives.

[INSERT FIGURE 9 AROUND HERE]

In the previous section, we had also predicted that the differential entropy of an equivalent Gaussian distribution (EGE) should have an effect on response latencies, of opposite direction to the effect of the negentropy of the distribution. However, including EGE as an additional predictor in the previous regressions did not show any additional significant effect, either in the regression including number of syntactic function and meanings ($F < 1$), or in the one using negentropy ($F < 1$).

In sum, we have seen that in this dataset, our negentropy measure shows similar effects to those of number of functions and meanings. Some of the explanatory power of the original count is lost when we use the negentropy instead of Kostić's original count. However, we consider that this is not a reason for concern, since the negentropy was calculated over a series of approximations using a small sample (500 occurrences) with little linguistic labeling. In contrast, the counts of number of functions and meaning were calculated by an exhaustive linguistic analysis across the whole CSL (Kostić, 1965). This predictivity is important in two directions: On the one hand, it provides a validation of the count provided by Kostić through un-supervised means. On the other hand, it provides an anchor at the neurophysiological level for the effects of counts calculated through linguistic analysis, and it verifies the predictions of the underlying neurophysiological level on the behavioral measures.

The lack of predictivity in these experiments of the overall width of of the receptive fields of the assemblies, measured through the EGE, could question our underlying neural model. A possible reason for this lack of predictivity could lie on the nature of the data: After all, the width of the assemblies for particular inflectional suffixes should not show a great degree of variation, since all of them can attach to exactly the same nouns. The effect of EGE should become more evident when looking at the variation present for morphemes that vary also in terms of their semantic content. The following section investigates this issue further, by analyzing the responses to a larger set of English stems, for which effects that we assumed are related to the width of the assemblies are present.

Analysis 2: Equivalent normal entropy and assembly width

In the previous section we have shown that the amount of competition between different assemblies that are candidates for activation given a particular form, measured as the the negentropy of the PDF of second order vectors, correlates negatively with average lexical decision latencies to Serbian inflectional affixes. Based on our neurophysiological model, we also predicted that time it takes one of the neural assemblies corresponding to a word to fire should also be related to their combined likelihood of receiving activation. Following our prediction, this should be related to the area of the representational space that is covered by the components of the Gaussian mixture, measured by the differential entropy of the equivalent normal distribution. If our assumptions are correct, this measure should relate to measures of the support of a word's morphological paradigm, such as the inflectional entropy measure (Baayen, in press; Baayen & Moscoso del Prado Martín, 2005; Moscoso del Prado Martín et al., 2004), and should correlate negatively with both lexical decision latencies and errors.

Method

We constructed a experimental list of 85 monomorphemic words that appeared, across all inflectional variants, at least 500 times (approximately 5 per million) in the British National Corpus (BNC[3]). For each word we extracted its surface frequency from the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995), and we computed inflectional entropy using the CELEX frequencies, following the method described by Moscoso del Prado Martín et al. (2004). We extracted from the English Lexicon Project's database (Balota et al., 2002) visual lexical decision error scores and (average and by-participant) reaction times for each of the words.

From the BNC, we selected a sample of 500 random occurrences of each word (in any inflectional variant). Each occurrence included the words located within a centered context window of seven words from the target (three words to each side). Using the same technique that we employed for Serbian, we constructed untransformed first-order co-occurrence vectors

[3] http://www.natcorp.ox.ac.uk

for all words that occurred in the corpus with a frequency equal or higher than one per two millions, using the 1000 most frequent type as context words. We used these first order vectors to compute second-order co-occurrence vectors for each of the 42,500 occurrences (85 items $\times$ 500 occurrences/item). The second-order vectors for the contexts were computed as the average of the first-order vectors of the words in the window (excluding the word itself).

The context vectors were subjected to a PCA (after centreing to zero, and scaling their components to unit variance). In order to speed the calculation the PCA rotation matrix, we randomly selected a subset of 30 occurrences for each of the target words, and the rotation matrix was computed on this smaller sample of 2,550 vectors. In this reduced set, the first six principal components accounted for 92% of the variance. The resulting rotation matrix was applied to the full set of 42,500 context vectors, and for each vector, the first six principal components were selected. Finally, to ensure that the similarity space between the resulting vectors is defined by the Euclidean distance, we normalized the vectors to having unit length. By this process we obtained a six-dimensional vector describing each of the 42,500 contexts.

We estimated the covariance matrix of the 500 contexts for each of the words. With the estimated covariance matrix, we calculated the differential entropy of the equivalent normal distribution for each of the words using (2). In addition, for comparison purposes, we fitted a mixture of Gaussians to each of the sets of 500 contexts using the methods described in the previous section, and we estimated its differential entropy and negentropy using (3) and (1).

Results and Discussion

Figure 10 illustrates the relationship between the inflectional entropy measure and the differential entropy of the Gaussian of equivalent covariance to its distribution of context vectors. Note that, although there is a significant positive correlation, both by parametric ($r = .34, p = .0014$) and non-parametric methods ($\rho_s = .28, p = .0085$), this accounts for at most 11% of the variance, which does not appear to support our hypothesis that both are measuring the same thing. However, the crucial question concerns not so much the direct relationship between both counts, but their relation with the lexical decision responses.

[INSERT FIGURE 10 AROUND HERE]

A multilevel regression fitted to the log reaction times, with participant as a random effect and log frequency and inflectional entropy as fixed effects, revealed significant main effects for both frequency ($F(1, 1304) = 14.1, p = .0002$) and inflectional entropy ($F(1, 1304) = 5.2, p = .0230$, after partialing out the effect of frequency). This regression did not provide any evidence for mixed-effects of participant by frequency or participant by inflectional entropy. A by-item regression on the log reaction times with log frequency and inflectional entropy independent variables confirmed the main effect of frequency ($F(1, 82) = 16.66, p = .0001$) and that of inflectional entropy ($F(1, 82) = 6.73, p = .0112$). A logistic regression to the number of correct and incorrect responses also revealed the same main effects of log frequency ($\chi^2_1 = 8.19, p = .0042$) and inflectional entropy ($\chi^2_1 = 5.40, p = .0202$). Both effects – frequency and inflectional entropy – had negative coefficients in the three regressions. According to parallel regressions using additional non-linear restricted cubic spline terms for the independent variables, no significant non-linear components were detected for any of the effects in any of the three regressions. These analysis ensure that, as reported by Moscoso del Prado Martín et al. (2004), the inflectional entropy measure has a facilitatory effect both on the response latencies and on the error scores.

In order to assess whether the EGE measure has a similar effect on the response latencies and errors, we added negentropy to the above regressions, after having partialed out the effect of inflectional entropy (given the weak correlation between both counts, we found it safe to include them simulateneously in the same regression).

In the multilevel by-participant regression, EGE did not have any significant effect ($F(1, 1303) = 2.0, p = .1559$) on the RTs on top of those of frequency and inflectional entropy. However, when the effect of EGE is considered before that of inflectional entropy, it is EGE that showed a significant effect ($F(1, 1303) = 4.4, p = .0352$) while that of inflectional entropy disappeared ($F(1, 1303) = 2.8, p = .0969$), indicating that both variables are capturing roughly the same part of the variance.

Adding EGE as an independent variable to the by-item regression on the RTs, after considering the contributions of frequency and inflectional entropy, revealed that while EGE

still had a significant effect ($F(1, 81) = 4.18, p = .0442$), the effect of inflectional entropy became only marginally significant ($F(1, 81) = 2.98, p = .0882$). Indeed, a fast backwards elimination of factors using Akaike's information criterion (Lawless & Singhal, 1978) on this regression recommended keeping EGE as an independent variable, and removing inflectional entropy as a predictor from the regression. After excluding inflectional entropy from the regression, the effect of EGE became even more clear ($F(1, 82) = 6.73, p = .0112$). We observed a similar pattern when we added EGE to the logistic regression on the error counts. After adding EGE into the model, the effect of inflectional entropy became un-significant ($\chi^2_1 = 1.82, p = .1769$) while that of EGE approached significance ($\chi^2_1 = 3.79, p = .0515$). Once again, the fast backward elimination of factors suggested deleting inflectional entropy from the model. After doing this, the effect of EGE reached full significance ($\chi^2_1 = 7.65, p = .0057$). As it was the case with the effects frequency and inflectional entropy, no additional non-linear component was present in the effects of EGE on response latencies and errors.

Finally, we assessed the contribution of negentropy by adding the term in the regressions after partialing out the effects of frequency and EGE. Negentropy did not add any significant contribution to the analyses on the RTs ($F(1, 1303) = 2.1, p = .1492$ by-participant, and $F < 1$ by-item), or error counts ($\chi^2_1 = .11, p = .7432$).

[INSERT FIGURE 11 AROUND HERE]

The conclusion of these analyses is that, although the correlation between inflectional entropy and EGE is relatively weak, they both appear to be capturing the same part of the variance of RTs and errors. The explanatory power of EGE seems to be, if anything, superior to that of inflectional entropy. Figure 11 summarizes the effects (as estimated in the regressions) of frequency (left column), EGE (middle column), and inflectional entropy (right column), on the RTs (top row) and error scores (bottom row). Note that while the magnitude of the effect of EGE on the RTs is only slightly larger than that of inflectional entropy, this difference becomes more marked in the error analyses, where the effect of EGE is clearly more pronounced.[4]

[4]The non-linearities in the graphs are due to the back transformation from the logarithm in the case of the reaction times, and the logit function in the case of the error scores. If those transformations are applied,

These results show that, as we predicted, the effect of inflectional entropy can be seen as a higher level parallel of the effect of the overall spread the distribution of meanings that would be predicted by a model based on neural assemblies. In contrast to the RTs to Serbian inflectional affixes, we did not observe any effects of negentropy on this dataset. This is due to the words in this experiment not offering any particular contrast in number of functions and meanings. They were all selected to be monomorphemic nouns, some of which could also have verbal conversions but, in general, there was no particularly great variation in the number of meanings. In principle, we would expect the effect of negentropy to show an additional contribution to the responses to sets of words that have been designed to contrast levels of ambiguity. For this purposes, we now turn to investigate the effects of homonymy and polysemy reported by Rodd et al. (2002).

Analysis 3: Polysemy and homonymy

Rodd et al. (2002) showed that a distinction should be made between polysemous words, having more than one related senses, and homonymic words, having more than one unrelated meanings. They found that in both visual and auditory lexical decision tasks, words that have many senses are recognized faster than words than have few senses and, at the same time, words that have many meanings are recognized slower than words that have few meanings. Beretta et al. (2005) and Pylkkänen et al. (in press) confirmed these results, and showed that the differences are related to differences in the M350 component in Magneto Encephalography. As we argued above, this distinction is analogous to the opposite effects of negentropy and EGE: The inhibitory effect of having many unrelated meanings is equivalent to the amount of competition between different assemblies, that we measured by means of the negentropy of its distribution of usages, while the amount of facilitation provided by related senses is indexed by the with of the equivalent Gaussian distribution (EGE). In this section we investigate in detail this relationship.

as was done to perform the analyses, the effects become linear.

Method

We selected from the 128 words used in Rodd et al. (2002)'s visual lexical decision experiment (Experiment 1) all 97 items for which we could find at least five hundred occurrences in the BNC and response latencies in the Balota et al. (2002) database. As this selection decreased the power of the original design (to the extent that both of the effects reported by Rodd and colleagues disappeared) we added 93 additional homonymic words for which we could also find lexical decision RTs from the Balota et al. database and 500 occurrences in the BNC. Of these additional 93 words, 47 were classified as homonyms (having more than one entry in the Oxford English Dictionary), while the remaining 46 were left uncontrolled, but matched for frequency with the homonymic ones. In this way, we have extended Rodd et al.'s original dataset to have a more continuous degree of variation between homonymy and polysemy, instead of the original purely orthogonal design. In total we have now 190 words, 92 of which are classified as homonyms and 98 are mostly non-homonymic.

From the BNC, we selected a sample of 500 random occurrences of each word (in any inflectional variant). Each occurrence included the words located within a centered context window of seven words from the target (three words to each side). As was done in the previous section, we constructed untransformed first-order co-occurrence vectors for all words that occurred in the corpus with a frequency equal or higher than one per two millions, using the 1000 most frequent type as context words. We used these first order vectors to compute second-order co-occurrence vectors for each of the 95,000 occurrences (190 items $\times$ 500 occurrences/item). The second-order vectors for the contexts were computed as the average of the first-order vectors of the words in the window (excluding the word itself).

The context vectors were subjected to a PCA (after centreing to zero, and scaling their components to unit variance). In order to speed the calculation the PCA rotation matrix, we randomly selected a subset of 20 occurrences for each of the target words, and the rotation matrix was computed on this smaller sample of 3,800 vectors. In this reduced set, the first six principal components accounted for 88% of the variance (with no additional component accounting for more than 5% of variance). The resulting rotation matrix was applied to the full set of 95,000 context vectors, and for each vector, the first six principal components were

selected. Finally, as was done in the previous analyses, we normalized the vectors to having unit length to ensure that the similarity space between the resulting vectors is defined by the Euclidean distance. By this process we obtained a six-dimensional vector describing each of the 95,000 contexts.

We estimated the covariance matrix of the 500 contexts for each of the words. With this covariance matrix, we calculated the differential entropy of the equivalent normal distribution for each of the words using (2). Using the same methods as in the two previous sections, we fitted a mixture of Gaussians to each of the sets of 500 contexts, and we estimated its differential entropy and negentropy using (3) and (1).

Results and Discussion

A multilevel regression fitted to the log reaction times, with participant as a random effect and log frequency, EGE and negentropy as fixed effects, revealed significant linear effects for frequency ($F(1, 3512) = 17.0, p < .0001$) and EGE ($F(1, 3512) = 37.3, p < .0001$, after partialling out the effect of frequency), and an effect of negentropy ($F(1, 3512) = 3.2, p < .0001$, after partialling out the contributions of frequency and EGE) that was significantly non-linear ($L(6, 7) = 3.86, p = .0494$). This regression did not provide any evidence for mixed-effects of participant by frequency or participant by inflectional entropy. A by-item regression on the log reaction times with log frequency, EGE and negentropy as independent variables confirmed the linear effects of frequency ($F(1, 186) = 10.22, p = .0001$) and the effect of EGE ($F(2, 186) = 11.16, p < .0001$), which had a significant non-linear component ($F(1, 186) = 4.03, p = .0461$). No effect of negentropy ($F(1, 185) = 1.61, p = .2063$, after partialling out the effects of frequency and EGE) was detected in this regression.

A logistic regression to the number of correct and incorrect responses also revealed the a main effect of log frequency ($\chi^2_2 = 18.56, p < .0001$), with a significant non-linear component ($\chi^2_1 = 4.67, p = .0307$) and a linear effect of EGE ($\chi^2_1 = 30.72, p < .0001$), without any significant contribution of negentropy ($\chi^2_1 = 0.00, p = .9945$).

[INSERT FIGURE 12 AROUND HERE]

Figure 12 illustrates the non-linearities observed in the effects of EGE and negentropy on the response latencies. Note that both effects seem to have opposite directions, and a clear attenuation of the effect in the higher part of their range. In addition, the effect of negentropy was much smaller and – after partialing the out contribution of EGE – only reached significance in the more sensitive multilevel regression on the RTs, but appeared to be too weak to show up in the by-item regression or in the error analyses.

A possible reason for the non-linear attenuation of the effects of EGE and negentropy, and for the relative unstability of the second one, comes from the fact that, in this dataset, both measures are mildly correlated with each other ($r = -.29, p < .0001$), combined with the smaller magnitude of the effect of negentropy in relation with that of EGE. The left panel in Figure 13 shows the correlation between the negentropy and EGE measures. This negative correlation might indeed account for the attenuation of the effect of negentropy on the multilevel regression and the attenuation of EGE and disappearance of negentropy in the by-item regression. We can test this hypothesis by decorrelating both variables. As their inter-correlation is moderately weak, and the effect of EGE is relatively strong compared to that of negentropy, we can discount from the EGE count that part of its variance that can be predicted by the negentropy through a linear regression. In this way we obtain a residualised count that is orthogonal to negentropy but still captures most of the variation of EGE, as it is shown in the right panel of Figure 13. If our hypothesis is correct, using this residualised count on the above regressions, would make both effects linear, and would make the inhibitory role of negentropy more stable, even in the less sensitive by-item regression.

[INSERT FIGURE 13 AROUND HERE]

We repeated the above regressions using the EGE residual measure instead of EGE. The multilevel regression revealed significant linear effects for frequency ($F(1, 3512) = 17.0, p < .0001$) and the residualized EGE ($F(1, 3512) = 28.8, p < .0001$, after partialling out the effect of frequency), and an effect of negentropy ($F(1, 3512) = 7.4, p = .0006$, after partialling out the contributions of frequency and residualized EGE). A significant non-linearity was still present in the effect of negentropy ($L(6, 7) = 3.89, p = .0486$). The by-item regresion revealed significant linear effects of frequency ($F(1, 186) = 9.89, p = .0019$), residualized EGE

($F(1,186) = 14.00, p = .0002$) and negentropy ($F(1,186) = 5.18, p = .0240$), without any significant non-linearity in the effects. Introducing the residualized measure in the logistic regression on the error counts, revealed a non-linear effect of frequency ($\chi^2_2 = 18.53, p < .0001$; non-linear component $\chi^2_1 = 4.63, p = .0314$) and a linear effect of residualized EGE ($\chi^2_1 = 28.27, p < .0001$), with no significant contribution of negentropy ($\chi^2_1 = 2.05, p = .1524$). However, fast backwards elimination of factors using Akaike's information criterion recommended that, although not significant according to the χ^2 test, keeping the negentropy as a predictor in the regression model still produced a significant improvent in the quality of the fit.

As we had predicted, these analyses show that the unstability in the effect of negentropy is indeed a by-effect of its relatively small magnitude and its correlation with the EGE measure. However, although less marked, the attenuation of the effect of negentropy for the upper tertile remained significant in the multi-level regression (see Figure 14). This nonlinearity may reflect that our count is overestimated in its upper range, as a result of the multiple approximations that were performed to estimate it. This can also be a consequence of having used equal size samples to estimate the underlying distribution on all words, independently of their different frequencies of occurrence. This might have led to an overestimation of the negentropy for low frequency as compared to high frequency ones. However had we used unequally sized samples, it would have become very difficult to disentangle the effect of frequency from that of negentropy.

[INSERT FIGURE 14 AROUND HERE]

Finally, in order to compare the inhibitory effect of word homonymy described by Rodd et al. (2002) and Beretta et al. (2005), we added the factor homonymy (homonymic vs. non-homonymic) to the above regressions, after partialing out the contribution of the other effects.

In the multilevel regression, homonymy still had a significant effect ($F(1,3511) = 4.9, p = .0269$) on the RTs after partialing out the effects of frequency, EGE and negentropy.[5] In

[5]Both the effect of negentropy and that of homonymy remain significant independently of the order in which they are entered in the regression.

the by-item regression the effect of homonymy was marginally significant when included in a regression that also includes negentropy ($F(1, 185) = 2.96, p = .0872$), while negentropy retained its full significance. The backwards elimination of factors recommended that both negentropy and homonymy should be kept in the regression model. Finally, including homonymy as a factor in the logistic regression on the error counts did not show any additional effect ($\chi^2_1 = .59, p = .4433$), and fast backwards elimination of factors also suggested to remove it from the regression. Figure 15 compares the contributions of EGE, negentropy, and homonymy on the RTs. For simplicity, we have estimated the three effects using the by-item regression (thus the linearity of the effect of negentropy). It can be observed that the contribution of the homonymy factor is smaller than that of negentropy (approximately 15 vs. 25 ms.). This reflects the advantage of having a continuous (although innaccurate) estimate. Both effects are small compared to the effect of EGE (approximately 45 ms.).

[INSERT FIGURE 15 AROUND HERE]

These analyses show that the effects of negentropy and homonymy are indeed related and negentropy seems to be a more solid predictor (i.e., the inclusion of negentropy into the model substantially weakens the contribution of the homonymy factor). However there is still some additional variance explained by homonymy on the RTs. Again, this possibly reflects the limitations of our estimation process.

General Discussion

This study addressed the link between neurophysiological theories of language processing, and the effects that have been observed in behavioral experiments. We have shown that information-theoretical measures in combination with Bayesian distribution fitting provide us with a powerful tool to investigate this link. We have used multidimensional probability distributions to characterize four basic properties of Pulvermüller (1999)'s neurophysiological theory of lexical/morphemic processing:

i Words and morphemes are processed by neural assemblies.

ii The assemblies are formed by Hebbian association.

iii A discrete number of assemblies develops as a result of linguistic experience.

iv Different candidate assemblies will compete.

We have demonstrated the power of this technique by predicting the effects of the morphological and semantic negihborhoods of words on the response latencies and error scores of three visual lexical decision datasets. This is a promising method that enables us to achieve the integration between the different levels of explanation that were anticipated by Marr (1982).

Information theory and lexical neighborhoods

Information theory has a long tradition of use in psychological research. As early as 1949, Miller and Frick showed that response sequences in behavioral experiments could be measured in information-theoretical terms (Miller & Frick, 1949). In the field of motor behavior, the classical Fitts' Law (Fitts, 1954; Fitts & Peterson, 1965) constitutes a prime example of the application of information-theory to psychological theories.[6] Information measures have been shown to correlate with accuracy in discrimination tasks in auditory, gustatory, and visual modalities (see Attneave, 1959 and Miller, 1956 for reviews on early applications of information theory to psychology, and Baddeley, Hancock, and Földiák, 2000, for a more recent survey). More recently, entropy has been found to correlate with response latencies in cognitive control tasks (Koechlin, Orly, & Kouneiher, 2003). Koechlin et al. found that the amount of information (i.e., entropy) conveyed by the instruction cues, context, and stimuli of several functional magnetic resonance imaging experiments, predicts the amount of activation observed in several areas of the lateral pre-frontal cortex. Most interestingly this includes Broca's area (Brodman's areas 44 and 45), which is long assumed to be also involved in language processing.

Coming to the domain of language, the model that we have presented is related to previous information-theoretical models. Most evidently, our definition of negentropy builds

[6] In fact, in MacKenzie's reformulation (MacKenzie, 1992), Fitts' Law actually corresponds to the Shannon-Hartley theorem (Shannon, 1949) stating the limit of the capacity of a continuous-time analog communication channel.

on Kostić's finding of the inhibitory effect on response latencies of the count of syntactic functions and meanings of a Serbian inflectional affix (Kostić, 1991; 1995; 2005; Kostić et al., 2003). Our study extends this approach in several aspects: On the theoretical side, the BIT model provides a anchor for the models developed by Kostić and his colleagues, by showing how these effects might arise from the properties of the neurophysiological system. Furthermore, in the behavioural level, our model integrates the effects reported by Kostić et al. for Serbian inflectional morphology, with those reported by Moscoso del Prado Martín et al. (2004) for Dutch inflectional and derivational morphology, and the effects reported by Rodd et al. (2002) on the semantic level. Finally, more on the technical side, the techniques that we have developed here also provide a method for automatically estimating Kostić (1965)'s counts in languages where resources as detailed as the Corpus of Serbian Language are not available.

Of particular similarity to ours, is the approach presented by McDonald and Shillcock (2001): They developed an information-theoretical measure of the predictability of the contexts in which a word appears: the Contextual Distinctiveness (CD) of a word. Their experiments showed that this magnitude is positively correlated with visual lexical decision RTs. Notice the similarity between CD and the EGE measure that we have developed here. Indeed, we believe that both of them consitute different approximations of the same measure. However, there are issues that differentiate our approach from that of McDonald and Shillcock. First, the CD measure does not take into consideration the presence of discrete assemblies that might compete with each other, as measured by the negentropy of the distribution. As a consequence, CD alone would not be capable of predicting the inhibitory effects of multiple unrelated meanings or morpho-syntactic functions. Second, CD is calculated on the basis of first order co-occurrence vectors, which are representations of the 'average' meaning of a word, instead of our explicit consideration of each particular occurrence of a word (using the second order vectors). In addition, McDonald and Shillcock's approach is not concerned with the explicit neurobiological properties of the underlying system, which constitute the major motivation for our measures. As in the case above, our model can be considered as a way of grounding the results of McDonald and Shillcock on the properties of neural assemblies. There is however, one point of disagreement between

the results presented here and those of McDonald and Shillcock. They observe that word frequency and CD are interchangeable as predictors of the lexical decision RTs, and they conclude that both magnitudes are reflecting the same effect (i.e., frequent words are recognised faster than infrequent ones because the former tend to appear in a wider variety of contexts then the later). However, as shown by Analyses 2 and 3, both the variability of the contexts in which a word occurs (measured by the EGE) and its frequency, have independent contributions to explaining the RTs and error scores. Our results are consistent with the independent effects that of word frequency and semantic variability (Jastrzembski, 1981). Therefore, our analyses support the separate consideration of the effects the word frequency effect and variability in usage of words.

We have addressed the questions of the origin of these neighborhood effects, and of the reasons why information-theoretical measures appear to be most succesful in characterizing them. We have seen how the processes of inter-assembly competition can be characterized by the negentropy of a probability distribution, that follows directly from the assumption of competition between assemblies. This is directly linked with the inhibitory effects of the number of syntactic functions and meanings of a Serbian inflectional morpheme (Kostić et al., 2003), and with the inhibitory effect of the degree of homonymy of an English word (Rodd et al., 2002). Simultaneusly, we have described the how the general width of the probability distribution – measured by the EGE – is related to the ease of activation of a particular set of assemblies. This measure is negatively correlated with response latencies and error scores, thus predicting the effects of semantically related morphological relatives described by Moscoso del Prado Martín et al. (2004), and of the number of related meanings (Jastrzembski, 1981; Rodd et al., 2002). It remains to be seen whether the effects of phonological and orthographic neighborhoods can also be replicated using this techniques, but the large amount of parallelism to morphological and semantic neighborhoods suggest that this might also be the case. In this respect, Luce and Large (2001) described the effects of phonological neighborhoods using a measure which is very related to the inflectional entropy measure reported in morphology.

Interestingly, in the relatively distant domain of spoken speech processing, Aylett (2000) presents a model that bears a surprising degree of similarity to ours. He developed a measure

of the Clarity of a particular speaker calculated on the probability distribution of the frequency formants $F1$ and $F2$ of a sample of vowels produced by that speaker. He showed that a speaker's Clarity correlates with the error scores produced by participants whose task was to auditorily recognize utterances by the same speaker. Aylett used the EM algorithm to fit the probability distribution as a finite mixture of Gaussians (with the number of components fixed to the number of English vowels). Although Aylett explicitly denies any relationship between his Clarity measure and entropy (p. 218), his measure is in fact identical (save for a change in sign) to the MonteCarlo estimator that we used for estimating the entropy the multidimensional distributions. He reports that the Clarity measure correlates positively with the ability of subjects to recognize words, however, this effect seems to be weak. Note that the Clarity measure is equivalent to a negentropy measure without normalizing for the EGE. Another iterpretation of the results would thus be that subjects are better at recognizing vowels whenever there is more information about them present in the signal. Of course, to measure this information (as the degree of clustering in the vowel space) one would also need to normalize for the general width of the distributions (through EGE). We believe that this parallelism is not merely coincidental: The mixture of Gaussians employed to model the probability distribution of the formants could well correspond to the underlying neural structures representing the phonemes.

But, is the BIT model Bayesian?

Bayesian inference techniques are currently gaining a prominent position in many sciences. These techniques formalize of the inferences that can be drawn from a given set of data, making explicit the set of assumptions that are made in the inference process. The general principle is to use Bayes' theorem to estimate the probabilities ("posterior probabilities" in Bayesian jargon) of each possible conclussion given a set of assumptions ("priors"), observed data ("evidence" or "likelihood"). Detailed introductions to the techniques of Bayesian statistics can be found in MacKay (2003) and Sivia (1996). In psychology, Bayesian statistics have been employed to describe a large amount of phenomena (cf., Oaksford & Chater, 1998). In addition, Körding and Wolpert (2004) have shown that Bayesian inference is also employed

by the human central nervous system in tasks involving sensory-motor integration, with evidence for explicit representation of prior and likelihood probability distributions.

In our approach we have used Bayesian inference tools to estimate the probability density functions of the distribution of usages or meanings of a word or inflectional affix. If we ignore the hyper-parameters of the inference model (i.e., the parameters that govern the prior assumptions on the distributions of the actual parameters of the model), our main prior assumption in estimating the distributions is that they will be mixtures of multidimensional Gaussians with a finite – but a-priori unknown – number of components with different amplitudes, centroids, and covariances. As we explained above, this assumption comes as a consequence of Pulvermüller's neurophysiological model: If the different meanings of words and morphemes are represented by partially-overlapping neural assemblies developed by Hebbian association, these should develop from the different 'bumps' of a multi-modal distribution in a multi-dimensional representation of the meanings encountered by experience. Once we have assumed that the underlying distribution is multi-modal, the assumption that each of the components will correspond to a Gaussian distribution follows from the Gaussian shapes of the receptive fields of the neurons that make up an assembly. Note however, that this last point is not crucial for justifying the normality of the components. One of the fundamental principles of Bayesian theory, the Maximum Entropy Principle (MaxEnt; cf., Sivia, 1996), estates that on the lack of any a-priory information on the shape of a probability distribution, one should assume the least informative probability distribution, which will be the one with the largest possible entropy. In this respect, the MaxEnt principle can be seen as an operationalization of the traditional Occam's razor. As it happens, the continuous probability distribution with the largest possible entropy is the Gaussian distribution (cf., Cover & Thomas, 1991), thus assuming that each component is a Gaussian constitutes the minimum assumption on the lack of additional knowledge.

Although we have used Bayesian inference for estimating the distributions, it could be argued that the BIT model is not in itself Bayesian. The model would then be seen as independent of the method of inference that is employed. However, as we have argued above, the assumption of an infinite mixture of Gaussians as a prior is itself a consequence of the underlying theory. The model is therefore Bayesian not only in the techniques used

for fitting the distributions but also in the more fundamental sense that it depends on the representation of an explicit prior distribution (the mixture of Gaussians) and the use of a likelihood function of the observations given that prior. It is important to highlight here that it is the actual neurophysiological properties of the system that consititute the initial prior. Although in our approach we have included all observations at the same time, in principle this process would be a gradual one by which the prior distribution would be updated with each new observation (which would be achieved by the Hebbian Long Term Potentiation and Long Term Depression processes acting on the neural synapses). Note here that this mechanism offers a natural way to study the evolution of the distribution during the development of the language abilities: in principle one could fit models with different degrees of language experience and make predictions on the conditions this should impose on behaviour. This is left for further research.

Relationship to distributed and localistic models

Our model coincides in its basic properties with the assumptions of distributed connectionist models (DCMs; Rumelhart & McClelland, 1986): It is based on the usage of distributed representations fully acquired through experience, and it relies on 'domain-general' properties of the cognitive system rather than on language-specific mechanisms. However, the differences between our approach and distributed connectionist models should not be overlooked.

The first difference is a question of scope and goals. On the one hand, DCMs investigate the types of information and forms of representation that are necesary for language processing. The use of these models has made – and still does – important contributions to our understanding of the lexical processing system at Marr's computational level of explanation. Indeed, our model is heavily indebted to DCM research on the assumptions it makes. On the other hand, our approach investigates the link between the cognitive processes that are targeted by DCMs, and an underlying neurophysiological theory. Consequently, the BIT approach is subject to a more stringent constraint on biological plausibility than DCMs are. For instance, issues like the biological plausibility of the learning algorithm are irrelevant when one's goal is to demonstrate the influence of statistical factors in language processing

or the importance of the similarity structure in the data. However, for the purposes of the BIT model, this is an issue of central importance since the goal is precisely to specify how do the underlying neural mechanisms account for the observed effects. The BIT approach should therefore be viewed as complementary to DCMs, rather than as an alternative.

Second, the BIT model does not attempt to simulate the processing of words or morphemes. Instead, in our approach we make predictions on the behavioural results following directly from the mathematical formalization of the underlying neurophysiological theory. As a result, in the BIT model there is no analog of measurements such as RTs. Instead we explicitly quantify the variables that should influence the RTs, which in the examples that we have presented are the negentropy and EGE of the distribution of meanings (or usages) of a morpheme. On the one hand, connectionist models can be described as high-level simulations of the psychological processes in question. The BIT model is, on the other hand, a direct mathematical model of an underlying neurophysiological theory. This might seem a minor terminological issue, however it is of significance when it comes to the interpretation of the sources of the underlying effects: the BIT model provides a direct rationale for the observed effects in terms of neural structures.

A third, and may be most salient aspect that distinguishes the BIT approach from DCMs is the usage of explicit and discrete symbolic structures – the neural assemblies represented as the components of the Gaussian mixture – in combination with the distributed representations that define the receptive fields for each assembly. This contrasts with the traditional view held by supporters of DCMs, where each unit participates in the representation of all entitites present in the system. DCMs have been criticized for the problems entailed by this committment to fully distributed representation: difficulty for interpreting the behavior of the system, and lack of ability to represent discrete complex structures (Fodor & Pylyshyn, 1988). However, proposers of the DCM framework have long been aware of these problems. As stated by Hinton (1991):

> "Most connectionist researchers are aware of the gulf in representational power between a typical connectionist network and a set of statements in a language such as predicate calculus. They continue to develop the connectionist framework

> not because they are blind to its current limitations, but because they aim to eventually bridge the gulf by building outwards from a foundation that includes automatic learning procedures and/or massively parallel computation as essential ingredients." (p. 2)

Indeed, some recent connectionist models have made use of 'locally-tuned' processing units with Gaussian receptive fields that are trained by Hebbian association (e.g., Moody & Darken, 1989; Westermann & Mareschal, 2004). This family of models is in fact very related to the statistical model that we have proposed here, with the exception that the number of Gaussians that are used to model the distribution is set a priory in the number of units in the 'hidden' layer.

Note here that, although the representational scheme is localistic in the sense that there is a discrete number of 'symbols' which correspond to directly interpretable entities (e.g., morphemes, word senses, etc.), different symbols can share many of the neurons of which they are made. This is in contrast with the classical notion of purely localist models, since in these models each symbol would univocally correspond to one entity, without the possibility of overlap among symbols. However, current research in localist connectionism suggests the possibility of two levels of representation. For instance, Page (2000) proposes two layers of representation, in which one represents the stimuli in a completely distributed manner (L_1 in Page's notation), and the other is fully localistic in the traditional sense (L_2). The localistic (L_2) units would correspond to the Gaussian components in our model, while the distributed representations in the L_1 layer would directly correspond to the underlying multidimensional patterns of activation in our model. A recent example of such type of models is the Bayesian Reader (Norris, in press). In this model, distributed representations of a word's orthography, are used as inputs to an additional localistic layer of units, each of which corresponds to a word. This type of localistic model, using underlyingly distributed representations is fully in line with the approach that we are proposing.

The BIT model introduces a physiologically motivated way of accomodating the combination of distributed and localist mechanisms that have been described in previous distributed and localist connectionist models. It also underlines the fact that, in their more recent

forms, localist and distributed connectionists models are becoming more and more similar in terms of their properties. On the one hand, interpretability and need for usage of complex structures calls for the presence of discrete localist representations. On the other hand, the role palyed by gradient-like properties and patterns of interference suggests the need for distributed patterns (cf., Page, 2000; Rumelhart & McClelland, 1986). In our approach, we rely on a neurophysiological theory that implies the co-existence of both mechanisms with one being the natural consequence from the other.

Is our approach tied to Pulvermüller's theory?

We have followed the predictions of Pulvermüller (1996, 1999, 2001) into behavioural measures. In this way, Pulvermüller's theory is an intrinsic part of the model that we have presented here. The core of our results relies on the truth of the underlying assumptions (such as the Gaussian mixture). In this sense our model consititutes a direct implementation of the theory's prediction. However, it should also be noted that we have also presented a methodology for predicting behavioral results from an underlying theory. Our method could thus be applied with different underlying assumptions (for instance assuming that the shape of the distributions should be different than the Gaussian mixtures employed here). The techniques that have been demonstrated in this study could be used to compare the posterior probabilities of different candidate theories in the light of the experimental data. For these purposes, the Bayesian framework offers simple and elegant ways of comparing different candidate theories.

Conclusion

In this study, we have illustrated how a combination of techniques from Bayesian Statistics and Information Theory, can be employed to link the results obtained by behavioral and neurobiological research on the human language processing system. Although further research is required to be able to explain a wider variety of psychological phenomena related to lexical processing, the current study contributes a promising new approach for understanding how words are represented and processed in the human brain, also providing a meeting point for

distributed connectionist and localist theories. We have shown how the presence of neural assemblies developed through Hebbian association, as proposed by Pulvermüller (1999), is sufficient to explain the effects of competition and facilitation between members of morphological and semantic neighborhoods that have been observed in behavioral experiments. Furthermore, our studies provide a grounding for the information-theoretical approaches to the study of lexical processing. Information-theory provides us with a very powerful tool to investigate language. In fact, language was one of the problems for which information-theory was explicitly developed, as evidenced the seminal study of Shannon (1948).

Acknowledgements

This research was funded by the European Commission through a Marie Curie Intra-European Fellowship (MC–EIF–010318) to the first author.

Appendix A: Entropy and Negentropy

In this Appendix we provide an overview of the concepts from Information Theory that are used in this paper. For a detailed discussion of the concepts related to differential entropy the reader should consult Chapter 9 of Cover and Thomas (1991). An in-depth discussion of the probabilistic concept of negentropy can be found in Brillouin (1956), and a more recent discussion of its use in ICA is provided by Hyvärinen (1998; 1999).

Entropy

The entropy (Shannon, 1948) of a random variable X over a discrete range of possible values $\{x_i\}$ is defined as the expectation of the logarithm of its inverse probability, that is:

$$H(X) = \sum_i P(X = x_i) \cdot \log_2 \frac{1}{P(X = x_i)} = -\sum_i P(X = x_i) \cdot \log_2 P(X = x_i). \qquad \text{(A-1)}$$

This measure represents the uncertainty on the value of X contained on its probability distribution $P(X)$. In terms of information transmission, this quantity represents the minimum number of bits per draw that would be necessary to transmit over a binary channel a sequence of events drawn according to $P(X)$.

The differential entropy (Shannon, 1948) is an extension of the concept of entropy for random variables defined over a continuous space. Given a continuous variable x defined over a space S with a probability density function $p(x)$, its differential entropy is defined by:[7]

$$h(p) = -\int_S p(x) \log_2 p(x) dx, \qquad \text{(A-2)}$$

Note here that, unlike the entropy in the discrete case, the differential entropy is not bound to have positive values (since the value of the probability density function can be greater than one, unlike the probability in the discrete case), and the magnitude is only defined for

[7] The base of the logarithm is only a factor of scale. In the domain of discrete variables binary logarithms have traditionally been aployed, as this results in the entropy being measured in bits, which are easily understandable units by the analogy to pulses in a digital system. However, in the continuous domain it is more common to use the natural logarithm (base e), which results in the quantities being measured in nats. Converting from nats to bits only involves scaling by a factor of $\log_2(e)$.

probability density functions for which the integral in (A-2) converges. As in the discrete case, the differential entropy is a measure of the uncertainty in the probability distribution $p(x)$. High values of $h(p)$ correspond to high uncertainty on the expected value of x. The definition of differential entropy introduced in (A-2) is also valid on multidimensional spaces, simply by substituting x for the corresponding vector $\mathbf{x}$ and integrating on a multidimensional space.

Negentropy

Negentropy (sometimes called normalized negative entropy, or negative Shannon-Jaynes entropy) is a probabilistic concept introduced by Brillouin (1956)[8] to describe the amount of organization present in a system. Note here that, while entropy is a measure of the uncertainty or disorder present in a system, the negentropy measures the amount of order or information in that same system. More recently, an operationalization of this probabilistic measure has become widely used for selection of components in Independent Component Analysis (Comon, 1994). Formally the measure provides an index of how much does a random variable deviate from normality.

The negentropy of a continuous probability distribution $p(x)$ is operationalized as the difference between a probability distribution and a Gaussian distribution with equal mean and covariance. In this way negentropy is measuring the amount of order that is present in the system, in relation to a situation of maximum disorder, which would be characterized by a Gaussian distribution (Comon, 1994). In Information-Theory, the difference between probability distribution is measured by the Kullback-Leibler divergence (also known as cross-entropy) between their probability density functions:

$$J(p) = KL(p||p_{\mathcal{N}}) = \int_S p(x) \log_2 \frac{p(x)}{p_{\mathcal{N}}(x)} dx, \tag{A-3}$$

which for our purposes can be reduced to:

$$J(p) = h(p_{\mathcal{N}}) - h(p) \tag{A-4}$$

[8] Although the term was first coined by Brillouin, the original concept in statistical physics can be traced back to Schrödinger (1944).

where S is the space over which the distributions are defined, p is a PDF, $h(p)$ is its differential entropy, $p_{\mathcal{N}}$ is a normal distribution with equal variance to that of p, and $h(p_{\mathcal{N}})$ is the differential entropy of that normal distribution. The definition of $J(p)$ in (A-4) can be intuitively interpreted as the reduction in disorder from $p_{\mathcal{N}}$ to p .

Note that all probability distributions for which the entropy is defined verify $h(p) \leq h(p_{\mathcal{N}})$, since $h(p_{\mathcal{N}})$ is the maximum possible entropy. Therefore, according to (A-4), unlike the differential entropy, the negentropy of a probability distribution is always greater than zero (being zero if and only if the original variable is itself a normal distribution).

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.

Andrews, S. (1997), 'The effects of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts', Psychological Bulletin & Review 4, 439–461.

Attneave, F. (1959), Applications of information theory to psychology: a summary of basic concepts, methods, and results, Holt, Rinehart, and Winston, New York.

Aylett, M. (2000), Modelling clarity of change in spontaneous speech, in R. Baddeley, P. Hancock & P. Földiák, eds, 'Information Theory and the Brain', Cambridge University Press, Cambridge, U.K., pp. 204–220.

Azuma, T. & Van Orden, G. C. (1997), 'Why safe is better than fast: The relatedness of a word's meaning affects lexical decision times', Journal of Memory and Language 36, 484–504.

Baayen, R., Feldman, L. & Schreuder, R. (2005), 'A principal components regression analysis of simple word recognition', Manuscript submitted for publication, Max Planck Institute for Psycholinguistics.

Baayen, R. H. (2005), Data mining at the intersection of psychology and linguistics, in A. Cutler, ed., 'Twenty-First Century Psycholinguistics: Four Cornerstones', Erlbaum, Hillsdale, N.J.

Baayen, R. H. & Moscoso del Prado Martín, F. (2005), 'Semantic density and past-tense formation in three Germanic languages', Language 81(3), 666–698.

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.

Baddeley, R., Hancock, P. & Földiák, P. (2000), Information Theory and the Brain, Cambridge

University Press, Cambridge, U.K.

Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., Simpson, G. B. & Treiman, R. (2002), 'The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 english words and nonwords.', Washington University, St. Louis, MO. <http://elexicon.wustl.edu/>.

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

Beretta, A., Fiorentino, R. & Poeppel, D. (2005), 'The effects of homonymy and polysemy on lexical access: and MEG study', Cognitive Brain Research 24(1), 57–65.

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

Borowsky, R. & Masson, M. E. J. (1996), 'Semantic ambiguity effects in word identification', Journal of Experimental Psychology: Learning Memory and Cognition 22, 63–85.

Brillouin, L. (1956), Science and Information Theory, Academic Press, New York.

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

Coltheart, M., Rastle, K., Perry, C., Langdom, R. & Ziegler, J. (2001), 'Effects of word imageability and age of acquisition on children's reading', Psychological Review 108(1), 204–256.

Comon, P. (1994), 'Independent component analysis - a new concept?', Signal Processing 36, 287–314.

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

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.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', Journal of the Royal Statistical Society 39(1), 1–38.

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 recog-

nition', Language and Cognitive Processes 20(1), 7–42.

Edelman, S. (in press), Bridging language with the rest of cognition: computational, algorithmic and neurobiological issues and methods, in 'Proceedings of the Ithaca EMCL Workshop', John Benjamins.

Fitts, P. M. (1954), 'The information capacity of the human motor system in controlling the amplitude of movement', Journal of Experimental Psychology 47, 381–391.

Fitts, P. M. & Peterson, J. R. (1964), 'Information capacity of discrete motor responses', Journal of Experimental Psychology 67, 103–112.

Fodor, J. A. & Pylyshyn, Z. W. (1988), 'Connectionism and cognitive architecture: a critical analysis', Cognition 28, 3–71.

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

Goldinger, S. D., Luce, P. A. & Pisoni, D. B. (1989), 'Priming lexical neighbors of spoken words: Effects of competition and inhibition', Journal of Memory and Language 28, 501–518.

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

Grainger, J. & Segui, J. (1990), 'Neighborhood frequency effects in visual word recognition: A comparison of lexical decision and masked identification latencies', Perception and psychophysics 47, 191–198.

Grossberg, S., Boardman, I. & Cohen, M. (1997), 'Neural dynamics of variable-rate speech categorization', Journal of Experimental Psychology: Human Perception and Performance 23, 483–503.

Hebb, D. O. (1949), The organization of behavior. A neuropsychological theory, John Wiley & Sons, New York.

Hinton, G. E. (1991), 'Preface to the Special Issue on Connectionist Symbol Processing', Artificial Intelligence 46, 1–4.

Holcomb, P. J., Grainger, J. & O'Rourke, T. (2002), 'An electrophysiological study of the effects of orthographic neighborhood size on printed word perception', Journal of Cognitive Neuroscience 14(6), 938–950.

Hyvärinen, A. (1998), New approximations of differential entropy for independent component

analysis and projection pursuit, in 'Advances in Neural Information Processing Systems', Vol. 10, The MIT Press, Cambridge, MA, pp. 273–279.

Hyvärinen, A. (1999), 'Fast and robust fixed-point algorithms for Independent Component Analysis', IEEE Transactions on Neural Networks 10(3), 626–634.

Jastrzembski, J. (1981), 'Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon', Cognitive Psychology 13, 278–305.

Kanerva, P., Kristofersson, J., & Holst, A. (2000), Random indexing of text samples for Latent Semantic Analysis, in L. R. Gleitman & A. K. Josh, eds, 'Proceedings of the 22nd Annual Conference of the Cognitive Science Society', Lawrence Erlbaum, Mahwah, NJ, p. 1036.

Kellas, G., Ferraro, F. R. & Simpson, G. B. (1988), 'Lexical ambiguity and the timecourse of attentional allocation in word recognition', Journal of Experimental Psychology: Human Perception and Performance 14, 601–609.

Koechlin, E., Ody, C. & Kouneiher, F. (2003), 'The architecture of cognitive control in the human prefrontal cortex', Science 302, 1181–1185.

Körding, K. P. & Wolpert, D. M. (2004), 'Bayesian integration in sensorimotor learning', Nature pp. 244–247.

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

Kostić, Đ. (2001), 'Kvantitativi opis strukture srpskog jezika – Korpus Srpskog Jezika ('Quantitative description of Serbian language structure – the Corpus of Serbian Language')', Institute for Experimental Phonetics and Speech Pathology & Laboratory of Experimental Psychology, University of Belgrade, Serbia and Montenegro <http://www.serbian-corpus.edu.yu/>.

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

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

Kostić, A. (2005), 'The effects of the amount of information on processing of inflected morphol-

ogy', Manuscript submitted for publication, University of Belgrade.

Kostić, A., Marković, T. & Baucal, A. (2003), Inflectional morphology and word meaning: orthogonal or co-implicative domains?, in R. H. Baayen & R. Schreuder, eds, 'Morphological structure in language processing', Mouton de Gruyter, Berlin, pp. 1–44.

Kraskov, A., Stögbauer, H. & Grassberger, P. (2004), 'Estimating mutual information', Physical Review E 69, 066138.

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.

Landauer, T., Laham, D., Rehder, B. & Schreiner, M. (1997), How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans., in 'Proceedings of 19th annual meeting of the Cognitive Science Society', Mawhwah, NJ., pp. 412–417.

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

Lowe, W. & McDonald, S. (2000), The direct route: Mediated priming in semantic space, in 'Proceedings of the 22nd Annual Conference of the Cognitive Science Society'.

Luce, P. A. & Large, N. R. (2001), 'Phonotactics, density, and entropy spoken word recognition', Language and Cognitive Processes 16(5/6), 565–581.

Luce, P. A. & Pisoni, D. B. (1998), 'Recognizing spoken words: the Neighborhood Activation Model', Ear & Hearing 19, 1–36.

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

Lund, K., Burgess, C. & Atchley, R. A. (1995), Semantic and associative priming in high-dimensional semantic space, in 'Proceedings of the 17th Annual Conference of the Cognitive Science Society', Erlbaum, Hillsdale, NJ., pp. 660–665.

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

MacKenzie, I. S. (1992), 'Fitts' law as a research and design tool in human computer interaction', Human-Computer Interaction 7, 91–139.

Marr, D. (1982), Vision: A computational investigation into the human representation and processing of visual information, Freeman & Co., San Francisco.

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'.

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

Miller, G. A. (1956), 'The magical number seven, plus or minus two: Some limits on our capacity for processing information', Psychological Review 63, 81–97.

Miller, G. A. & Frick, F. C. (1949), 'Statistical behavioristics and sequences of responses', Psychological Review 56, 311–324.

Miller, R. & Wickens, J. R. (1991), 'Corticostriatal cell assemblies in selective attention and in representation of predictable and controllable events: a general statement of corticostriatal interplay and the role of striatal dopamine', Concepts in Neuroscience 2, 65–95.

Moody, J. & Darken, C. (1989), 'Fast learning in networks of locally tuned processing units', Neural Computation 1, 289–303.

Moscoso del Prado Martín, F. & Baayen, R. H. (2005), 'Breaking the tyranny of learning: a broad-coverage distributed connectionist model of visual word recognition', Manuscript, MRC–Cognition and Brain Sciences Unit.

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. (2005), 'Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch', Journal of Memory and Language 53, 496–512.

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

Neal, R. M. (1991), Bayesian mixture modeling by Montecarlo simulation, Technical Report CRG-TR-91-2, Department of Computer Science, University of Toronto.

Neal, R. M. (1998), Markov chain sampling methods for Dirichlet process mixture models., Technical Report No. 9815, Department of Statistics, University of Toronto.

Neal, R. M. (2004), 'Software for flexible Bayesian modeling and Markov chain sampling (release of 10-11-2004)', Department of Statistics, University of Toronto, Canada. <http://www.cs.toronto.edu/~radford/fbm.software.html>.

Norris, D. (2005), How do computational models help us build better theories?, in A. Cutler, ed., 'Twenty-First Century Psycholinguistics: Four Cornerstones', Erlbaum, Hillsdale, N.J., pp. 331–346.

Norris, D. (in press), 'The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process', Psychological Review.

Oaksford, M. & Chater, N. (1998), Rational Models of Cognition, Oxford University Press, Oxford, U.K.

Oja, E. (1982), 'A simplified neuron model as a principal component analyzer', Journal of Mathematical Biology 15, 267–273.

Page, M. (2000), 'Connectionist modelling in psychology: a localist manifesto', Behavioral and Brain Sciences 23(4), 443–467.

Pulvermüller, F. (1996), 'Hebb's concept of cell assemblies and the psychophysiology of word processing', Psychophysiology 33, 317–333.

Pulvermüller, F. (1999), 'Words in the brain's language', Behavioral and Brain Sciences 22, 253–336.

Pulvermüller, F. (2001), 'Brain reflections of words and their meaning', Trends in the Cognitive Sciences 5, 517–524.

Pulvermüller, F. (2002), 'A brain perspective on language mechanisms: from discrete neuronal ensembles to serial order', Progress in Neurobiology 67, 85–111.

Pulvermüller, F. (2003), The Neuroscience of Language, Cambridge University Press, Cambridge (U.K.).

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.

Pylkkänen, L., Llinás, R. & Murphy, G. L. (in press), 'The representation of polysemy: MEG evidence', Journal of Cognitive Neuroscience.

Rodd, J., Gaskell, M. G. & Marslen-Wilson, W. D. (2002), 'Making sense of semantic ambiguity: semantic competition and lexical access', Journal of Memory and Language 46, 245–266.

Rodd, J., Gaskell, M. G. & Marslen-Wilson, W. D. (2004), 'Modelling the effects of semantic ambiguity in word recognition', Cognitive Science 28, 89–104.

Rolls, E. T. & Deco, G. (2001), The computational neuroscience of vision, Oxford University Press, Oxford, U.K.

Rumelhart, D. E. & McClelland, J. L., eds (1986), Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations, MIT Press, Cambridge, Mass.

Schone, P. & Jurafsky, D. (2001), Knowledge free induction of inflectional morphologies, in 'Proceedings of the North American Chapter of the Association for Computational Linguistics NAACL-2001'.

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

Schrödinger, E. (1944), What is Life?, Cambridge University Press, Cambridge, UK.

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

Schütze, H. (1994), Towards connectionist lexical semantics, in S. D. Lima, R. L. Corrigan & G. K. Iverson, eds, 'The Reality of Linguistic Rules', Vol. 26 of Studies in Language Companion Series, John Benjamins Publishing Company, Amsterdam, PA., pp. 171–191.

Schütze, H. (1995), Distributional part-of-speech tagging, in 'EACL 7', pp. 251–258.

Schütze, H. & Pedersen, J. O. (1997), 'A cooccurrence-based thesaurus and two applications to information retrieval', Information Processing & Management 33(3), 307–318.

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

Shannon, C. E. (1949), 'Communication in the presence of noise', Proceedings of the Institute of Radio Engineers 37(1), 10–21.

Shtyrov, Y. & Pulvermüller, F. (2002), 'Memory traces for inflectional affixes as shown by the mismatch negativity', European Journal of Neuroscience 15(6), 1085–1091.

Sivia, D. S. (1996), Data Analysis: A Bayesian Tutorial, Oxford University Press, Oxford, U.K.

Slotnik, S. D., Moo, L. R., Kraut, M. A., Lesser, R. P. & Hart, J. (2002), 'Interactions between thalamic and cortical rythms during semantic memory recall in human', Proceedings of

the National Academy of Sciences U.S.A. 99(9), 6440–6444.

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

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

Van Hulle, M. M. (2005a), 'Edgeworth approximation of multivariate differential entropy', Neural Computation 17(9), 1903–1910.

Van Hulle, M. M. (2005b), Multivariate Edgeworth-based entropy approximation, in 'Proceedings of the 2005 IEEE Workshop on Machine Learning for Signal Processing', Mystic, CT.

Vitevitch, M. S. & Luce, P. A. (1999), 'Probabilistic phonotactics and neighborhood activation in spoken word recognition', Journal of Memory and Language 40, 374–408.

Westermann, G. & Mareschal, D. (2004), 'From parts to wholes: mechanisms of development in infant visual object processing', Infancy 5(2), 131–151.

List of Figure Captions

Figure 1: Representation of a possible distribution of occurrences of words in a hypothetical two-dimensional space in which word forms and word meanings could be coded each with a single real number. Each point in the scatter correspond to the occurrence of a particular word form with a particular meaning.

Figure 2: Probability density function corresponding to the distribution of the points in Figure 1. Each of the 'bumps' in the distribution corresponds to a Gaussian component of the mixture model. Neural assemblies would be formed around these areas.

Figure 3: Effect of conditioning the probability distribution from Figure 2 to a particular word form.

Figure 4: Illustration of the distributions employed to estimate negentropy and EGE. The black line plots the density function of a (unidimensional) Gaussian mixture with five components, and the grey line corresponds to the density of a Gaussian distribution with equal mean and variance.

Figure 5: Relationship between negentropy (horizontal axis) and number of syntactic functions and meanings following Kostić (1965) (vertical axis) for the Serbian masculine and feminine nominal suffixes used in Analysis 1. Note that the correlation is significant both by parametric (Pearson) and non-parametric methods (Spearman).

Figure 6: Explanatory power of the frequency of the inflectional suffix on average lexical decision RTs to Serbian inflected nouns. The effect as been estimated by a linear model including log frequency as a fixed effect and word gender (masculine vs. feminine) as a random effect, to account for the fact that the RTs to nouns of different genders collected in two experiments (Kostić et al., 2003). Note that by themselves, gender and frequency account for up to three quarters of the RT variance.

Figure 7: Comparison of the effects of the (log) number of syntactic functions and meanings (left panel) and negentropy (right panel) on explaining the RT residuals from the

regression (Figure 6) using (log) frequency as a fixed effect and gender as a random effect. Notice the similar predictive power of both measures.

Figure 8: Combined explanatory power of inflectional suffix frequency and number of syntactic functions and meanings on average lexical decision RTs to Serbian inflected masculine and feminine nouns. The effect as estimated by a linear model including log frequency and log number of syntactic functions and meanings as a fixed effects and word gender (masculine vs. feminine) as a random effect. Note that adding number of syntactic functions and meanigns to the regression increases its explanatory power in up to 15% over the variance explained by frequency (see Figure 6).

Figure 9: Combined explanatory power of inflectional suffix frequency and negentropy on average lexical decision RTs to Serbian inflected masculine and feminine nouns. The effect as estimated by a linear model including log frequency and log number of syntactic functions and meanings as a fixed effects and word gender (masculine vs. feminine) as a random effect. Note that substituting number of syntactic functions and meanigns by negentropy in the regression decreases its explanatory power by up to 5% (see Figure 8), but still constitutes up a 10% over frequency (see Figure 6).

Figure 10: Correlation between EGE (horizontal axix) and inflectional entropy (vertical axis) in the dataset of Analysis 2. The inflectional entropy measure has been calculated using the method described in Moscoso del Prado Martín et al. (2004).

Figure 11: Summary of the effects found in Analysis 2 on RTs (top row), and error scores (bottom row). The left column illustrates the effects of word frequency, and the middle and right columns respectively show the effects of EGE and inflectional entropy after partialling our the contribution of word frequency.

Figure 12: Summary of the non-linear effects found on RTs in Analysis 3. The left panel illustrates the effect of EGE (after partialling out the effect of word frequency), and the right panel shows the effects of negentropy entropy after partialling our the contributions of word frequency and EGE. Notice that, in both cases, the effects seem to be attenuated in the higher ranges of the counts.

Figure 13: Illustration of the process of marginalization applied on the EGE measure on Analysis 3. The left panel shows the weak correlation between EGE and negentropy before marginalization. The right panel shows how, once any contribution of negentropy to the EGE count has been removed, the correlation dissappears, but the modification to the joint distribution of the variables is minimal.

Figure 14: Effect of negentropy on the RTs of Analysis 3 after partialling out the contributions of word frequency and residualized EGE. Note that, although the atteanuation of the effect in the higher range of the negentropy measure is now weaker (compare to the right panel in Figure 12), there is still a significant nonlinearity in the effect.

Figure 15: Summary of the effects found on RTs in Analysis 3. The left panel illustrates the effect of EGE (after partialling out the effect of word frequency), the middle and right panels show the effects of negentropy entropy, and homonymy (after partialling our the contributions of word frequency and EGE).

Figure 1:

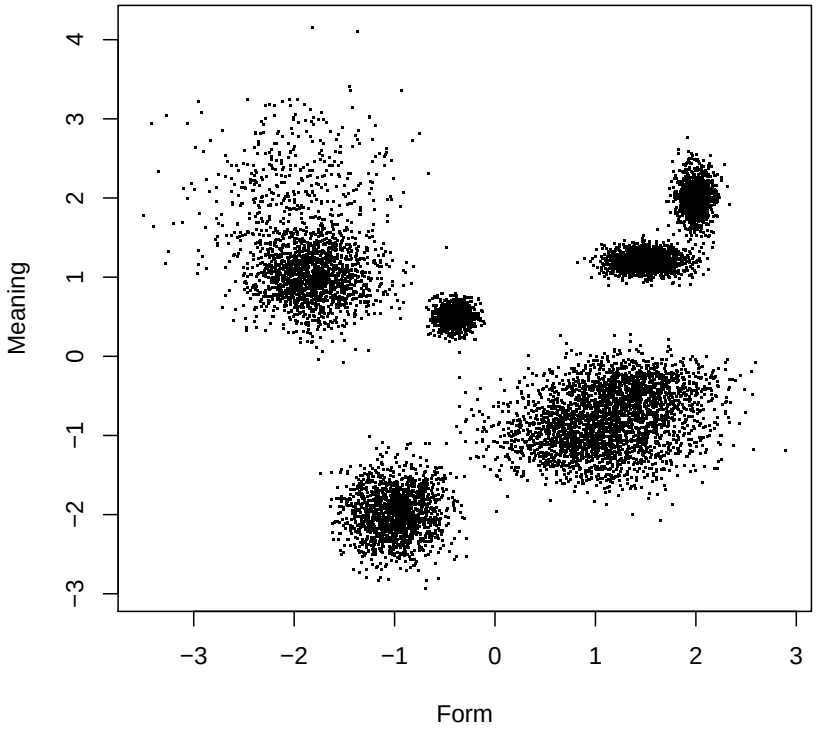


Figure 2:

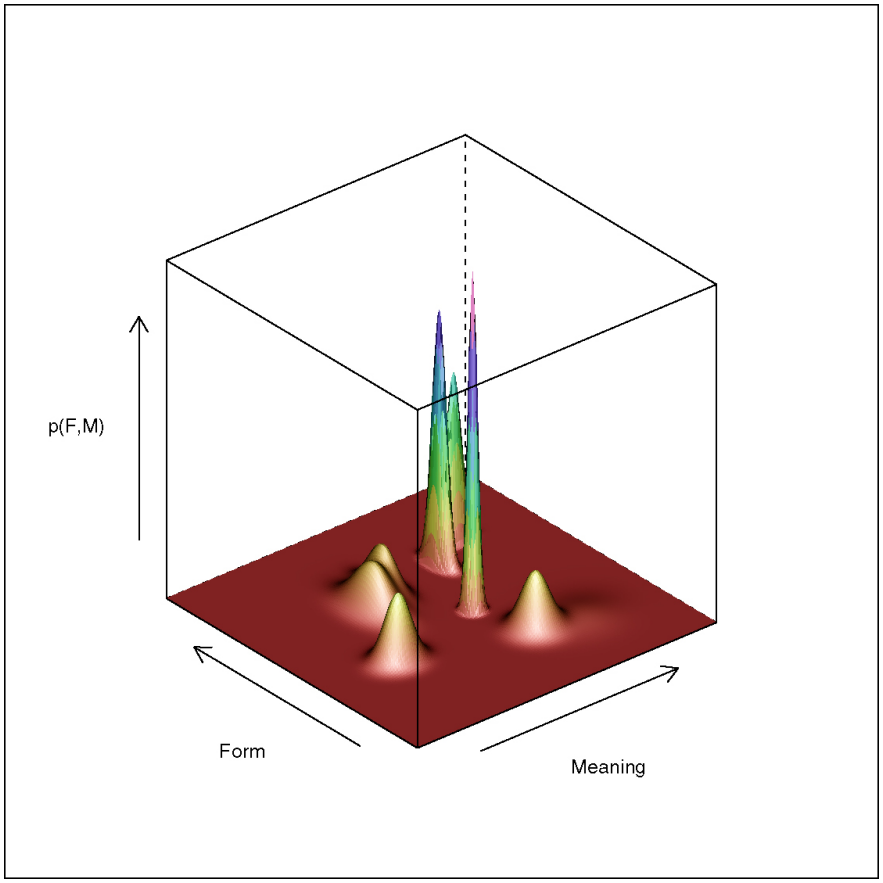


Figure 3:

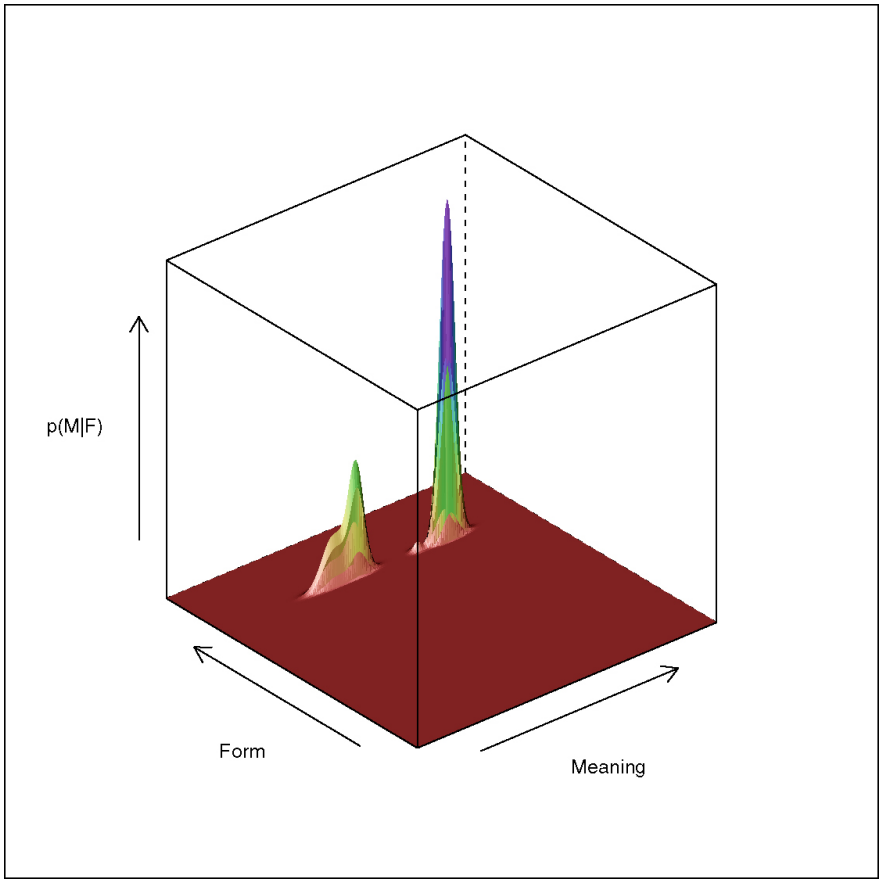


Figure 4:

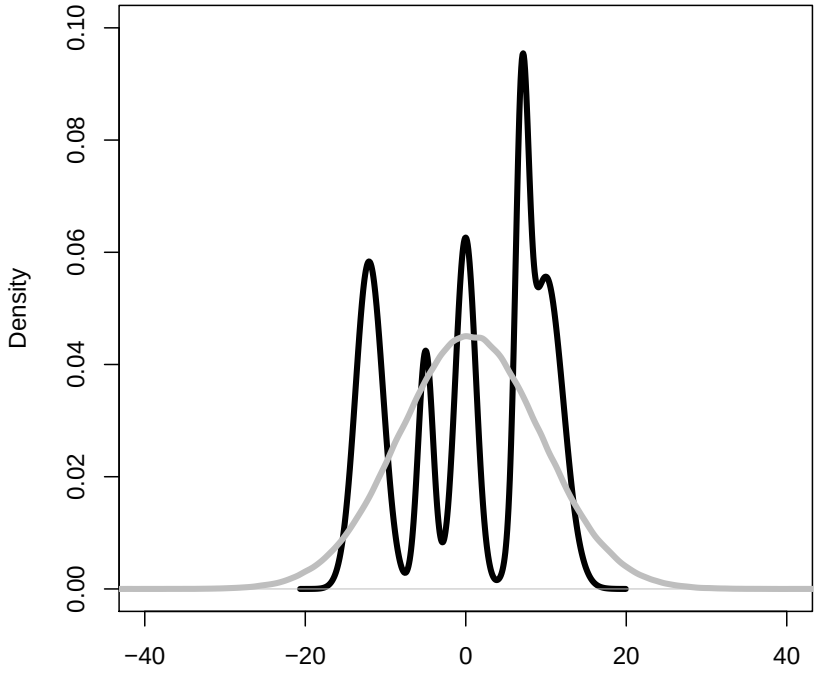


Figure 5:

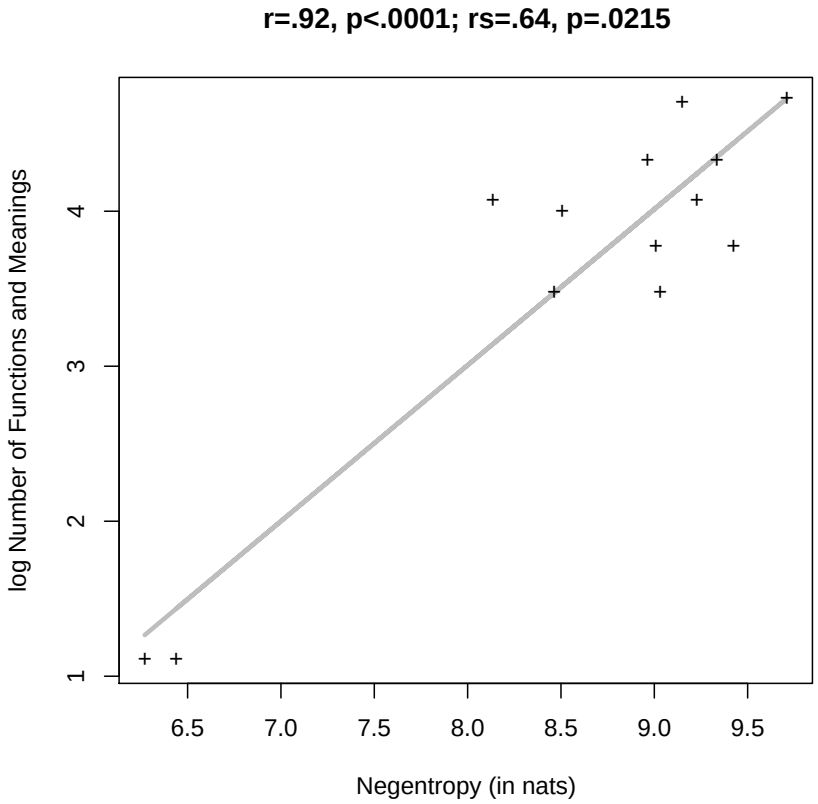

r=.92, p<.0001; rs=.64, p=.0215
log Number of Functions and Meanings
1
2
3
4
6.5
7.0
7.5
8.0
8.5
9.0
9.5
Negentropy (in nats)

Figure 6:

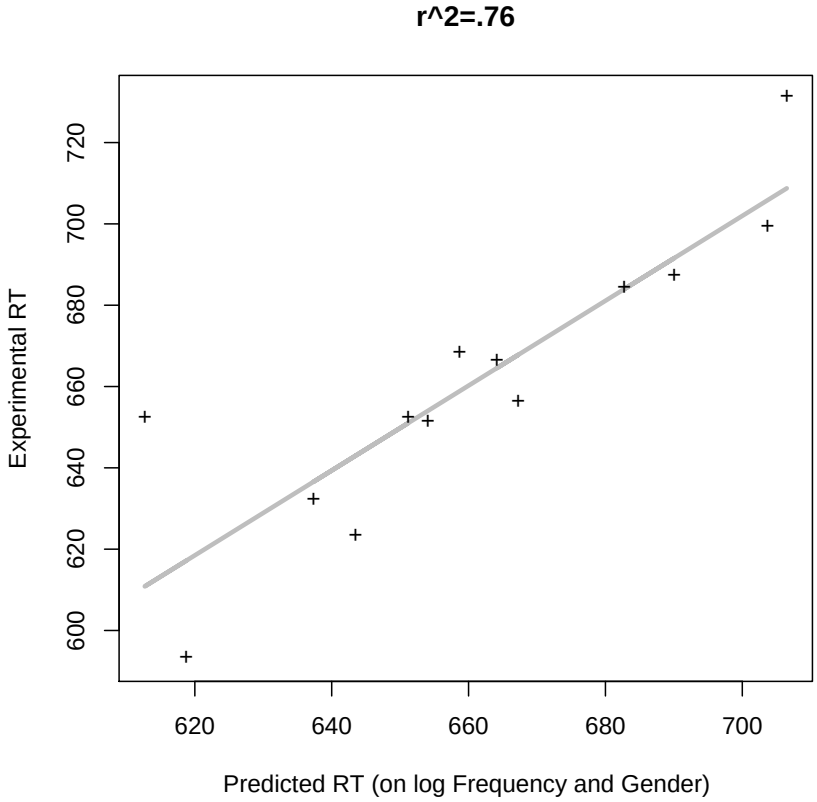

r^2=.76
Experimental RT
600
620
640
660
680
700
720
620
640
660
680
700
Predicted RT (on log Frequency and Gender)

Figure 7:

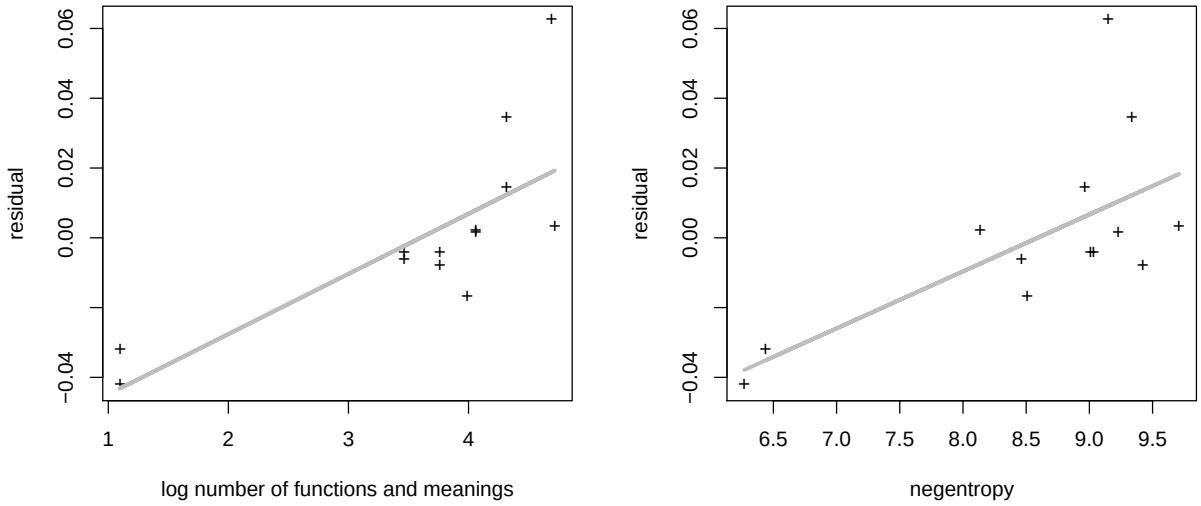

residual
0.06
0.04
0.02
0.00
−0.04
1
2
3
4
log number of functions and meanings
residual
0.06
0.04
0.02
0.00
−0.04
6.5
7.0
7.5
8.0
8.5
9.0
9.5
negentropy

Figure 8:

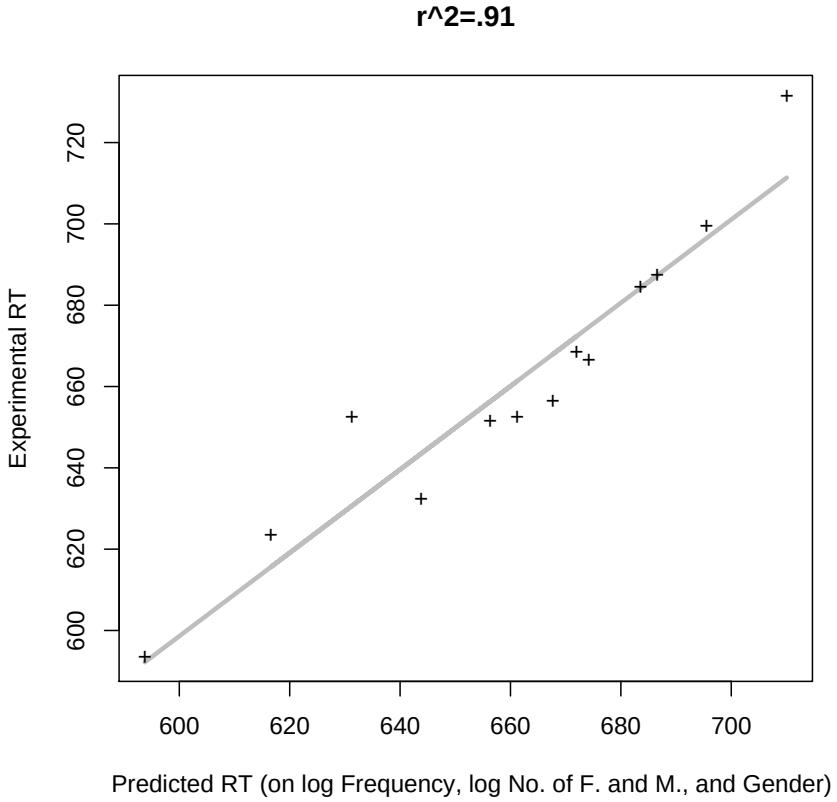

r^2=.91
Experimental RT
600
620
640
660
680
700
720
Predicted RT (on log Frequency, log No. of F. and M., and Gender)

Figure 9:

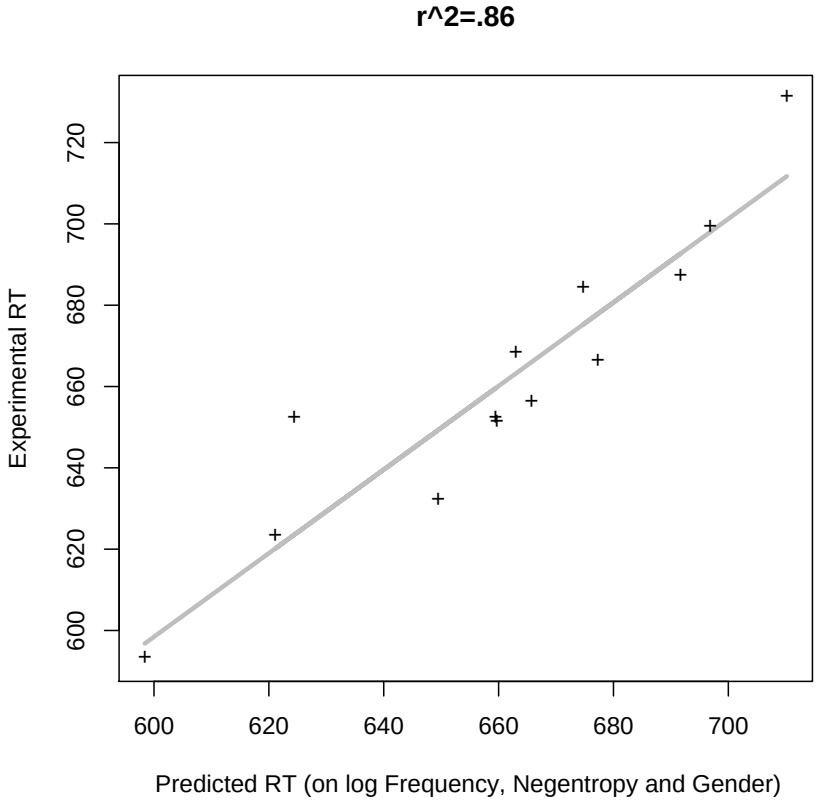

r^2=.86
Experimental RT
600
620
640
660
680
700
720
Predicted RT (on log Frequency, Negentropy and Gender)

Figure 10:

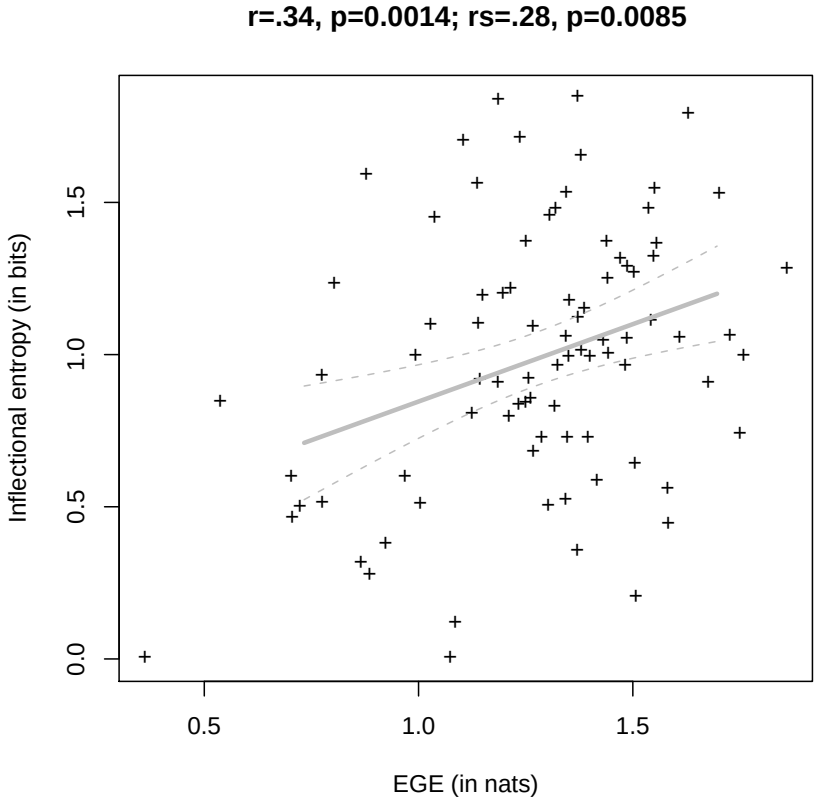

r=.34, p=0.0014; rs=.28, p=0.0085
Inflectional entropy (in bits)
0.0
0.5
1.0
1.5
0.5
1.0
1.5
EGE (in nats)

Figure 11:

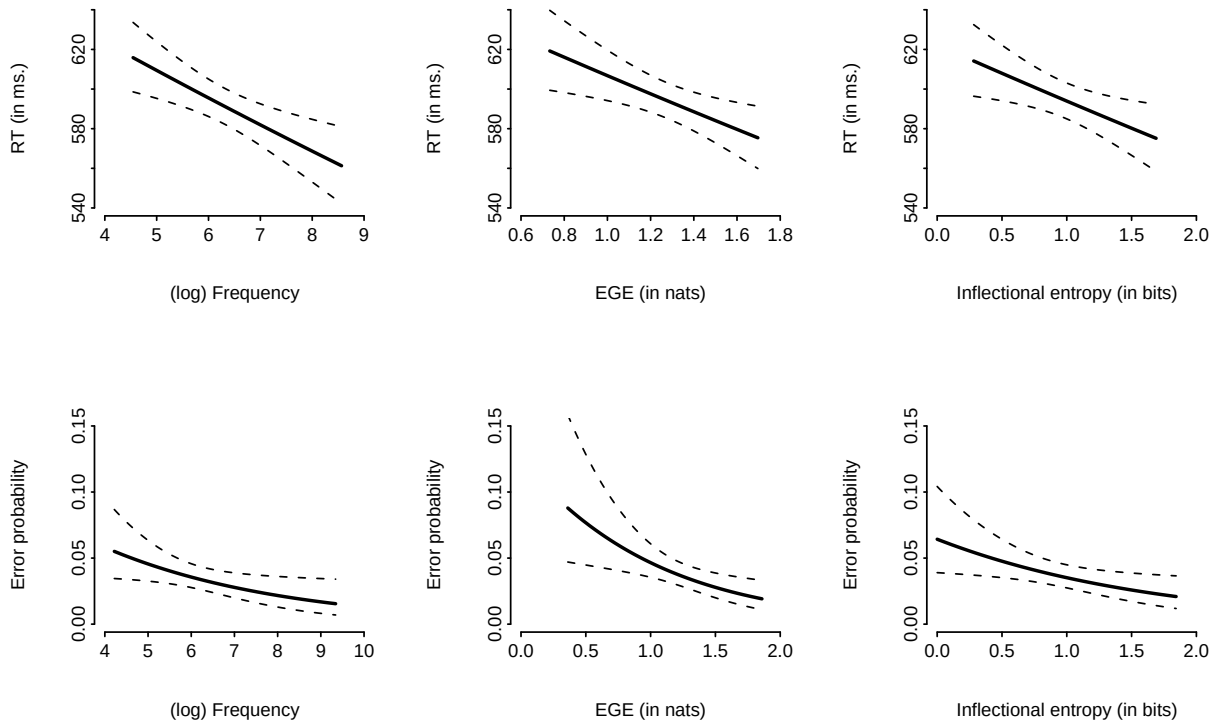

RT (in ms.)
540
580
620
4
5
6
7
8
9
(log) Frequency
RT (in ms.)
540
580
620
0.6
0.8
1.0
1.2
1.4
1.6
1.8
EGE (in nats)
RT (in ms.)
540
580
620
0.0
0.5
1.0
1.5
2.0
Inflectional entropy (in bits)
Error probability
0.00
0.05
0.10
0.15
4
5
6
7
8
9
10
(log) Frequency
Error probability
0.00
0.05
0.10
0.15
0.0
0.5
1.0
1.5
2.0
EGE (in nats)
Error probability
0.00
0.05
0.10
0.15
0.0
0.5
1.0
1.5
2.0
Inflectional entropy (in bits)

Figure 12:

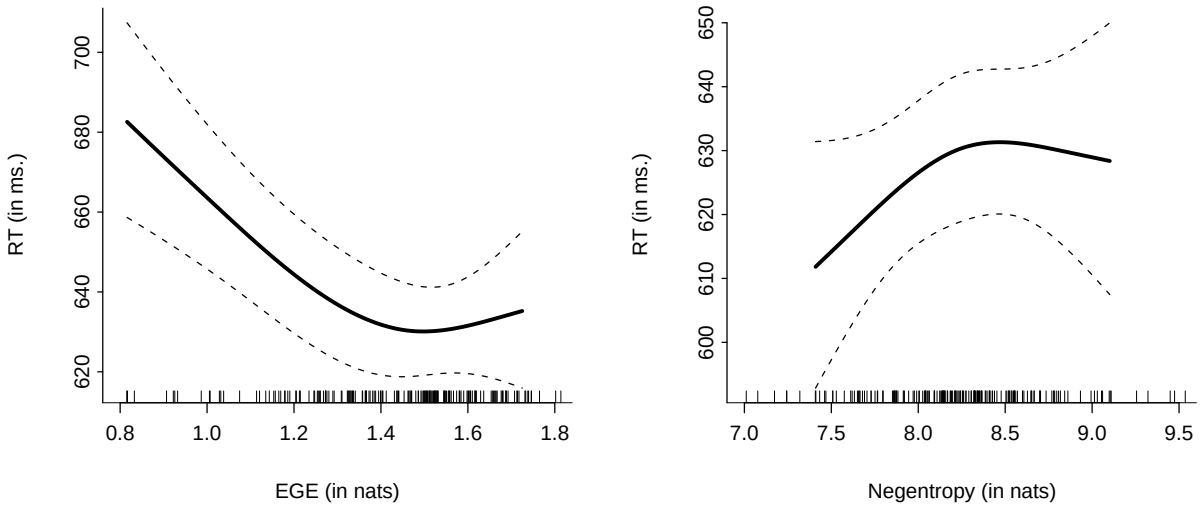

RT (in ms.)
620
640
660
680
700
0.8
1.0
1.2
1.4
1.6
1.8
EGE (in nats)
RT (in ms.)
600
610
620
630
640
650
7.0
7.5
8.0
8.5
9.0
9.5
Negentropy (in nats)

Figure 13:

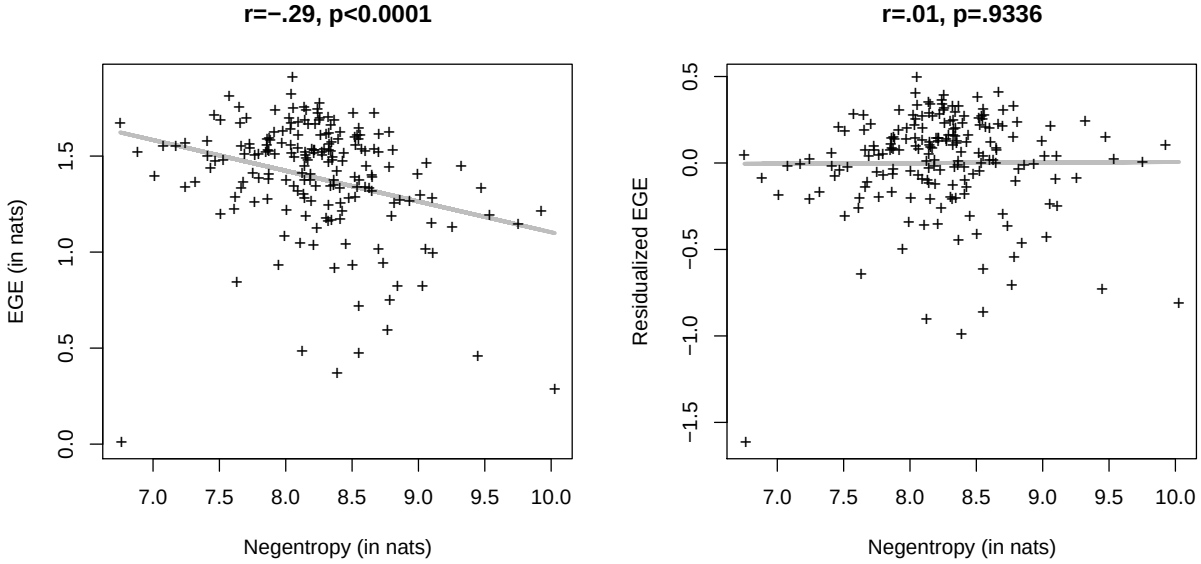

r=−.29, p<0.0001
EGE (in nats)
0.0
0.5
1.0
1.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
Negentropy (in nats)
r=.01, p=.9336
Residualized EGE
−1.5
−1.0
−0.5
0.0
0.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
Negentropy (in nats)

Figure 14:

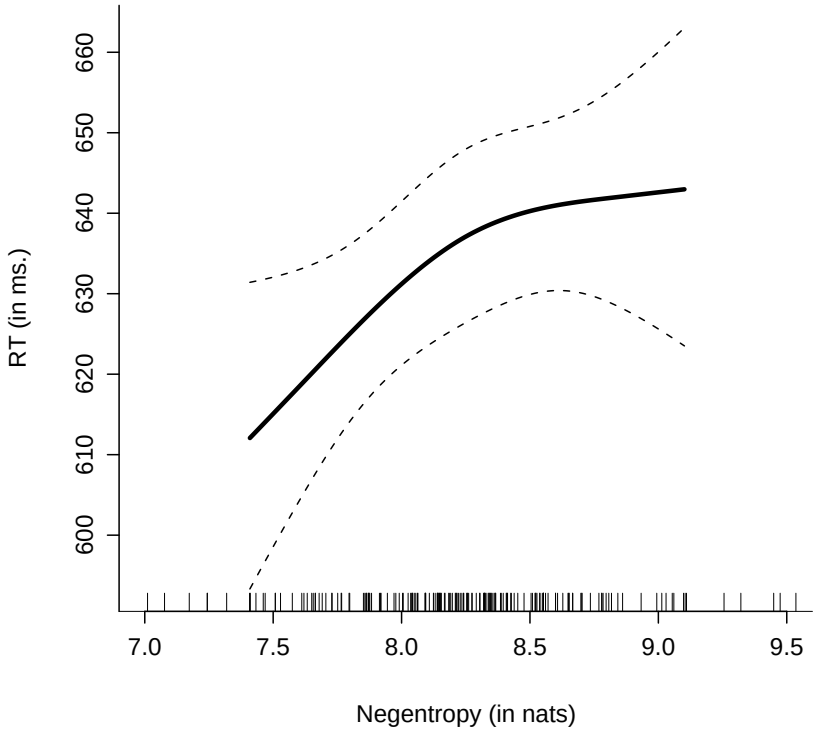
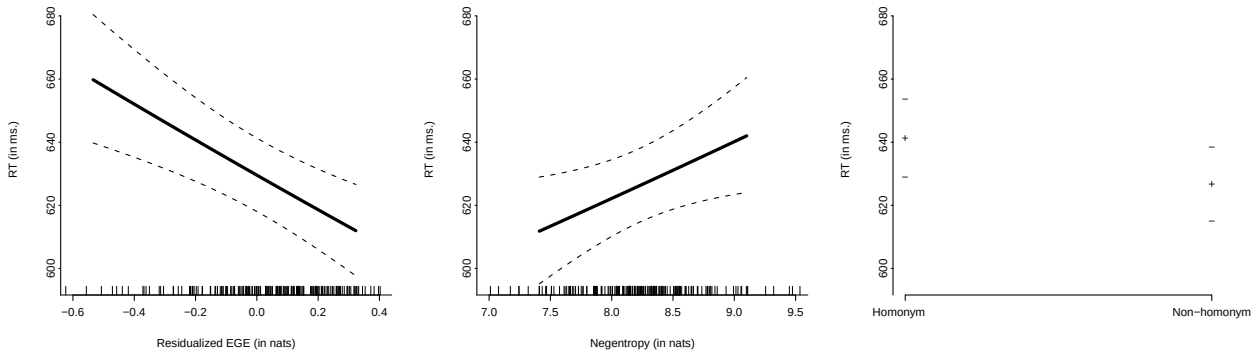

RT (in ms.)
600
610
620
630
640
650
660
7.0
7.5
8.0
8.5
9.0
9.5
Negentropy (in nats)

Figure 15:


RT (in ms.)
600
620
640
660
680
−0.6
−0.4
−0.2
0.0
0.2
0.4
Residualized EGE (in nats)
RT (in ms.)
600
620
640
660
680
7.0
7.5
8.0
8.5
9.0
9.5
Negentropy (in nats)
RT (in ms.)
600
620
640
660
680
Homonym
Non−homonym