This is an Accepted Manuscript of an article published by Taylor & Francis in Language, Cognition and Neuroscience on 2 Apr 2021, available online:

https://www.tandfonline.com/doi/full/10.1080/23273798.2021.1909083?src=

We Probably Sense Sense Probabilities

Dušica Filipović Đurđević*[1, 2, 3] and Aleksandar Kostić[2]

[1]*Department of Psychology, Faculty of Philosophy, University of Belgrade, Belgrade, Serbia*

[2]*Laboratory for Experimental Psychology, Faculty of Philosophy, University of Belgrade, Belgrade, Serbia*

[3]*Laboratory for Experimental Psychology, Faculty of Philosophy, University of Novi Sad, Novi Sad, Serbia*

* Correspondence:

Dr Dušica Filipović Đurđević

Department of Psychology

Faculty of Philosophy

Čika Ljubina 18-20

11000 Beograd

Serbia

email: dusica.djurdjevic@f.bg.ac.rs

https://orcid.org/0000-0001-5044-5428

Web of Science ResearcherID: A-8589-2011

Scopus Author ID 25823080300

We Probably Sense Sense Probabilities

In this paper, we demonstrate the effects of Information Theory measures on the processing of polysemous nouns and reveal that the sensitivity to multiple related senses can be learned from the linguistic context. We collected large-scale data and applied a correlation design to show that an increase in sense uncertainty (or sense diversity) is followed by a faster visual lexical decision. The facilitatory effect of sense uncertainty was revealed by the predictive power of entropy, followed by the additional analysis, which revealed that both the number of senses and the balance of sense probabilities affected processing. For the first time, the balance of sense probabilities was described via redundancy to demonstrate the effect of the numerical description of the balance of sense probabilities. Finally, we crossed distribution semantics and discrimination learning to show that polysemy effects can arise as a consequence of the principles of error-driven learning.

Keywords: entropy; lexical ambiguity; naïve discrimination learning; the number of senses; polysemy; redundancy; sense probabilities; visual lexical decision

Introduction

Polysemy is a form of lexical ambiguity in which a single word is linked to multiple related senses (e.g., *scientific paper* and *folding paper*), unlike homonymy where a single word is linked to multiple unrelated meanings (e.g., *river bank* and *financial bank*), and unlike unambiguous words in which a single word is linked to a single meaning/sense. The majority of research has suggested that, as compared to unambiguous words, polysemes take less time to recognise in a lexical decision task (Beretta et al., 2005; Klepousniotou, 2002; Pylkkänen et al., 2006; Rodd et al., 2002).

The true meaning/sense of an ambiguous word is typically resolved by sentential or pragmatic context. However, in a typical lexical decision experiment, in the absence of the disambiguating context, the word could refer to any of its meanings/senses. Put differently, when a word is presented in isolation, there is some uncertainty of its true meaning/sense. Therefore, in this paper, we aim to apply Information Theory measures to describe polysemy as sense uncertainty (but also sense diversity), as previously suggested by Gilhooly and Logie (1980), and, for the first time, to explicitly test for the processing effects of this description. Crucially, for the first time, we will demonstrate that recognition time is affected by the balance of sense probabilities, as expressed via continuous measure of redundancy (instead of comparing categories of balanced vs imbalanced words as done previously). Finally, we will look into the observed results from the point of view of discrimination learning.

Previous research on lexical ambiguity

Numerous empirical studies demonstrated that word ambiguity affects the processing of isolated words in lexical decision and naming tasks (Borowsky & Masson, 1996; Hino & Lupker, 1996; Hino et al., 2002; Millis & Button, 1989). These studies revealed that ambiguous words are processed faster and more accurately than words with only one meaning. However, investigations by Rodd, Gaskell, and Marslen-Wilson (2002; Rodd, 2004; but also Frazier & Rayner, 1990) demonstrated that ambiguity advantage applied only to polysemous words, that is, words with multiple related senses. A typical polysemous word is *paper*, which can denote the material used for writing (*writing paper*), the daily publication (*daily paper*), the scientific publication (*submitted paper*), and so forth (Lyons, 1977). The observed difference between homonymy and polysemy, as well as the processing advantage of polysemous words, was mirrored in the M350 component of magneto-encephalogram (Beretta et al., 2005; Pylkkanen et al., 2006) and N400 component of electroencephalogram

(Klepousniotou et al., 2012; MacGregor et al., 2015). Behavioral studies conducted in Serbian (which is the testbed language in the current research) also confirmed the processing disadvantage of homonymous words (Filipović Đurđević, 2019), as well as processing advantage of polysemous words and a negative correlation between the number of dictionary senses and processing latencies in a visual lexical decision task (Filipović Đurđević & Kostić, 2008).

Numerous studies that followed revealed a complex relation between processing time and lexical ambiguity (as reviewed by Eddington & Tokowicz, 2015 and Rodd, 2018). The pattern of results was highly influenced by the relatedness among the meanings/senses of the word, the task at hand, and the balance of meaning/sense frequencies.

The relatedness of the meanings/senses

Although Rodd et al. (2002) explicitly pointed to the difference between homonymy and polysemy, there were earlier results that implied the relevance of this distinction. For example, Frazier and Rayner (1990) recorded longer fixation time for homonyms and shorter fixation time for polysemous words in sentence reading task. Azuma and Van Orden (1997) showed a negative correlation between estimated relatedness of word meanings and processing time of a given word presented in isolation. In one of the first studies dedicated to the processing of ambiguous words, Jastrzembski (1981) came to a similar conclusion by performing the etymological analysis of word meanings. In a lexical decision task, he demonstrated that in addition to the total number of meanings given in the dictionary, processing time was affected by the size of the dominant cluster. This way, by using different approaches, two authors anticipated the later results of Rodd and colleagues (2002). Klepousniotou (2002) came to similar conclusions by using a cross-modal task in the procedure of sentence priming. Results of the experiments showed that the facilitation effect of prime was connected with polysemous words. Moreover, the facilitation effect was larger for polysemous words with high sense overlap (metonymy). The observed processing advantage of words with high relatedness among senses was confirmed in a lexical decision task, as well (Klepousniotou & Baum, 2007). During the last decade, researchers embraced the idea of also dividing the polysemy into categories based on the relatedness among the senses (Brocher et al., 2018). They started to separately investigate the processing of irregular polysemy (i.e., polysemous words with related senses which cannot be mutually linked via some rule; e.g., metaphor: *horn*) and regular polysemy (i.e., polysemous words with systematically related, mutually derivable senses; e.g., metonymy: *lamb*). It has been observed that the processing advantage was larger for regular as compared to irregular polysemy (Klepousniotou et al., 2012; Klepousniotou et al., 2008). Along the same line, attempts were made to describe the level of sense/meaning overlap by using various quantitative measures of the relatedness of contexts in which the word appears (Filipović Đurđević et al., 2009; Hoffman et al., 2013; Hoffman & Woolams, 2015; Jones et al., 2012; McDonald & Shillcock, 2001).

The task effect

The processing differences that arose as the consequence of the relatedness of word meanings were additionally modified by the task in which the ambiguous words were presented. The asymmetry in processing of polysemy and homonymy (i.e., polysemy advantage and homonymy disadvantage) was typically observed when ambiguous words were presented in isolation and relied on simple word recognition (lexical decision task, naming, cross-modal priming; Klepousniotou, 2002; Lichacz et al., 1999; Rodd et al., 2002). However, the observed pattern would change if the task elicited the activation of the single meaning/sense. For example, polysemy advantage would turn into polysemy disadvantage in semantic categorisation task (Hoffman & Woollams, 2015; Pexman et al., 2004; Piercey & Joordens, 2000). Similarly, polysemy effects could be absent or even reversed if the ambiguous word was preceded by the biasing context (Duffy et al., 1988; Foraker & Murphy, 2012).

The balance of meaning/sense frequencies

The issue of probabilities of individual meanings/senses has been addressed previously, mostly in studies that recorded eye movements in sentence reading (Duffy et al., 1988; Foraker & Murphy, 2012; Frazier & Rayner, 1990) and priming studies (Klepousniotou et al., 2012; Onifer & Swiney, 1981; Swiney, 1979). This line of research presented evidence that early in processing, multiple meanings of the ambiguous words are activated at the same time, with the level of activation being proportional to the meaning frequency (Onifer & Swiney, 1981; Seidenberg et al., 1982; Simpson & Burgess, 1985; Swiney, 1979). The effect of meaning/sense frequencies has been addressed by the name of the dominance effect. More precisely, this effect refers to the finding that the dominant meaning/sense of the ambiguous word is the easiest to evoke. This has been concluded based on two lines of observations. On the one hand, eye fixations on the target ambiguous word are longer when a balanced homonym is presented in the neutral context as compared to the biased homonym, or the unambiguous word, whereas biased homonym and unambiguous word are fixated equally long. On the other hand, eye fixations on the following disambiguating region, which is referring to the subordinate meaning, are longer for biased words, thus indicating that it is harder to disengage/recover from the dominant meaning. This has been observed with homonyms in multiple studies (Duffy et al., 1988; Frazier & Rayner, 1990; Rayner & Duffy, 1986; Rayner & Frazier, 1989). More recently, these findings were mirrored in neuroimaging studies that applied the same design and showed that the activation in the left inferior frontal gyrus and left posterior temporal gyrus was higher for ambiguous than for unambiguous words (Vitello et al., 2014; see also Maciejewski & Klepousniotou, 2020 for a different design). In accordance with the results from behavioural studies, the activation in these areas was also higher for biased compared to balanced homonyms, indicating that recovering from the activation of the dominant but inadequate meaning (as the context supported the subordinate meaning) is demanding in terms of processing effort.

However, when it comes to polysemy, the findings have not been as consistent. For example, Foraker and Murphy (2012) observed the same effect that has previously been observed with homonyms, whereas other studies failed to observe the dominance effect in neutral context for polysemous words (Brocher et al., 2016; Brocher et al., 2018; Frazier & Rayner, 1990; Frisson & Pickering, 1999). Therefore, polysemy appears to be less apprehended when it comes to the effect of sense frequencies.

Additionally, investigations of the processing effects of meaning frequencies have had some important limitations. Although these studies addressed the issue of meaning/sense probabilities, they did so by comparing biased (non-equibiased) and balanced (equibiased) ambiguous words in a factorial design (e.g., Brocher et al., 2018; Klepousniotou & Baum, 2007; Klepousniotou et al., 2012). In other words, they described the balance of meaning/sense probabilities by using two categories. However, the balance of meaning/sense probabilities is a matter of degree. In order to elaborate the claim that the activation of the meaning/sense is proportional to its frequency, a functional relationship between the two variables must be determined. In order to describe such a relationship, both variables need to be described in terms of continuous measures, which was not the case in the majority of the previous research. Studies that did so are very rare and directed towards homonymy processing. For example, Armstrong, Tokowicz, and Plaut (2012) showed that the relative frequency of the dominant meaning (β) was inversely proportional to visual lexical decision latencies of homonymous words (see also Rodd et al., 2002; Twilley et al., 1994). To the best of our knowledge, the functional relation between the balance of sense probabilities and the recognition time of polysemous words has never been demonstrated, as also pointed out by Eddington and Tokowicz (2015, p. 31).

Theoretical accounts of the effect of balance of sense/meaning frequencies

Lexical ambiguity effects are typically interpreted in the light of two word recognition frameworks (for an illustrative and concise review, see Maciejewski & Klepousniotou, 2020). On the one hand, there is the *semantic competition account,* which stems from the tradition of the connectionist models of reading and rests upon the assumption that ambiguous words are characterised by multiple form-to-meaning mapping (Armstrong & Plaut, 2008; 2016; Kawamoto, 1993; Rodd et al., 2004; Rodd, 2020). The existence of separate representations competing for activation is considered to be the source of the ambiguity disadvantage, as observed with homonyms in the majority of the tasks and with polysemes in tasks that are more engaging. Polysemy advantage, as observed in early processing, is the consequence of the blend state of activations of overlapping semantic representations. On the other hand, according to the *decision making account*, ambiguity effects are not bound to semantic processing but arise later in the process of the response selection and as a consequence of the task demands (Hino et al., 2006; Pexman et al., 2004). Recently, a study by Maciejewski and Klepousniotou (2020) shed new light in favour of the semantic competition account. However, understanding the nature of the effects of the relatedness of meanings/senses and the balance of sense/meaning frequencies in the context of different task demands remains the core issue in lexical ambiguity research. One of the first steps of advancing our understanding of lexical ambiguity effects would imply the understanding of the way the number of senses, the balance of sense probabilities, and the relatedness among the senses are incorporated in the network of lexical knowledge and how such network is acquired during the course of experience with language. One path in accomplishing this goal is to approach lexical ambiguity from the perspective of basic learning principles.

Error-driven learning

According to the usage-based accounts, the structure of language is rooted in the processes upon which the language relies (Bybee, 2007; 2010). For example, grammar is considered to represent the cognitive organisation of one's experience with language (Bybee, 2006), and the same easily applies to other levels of language, as well. The crucial aspect of the experience with language is the recording of the traces of that experience, in other words – the process of learning.

The simplest form of learning, which enables acquiring the information on multiple stimuli is classical conditioning, according to which the stimuli that repeatedly co-occur are being associated (Pavlov, 1927). However, Rescorla (1968; 1988) demonstrated that it is not the mere association that accounts for the behavioral change incurred by the learning process. Instead, an organism learns the relations among the stimuli – conditioned stimulus (CS) becomes the *cue* that is able to predict the unconditioned stimulus (US) as the *outcome*. Even more, during the course of learning, it gradually learns to discriminate the stimuli which are good from the stimuli which are bad at predicting the outcome. *Discriminative learning* (Rescorla, 1988) is the process of becoming familiar with the structure of the environment by learning the informative value of various cues for various outcomes. Crucially, in order for the stimulus to become predictive of a certain outcome, an organism needs to be *surprised* by the experienced cue-outcome co-occurrence, i.e., it needs to make the prediction error (Kamin, 1969). For example, in order for the dog to learn that the sound of the metronome announces the serving of the food, the dog previously needs to be ignorant of the metronome-food relation, and hence make the wrong prediction that metronome is not saying anything about the food (or even announces the lack of food). Hence, the learning unrolls in the sequence of a) predictions that are inferred based on the cue-outcome associations, b) being surprised after making the wrong predictions, and c) the changing of the association strengths among the cues and outcomes to compensate for the error. The basic process is that of updating the cue-outcome association values:

$$w_{ij}^{t+1} = w_{ij}^{t} + \Delta w^{t} \qquad (1)$$

This process follows the simple rule formalised as the Rescorla-Wagner equation (Rescorla & Wagner, 1972):

$$\Delta w^{t} = \begin{cases} a)\ 0 & ,\ if\ ABSENT(C_i, t) \\ b)\ \alpha_i \beta\left(\lambda - \sum_{PRESENT(C_k,t)} w_j\right), & if\ PRESENT(C_i, t)\ \ PRESENT(O, t) \\ c)\ \alpha_i \beta\left(1 - \sum_{PRESENT(C_k,t)} w_j\right), & if\ PRESENT(C_i, t)\ \ PRESENT(O, t) \end{cases}$$

(2)

In equation (2), which is also a special case of the delta rule (Widrow & Hoff, 1960), α_i is the cue salience parameter, β denotes the learning rate, and λ is the maximum association value (actually without upper limit). These parameters are frequently set in advance to αi =0.1, β =.1, and λ=0.1. According to this rule, a) association values are unchanged if the cue is absent, b) they are increased if both cue and outcome are present, and c) they are decreased if the cue is present, whereas the outcome is absent. This simple principle brings about highly complex dynamics, as the presence or absence of one outcome does not affect only one cue-outcome association but leads to recalculations across the system (the presence of one outcome given the cue is simultaneous with the absence of all other outcomes). This unfolds through the process of cue competition (but also outcome competition), which is perpetual, thus making all knowledge a dynamic system of the associations among cues and outcomes (for detailed elaboration on the nature of the process, see Hoppe et al., 2020, and also Tomaschek, 2020).

This approach has recently been revived in psychology (Ramscar et al., 2013; Ramscar & Port, 2016; Ramscar et al., 2010). Baayen and colleagues (2011) were the first to build Naïve Discriminative Reader (NDR; Baayen et al., 2011), a simple computational model that learns the direct mappings from the input cues (typically bigrams that constitute the word) to the outcomes that represented the meaning. However, outcomes typically consisted of local representations, which only served as the pointers to some true representation in the semantic space, and was named *lexome*, to separate the pointer from its content. Milin et al. (2017) further developed the idea by keeping the original network and naming it the G2L network. *G* stood for the grapheme cues and *L* for the lexomes at the outcome level. The crucial novelty of their work was the introduction of another network, in which the lexomes served both as cues and as outcomes, hence named the L2L network. Mapping of lexomes to lexomes was effectively analogous to representing the lexome meaning via its location in the semantic space defined by other lexomes. Hence, this approach resembled the strategy developed within the framework of distributional semantics (Landauer & Dumais, 1997; Lund & Burgess, 1996) – representing word meaning via location in the semantic space based on co-occurrence with the words that were pre-selected to represent the dimensions of that space. Milin et al. (2017) pointed out that the rows of the L2L network can be seen as the semantic vectors. Numerous research demonstrated that different measures derived from the two networks – the measures that capture various aspects of network dynamics – successfully predicted various psycholinguistic effects (Arnon & Ramscar, 2012; Arnold et al., 2017; Baayen et al., 2011; Baayen et al., 2013; Baayen et al., 2016; Filipović Đurđević & Milin, 2019; Hendrix, Bolger, & Baayen, 2017; Milin et al., 2017; Nixon, 2020; Tomaschek et al., 2021). However, to the best of our knowledge, lexical ambiguity has not been previously investigated within the framework of discriminative learning.

Current study

We aim to investigate the effect of the continuous measure of the balance of sense frequencies on visual lexical decision latencies of polysemous nouns and to portrait it in the light of discriminative learning, using Serbian as the testbed language.

Having in mind the differences in the processing of homonymy and polysemy (Rodd et al., 2002), in this study, we will limit ourselves to the investigation of *polysemous* words. Moreover, in order to avoid confounding of polysemy and part of speech effects, we will focus on polysemous *nouns* (i.e., words that were nouns in all polysemous interpretations). In doing so, we are also following Eddington and Tokowicz (2015, p. 31), who pointed out that previous investigations of lexical ambiguity typically mixed noun-noun, noun-verb, and other types of between-part-of-speech ambiguity (e.g., *a twist*/*to twist*). They suggested that part of speech should be either systematically investigated or controlled, and we decided to control for it.

With different task effects in mind (as reviewed in Eddingtnon & Tokowicz, 2015; Rodd, 2018), in the current research, we will focus on *visual lexical decision task*, as polysemy advantage has been typically observed in this task.

Crucially, we set our main goal starting from the observation that the majority of the previous studies focused on the effect of the balance of sense/meaning frequencies by contrasting balanced and unbalanced (or equibiased and non-equibiased) ambiguous words in a factorial design. Although factorial designs hold numerous advantages, when it comes to psycholinguistic experiments, it has been argued that regression design should be preferred (see Baayen, 2010 for a discussion on the strengths of this approach). Therefore, in this study, we set the aim of describing the *functional relation* between the balance of sense probabilities and the processing latencies.

Finally, we aim to contribute to the efforts of understanding the nature of the observed effects by testing whether the structure that is described by the attested Information Theory measures can emerge through the process of *discriminative learning* of mappings from word form to meaning – meaning being operationalised through the context of co-occurring words.

Polysemy as sense uncertainty

Traditionally, polysemy has been described in terms of the number of senses. For example, a word with three senses would be considered to be more ambiguous than the word with two senses, as illustrated in Figure 1A and Figure 1B. Indeed, it has been demonstrated that an increase in the number of senses was followed by a decrease in processing latencies (Filipović Đurđević & Kostić, 2008; Rodd et al., 2002). However, the level of ambiguity of the polysemous words (and any type of lexical ambiguity in general) is not dependent only on the number of senses (or meanings). It is also influenced by the relative frequencies of individual senses, that is, by the balance of sense probabilities. For example, a word with one sense which is much more frequent than the remaining senses (the so-called dominant sense of the unbalanced polyseme; Figure 1B) would be less ambiguous as compared to a word with the same number of senses that occur with mutually similar frequencies (balanced polyseme; Figure 1C).

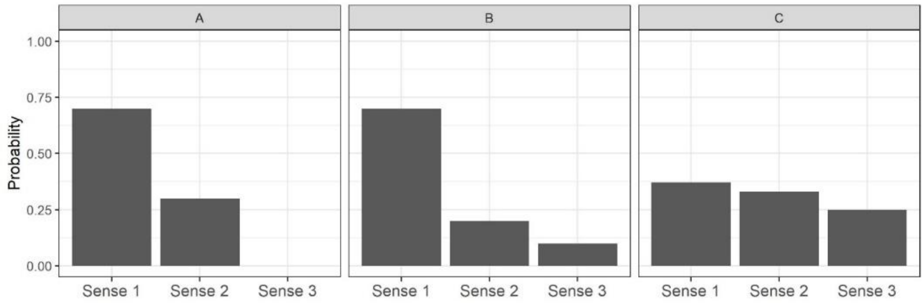

Figure 1. Illustration of the influence of the number of senses and balance of sense probabilities on the overall level of lexical ambiguity (i.e., sense uncertainty of the given word). The sense uncertainty grows from left to right, following an increase in the number of senses (A to B) and an increase in the balance of sense probabilities (i.e., a decrease in redundancy; B to C).

Therefore, the number of senses and the balance of sense probabilities in conjunction contribute to the uncertainty of the word's true meaning. This uncertainty can be jointly quantified by applying the Information Theory measure of the *entropy* of sense probability distribution, higher entropy mirroring the higher level of uncertainty, i.e., greater ambiguity (Shannon, 1948; as proposed by Gilhooly & Logie, 1980).

Entropy (*H*) is calculated over the distribution of probabilities of a given event (p_i) by summing the products of probabilities and log-probabilities of each event (3).

$$H(w) = -\sum_{i=0}^{n} p_i \cdot \log p_i \qquad (3)$$

In (3), *H* stands for the entropy of the polysemous word *w*, the index *i* points to different senses of word *w*, p_i denotes proportion (relative frequency) of the given sense of *w*, and *n* points to the number of senses of *w*. This measure can be interpreted as an index of the level of word ambiguity (or degree of ambiguity [U] as suggested by Gilhooly & Logie, 1980).

When compared to the number of senses that has been traditionally applied in polysemy research, the added information that is included in entropy concerns the balance of sense probabilities. The entropy of sense probability distribution can be interpreted as the uncertainty of senses. It is influenced by the number of senses in such a way that a larger number of senses leads to larger entropy, that is, a larger degree of uncertainty of the true sense of the word (with *logN(w)* being theoretical maximum, as illustrated by the black line in the first plot in Figure 2). However, it is also influenced by the balance of sense probabilities in that words with balanced probabilities of senses carry greater sense uncertainty, i.e., larger entropy. Words with the dominant sense, i.e., unbalanced sense frequencies, carry less uncertainty of the true sense of the word. This added information can be described independently of the number of senses (Cover & Thomas, 1991; Mac Kay, 2003; Shannon, 1948). The Information Theory measure that quantifies the balance of probabilities is called *redundancy* (4).

$$T(w) = 1 - H(w)/\log N(w) \quad (4)$$

In (4), *T(w)* denotes redundancy of the polysemous word *w*, *H(w)* denotes the entropy, and *N(w)* marks the number of senses of the word *w*. Balance among sense probabilities is mirrored in low redundancy values, i.e., in high uncertainty of the word’s true sense (Figure 2, the second panel in the upper row). By definition, this measure ranges from zero to one, denoting the distance from the distribution with maximum entropy (with equally probable events). Therefore, a word with balanced sense probabilities would have a redundancy score closer to zero, while a word with imbalanced sense probabilities would have a redundancy score closer to one. A redundancy of 0 would indicate a word with perfectly equal probabilities of senses. Taking only the number of senses into account would imply the tacit assumption of such perfect balance among sense probabilities. It should be noted that the newly introduced concept of balance of sense probability distribution does not only capture the difference between words with or without the dominant sense. In addition to that, it captures the fine-grained differences in probabilities across all existing senses (Berger & Parker, 1970).

Based on (4), it can also be shown that entropy – *H(w)* can be decomposed to the number of senses – *N(w)* and redundancy – *T(w)*, as illustrated in (5).

$$H(w) = \log N(w)\,(1 - T(w)) \quad (5)$$

This equation (5) perfectly illustrates the point that the uncertainty of the word's true sense (as quantified by the entropy) is influenced both by the number of senses and the balance of sense probabilities. This is also illustrated empirically by the high correlation coefficients both between the number of senses and entropy and between the redundancy and entropy (see the first two panels in Figure 2). At the same time, although the number of senses and redundancy are also related, their correlation coefficient is moderate (see the third panel in Figure 2).

As illustrated by the rightmost panel in the upper row and the first two panels in the bottom row in Figure 2, the measures we have focused on are all related to the values of the probability of the most frequent sense, i.e., the dominance (*D*; Berger & Parker, 1970), which has also been proposed as the measure of ambiguity (Rodd et al., 2002), also named β (for *biggest*) by Armstrong et al. (2012). This finding is perfectly expected, knowing that it arises as a consequence of the very calculation of the sense probabilities: an increase in the redundancy (i.e., a decrease in the balance) implies an increase in the dominant sense frequency, whereas an increase in the number of senses implies a decrease in the probability of the most frequent sense.

One similar ambiguity measure has also been proposed to combine the focus on the dominant meanings and the information theory (Twilley et al., 1994). This measure, named *B* (for *balance estimate*), is simply the entropy of the two most frequent meanings (the two highest relative frequencies previously being normalised to sum up to 1). The smaller the difference between the relative frequencies of the first and the second most frequent meaning, the higher the *B* (*B* reaching 1 for equal relative frequencies of the two meanings). Unlike the dominance, *B* is related neither to the number of senses nor to the entropy of the entire distribution. It is only moderately related to the redundancy and the dominance (see the last two panels in Figure 2).

With all this in mind, and knowing that the processing of polysemous words has thus far been only linked to the number of senses, we will conduct two analyses to separately test for the predictive power of the two sides of equation (5). In the first analysis, we will test whether the entropy of sense probability distribution can account for the processing latencies of polysemous words (as previously proposed by Gilhooly and Logie, 1980). In the second analysis, we will test whether the redundancy can account for the processing latencies of the polysemous words over and above the known effect of the number of senses, thus showing that balance of sense probabilities accounts for lexical decision latencies over and above the attested effect of the number of senses (Filipović Đurđević & Kostić, 2008; Rodd et al., 2002). In addition to this main analysis, we will also look at the relative contribution of two alternative measures: dominance (*D*, or β; Armstrong et al., 2012; Rodd et al., 2002) and *B* (Twilley et al., 1994).

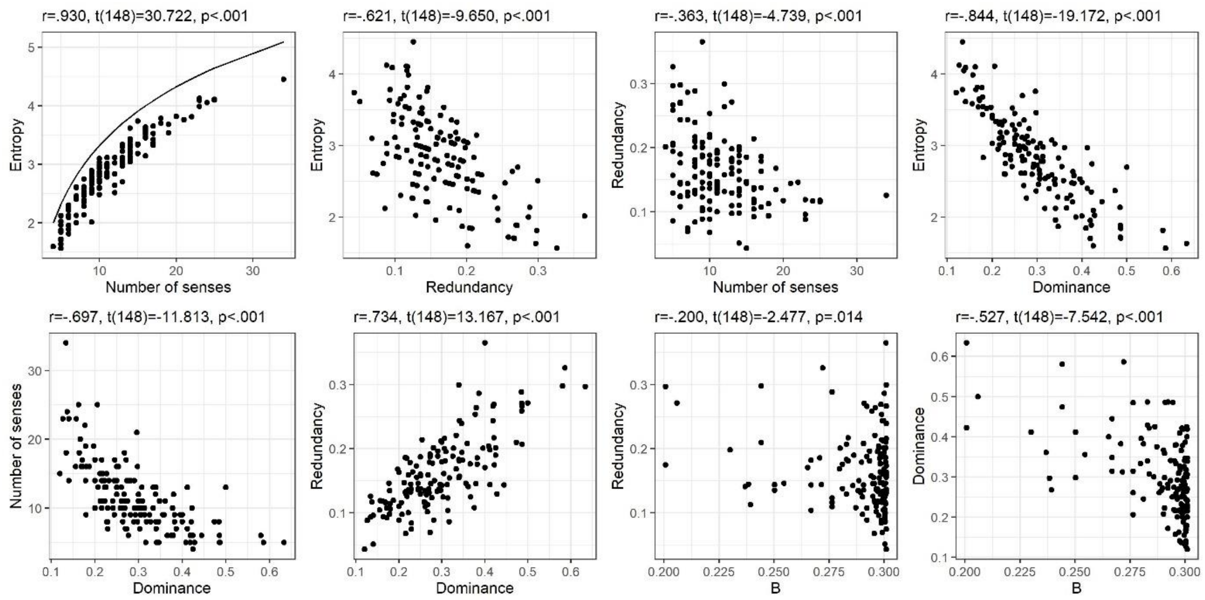

Figure 2. Relation among the number of senses, the entropy, the redundancy, the dominance, and B, based on the data collected in the study conducted by Filipović Đurđević and Kostić (2017).

Polysemy and discriminative learning

In order to deepen the understanding of the functional relations between processing effort indicators and Information Theory based descriptions of various language systems, some researchers took the approach of relating Information Theory measures and learning (Baayen et al., 2011; Filipović Đurđević & Gatarić, 2018; Filipović Đurđević & Milin, 2019). We will take that approach here to try and investigate whether the complex structure captured by Information Theory measures can emerge through the process of learning. More specifically, we will try to test whether the measures of uncertainty, especially the redundancy of sense frequency distribution can be accounted for by the principles of error-driven learning, namely discriminative learning (Rescorla, 1988).

The general challenge with simulating the learning of ambiguous words concerns the one-to-many mapping from orthography to semantics. The additional challenge concerns the modulating effect that the relatedness of meanings/senses has on the direction of the ambiguity effect (as discussed earlier – the number of unrelated meanings is positively correlated with lexical decision latencies of homonyms, whereas the number of related senses is positively correlated with lexical decision latencies of polysemes; Rodd et al. 2002). We will test whether the complex one-to-many mappings of polysemous words can be incorporated in the discrimination learning framework in such a way to allow for the dynamics of cue and outcome competition to give birth to the network complexity which will resemble the complexity found in polysemous words. Finally, the challenge we set here is to incorporate these dynamics in a single network. We will call this network G2L network, but it will not be identical to the original G2L network as introduced by Milin et al. (2017). In building our G2L network, the bigrams that constitute the polysemous word form are the obvious candidates for the input cues. However, the outcome lexomes will not serve only as pointers to meaning in the semantic space, which is to be represented in a separate network, as is the case in a typical G2L network (Milin et al., 2017). Although the lexomes of our network will not contain the meaning in itself, we will try to use them to implicitly code for the relatedness of the multiple senses associated with the same set of orthographic cues. This way, we will try to capture the multitude of outcomes and the relatedness among them.

Given that the complexity of polysemous words resides precisely at the level of their meaning, we aim to introduce a new way of specifying the model outcomes in order to capture exactly this. We propose to use the tools of distributional semantics as the means to accomplish this goal (Landauer & Dumais, 1997; Lund & Burgess, 1996; but lately also Bojanowski et al., 2016; Mikolov, Sutskever et al., 2013; Mikolov, Chen et al., 2013). Distributional semantics uses vector-based semantic representations or word embeddings to quantitatively describe the meaning of the word by relying on the statistics of that word co-occurring with other words or within a given document or a topic. This approach was successful in accounting for multiple semantic phenomena (such as predicted TOEFL responses; Landauer & Dumais, 1997) and has been developing very fast during the last two decades. Although there are many available state-of-the-art ready-to-use vector representations, they are not of much use in the study of lexical ambiguity. These vectors are built in such a way that one word is represented by a single vector. This implies that the variety of lexical meanings/senses are collapsed in a single vector, i.e., a single location in the semantic space. In order to preserve the information on the variability of ambiguous words in the semantic space, Schütze (1998) introduced the so-called first-order and second-order semantic vectors. Here, each vector represents a single occurrence of the word. Therefore, instead of building one vector per word, in the study of lexical ambiguity, many vectors are built per word, the number of vectors being equal to the number of occurrences of the given word in the corpus. The first step in the process of building the first order co-occurrence vectors is simply noting whether any of the preselected context words (the dimensions of the vectors) occurred in the vicinity of the target polysemous word (e.g. three positions to the left and three positions to the right). This produces a vector containing many zeros and several ones. This sparse data problem is resolved in the following steps of the process (see also Filipović Đurđević et al., 2009 for the application of the full process). However, this is exactly what we will use to define the outcomes of our G2L network. For each occurrence of the polysemous word, we will map its form to multiple outcomes. Firstly, the word will be mapped to a lexome marking the lemma, or the base form of the word. This is done having in mind that the individual occurrences of our target polysemous words will be presented by its inflected forms. Additionally, each word will be mapped onto each of the context words with which it co-occurred (marked by value *one* in the first-order vector). By doing so, we will instantiate the dynamics of competition among the outcomes. Our hypothesis is that the relatedness among senses will be captured by the fact that related senses are, in fact, reducible to the related contexts in which the word occurs. Therefore, the related senses would share some of the outcome lexomes (context words), thus enabling us to model the effect of relatedness among the senses without explicitly introducing the mechanism for it, and without introducing different mechanisms for related senses and unrelated meanings (e.g. facilitation/inhibition).

To summarise, we propose to cross distributional semantics with a discrimination learning approach aiming at two goals: 1) to try and operationalise the relatedness of senses of polysemous words (the sense overlap) as the overlap of the contexts in which the polysemous words appear, and 2) to try and demonstrate that such complex one-to-many mapping from form to meaning can be learned via discrimination learning. With respect to sense overlap, we hypothesise that polysemous words will tend to co-occur with the same words across different contexts and that the level of such repetition will serve as the index of the relatedness among senses.

Our starting point were the first-order co-occurrence vectors (one vector per occurrence) created for 150 polysemous words, which were presented in the experiment (Schütze, 1998). They were built using the software developed by Filipović Đurđević, Đurđević, and Kostić (2009). A separate vector was created for each individual occurrence of the target word. Target words were all existing inflected forms of the polysemous words from our set (e.g., for the stimulus *momak* (guy), target words would be: *momak, momka, momku, momkom, momci, momcima, momke*). As context words, we selected from the Frequency Dictionary of Contemporary Serbian Language (Kostić, 1999) 1000 most frequent nouns, 1000 most frequent adjectives, and 1000 most frequent verbs. After cleaning for duplicates (due to homography), we were left with a total of 2383 context words. The vectors were built by moving a seven-point window (centred on the target word) through the Ebart media database consisting of approximately 65 million words of mostly journal articles (http://www.arhiv.rs). Each vector was an array of numbers 0 and 1, value one being placed in the column corresponding to the context word that was found in the proximity of the target word (more precisely three positions to the left or three positions to the right, i.e., within the -/+ 3 context window). The example presented in Table 1a illustrates the process of the moving seven-point window being centred on the target word *momka*. Two of the pre-selected context words (in our case, most frequent nouns, adjectives, and verbs) are found within the window (coded here as *context1166* and *context1963*). The first row in Table 1b represents the first-order co-occurrence vector for the target word *momka*. Values of this vector are 1 for context words that were found within the window and 0 for context words that did not co-occur with the target word in this instance. In total, we built 1164207 first-order co-occurrence vectors (obtained a matrix containing 1164207 rows and 2383 columns).

Based on the first-order co-occurrence vectors, we defined the cues and the outputs for the ndl simulation. The bigrams that constitute each of the target words (all inflected forms of 150 polysemous lemmas) were taken as the input, i.e., as cues. For example, as also illustrated in Table 1c, input for the word *momak* would be *#m, mo, om, ma, ak, k#* (# marks the beginning and the end of the word). To specify the output, we formed an array consisting of the lemma, which is associated with the inflected target word, followed by the list of all of the context words that were assigned value 1 in the first-order co-occurrence vector for the given target word, i.e., which co-occurred with the target word in that particular instance (see Table 1c for illustration).

Table 1. An illustration of the process of creating cues and outcomes based on first-order vectors.

a) One hypothetical occurrence of the word *momka* (found at the centre of the 7-point moving window) in the text ("word" denotes any word in the text; "context1166" and "context1963" represent codes for the actual words that were preselected to serve as the context words).

…	word	word	*context1166*	word	**momka**	*context1963*	word	word	word	…

b) First-order vectors for inflected forms of lemma *momak*.

Target word	*context523*	⋮	*context700*	⋮	*context1166*	⋮	*context1472*	⋮	*context1543*	⋮	*context1780*	⋮	*context1963*	⋮
momka	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**
momka	0	0	0	0	1	0	0	0	0	0	0	0	0	0
momka	1	0	0	0	0	0	1	0	1	0	0	0	0	0
momkom	1	0	0	0	0	0	0	0	0	0	1	0	0	0
momcima	0	0	1	0	0	0	0	0	0	0	1	0	1	0

c) Examples of cues and outcomes built from first-order vectors of inflected forms of lemma *momak*.

Target word	*Input bigrams (Cues)*	*Output (Outcomes)*
momka	**#m, mo, om, mk, ka, a#**	**momak_context1166_context1963**
momka	#m, mo, om, mk, ka, a#	momak _context1166
momka	#m, mo, om, mk, ka, a#	momak _context523_context1472_context1543
momkom	#m, mo, om, mk, ko, om, m#	momak _context523_context1780
momcima	#m, mo, om, mc, ci, im, ma, a#	momak _context700_context1780_context1963

For estimating the weights between cues (bigrams) and outcomes (lexomes), the Rescorla-Wagner equation (with learning parameters set to default values) was applied in an iterative learning algorithm using R statistical software (R CoreTeam, 2017) by relying on the **ndl** package (Arppe et al., 2015). The algorithm iterated randomly through pairs of inflected forms and their associated sets consisting of the lemma and the context words which they co-occurred at the given occasion (these pairs are illustrated in Table 1c). Based on the estimated strengths of cue-outcome associations, we calculated the values of three ndl-based measures: G2L Activations, G2L Diversity, and G2L Form Prior. By doing so, we asked three questions, as illustratively listed by Tomaschek (2020): a) how strong the presented set of cues activates the outcomes, b) how many outcomes these cues are related to, and c) how many cues the outcomes are related to.

G2L Activations were calculated by summing the cue-outcome weights for the active set of cues (set of bigrams contained in the given word form), for the lemma associated with the given word form, and for all the context words:

$$a_j = \sum_{i \in C} w_{ij} \quad (6)$$

For example, for the word form *momka* (Table 7c), we would calculate the activation by summing the association weights for the cues *#m, mo, om, mk, ka*, *a#*, and for the outcomes *momak*, and the context words. By doing so, we are capturing the strength of activation at the level of outcomes, which is elicited by the presentation of the given set of cues (bigrams from the word form), or how convincing the evidence presented in the input is. This measure was first described by Baayen et al. (2011) and has been applied in multiple research revealing that it was correlated with N-gram frequencies, morphological family size, paradigmatic/syntagmatic relative entropy, etc. (Baayen, Hendrix, Ramscar, 2013; Filipović Đurđević & Gatarić, 2018, Filipović Đurđević & Milin, 2019; Tomaschek et al., 2019). G2L Activation was typically observed to correlate negatively with processing latencies and positively with various indicators of lexicality. Therefore, in addition to being negatively correlated with lexical decision latencies, we would expect it to be positively correlated with the number of senses and the entropy and negatively correlated with the redundancy.

G2L Diversity was derived as the 1-norm, i.e., the sum of the absolute values of all the weights associated with the active set of cues (bigrams contained in the given word form):

$$D_C = \sum_{i \in C, j} \left| w_{ij} \right| \quad (7)$$

Therefore, for the presented cues, all of their weights are taken into account. Given that some weights have a negative, whereas some weights have a positive value, summing their absolute values reveals a different type of information as compared to summing their values. It reveals the extent of association weights being different from zero or the level of variability among the weights. Hence, Diversity reveals how much activity in the network is elicited by the presented cues, regardless of how convincing the evidence is. Milin et al. (2017) also describe Diversity as the measure of the amount of competition among the lexomes and is typically associated with inhibitory effect on processing latencies (e.g., in lexical decision task; Filipović Đurđević & Milin, 2019). Hence, we expect it to be directly proportional to lexical decision latencies in our experiment, as well. However, making predictions with respect to its relation with measures of polysemy is not straightforward. On the one hand, based on the fact that polysemous words have multiple interpretations and appear in various contexts, we would predict words with more sense uncertainty to obtain higher values of G2L Diversity. On the other hand, the multiple senses of polysemous words are mutually related, which could lead to reduced competition.

G2L Form Prior is also calculated as the 1-norm, but instead of summing row-wise, it is applied to the column and relies on the summing of the absolute values of all the weights associated with a given outcome:

$$P_O = \sum_i |w_{iO}| \quad (8)$$

We calculated it for each lemma as the outcome. This measure situates the lexome in the orthographic space and is considered to be a measure of prior availability of the lexome, showing how well the lexome is learned. It is typically correlated with word frequency and similar indicators of lexicality. We would therefore expect a positive correlation between G2L Form Prior and the sense uncertainty (as well as with the number of senses and entropy, and a negative correlation with redundancy) and to exhibit a facilitatory effect on lexical decision latencies.

Summary

To summarise, we presented Serbian polysemous nouns in a large scale visual lexical decision task with the goal of demonstrating that the entropy of sense probability distributions represented an adequate description of polysemy, as suggested previously by Gilhooly & Logie (1980). Next, having in mind the close relation of the entropy and the number of senses, we aimed to demonstrate that the cognitive system was sensitive not only to the number of senses but to sense probabilities, as expressed via continuous measure of the redundancy as well. Finally, we looked at polysemy from the point of view of discriminative learning and suggested that the effects of sense uncertainty could arise as the consequence of the graphemic cues competing for multiple partially overlapping outcomes. In order to do so, we investigated the relation of ndl-based measures both with the sense uncertainty measures and with the processing latencies.

Method

Participants

One-hundred-and-eighty-four students from the Department of Psychology at the Faculty of Philosophy, University of Novi Sad, and the Department of Psychology at the Faculty of Philosophy, University of Belgrade took part in the experiment either for partial course credit or as volunteers. All participants were native speakers of Serbian with normal or corrected-to-normal vision, who signed the informed consent form prior to participating in the study. The study was approved by the Ethical Committee at the Department of Psychology at the Faculty of Philosophy, University of Novi Sad. The number of participants was the largest that we were able to collect, as we deliberately aimed at collecting the largest possible sample. This decision was made based on our attempt to estimate the value of the coefficient for the variable, the effect of which has previously never been tested. In order to estimate the size of the effect as reliably as possible, we aimed to collect data from as many participants as we were able to recruit in the given time.

Materials and Design

All stimuli and their relevant descriptions were taken from the previously conducted norming study (Filipović Đurđević & Kostić, 2017). The stimulus set comprised 150 polysemous nouns (that were the subject of the analyses), ten unambiguous nouns, and 160 pseudo-words that were constructed by replacing a single grapheme/phoneme of the existing word ("hard" pseudo-words). The presented pseudo-words did not contain illegal bigrams and compared to words, they were of similar length in letters, summed bigram frequencies, and orthographic neighbourhood (OLD20), as illustrated in Table 2. Additionally, 20 polysemous nouns and 20 pseudo-words that were not included in the analyses were presented in the practice sessions.

Table 2. Descriptive statistics of the stimuli.

		M	*min*	*max*	*SD*
Length in letters	Words	5.038	4	6	.655
	Pseudo-words	5.068	4	7	.672
Summed bigram frequency	Words	195099.0	42124	408095	85575.55
	Pseudo-words	188047.4	21342	407015	86370.79
OLD20	Words	1.564	1	2.15	.260
	Pseudo-words	1.741	1	2.90	.299
Lemma frequency	Words	102.023	5	499.5	107.391
Familiarity (1-7)	Words	6.279	3.71	6.95	.484
Concreteness (1-7)	Words	4.926	1.52	7	1.542
Number of senses	Words	11.533	4	34	4.999
Entropy	Words	2.864	1.565	4.448	.607
Redundancy	Words	.162	.043	.365	.058
Dominance	Words	.294	.12	.63	.099

The dependent variable was reaction time, whereas the set of fixed predictors consisted of several numerical control variables and three critical numerical predictors. The control variables were word length in letters, the number of orthographic neighbours (Coltheart's N), (log) lemma frequency (Kostić, 1999), word familiarity operationalised as subjective frequency, and word concreteness. The critical predictors were either entropy of sense probability distribution, or the number of senses and the redundancy of sense probability distribution, as estimated in the previous norming study (Filipović Đurđević & Kostić, 2017). In that study, the estimation of the number of senses was achieved by applying the total meaning metric proposed by Azuma (1996).

For example, Table 3 lists probabilities (relative frequencies) of individual senses of the Serbian word *momak* (guy), obtained in Filipović Đurđević and Kostić's (2017) norming study. These relative frequencies are based on the number of participants listing a given sense in a sense generation task.

Table 3. Distribution of relative frequencies and average sense familiarity ratings of individual senses of the word *momak* (*guy*; Filipović Đurđević & Kostić, 2017).

momak (guy)	*Number of participants listing a sense (F)*	*Relative frequencies of senses (p=F/Total)*	*Sense familiarity rating (M)*
young male	16	.42	6.94
partner, lover	13	.34	6.88
unmarried male	6	.16	6.18
a mocking expression	1	.03	5.00
waiter	1	.03	4.65
servant	1	.03	3.00
Total:	17		

The entropy of sense probability distribution of the word *momak* would be calculated as follows:

$$H_{momak} = -0.42 \cdot \log_2 0.42 - 0.34 \cdot \log_2 0.34 - 0.16 \cdot \log_2 0.16 - 0.03 \cdot \log_2 0.03$$
$$- 0.03 \cdot \log_2 0.03 - 0.03 \cdot \log_2 0.03$$
$$= 0.526 + 0.529 + 0.420 + 0.138 + 0.138 + 0.138 = 1.890$$

The word *momak* has six senses, and the redundancy of the sense probability distribution would be calculated as:

$$T_{momak} = 1 - 1.89/\log_2 6 = 1 - 1.89/2.59 = 1 - 0.73 = 0.27$$

Finally, ndl-based measures of G2L Activation, G2L Diversity, and G2L Form Prior were calculated following the procedure described in the introduction and the equations (6-8 respectively).

Procedure

Stimuli were presented in a visual lexical decision task by using Open Sesame experimental software (Mathôt et al., 2012). Each trial started with the presentation of the fixation point that remained on the screen for 1000ms, followed by a blank screen for 500ms. After that, the stimulus was presented to remain on the screen until the participant's response or until the end of the time-out interval of 2000ms. Participants responded with a mouse-button press. Stimuli were presented in random order that was generated separately for each participant. The experiment was divided into two blocks, and stimuli were counterbalanced across the blocks. At the beginning of each block, 20 practice trials resembling the overall design were presented that were not later included in the analyses.

Results

The data were analysed in **R** statistical software (R Core Team, 2017) by using **lme4** (Bates, et al., 2015), **lmertest** (Kuznetsova et al., 2017), **RePsychLing** (Bates et al., 2015; Matuschek et al., 2017), **gbm** (Greenwell et al., 2020), and **ggplot2** (Wickham, 2016) packages. Prior to analysis, we made sure that participants and items with above 25% error rate were excluded. This led to the exclusion of one participant and four items (*sfera*, *berba*, *pisak*, *patent*). Inspection of reaction time distribution revealed several outliers corresponding to very fast responses that probably occurred by mistake. Therefore, we removed a total of six data points with RT<300ms. The comparison of words and pseudo-words revealed no difference in processing accuracy (words: 95.41% correct responses; pseudo-words: 95.72% correct responses). However, as expected, words were responded to more quickly than pseudo-words ($M[RT_{words}]$ = 632.800 ms (151.924); $M[RT_{pseudo\text{-}words}]$ = 707.386 ms (163.542); β=-74.586, t=-57.7, p<.001).

Measures of sense uncertainty

We first looked at bivariate Pearson correlation coefficients between by-item average reaction time and each of our critical predictors (Figure 3). This quick inspection revealed that almost all of our variables of interest were moderately correlated with processing latencies. The exception was *B*, which was not related to reaction time (r=-.037, t(144)=-.444, p=.658), and hence will not be further analysed. The entropy and the number of senses were negatively correlated with reaction time, whereas the redundancy and the dominance were positively correlated. These correlations reveal that an increase in uncertainty was always followed by a decrease in reaction time.

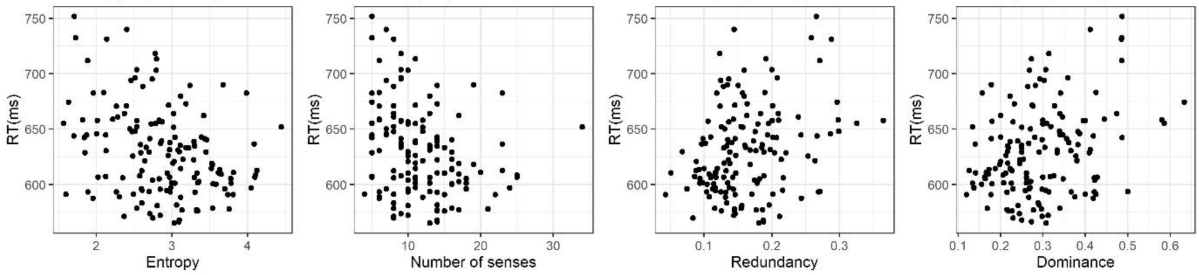

Figure 3. By-item correlation between entropy (leftmost panel), number of senses (second panel), redundancy (third panel), and dominance (rightmost panel) with reaction time.

We also applied Generalized Boosted Regression to estimate the relative contribution of multiple estimates of sense uncertainty both in relation to each other and in relation to control predictors. As depicted in Figure 4, familiarity and frequency have the strongest influence on the prediction of processing latencies, followed by redundancy and entropy. The relative contribution of the number of senses is fairly weak in the context of the rest of the variables, and the dominance (included here for the comparing purposes) is between the entropy and the number of senses.

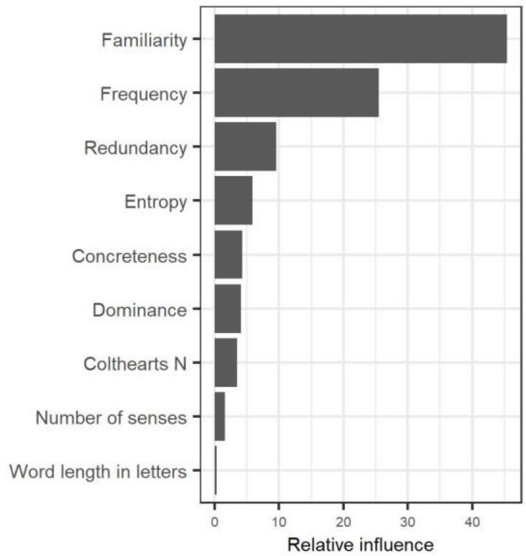

Figure 4. The relative importance of predictors as estimated by Generalized Boosted Regression Modelling.

In the next step, we tested whether entropy, number of senses, and redundancy could account for the processing time variation over and above the effects of control variables. Following Baayen and Milin (2010; but also Brysbaert & Stevens, 2018), reaction time was inversely transformed to approach the normal distribution. Next, we checked for collinearity in our predictor set by calculating the Kappa coefficient (Belsley et al., 1980). Given that collinearity was very high (68.558), we performed principal component analysis on control predictors (word length in letters, orthographic neighbourhood, (log) lemma frequency, word familiarity, and concreteness) and included all of the mutually orthogonal principal components as predictors in order to account for variability incurred by the control variables. The pairwise correlations between the five control variables and each of the five principal components are presented in Table 4. This led to a dramatic decrease in collinearity within the full set of predictors (Kappa coefficient was reduced to 9.392). Finally, the predictors were transformed into z-scores, as suggested by Gelman and Hill (2007).

Table 4. Correlations of each of the five principal components and the five control variables they replace (only significant correlations are presented).

	PC1	PC2	PC3	PC4	PC5
Word length	-.69	-.38	.25	-.57	
Familiarity	.62	-.64	.14		.44
Concreteness	.22	-.50	-.79	-.14	-.22
Lemma frequency	.61	-.19	.41	-.17	-.37
Coltheart’s N	.51	.66	-.22	-.47	.18

We continued by fitting the linear mixed-effects models to inversely transformed reaction time using the **lme4** package (Bates et al., 2015). The goal was to find the model that is simultaneously the best fit to the observed data and the most parsimonious way to account for the observed variability in the data. In order to provide the best fit, we tested for nonlinearities in the effects of all of our predictors (by including the quadratic component), as well as for all possible two-way and three-way interactions. In order to ensure parsimony, i.e., to make sure that only the predictors that are justified by the data are left in the model, we started from the most complex model and gradually built simpler models by excluding one parameter at the time. For each version of the simpler model, we compared its fit with the fit of the more complex model: we would leave the critical parameter in the model only if its exclusion led to a significant reduction in the fit indices. Following Barr, Levy, Scheepers, and Tily (2013), we also started with the maximum random structure. However, the model of such complexity was not able to converge. Therefore, we adopted the strategy of gradually decreasing the complexity of the random effects by eliminating the components with the least variance by applying the **RePsychLing** package (Bates et al., 2015; Matuschek et al., 2017). The final model included a random structure that contained both control and critical predictors that did not prevent the model from converging and were justified by the data (as determined by comparing AIC and log-Likelihood). In the end, the final models were refitted by excluding the data points that were beyond the -/+ 2.5 SD units range of residuals. This did not produce any change in the observed pattern of effects. We reported the refitted models. For the critical predictors (entropy, number of senses, and redundancy), we also reported the comparisons of the models with and without the predictor of interest. In these comparisons, we first looked at the significance of the chi-square test to make sure that we included the predictor that accounted for the significant portion of the additional variance. Next, we looked at AIC and BIC values, aiming for a model with lower AIC and at least unchanged BIC values. This decision was made based on BIC being more penalising on additional parameters and our goal of detecting the novel effects. Therefore, we will set the goal for the model to increase the fit (smaller AIC) and not jeopardise parsimony (equal or smaller BIC).

Entropy

The model that included entropy as the critical predictor revealed that processing latencies were influenced by the entropy of the sense probability distribution even after controlling for the effects of word length, orthographic neighbourhood size, (log) lemma frequency, word familiarity, and word concreteness, as represented by the five principal components (Table 5). An increase in entropy was followed by a decrease in processing time (Figure 5). The significant non-linear component of this effect revealed that the facilitation became attenuated for high values of entropy. As -/+95% confidence intervals in Table 5 show, the directions of the fixed effects seem to be reliably estimated. Crucially, the inclusion of entropy in the model significantly contributed to the model fit, i.e., it was justified by the data, as demonstrated by model comparison (Table 6).

Table 5. The coefficients from the linear mixed-effects model fitted to inversely transformed reaction time (-1000/RT) with entropy as the critical predictor.

Random Effects:

	Variance	St. deviation
Participants (Intercept adjustments)	.033	.182
Order of trial presentation by participant (slope adjustments)	.002	.039
PC1 by participant (slope adjustments)	<.001	.019
PC5 by participant (slope adjustments)	<.001	.011
Entropy by participant, linear (slope adjustments)	2.822	1.68
Entropy by participant, nonlinear (slope adjustments)	.002	.041
Items (Intercept adjustments)	.004	.063
Residual	.063	.251

Fixed effects:						
	Estimate	Std. Error	t value	-95%CI	+95%CI	Pr(>|t|)
			-			
Intercept	-1.235	.100	12.405	-1.43	-1.04	<.001
Order of trial presentation	.003	.003	.765	-.004	.009	.445
PC1	-.050	.006	-8.531	-.061	-.038	<.001
PC2	.042	.006	7.415	.031	.053	<.001
PC3	-.027	.006	-4.577	-.038	-.015	<.001
PC5 (linear)	-1.795	.903	-1.989	-3.564	-.026	.049
PC5 (nonlinear)	2.061	.896	2.301	.305	3.817	.023
Entropy (linear)	-.285	.069	-4.114	-.421	-.149	<.001
Entropy (nonlinear)	.045	.012	3.746	.021	.068	<.001

We also tested whether the dominance accounted for processing latencies variance over and above the entropy and vice versa. The former was tested by comparing the model that contains the control variables and the dominance (Model 2a in Table 6), and the model that contains the control variables, the dominance, and the entropy (Model 2b in Table 6). The comparison revealed that inclusion of the entropy significantly contributed to the model fit and that it was justified by the data (AIC values smaller for Model 2b compared to Model 2a; larger BIC values are not diagnostic in this case, as having both entropy and dominance in the model is clearly overfitting). Inclusion of the dominance over and above the entropy was not justified (as revealed by the comparison of the Model 2 and Model 2b, not presented here, but visible from comparing AIC and BIC values).

Table 6. The comparison of the model that included control variables and entropy (the model presented in Table 5) and the model that included only control variables. It is followed by the comparison of the model that included control variables and dominance and the model that included control variables, dominance, and entropy.

	Df	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>χ^2)
Model 1 (Control variables)	16	8087.7	8218.2	-4027.9	8055.7			
Model 2 (Model1+Entropy)	18	8071.7	8218.5	-4017.8	8035.7	20.012	2	<.001
Model 2a (Model1+Dominance)	17	8080.2	8218.9	-4023.1	8046.2			
Model 2b (Model2 + Dominance)	19	8073.0	8228.0	-4017.5	8035.0	11.193	2	.004

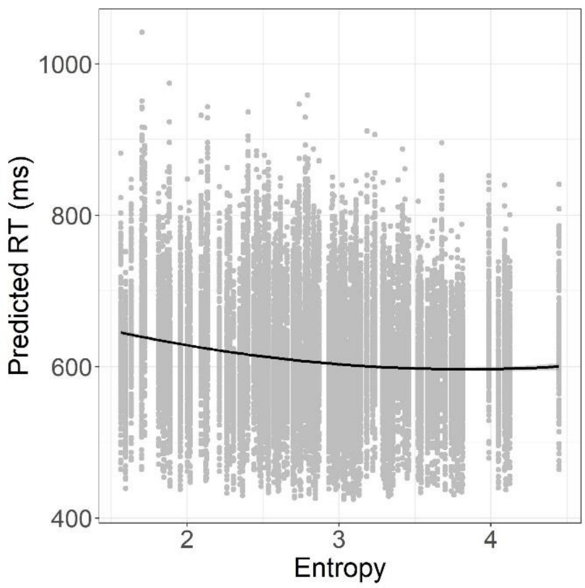

Figure 5. The effect of entropy of sense probability distribution on processing latencies in visual lexical decision task.

Analysis of accuracy (Table 7) confirmed the findings from the analysis of processing latencies. After controlling for the effects of the same covariates, a significant nonlinear effect of entropy was observed. An increase in entropy was followed by an increase in the probability of accurate response, the effect being attenuated for high values of entropy.

Table 7. The coefficients from the generalised linear mixed-effects model fitted to response accuracy (1 – accurate; 0 – inaccurate) with entropy as the critical predictor.

Random Effects:

	Variance	St. deviation
Participants (Intercept adjustments)	0.381	0.618
Order of trial presentation by participant (slope adjustments)	0.037	0.193
PC1 by participant (slope adjustments)	0.022	0.15
PC5 by participant (slope adjustments)	<.001	0.001
Entropy by participant, linear (slope adjustments)	<.001	0.015
Entropy by participant, nonlinear (slope adjustments)	<.001	0.011
Items (Intercept adjustments)	0.037	0.608

Fixed effects:

	Estimate	Std. Error	z value	-95%CI	+95%CI	Pr(>|z|)

Intercept	-.059	1.095	-.054	-2.205	2.087	.957
Order of trial presentation	-.183	.037	-4.983	-.254	-.111	<.001
PC1	.176	.066	2.686	.0483	.305	.007
PC2	-.434	.064	-6.814	-.559	-.309	<.001
PC3	.241	.068	3.528	.107	.374	<.001
PC5 (linear)	6.678	13.014	.513	-18.830	32.185	.608
PC5 (nonlinear)	-23.226	9.990	-2.325	-42.804	-3.646	.020
Entropy (linear)	2.456	.774	3.171	.938	3.973	.002
Entropy (nonlinear)	-.368	.134	-2.751	-.630	-.106	.006

Number of senses and redundancy

The model that included the number of senses and the redundancy as critical predictors revealed that both components of entropy were related to processing latencies even after controlling for the effects of the control variables, as represented by the five principal components (Table 8). We observed a non-linear effect of the number of senses (Figure 6; left panel). The number of senses was inversely related to reaction time, the facilitation being attenuated for high values of the number of senses. Additionally, we observed an effect of redundancy and the interaction between redundancy and PC1 (the principal component that accounted for the largest proportion of variance in control variables). This interaction revealed that the inhibitory effect of redundancy was strongest for low PC1 values, i.e., for longer, less familiar, and less frequent words. As the values of PC1 increased, the effect of redundancy decreased (Figure 6; right panel). Importantly, based on the -/+95% confidence intervals presented in Table 8, we believe that the directions of the fixed effects were reliably estimated. Finally, by applying the model comparisons, we demonstrated that the effect of the number of senses was significant over and above the effect of the control variables (Table 9).

Crucially, the effect of the redundancy was significant over and above the joint effects of the control variables and the number of senses. Finally, the interaction between redundancy and PC1 was also justified by the data.

Table 8. The coefficients from the linear mixed-effects model fitted to inversely transformed reaction time (-1000/RT) with the number of senses and redundancy as critical predictors.

Random Effects:

	Variance	St. deviation
Participants (Intercept adjustments)	.033	.182
Order of trial presentation by participant (slope adjustments)	.002	.040
PC1 by participant (slope adjustments)	<.001	.019
PC5 by participant (slope adjustments)	<.001	.011
Number of senses (linear) by participant (slope adjustments)	1.826	1.351
Number of senses (nonlinear) by participant (slope adjustments)	.011	.106
Items (Intercept adjustments)	.004	.059
Residual	.063	.251

Fixed effects:

	Estimate	Std. Error	t value	-95%CI	+95%CI	Pr(>|t|)
Intercept	-1.6	.030	- 52.578	-1.660	-1.54	<.001
Order of trial presentation	.003	.003	.765	-.004	.009	.445
PC1	-.051	.006	-9.141	-.062	-.040	<.001

PC2	.042	.005	7.978	.032	.052	<.001
PC3	-.025	.006	-4.509	-.036	-.014	<.001
PC5 (linear)	-1.547	.857	-1.805	-3.227	.133	.073
PC5 (nonlinear)	2.043	.841	2.431	.396	3.691	.016
Number of senses (linear)	-.012	.004	-2.982	-.019	-.004	.003
Number of senses (nonlinear)	.0004	.00001	3.218	.0002	.001	.002
Redundancy	.020	.006	3.404	.009	.032	.001
PC1 :Redundancy	-.017	.005	-3.190	-.028	-.007	.002

Table 9. The comparisons of the nested models: the simplest model contains only control variables while successive models include one additional variable each: number of senses (N), redundancy (T), and redundancy by PC1 interaction.

	Df	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>χ^2)
Model 1 (Control variables)	16	8088.8	8219.3	-4028.4	8056.8			
Model 3 (Model 1 + N)	18	8078.8	8225.6	-4021.4	8042.8	14.034	2	<.001
Model 4 (Model 3 + T)	19	8065.3	8220.3	-4013.6	8027.3	15.475	1	<.001
Model 5 (Model 4 + T:PC1)	20	8056.8	8220	-4008.4	8016.8	10.449	1	.001

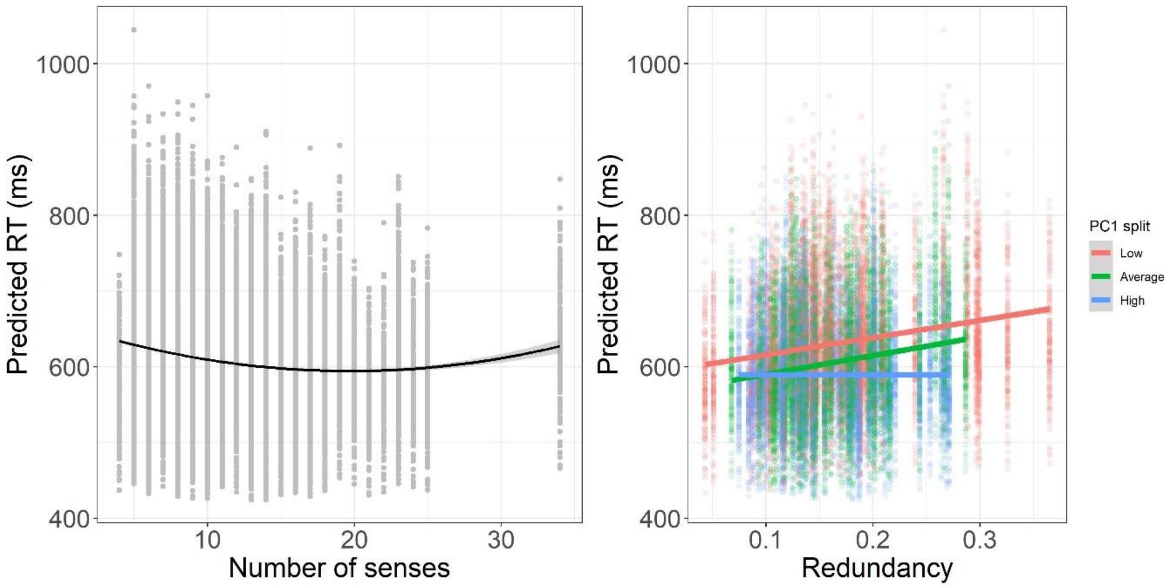

Figure 6. The effect of the number of senses (left panel) and the balance of sense probabilities – redundancy (right panel; low levels of redundancy refer to balanced polysemes) on processing latencies in visual lexical decision task; the right panel depicts the interaction of redundancy and the first principal component encapsulating control predictors (blue line – high values of PC1 refer to short, frequent and familiar words).

The analysis of accuracy data (Table 10) confirmed the effect of the number of senses as observed in the analysis of processing latencies. Upon controlling for the effects of the same covariates, an increase in the number of senses was followed by higher accuracy, and the effect was again attenuated for high values of the number of senses. However, unlike with reaction time data, although the direction of the coefficient was in the expected direction, the effect of redundancy was not significant in the analysis of accuracy.

Table 10. The coefficients from the generalised linear mixed-effects model fitted to response accuracy (1 – accurate; 0 – inaccurate) with the number of senses and redundancy as critical predictors.

Random Effects:

	Variance	St. deviation
Participants (Intercept adjustments)	.381	.617
Order of trial presentation by participant (slope adjustments)	.037	.193
PC1 by participant (slope adjustments)	.023	.150
PC5 by participant (slope adjustments)	<.001	<.001
Number of senses (linear) by participant (slope adjustments)	<.001	.009
Number of senses (nonlinear) by participant (slope adjustments)	<.001	.012
Items (Intercept adjustments)	.343	.586

Fixed effects:

	Estimate	Std. Error	z value	-95%CI	+95%CI	Pr(>|z|)
Intercept	2.575	.3191	8.071	1.950	3.200	<.001
Order of trial presentation	-.183	.037	-4.987	-.255	-.111	<.001
PC1	.173	.064	2.685	.047	.299	.007
PC2	-.430	.061	-6.998	-.551	-.310	<.001
PC3	.243	.067	3.623	.111	.374	<.001
PC5 (linear)	7.276	9.256	.786	-10.865	25.418	.432
PC5 (nonlinear)	-23.377	9.984	-2.341	-42.945	-3.809	.019
Number of senses (linear)	.185	.045	4.126	.097	.273	<.001

Number of senses (nonlinear)	-.006	.001	-3.941	-.008	-.003	<.001
Redundancy	-.081	.070	-1.152	-.219	.057	.249
PC1 :Redundancy	.078	.064	1.224	-.047	.203	.221

NDL-based measures

We analysed the ndl-based measures in several steps. First, we rank-transformed their values, as suggested by previous studies (Milin et al., 2017; Filipović Đurđević & Milin, 2019). Next, we looked at the bivariate correlations of ndl-based measures and lexical descriptors (Table 11). We found that higher G2L Activation was associated with frequent, familiar, abstract, shorter words with fewer orthographic neighbours. On the other hand, G2L Diversity was lower for frequent, familiar, abstract words. Finally, G2L Form Prior was higher for frequent, abstract words with more orthographic neighbours. Also, we found that G2L Activation and G2L Diversity were significantly negatively correlated: ($r=-.434$, $t(148)=-5.864$, $p<.001$).

Importantly, we observed a significant correlation between our measures of polysemy and ndl-based measures (Table 11). As also illustrated in Figure 7, the redundancy was negatively correlated with G2L Activation and G2L Form Prior, indicating that words with balanced sense probabilities lead to higher activation of their associated meaning and are better entrenched in the network. Similarly, G2L Form Prior was positively correlated with the entropy and the number of senses, indicating along the same line that words with the higher number of senses and generally words with higher sense uncertainty are better entrenched in the network. Finally, both the entropy and the number of senses were negatively correlated with G2L diversity, indicating that a higher number of senses is associated with less competition among the outcome lexomes. We found this result unexpected, and we will return to it in the next step of the analysis. However, although unexpected, it was consistent with the marginally significant positive correlation between redundancy and G2L Diversity (again indicating counterintuitively that words with balanced sense probabilities lead to less competition).

Table 11. Significant bivariate correlations ($r>.18$, $p<.05$) among lexical descriptors and ndl-based measures.

	G2L Activation	G2L Diversity	G2L Form Prior
Word length in letters	-.15º		
Coltheart's N	-.19*		.22**
Lemma Frequency	.26**	-.45***	.34***
Familiarity	.18*	-.35***	
Concreteness	-.23**	.29***	-.18*
Entropy		-.20*	.30**
Number of senses		-.20*	.26**
Redundancy	-.22**	.16 º	-.18*

*** p<.001; **p<.01; *p<.05; ºp<.1

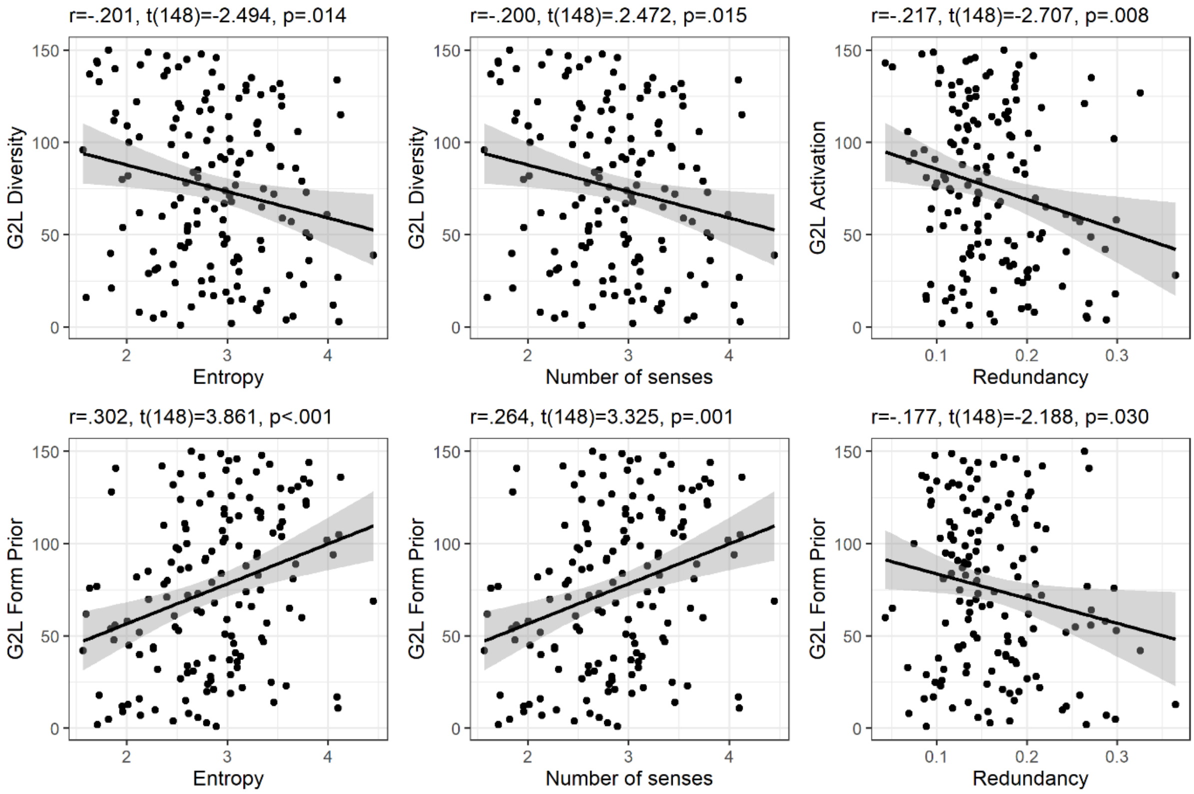

Figure 7. The relation between polysemy measures and ndl-based measures (based on significant bivariate correlation coefficients).

In the next step, we wanted to test whether the observed correlations between polysemy measures, on the one hand, and ndl-based measures, on the other hand, would remain upon taking into account the control variables. In other words, we wanted to make sure that the observed relationship was not a by-product of the mediation provided by the more basic lexical descriptors. We, therefore, built the models with ndl-based measures as the criterion variables and the polysemy measures as the critical predictors while including the control predictors in the model, as well. To avoid the collinearity issues, we included the five principal components that we used in the previous analysis (Table 4). Our results revealed that after controlling for the effects of more basic lexical descriptors, the redundancy contributes to G2L Activation (Table 12), and the entropy contributes to G2L Form Prior (Table 13). The values of the coefficients confirm that words with balanced sense probabilities lead to more activation at the outcome level and that words with higher sense uncertainty, in general, are more entrenched in the network, i.e., more readily available. Interestingly, neither the surprising negative correlation between the entropy/number of senses and G2L Diversity nor the positive correlation between the redundancy and G2L Diversity did not survive the inclusion of the control variables.

Table 12. Coefficients from linear regression fitted to G2L Activation.

	Estimate	SE	t	Pr(>|t|)
(Intercept)	109.476	15.391	7.113	<.001
PC1	5.839	2.494	2.341	.021
PC2	-4.692	2.872	-1.634	.105
PC3	14.346	3.267	4.392	<.001
PC4	13.700	4.211	3.253	.001
PC5	-2.868	4.722	-0.607	.545
Number of senses	-0.973	0.728	-1.337	.184
Redundancy	-140.636	61.321	-2.293	.023

Table 13. Coefficients from linear regression fitted to G2L Form Prior.

	Estimate	SE	t	Pr(>|t|)
(Intercept)	39.79	16.834	2.364	.019
PC1	7.858	2.552	3.08	.002
PC2	4.419	2.907	1.52	.131
PC3	9.151	3.36	2.724	.007
PC4	-6.128	4.249	-1.442	.151
PC5	-4.777	4.857	-.984	.327
Entropy	12.468	5.769	2.161	.032

Upon establishing that polysemy measures are related to the measures derived from the ndl simulation (even upon controlling for lexical variables), we wanted to test whether ndl-based measures would account for the processing latencies that were recorded in the visual lexical decision experiment. We started again by looking into bivariate correlations. As illustrated in Figure 8, we observed that all of our attested ndl-based measures were significantly correlated with lexical decision latencies. As expected, G2L Activation and G2L Form Prior were negatively correlated with reaction time, whereas G2L Diversity and the observed reaction time were positively correlated. Therefore, words with more activation at the outcome level with less competition among the outcomes, which are more entrenched in the network, are responded to more quickly in the lexical decision task, as also observed earlier (Filipović Đurđević & Milin, 2019).

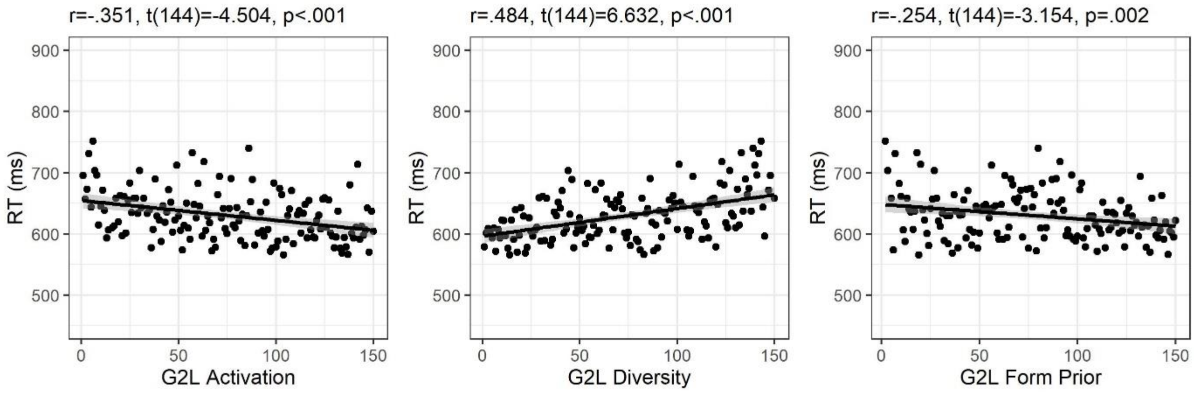

Figure 8. Bivariate correlations between reaction time and rank-transformed ndl-based measures: G2L Activation (leftmost panel), G2L Diversity (middle panel), and G2L Form Prior (rightmost panel).

To be able to evaluate the relative contribution of ndl-based predictors to accounting for the processing latencies, as compared to the relative contribution of other lexical descriptors, we performed Generalized Boosted Regression Modelling. As depicted in Figure 9, out of the three ndl-based measures we derived, G2L Diversity has the largest relative contribution to the prediction of lexical decision latencies, next to the effect of word familiarity, which was the most influential among the predictors. It is followed by the effect of frequency, G2L Activation, and G2l Form Prior being at the middle range of relative influence scale.

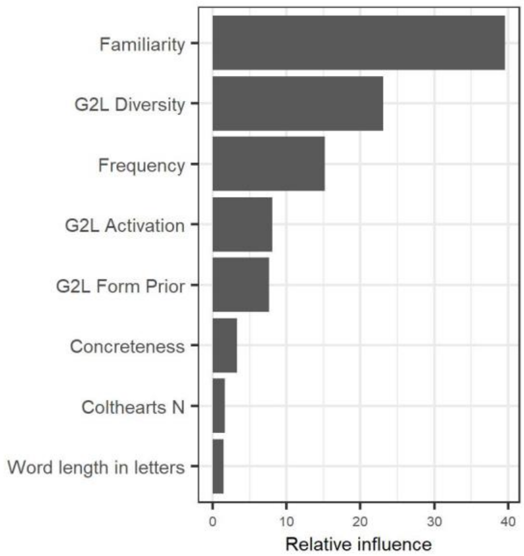

Figure 9. The relative importance of ndl-based predictors as estimated by Generalized Boosted Regression Modelling.

Finally, we conducted linear mixed-effect regression to transformed reaction time, with ndl-based measures as the critical predictors. We adopted the structure of the effects of the control variables that we have previously determined (Table 5) and included G2L Activation, G2L Diversity, and G2L Form Prior in the predictor set. As presented in Table 14 and Figure 10, upon controlling for the effects of the control variables, we observed significant facilitatory effect of G2L Activation and G2L Form Prior, as well as the significant inhibitory effect of G2L Diversity. Additionally, there was a significant interaction between G2L Form Prior and G2L Diversity, which indicated that the effect of G2L Diversity was less pronounced for words with high G2L Form Prior. This finding revealed that words with more activation of the outcome lexomes, as well as words that are better learned or more entrenched in the network, and the words with less competition among the outcomes are responded to more quickly. Additionally, the inhibitory effect of the competition was attenuated for the well-entrenched words.

Table 14. The coefficients from the linear mixed-effects model fitted to inversely transformed reaction time (-1000/RT) with ndl-based measures as the critical predictors.

Random Effects:

	Variance	St. deviation
Participants (Intercept adjustments)	.033	.182
Order of trial presentation by participant (slope adjustments)	.002	.039
PC1 by participant (slope adjustments)	<.001	.018
PC5 by participant (slope adjustments)	<.001	.011
G2L Activation by participant (slope adjustments)	<.001	.008
G2L Diversity by participant (slope adjustments)	<.001	.006
G2L Form Prior by participant (slope adjustments)	<.001	<.001
Items (Intercept adjustments)	.004	.062
Residual	.063	.251

Fixed effects:

	Estimate	Std. Error	t value	-95%CI	+95%CI	Pr(>|t|)
Intercept	-1.670	.0145	-115.353	-1.698	-1.641	<.001
Order of trial presentation	.002	.003	.749	-.004	.009	0.455
PC1	-.043	.006	-6.843	-.055	-.030	<.001
PC2	.034	.006	6.018	.023	.045	<.001
PC3	-.014	.007	-2.069	-.028	-.001	.040
PC5 (linear)	-1.911	.897	-2.127	-3.672	-.150	.035
PC5 (nonlinear)	1.535	.884	1.736	-.198	3.267	.085

G2L Activation	-.018	.006	-2.846	-.030	-.005	.005
G2L Diversity	.015	.007	2.026	.001	.029	.045
G2L Form Prior	-.017	.006	-2.81	-.029	-.005	.006
G2L Form Prior : G2L Diversity	-.015	.006	-2.643	-.025	-.004	.009

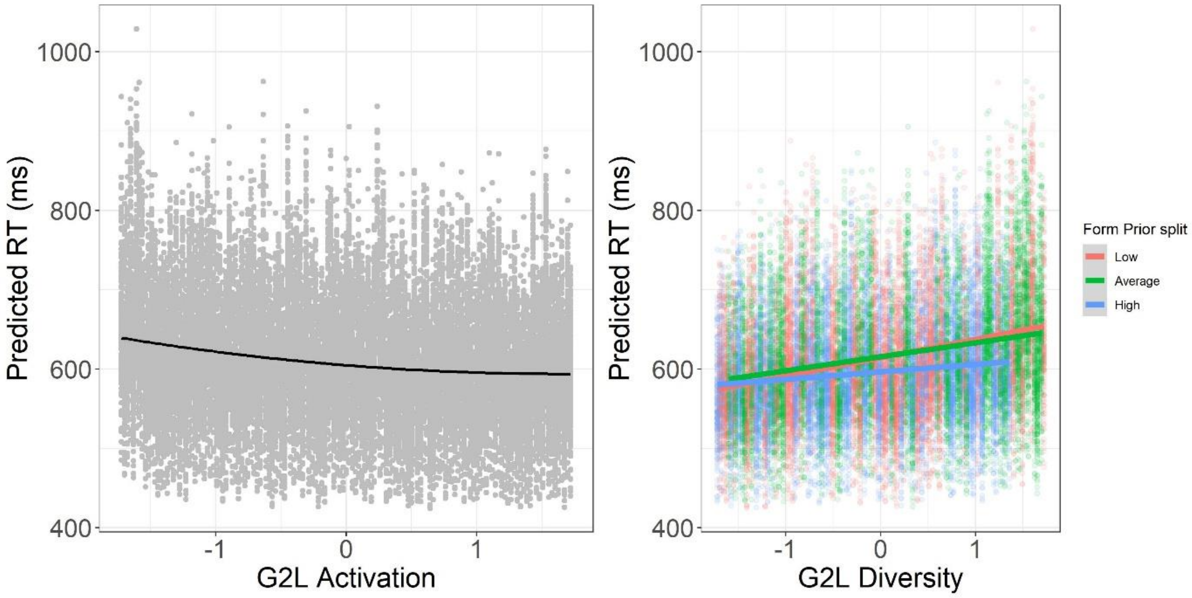

Figure 10. The effect of G2L Diversity; the left-hand panel depicts the interaction between G2L Diversity and the first principal components (high values: short, frequent, familiar words); the right-hand panel depicts the interaction between G2L Diversity and G2L Form Prior.

General discussion

We demonstrated that the balance of sense probabilities affected the lexical decision of isolated polysemous nouns over and above the previously demonstrated effect of the number of senses. Polysemous words with a higher number of senses and more balanced sense probabilities (i.e., lower redundancy) were processed faster. Accordingly, when combined in a unique measure of sense uncertainty, as quantified via Information Theory measure of entropy (as suggested by Gilhooly & Logie, 1980), the number of senses and the redundancy jointly affected processing. Our results, thus, suggested that processing latencies decreased with an increase in variability of related senses, which can be denoted by a given polysemous word. Because we conducted a large-sample study, we believe that we have achieved a reliable estimation of the effects, particularly in terms of their directions.

To the best of our knowledge, this is the first study that demonstrated processing effects of sense uncertainty quantified as the entropy of sense probability distribution and the first study to operationalize balance of sense probabilities as a continuum expressed by the redundancy of sense probability distribution. Albeit several large-scale norming studies provided data on the entropy of ambiguous words – so-called *U* for uncertainty (Gilhooly & Logie, 1980; Twilley & Dixon, 2000), in spite of our efforts, we have not been able to find an explicit empirical test of the processing effect of this variable. Similarly, although it is generally accepted that sense frequencies are of great importance for the processing of lexical ambiguity, the conclusions were always based on comparisons of balanced and unbalanced words (e.g., Brocher et al., 2018; Duffy et al., 1988; Frazier & Rayner, 1990; Grindrod et al., 2014; Klepousniotou & Baum, 2007; Klepousniotou et al., 2012; MacGregor et al., 2015; Rayner & Duffy, 1986; Rayner & Frazier, 1989), whereas we provided a test for the effect of redundancy as the continuous measure of the balance of sense frequencies. We would like to note that the measure of redundancy can be used in alternation with another measure of the balance of probabilities, called the *entropy ratio* (Batty et al., 2014) or *equitability* (Lloyd & Ghelardi, 1964). It is simply a complement to 1 of redundancy (calculated as the ratio between the observed and the maximum entropy), and the only difference between the equitability and the redundancy is in the direction of the effect: more balanced probabilities imply lower redundancy but a higher entropy ratio. Importantly, we also showed that the measure of entropy outperformed the dominance (probability of the most frequent sense; D; β; Armstrong et al., 2012; Rodd et al., 2002), whereas B (entropy of the two most frequent senses; Twilley et al., 1994) was not significantly correlated with processing latencies. This clearly demonstrated that it is the full distribution of senses and their probabilities that affects lexical decision latencies of the polysemous words, and not only one or two most frequent senses.

Our findings go in line with some similar demonstrations, both generally in language processing and specifically in the processing of lexical ambiguity. Generally, a growing body of psycholinguistic research has demonstrated that entropy of various language subsystems predicted processing latencies (Ackerman & Malouf, 2013; Baayen et al., 2006; Baayen et al., 2011; Balling & Baayen, 2012; Filipović Đurđević & Milin, 2019; Frank, 2010, 2013; Hale, 2001, 2003, 2006; Kemps et al., 2005; Levy, 2008; Milin et al., 2009; Moscoso del Prado Martìn et al., 2004; Tabak et al., 2005; Wurm et al., 2006). For example, it has been shown that inflectional entropy, which is calculated over the probability distribution of word's inflected variants (e.g., *walk-ø*, *walk-s*, *walk-ed*, *walk-ing*) was negatively correlated with reaction times (Moscoso del Prado Martín et al., 2004; Baayen et al., 2006). Along the same line, entropy and entropy-related effects have been recorded in the processing of syntactic functions and meanings (Kostić et al., 2003), inflectional paradigms (Milin et al., 2009), phonology (Luce & Large, 2001), and so forth. Given that entropy can be interpreted both as an index of uncertainty (as within Information Theory framework; Shannon, 1948) and as an index of diversity (e.g., as frequently used in biology and geography; Batty et al., 2014; Berger & Parker, 1970), we can also discuss our findings in terms of the wide group of research that aim to describe the semantic space quantitatively. More specifically, our findings go in line with the results of similar studies, which demonstrated that various indices of semantic richness facilitate recognition of polysemous words. For example, by applying the distributional semantic approach, McDonald and Shillcock (2001), as well as Adelman, Brown, and Quesada (2006), drew attention to the importance of contextual diversity (but see also Cevoli et al., 2020, and Hollis, 2020), followed by many others (Goh et al., 2016; Hoffman et al., 2013; Jones et al., 2012; Shaoul & Westbury, 2010). Hoffman and Woollams (2015) demonstrated faster processing in a visual lexical decision task for polysemous words with high semantic diversity (a measure of contextual variation based on cosine distances between context vectors in which a word occurred). Similarly, for Serbian words, Filipović Đurđević et al. (2009) demonstrated that processing time of isolated polysemous nouns was inversely proportional to their general semantic variability, that is, the variability of linguistic contexts in which they occur (general spread of vectors in semantic space defined by context words). All of these findings suggest that early word recognition (as the one in the visual lexical decision task) benefits from the large and more balanced spread across related semantic contexts. However, it should be noted that the effects are quite the opposite if the semantic contexts are not related, as is the case with the meanings of homonymous words. This was early observed in multiple studies that applied factorial design to show that unbalanced homonyms were processed faster than balanced homonyms (Duffy et al., 1988; Klepousniotou & Baum, 2007; Klepousniotou et al., 2012). The studies which applied correlational design mostly made the same observation. Although Rodd et al. (2002) failed to find a significant effect of the dominance score (*D*), Armstrong et al. (2012) revealed that recognition latencies were directly proportional to the relative frequency of the dominant meaning of homonyms (which they named β and which is analogue to *D*; see also Rice et al., 2019). They also reported that their measure performed almost identically to *B*, proposed by Twilley et al. (1994). Filipović Đurđević (2019; and also Mišić & Filipović Đurđević, submitted) found that recognition latencies were inversely proportional to the redundancy of the meaning distribution of the homonym. All of these findings testify in favour of the advantage of balanced sense frequency distributions (polysemy) and the disadvantage of balanced meaning frequency distribution (homonymy) in visual lexical decision task. Also, they point to an additional difference between polysemy and homonymy. Whereas homonymy processing is equally well predicted by the dominance (Armstrong et al., 2012) and by the *B* (Twilley et al., 1994), polysemy processing is related to the dominance, but not to *B*. Whereas we believe that this finding informed us of the importance of the full range of sense probabilities for polysemy processing, we would not conclude that it informed us regarding this matter in homonymy processing. This discrepancy could stem from the fact that the number of senses of polysemous words is higher than the number of meanings of homonymous words. Consequently, the first two meanings often cover the majority of meaning probabilities of homonymous words (as also found by Armstrong et al., 2012), whereas in our set of polysemous words, the sum of the largest sense probabilities covered on average approximately 50% of the sense probabilities (M=50.14%; min=24.00%; max=86.21%; sd=14.42). Our attempt to test this hypothesis by looking at the correlation between B and reaction time only for words with the low number of senses was not successful due to the small number of such words in our dataset. Another possible confound in comparing homonymy and polysemy in terms of the number of meanings/senses is the fact that relatedness among senses could lead to an overestimate of the number of senses. When categorizing responses, it is easy to form distinct categories for unrelated meanings (e.g., river bank vs money bank), whereas categorizing highly related senses sometimes can lead to creating a novel category for the descriptions which are between the related categories. For example, in the research reported here, there is a risk of the number of senses being overestimated for some words. This is visible in the high nonlinearity of the effect of the number of senses and also in the fact that the number of senses was the only predictor that led to an increase in BIC values. We have tested the models on a reduced set from which we excluded the words with extreme values of the number of senses to make sure that the effects remain the same. However, there is a risk of the number of senses effect being confounded with the effect of sense relatedness, which should be the subject of future research. We would expect the effect of redundancy to be less affected by this confounding, as the correlation between the redundancy and the number of senses is only moderate.

Our findings can be interpreted in light of the current models of lexical ambiguity. From the point of view of the semantic competition accounts (Armstrong & Plaut, 2008; 2016; Kawamoto, 1993; Rodd et al., 2004; Rodd, 2020), the facilitatory effect of sense diversity/uncertainty (many balanced senses) in visual lexical decision task would be attributed to fast accumulation of activation due to the absence of competition among the related senses. However, our findings as such are not able to distinguish between the two categories of current models of lexical ambiguity, as the facilitatory effect of sense diversity/uncertainty is not incompatible with the decision making account either (Hino et al., 2006; Pexman et al., 2004). Given that error-driven learning is pertinent with many PDP models, by stating that our results are in accordance with such models can be seen as the first step towards our goal of interpreting the sense uncertainty effects in the light of the principles of learning.

Our ultimate goal was to explore whether the complexity that we observed at the level of language – in our case the uncertainty of senses of polysemous words – could arise as the consequence of the principles of the discriminative learning (Rescorla, 1988). During this, we have obtained several important insights. First, our results showed that the three measures that we derived from the network of cue-outcome association strengths were correlated with our polysemy measures, thus showing that our network has captured the important features of our stimuli. Entropy and the number of senses were correlated with G2L Form Prior, and G2L Diversity, whereas the redundancy was correlated with G2L Activation and G2L Form

Prior. We also showed that upon controlling for the contribution of the control variables (such as word length, frequency, familiarity, etc.), there was a unique variance shared between ndl-based measures and the polysemy measures. This was the case for two pairs of our measures, redundancy and G2L Activation on the one hand, and the entropy and the G2L Form Prior, on the other hand. As predicted, the G2L Activation was inversely related with redundancy, indicating that for words with more balanced sense probabilities, more activation was incurred by the presentation of the bigram cues that constitute the word form. In other words, the presentation of the word form provided more bottom-up evidence in case of words with balanced sense frequencies. This finding was in accordance with the results of our experiment. Also, as predicted, the G2L Form Prior was positively correlated with the entropy and the number of senses, and negatively correlated with redundancy. This indicated that words with higher sense uncertainty (have a greater number of senses and have senses with balanced probabilities) were more readily available and better entrenched in the network. In addition to being in accordance with the results of the experiment, this finding is also easy to grasp intuitively – words that have more senses which are all frequently encountered are better learned as compared to words that we experience in a less variable way. However, opposite to our prediction, the G2L Diversity was inversely related to the entropy and the number of senses, suggesting that the presentation of bigram cues led to more competition at the outcome level. Given that the two correlation coefficients are almost identical, knowing that entropy and the number of senses are highly collinear, and having in mind that redundancy and G2L Diversity were not correlated, we would conclude that the crucial relation here is the one between the number of senses and the G2L Diversity. We find this interesting, as it could indicate an important profile of relations: redundancy – G2L Activation, number of senses – G2L Diversity, entropy/redundancy/number of senses – G2L Form Prior. Additionally, in terms of providing activation and support, as well as prior availability, both measures of uncertainty acted as predicted – more uncertainty led to more activation and availability. However, in terms of the number of senses, more senses seemed to lead to more competition. This finding would be perfectly expected for homonyms (having unrelated meanings and appearing in unrelated contexts). In the case of polysemy, it was contrary to our prediction that the overlapping contexts would cancel out the competition. This could indicate that the competition among the outcomes being predicted by orthographic cues takes place regardless of relatedness of the senses/meanings, but this claim certainly awaits further investigation. However, if the future investigations confirm this tendency, discriminative learning pathway could lead to solutions of several problems in the area of lexical ambiguity research. First, the concept of the discrete sense/meaning categories could be replaced by mapping of orthography to the co-occurring contexts. This way, the theoretical problems associated with sense/meaning categorization, as well as the laborious norming studies, could be avoided. Second, the opposing processing effects of the sense/meaning relatedness on lexical recognition could come for free.

Finally, we observed that all of the ndl-based measures were also correlated with reaction time. We showed that, even upon taking into account the control variables, all of them significantly contributed to lexical decision latencies. Words with greater G2L Activation were responded to more quickly, as well as words with higher G2L Form Prior, whereas word with higher G2L Diversity took more time to be responded to. This is an expected pattern of results, and the one which was observed in multiple studies (Arnold et al., 2017; Baayen et al., 2011; Baayen et al., 2013; Baayen et al., 2016; Filipović Đurđević & Milin, 2019; Hendrix et al., 2017; Milin et al., 2017; Nixon, 2020; Tomaschek et al., 2021). It indicates that fast lexical decision is obtained for words that provide lots of activation from the input, that we have lots of experience with and are thus well-entrenched, and words that do not provoke lots of competition at the outcome level.

By looking jointly at the insights that we gained, we could conclude that some aspects of the polysemy effect certainly can be simulated using Rescorla-Wagner equations (Rescorla & Wagner, 1972), and hence can be interpreted within the discriminative learning framework (Ramscar et al., 2010; Ramscar & Port, 2016; Rescorla, 1988). More specifically – these effects can arise within the network of simple mappings from orthographic cues to co-occurring words if crossed with the updating rule of Rescorla and Wagner. Although simple at a glance, the mapping and the rule allow for immensely complex dynamics (as elaborated by Hoppe et al., 2020). We believe that we have made the first step towards showing that these dynamics can resemble the dynamics that we find in the phenomenon of lexical ambiguity.

Importantly, having in mind the decisions that we made prior to the study with the purpose of controlling for confounding variables, we need to stress that, strictly speaking, our findings should be viewed in the light of the lexical decision to polysemous nouns. We believe that the same findings would be observed with other parts of speech, such as verbs, adjectives, or any type of words that can point to multiple related contents, but this remains an open question awaiting further empirical and computational investigation.

Conclusion

In this paper, we started by describing one case of complexity in language in terms of information theory – by describing polysemy as sense uncertainty. We then documented that the cognitive system is sensitive to this language complexity by conducting a large-scale lexical decision study. Finally, we demonstrated that the observed sensitivity to the described complexity in language could arise in the process of discrimination learning.

By introducing redundancy as a way to quantify the continuum of the balance of sense frequencies and by demonstrating its processing effects, we believe we have enabled further testing of different model predictions against each other. Additionally, by demonstrating the processing effects of sense uncertainty quantified as the entropy of sense probability distribution (as suggested by Gilhooly & Logie, 1980), we enabled a more efficient accounting for the variance introduced by lexical ambiguity. As it entails both factors contributing to the ambiguity level, the entropy measure should be a more sensitive measure of lexical ambiguity, as compared to the number of senses. Importantly, by making our observations on a set of strictly polysemous nouns, we enabled clear comparisons with the effects of the same descriptions within other categories of words (see Eddington & Tokowicz, 2015). Finally, by describing lexical ambiguity in terms of Information Theory measures, we added to the exponentially growing body of research that demonstrates the immense sensitivity of the cognitive system to the probabilistic structure of the environment in general (Chater & Oaksford, 1999) and language in particular (Milin et al., 2009; Milin, Kuperman et al., 2009). Such measures enable the formulating of functional dependencies in terms of Marr's computational level of analysis (1980) and provide insight into constraints of cognitive processing that serve as important guides in testing and formulating models and theories. Finally, by applying a computational model based on the principle of error-driven, discrimination learning to the case of polysemy, we attempted to broaden both the field of lexical ambiguity research and the field of error-driven learning.

Acknowledgements

This research is funded by Ministry of Education, Science and Technological Development of Republic of Serbia (grant number: 179033, and 179006).

Disclosure of interest

The authors have no conflicts of interest to report.

Data availability statement

The full dataset with all of the relevant variables can be found at:

https://osf.io/x8fdu/?view_only=158a0530c389477bb31387686ef7b580.

References

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language, 89*, 429–464. https://doi.org/10.1353/lan.2013.0054

Adelman, J., Brown, G., Quesada, J. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17 (9), 814. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Armstrong, B. C. (2012). *The Temporal dynamics of word comprehension and response selection: Computational and behavioral studies.* (Doctor of Philosophy), Carnegie Mellon University.

Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates.

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition, and Neuroscience*, *1(*7), 1-27. http://dx.doi.org/10.1080/23273798.2016.1171366

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior Research Methods, 44*, 1015–1027. http://dx.doi.org/10.3758/s13428-012-0199-8

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. PloS one, 12(4), e0174623. https://doi.org/10.1371/journal.pone.0174623

Arnon, I. & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*(3), 292–305. https://doi.org/10.1016/j.cognition.2011.10.009

Arppe, A., Hendrix, P., Milin, P., Baayen, R.H., Sering, T., & Shaoul, C. (2015). *ndl: Naive Discriminative Learning. R package version 0.2.17.* http://CRAN.R-project.org/package=ndl

Azuma, T. (1996). Familiarity and relatedness of word meanings: Ratings for 110 homographs. *Behavior Research Methods, Instruments, and Computers*, 28(1), 109–124. https://doi.org/10.3758/BF03203645

Azuma, T. and Van Orden, G. C. (1997). Why SAFE is better than FAST: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language, 36*, 484–504. https://doi.org/10.1006/jmla.1997.2502

Baayen, R.H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, 5.1, 149-157. https://doi.org/10.1075/ml.5.1.06baa

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55*(2), 290–313. https://doi.org/10.1016/j.jml.2006.03.008

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. Language and Speech, 56, 329–347. https://doi.org/10.1177/0023830913484896

Baayen, R. H. & Milin, P (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12-28. https://doi.org/10.21500/20112084.807

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naïve discriminative learning. *Psychological Review, 118*, 438–482. https://doi.org/10.1037/a0023851

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience, 31*(1), 106–128. https://doi.org/10.108/3273798.2015.1065336

Balling, L., & Baayen, R. H. (2012) Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*, 80–106. https://doi.org/10.1016/j.cognition.2012.06.003

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. Available from arXiv:1506.04967 (stat.ME).

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. http://dx.doi.org/10.18637/jss.v067.i01

Batty, M., Morphet, R., Masucci, P., Stanilov, K. (2014). Entropy, complexity, and spatial information. *Journal of Geographical Systems, 16*, 363–385. https://doi.org/10.1007/s10109-014-0202-2

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*, Wiley Series in Probability and Mathematical Statistics. https://doi.org/10.1002/0471725153

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: an MEG study. *Cognitive Brain Research, 24*(1), 57–65. http://dx.doi.org/10.1016/j.cogbrainres.2004.12.006

Berger, W. H., & Parker, F. L. (1970). Diversity of Planktonic Foraminifera in Deep-Sea Sediments. *Science, 168*(3937), 1345–1347. http://dx.doi.org/10.1126/science.168.3937.1345

Bojanowski*, P., Grave*, E., Joulin, A., Mikolov T. (2016). Enriching Word Vectors with Subword Information. arXiv:1607.04606

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. Journal of experimental psychology: Learning, Memory, and Cognition, 22, 63–85. http://dx.doi.org/10.1037/0278-7393.22.1.63

Brocher, A., Foraker, S., & Koenig, J.-P. (2016). Processing of irregular polyemes in sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(11), 1798-1813. https://doi.org/10.1037/xlm0000271

Brocher, A., Koenig, J. P., Mauner, G., & Foraker, S. (2018). About sharing and commitment: The retrieval of biased and balanced irregular polysemes. *Language, Cognition and Neuroscience, 33*, 443–466. https://doi.org/10.1080/23273798.2017.1381748

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition, 1*(1), 9. DOI: http://doi.org/10.5334/joc.10

Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language, 82*(4), 711-733. http://www.jstor.org/stable/4490266

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford University Press.

Bybee, J. (2010). *Language, usage, and cognition*. Cambridge University Press.

Cevoli, B., Watkins, C. & Rastle, K. (2020). What is semantic diversity and why does it facilitate visual word recognition?. Behavior Research Methods https://doi.org/10.3758/s13428-020-01440-1

Chater, N. & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in cognitive sciences, 3*(2), 57-65. https://doi.org/10.1016/s1364-6613(98)01273-x

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons. http://dx.doi.org/10.1002/0471200611

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language, 27*, 429–446. https://doi.org/10.1016/0749-596X(88)90066-6

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: the current state of the literature. *Psychonomic Bulletin & Review, 22*(1), 13–37. http://doi.org/10.3758/s13423-014-0665-7

Faul, F., Erdfelder, E., Lang, AG., & Buchner, A. (2007). *Behavior Research Methods, 39*, 175. https://doi.org/10.3758/BF03193146

Filipović Đurđević, D. (2019). Balance of meaning probabilities in processing of Serbian homonymy. *Primenjena psiihologija, 12*(3), 283-304. DOI: https://doi.org/10.19090/pp.2019.3.283-304

Filipović Đurđević, D., Đurđević, Đ., & Kostić, A. (2009). Vector based semantic analysis reveals absence of competition among related senses. *Psihologija, 42*, 95–106. http://dx.doi.org/10.2298/PSI0901095F

Filipović Đurđević, D. & Gatarić, I. (2018). Simultaneous effects of inflectional paradigms and classes in processing of Serbian verbs. *Psihologija, 51*(3), 259–288. https://doi.org/10.2298/PSI170811015F

Filipović Đurđević, D. i Kostić, A. (2008). The effect of polysemy on processing of Serbian nouns. *Psihologija, 41*(1), 69-86. https://doi.org/10.2298/PSI0801059F

Filipović Đurđević, D., & Kostić, A. (2017). Number, Relative Frequency, Entropy, Redundancy, Familiarity, and Concreteness of Word Senses: Ratings for 150 Serbian Polysemous Nouns. In S. Halupka-Rešetar and S. Martínez-Ferreiro (Eds.) *Studies in Language and Mind 2* (pp.13-77). Filozofski fakultet u Novom Sadu. http://digitalna.ff.uns.ac.rs/sadrzaj/2017/978-86-6065-446-7

Filipović Đurđević, D., & Milin, P. (2019). Information and Learning in Processing Adjective Inflection. *Cortex, 116*, 209-227. https://doi.org/10.1016/j.cortex.2018.07.020

Foraker, S., & Murphy, G. L. (2012). Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language, 67*(4), 407–425. http://dx.doi.org/0.1016/j.jml.2012.07.010

Frank, S. L. (2010). Uncertainty reduction as a measure of cognitive processing effort. *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 81–89). Association for Computational Linguistics.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science, 5*, 475–494. https://doi.org/10.1111/tops.12025

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs multiple senses. *Journal of Memory and Language, 29*(2), 181–200. http://dx.doi.org/10.1016/0749-596X(90)90071-7

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1366–1383. https://doi.org/10.1037/0278-7393.25.6.1366

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gilhooly, K.J. & Logie, R.H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4), 395-427. https://doi.org/10.3758/BF03201693

Goh, W.D., Yap, M.J., Lau, M.C., Ng, M.M.R., & Tan, L-C. (2016). Semantic Richness Effects in Spoken Word Recognition: A Lexical Decision and Semantic Categorization Megastudy. *Frontiers in Psychology*, *7*, 976. https://doi.org/10.3389/fpsyg.2016.00976

Greenwell, B., Boehmke, B., Cunningham, J. & GBM Developers. (2020). *gbm: Generalized Boosted Regression Models. R package version 2.1.8*. https://CRAN.R-project.org/package=gbm

Grindrod, C. M., Garnett, E. O., Malyutina, S., & den Ouden, D. B. (2014). Effects of representational distance between meanings on the neural correlates of semantic ambiguity. *Brain and Language, 139*, 23–35. https://doi.org/10.1016/j.bandl.2014.10.001

Hale, J. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the second conference of the North American chapter of the Association for Computational Linguistics, volume 2*, (pp. 159–166). Association for Computational Linguistics.

Hale, J. (2003). The information conveyed by words. *Journal of Psycholinguistic Research, 32*, 101–123. https://doi.org/10.1023/A:1022492123056

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 643–672. https://doi.org/10.1207/s15516709cog0000_64

Hendrix, P., Bolger, P., & Baayen, H. (2017). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 43(1), 128–149. https://doi.org/10.1037/a0040332

Hino, Y., and Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: an alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1331–1356. http://dx.doi.org/10.1037/0096-1523.22.6.1331

Hino, Y., Lupker, S. J., and Pexman, P. M. (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks: interactions between orthography, phonology, and semantics. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 28*, 686–713. https://doi.org/10.1037/0278-7393.28.4.686

Hino, Y., Pexman, P. M., & Lupker, S. J. (2006). Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding? *Journal of Memory and Language, 55*(2), 247–273. http://dx.doi.org/10.1016/j.jml.2006.04.001

Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language, 114*, October 2020, 104146. https://doi.org/10.1016/j.jml.2020.104146

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45*, 718–730. https://doi.org/10.3758/s13428-012-0278-x

Hoffman, P., & Woollams, A.M., (2015). Opposing Effects of Semantic Diversity in Lexical and Semantic Relatedness Decisions. *Journal of Experimental Psychology: Human Perception and Performance, 41*(2), 385–402. http://dx.doi.org/10.1037/a0038995

Hoppe, D. B., Hendriks, P., Ramscar, M., & van Rij, J. (2020). An Exploration of Error-Driven Learning in Simple Two-Layer Networks From a Discriminative Learning Perspective. https://doi.org/doi.org/10.31234/osf.io/py5kd.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology, 13,* 278–305. https://doi.org/10.1016/0010-0285(81)90011-6

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 66*(2), 115–124. https://doi.org/10.1037/a0026727

Kamin, L. (1969). Predictability, surprise, attention, and conditioning. In R. A. Camp-bell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton-Century-Crofts.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language, 32,* 474–516. https://doi.org/10.1006/jmla.1993.1026

Kemps, R., Wurm, L., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005) Prosodic cues for morphological complexity: Comparatives and agent nouns in Dutch and English. *Language and Cognitive Processes, 20,* 43–73. https://doi.org/10.1080/01690960444000223

Klepousniotou, E. (2002). The processing of lexical ambiguity: homonymy and polysemy in the mental lexicon. *Brain & Language*. 81(1-3), 205-23. https://doi.org/10.1006/brln.2001.2518

Klepousniotou, E., & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics, 20,* 1–24. https://doi.org/10.1016/j.jneuroling.2006.02.001

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language, 123*(1), 11–21. http://dx.doi.org/10.1016/j.bandl.2012.06.007

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1534–1543. https://doi.org/10.1037/a0013012

Kostić Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika*. Beograd: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju.

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

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lichacz, F. M., Herdman, C. M., Lefevre, J.-A., & Baird, B. (1999). Polysemy effects in word naming. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 53*(2), 189–193. https://doi.org/10.1037/h0087309

Lloyd, M. and Ghelardi, R.J. (1964). A Table for Calculating the “Equitability” Component of Species Diversity. *Journal of Animal Ecology, 33*. 217-225. https://doi.org/10.2307/2628

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes, 16*, 565–581. https://doi.org/10.1080/01690960143000137

Lyons, J. (1977). *Semantics: Volume 2*. Cambridge University Press.

MacGregor, L. J., Bouwsema, J., & Klepousniotou, E. (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. *Neuropsychologia, 68*, 126–138. https://doi.org/10.1016/j.neuropsychologia.2015.01.008

Maciejewski, G., & Klepousniotou, E. (2020). Disambiguating the ambiguity disadvantage effect: Behavioral and electrophysiological evidence for semantic competition. *Journal of Experimental Psychology: Learning, Memory, and Cognition. Advance online publication*. https://doi.org/10.1037/xlm0000842

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

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314-324. https://doi.org/10.3758/s13428-011-0168-7

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., and Bates, D. (2017). Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language*, 94, 305 - 315. http://dx.doi.org/10.1016/j.jml.2017.01.001

Marr, D. (1980). *Vision: A Computational Approach*. Freeman & Co.

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. http://dx.doi.org/10.1177/00238309010440030101

MacGregor, L. J., Bouwsema, J., & Klepousniotou, E. (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. *Neuropsychologia, 68*, 126–138. https://doi.org/10.1016/j.neuropsychologia.2015.01.008

Milin, P., Feldman, L.B., Ramscar, M., Hendrix, P., Baayen, R.H. (2017). Discrimination in lexical decision. *PLoS ONE 12*(2): e0171935. https://doi.org/10.1371/journal.pone.0171935

Milin, P., Filipović Đurđević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language, 60,* 50–64. https://doi.org/10.1016/j.jml.2008.08.007

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

Millis, M. L., & Button, S. B. (1989). The effect of polysemy on lexical decision time: now you see it, now you don't. *Memory and Cognition, 17,* 141–147. https://doi.org/10.3758/BF03197064

Mišić, K., & Filipović Đurđević, D. Dynamics of semantic ambiguity effects – predictive power of the SSD account. Submitted.

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. https://doi.org/10.1016/j.cognition.2003.10.015

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition, 197*, 104081. https://doi.org/10.1016/j.cognition.2019.104081

Onifer, W., & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: effects of frequency of meaning and contextual bias. *Memory & Cognition, 9*, 225–236. https://doi.org/10.3758/BF03196957

Pavlov, I. P. (1927). *Conditioned reflexes.* Oxford University Press.

Pexman, P. M., Hino, Y., & Lupker, S. J. (2004). Semantic ambiguity and the process of generating meaning from print. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1252–1270. https://doi.org/10.1037/0278-7393.30.6.1252

Piercey, C. D., & Joordens, S. (2000). Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition, 28*, 657–666. https://doi.org/10.3758/BF03201255

Pylkkänen, L., Llinas, R. & Murphy, G. (2006). Representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, 18(1), 1–13. https://doi.org/10.1006/jmla.2001.2779

Ramscar, M., Dye, M., & Klein, J. (2013) Children value informativity over logic in word learning. *Psychological Science, 24*(6), 1017-1023. https://doi.org/10.1177/0956797612460691

Ramscar, M., & Port, R. (2016) How spoken languages work in the absence of an inventory of discrete units. *Language Sciences, 53*(A), 58-74. https://doi.org/10.1016/j.langsci.2015.08.002

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*, 909 – 957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition, 14*, 191–201. https://doi.org/10.3758/BF03197692

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 779–790. https://doi.org/10.1037/0278-7393.15.5.779

Rescorla, R. A. (1968). Probability of shock in the presence and absence of cs in fear conditioning. *Journal of Comparative and Physiological Psychology, 66*(1), 1–5.

Rescorla, R. A. (1988). Pavlovian Conditioning: It's Not What You Think It Is. *The American psychologist, 43,* 151-60. https://doi.org/10.1037//0003-066x.43.3.151

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

Rice, C. A., Beekhuizen, B., Dubrovsky, V., Stevenson, S., & Armstrong, B. C. (2019). A comparison of homonym meaning frequency estimates derived from movie and television subtitles, free association, and explicit ratings. *Behavior Research Methods, 51*, 1399–1425. https://doi.org/10.3758/s13428-018-1107-7

Rodd, J. M. (2004). The effect of semantic ambiguity on reading aloud: A twist in the tale. *Psychonomic Bulletin & Review, 11*, 440–445. https://doi.org/10.3758/BF03196592

Rodd, J. M. (2018). Lexical ambiguity. In S.-A. Rueschemeyer, & M. G. Gaskell (Eds.) *Oxford handbook of psycholinguistics* (2nd Edition, pp. 120-144). Oxford University Press.

Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science, 15*(2), 411–427. https://doi.org/10.1177/1745691619885860

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *4*6, 245–266. https://doi.org/10.1006/jmla.2001.2810

Rodd, J. M., Gaskell, M. G. & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science, 28*, 89-104. https://doi.org/10.1016/j.cogsci.2003.08.002

Schultz, W. (1998) Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1– 27. https://doi.org/10.1152/jn.1998.80.1.1

Schultz, W., Dayan, P. & Montague, P.R. (1997) A neural substrate for prediction and reward. *Science, 275*, 1593– 1599. https://doi.org/10.1126/science.275.5306.1593

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M. & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology, 14*, 489-537. https://doi.org/10.1016/0010-0285(82)90017-2

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

Shaoul, C., and Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods, 42*, 393–413. https://doi.org/10.3758/BRM.42.2.393

Simpson, G. B., & Burgess, C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 28–39. https://doi.org/10.1037/0096-1523.11.1.28

Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 645–660.

Tabak, W., Schreuder, R., & Baayen, R. H. (2005). Lexical statistics and lexical processing: semantic density, information complexity, sex, and irregularity in Dutch. In M. Reis, & S. Kepser (Eds.), *Linguistic Evidence* (pp. 529–555). Mouton. https://doi.org/10.1515/9783110197549.529

Tomaschek, F. (2020, November 20). *The wizard and the computer: An introduction to preprocessing corpora using R*. https://doi.org/10.31234/osf.io/jsv38

Tomaschek, F., Plag, I., Ernestus, M., and Baayen, R. H. (2021). Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics, 57*(1), 123-161. https://doi.org/10.1017/S0022226719000203

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition, 22*, 111–126. https://doi.org/10.3758/BF03202766

Vitello, S., Warren, J. E., Devlin, J. T., & Rodd, J. M. (2014). Roles of frontal and temporal regions in reinterpreting semantically ambiguous sentences. *Frontiers in Human Neuroscience, 8*, 1-14. https://doi.org/10.3389/fnhum.2014.00530

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer.

Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, R. H. (2006) Dynamics of the Auditory Comprehension of Prefixed Words: Cohort Entropies and Conditional Root Uniqueness Points. *The Mental Lexicon, 1*, 125–146. https://doi.org/10.1075/ml.1.1.08wur