Semantic richness effects in isolated spoken word recognition: Evidence from Massive Auditory Lexical Decision

Filip Nenadić[1,2], Ryan G. Podlubny[2], Daniel Schmidtke[3], Matthew C. Kelley[4], and Benjamin V. Tucker[2]

[1]Department of Psychology, Singidunum University, Serbia

[2]Department of Linguistics, University of Alberta, Canada

[3]Department of Linguistics and Languages, McMaster University, Canada

[4]Department of Linguistics, University of Washington, USA

Author Note

- Corresponding author: Filip Nenadić, Reljkovićeva 67, 21131 Petrovaradin, Serbia.

- This project was supported in part by the Social Sciences and Humanities Research Council of Canada [Insight Grant 435-2014-0678] and a grant from the Kule Institute for Advanced Studies both to the last author.

- An earlier version of this analysis was reported at the Acoustics Week in Canada 2019 (https://jcaa.caa-aca.ca/index.php/jcaa/article/view/3335).

- The data and the analyses scripts are available at: https://doi.org/10.7939/r3-7ejm-yf30.

- This study was not preregistered.

Abstract

While known to influence visual lexical processing, the semantic information we associate with words has recently been found to influence auditory lexical processing as well. The present work explores the influence of *semantic richness* in auditory lexical decision. Study 1 recreates an experiment investigating semantic richness effects in concrete nouns (Goh et al., 2016, FRONT PSYCHOL 7). In Study 2 we expand the stimulus set from 442 to 8,626 items, exploring the robustness of effects observed in Study 1 against a larger dataset with increased diversity in both word class and other characteristics of interest. We also utilize generalized additive mixed models to investigate potential non-linear effects. Results indicate that semantic richness effects become more nuanced and detectable when a wider set of items belonging to different parts of speech are examined. Findings are discussed in the context of models of spoken word recognition.

Keywords: semantic richness, spoken word recognition, auditory lexical decision

Semantic richness effects in isolated spoken word recognition: Evidence from Massive Auditory Lexical Decision

When considering the task of spoken word recognition, the breadth of factors which may influence a listener's performance has been drawing increased attention. Specifically, the importance of so-called *semantic richness* has become increasingly apparent. Semantic richness refers to the qualities associated with a word or concept that contribute to the word's overall meaning. Higher degrees of semantic richness have consistently been connected to facilitation of processing in visual word recognition (see, e.g., L. Buchanan et al., 2001; Kuperman et al., 2014; Pexman et al., 2002; Schwanenflugel & Shoben, 1983; Yap et al., 2015; Yap et al., 2012; Yap et al., 2011). That is to say, items with higher semantic richness are, for example, typically recognized more quickly during reading and picture-based tasks than those with lower levels of semantic richness. The foci of this paper are the relatively less-investigated semantic richness effects in spoken word recognition.

Importantly, Pexman et al. (2013) explain that semantic richness is not a unitary construct, but instead the product of a large number of various measures combined. Variable degrees of word concreteness, imageability, semantic feature network size or complexity, semantic neighborhood size or density, and others all come together as contributors to a word's greater semantic richness. This type of richness, therefore, involves the combination of any number of related semantic contributors, where these specific operationalizations are known to impact participant performance in various behavioral tasks.

When considering the influence of semantic richness in the visual modality specifically, semantic richness variables are known to predict participant performance when words are presented in isolation (i.e., extremely limited context). One of the earliest and best established findings in this area involves words with higher degrees of concreteness being responded to more quickly than relatively abstract words (see, e.g., Schwanenflugel, 2013; Schwanenflugel et al., 1988; Schwanenflugel & Shoben, 1983). Another example can

be found through a measure developed by McRae et al. (2005) which expresses the number of semantic features a word has. This measure is described as a good predictor of visual lexical decision and word naming (Pexman et al., 2002), and later as facilitating self-paced reading and semantic categorization (Pexman et al., 2003). Beyond word concreteness and number of semantic features, other factors of semantic richness have also been observed to predict participant behavior, such as number of semantic neighbors and contextual dispersion (Pexman et al., 2008), semantic neighborhood density (L. Buchanan et al., 2001), arousal, which is a measure expressing the extent to which a word is calming or exciting, and valence, which is a measure expressing the extent to which a word is negative or positive (Kuperman et al., 2014).

Semantic richness effects are found in a range of tasks and with a range of measures. For example, Yap et al. (2011) find that higher word ambiguity, expressed as the number of senses in the WordNet database (Miller et al., 1990), leads to faster visual lexical decision, but slower semantic categorization of a given word. Effects driven by semantic richness have even been noted to aid bilinguals' word learning (Kaushanskaya & Rechtzigel, 2012), for decreasing the tip-of-the-tongue effect (Gianico-Relyea & Altarriba, 2012), and to facilitate word remembering and retrieval (Hargreaves et al., 2012). An overview of the effects related to various semantic richness variables in multiple behavioral tasks is provided by Yap et al. (2012); however, semantic richness effects are not restricted to responses in behavioral experiments. Semantic richness effects have also been recorded through other online measures such as event-related potentials and functional magnetic resonance imaging (e.g., Barber et al., 2013; Pexman et al., 2007; Rabovsky et al., 2012; Taler et al., 2013; Van Petten, 2014).

Thus, it is understood that semantic richness effects are reliable and well-established in their contribution to *visual* word processing. Indeed, similar effects have been recognized in a wide variety of visual behavioral tasks and using a large number of different online and offline measures. Very few studies, however, explore the role of semantic richness in *spoken*

word recognition (Goh et al., 2016). Through these works, while limited, there is some evidence to support an influence of semantic richness in spoken word recognition as well. Therefore, in the overview to follow, we explore semantic richness as captured in the auditory modality, focusing on the auditory lexical decision or similar tasks, in order to provide comparable context.

Semantic richness in the auditory modality

When comparing visual and auditory lexical decision, certain unavoidable differences across modalities are likely to influence participant performance. We highlight the most notable differences here. (1) Time—In the visual modality stimuli are typically presently wholly and instantaneously, where in auditory presentation signals must unfold over time. (2) Variability—Acknowledging minor differences that may come with font changes, orthographic characters in most languages are relatively consistent in reproduction. In the auditory modality, it has been argued that multiple productions of an intended lexical item are never acoustically identical, even when produced by the same speaker (e.g., Podlubny et al., 2015). Indeed, Ladefoged (1990) argued that the logical extreme of the anthropophonic perspective in Lindblom (1990) is that all possible speech articulations should be considered their own distinct speech sound categories. (3) Access to the signal—in a situation where one misses any portion of a stimulus in a reading task (i.e., the visual modality), this missed information remains available to the subject because graphemes do not typically disappear as a reader moves through the material (but see Rayner et al., 1981, for an example of a task designed specifically to do so). However, in auditory presentation, any missed information is immediately no longer available to the listener as a signal progresses.

Additionally, certain findings are known not to replicate across modalities. One clear example of cross-modal differences is available through phonological neighborhood density (PND). For English words, higher PND seems to reliably facilitate visual word

recognition, but inhibit spoken word recognition Tucker et al., 2019. Another example can be seen in effects related to phonological uniqueness point, as Ferrand et al. (2018) note an opposite effect in the visual versus the auditory modality (for an extended discussion and results of differences across modalities, see Ferrand et al., 2018; Tucker et al., 2019). Considering known cross-modal differences and in light of contrasting experiences that language users are likely to have with written vs. spoken words, it seems possible that semantic characteristics of words may exert differing influences in word processing across the two modalities (see also Westbury & Moroschan, 2009).

Some of the earliest studies investigating the effects of semantic characteristics of words in the auditory modality were performed by Wurm and colleagues (Wurm, 1996; Wurm et al., 2004). The authors found significant effects for the connotative dimensions of evaluation, activity, and potency on spoken word recognition (see Osgood & Suci, 1955; Osgood et al., 1957, and later studies citing these two sources for a detailed explanation of these dimensions). Elsewhere, Wurm and Vakoch (2000) developed dimensions of noun danger and usefulness and tested them in an auditory lexical decision study. The authors reported an interaction where increased word danger values were correlated with reduced response latency for low usefulness words, but increased response latency for high usefulness words. If a word was associated with a low danger value, high usefulness words were responded to much faster than were low usefulness words. Similar findings were also recorded in later studies while registering an effect of word animacy as well (Wurm et al., 2007), or by using event-related potentials to measure the effects of danger and usefulness (Kryuchkova et al., 2012).

Tyler et al. (2000) investigate how word imageability (a term the authors use interchangeably with concreteness) affects spoken word recognition, again in an auditory lexical decision task, and find that participants are significantly faster when responding to high-imageability words. Comparable results were found in a later fMRI study from Zhuang et al. (2011). Sajin and Connine (2014) showed that words with a higher number

of semantic features are responded to more quickly than those with fewer semantic features for more information about semantic features see McRae et al., 2005, also described in more detail below; similar results were found by Devereux et al. (2016). Notably, the influence of number of semantic features on spoken word recognition appears to be even stronger when noise is added to the signal (Sajin & Connine, 2014) or when the speech is accented (Sajin & Connine, 2017), where listeners are forced to increasingly rely on prior experience and less on acoustic information within the signal itself.

Goh et al. (2016) provide likely the most widely encompassing investigation of semantic richness effects and their influence on isolated spoken word recognition, which is the study most closely related to the present work. This work employed an auditory lexical decision task and a semantic categorization study to explore the effects of six semantic richness variables (alongside other lexical control predictors). The authors selected 468 nouns embodying high levels of concreteness as stimuli. A hierarchical linear regression analysis of the lexical decision data showed that higher concreteness (within the small range of concreteness captured in the stimuli), valence, and number of semantic features were all associated with shorter response latencies. Additionally, a quadratic effect of valence was recorded, indicating that extremely positive and negative words are associated with faster responses, and more neutral words exhibit less of this effect. Arousal, semantic neighborhood density, and semantic diversity were not significant predictors of response latency. Barring some differences in effect sizes, results from the semantic categorization task were generally in line with those of the lexical decision task; however, one minor exception involves number of morphemes (a lexical control predictor), the influence of which was significant in the auditory lexical decision but not in the semantic categorization task.

Semantic richness and models of spoken word recognition

Despite mounting evidence supporting a role of semantic characteristics in spoken word recognition, models of spoken word recognition for the most part do not account for it. Instead, the process of lexical access in models of spoken word recognition most often begins with the processing of (pseudo)acoustic features or phonemes and ends with detection of the correct form (unit/word) in the mental lexicon, while effectively eschewing any meaning activation (see also Gaskell & Marslen-Wilson, 2002, for additional critique). The mental lexicon is then usually represented as a semantically unconnected list of words—words are often strings of phonemes, related to each other by form only as in, e.g., Hannagan et al., 2013; Luce, 1986; Luce et al., 2000; Luce and Pisoni, 1998; Marslen-Wilson, 1987; Marslen-Wilson and Tyler, 1980; McClelland and Elman, 1986; Norris, 1994; Norris and McQueen, 2008; ten Bosch et al., 2015; You and Magnuson, 2018.

Most models of spoken word recognition were therefore not necessarily conceived with semantics in mind. However, certain models can potentially provide a means to explain and simulate the role of semantics in the process of spoken word recognition with minor modification. For example, TRACE (McClelland & Elman, 1986) and DIANA (Ten Bosch et al., 2022; ten Bosch et al., 2015) both posit a feedback mechanism designed to account for word frequency effects. Such a mechanism could be employed to account for the effects of other variables as well, including those related to a word's semantic richness. Although the mental lexicon as employed in such models would remain a list of items with numbers assigned to them (e.g., frequency or semantic weights), the effects of word characteristics related to meaning and/or the meaning contexts in which they occur could be captured and simulated.

In comparison, Shortlist B (Norris & McQueen, 2008) does not include such a feedback mechanism, though the prior probability of a word can also be affected by semantic information in this model. Shortlist B is a Bayesian model that calculates probabilities of outcomes as the signal unfolds, and expectations (expressed through prior

probabilities that are governed by word frequency in the default model setup) of these outcomes may affect model estimates and final results. The authors recognize that the prior probability of a word could be influenced by factors such as contextual, syntactic, or semantic characteristics of words or groups of words, in addition to word frequency. Therefore, these influences, as in TRACE or DIANA, could be represented by a single number describing their collective effect on prior word probability; though, investigation of such influence was outside the scope of original Shortlist B simulations. To the best of our knowledge, simulations investigating semantic effects were not performed using either TRACE or DIANA.

While the above-mentioned approaches could account for the role of semantic information, we note that conflating various streams of information as a single score would make it difficult to investigate their individual contributions. Certain model extensions would be required to allow for exploring separate effects of frequency and various semantic or syntactic factors within these models. Furthermore, even though this method would likely provide a useful predictor in modelling, any such measure could not capture the interrelated structure and semantic relations between items. Thus, any representation of the mental lexicon would in itself remain a list of unconnected items.

Three models of spoken word recognition, however, explicitly recognize the importance of semantic characteristics of words and attempt to model their effects. The Distributed Cohort Model (DCM; Gaskell & Marslen-Wilson, 1997, 2002) incorporates a direct mapping of input to both form and meaning. Initially, each word in the DCM's mental lexicon was represented by values on a number of binary semantic feature vectors. Eleven arbitrary semantic vectors were used as a proof of concept, as it is impossible to generate an exhaustive list of semantic features. For the same reason, semantic feature vectors were later replaced by vectors based on word co-occurrence statistics (Gaskell & Marslen–Wilson, 1999). The Discriminative Lexicon (Baayen et al., 2019) also includes a direct connection between acoustic and semantic information. This approach to

representing the mental (discriminative) lexicon is similar to the latter implementation of the DCM; in this context, (Baayen et al., 2019) suggest a matrix of semantic feature vectors based on frequency of co-occurrence with other units in a language corpus. Lastly, the recently developed EARSHOT model (Magnuson et al., 2020) uses random, sparse pseudo-semantic feature vectors. This setup is reminiscent of the early versions of the DCM, although the number of pseudo-semantic vectors is increased to 300.

While these three models are examples of models that improve upon other frameworks and include semantic relationship between words, it is unclear at present whether or not these approaches can simulate the effects of more subjective, rating-based semantic characteristics (e.g., word valence or arousal). One might extract clusters of words with similar semantic vectors aiming to describe the semantic neighborhood density of a word, or words which may share a similar meaning context or syntactic environment. In none of these frameworks, however, do we see a ready-made way to estimate other important semantic characteristics of a word from such groupings—such as whether it induces positive or negative emotions—without employing a human interpreter or some other pre-existing, external ratings. Although there is no reason to believe that the information we collect from human ratings is somehow stored differently in the mental lexicon in comparison to measures extracted from language corpora, this technical limitation would still make these models of spoken word recognition unable to include rating-based measures and therefore simulate their effects captured in behavioral tasks without some extension. Note, however, that the more recent advances in machine learning enable estimation or approximation of semantic characteristics that have historically been assessed through human ratings by means of unsupervised sentiment analysis (Hu et al., 2013) or by analyzing word co-occurrence (Hofmann et al., 2018; Hollis & Westbury, 2016; Hollis et al., 2017). We return to this possibility in the General Discussion. On the other hand, models that use semantic feature vectors, e.g., the original structure of the DCM and EARSHOT, could theoretically incorporate features such as “is dangerous” or “can be

touched". But, as we have seen above, implementing this architecture with actual semantic features has proven to be very difficult.

The present work

Given the recently noted replicability crisis (Open Science Collaboration, 2015) and the limited number of studies exploring the effects of semantic richness in the auditory modality, we first attempt to replicate the auditory lexical decision experiment and analytical methods described by Goh et al. (2016). Open material archives allow researchers to test the robustness of findings described in other studies; in Study 1 we use stimuli recorded as part of the Massive Auditory Lexical Decision project (MALD; Tucker et al., 2019) to build our similarly targeted auditory lexical decision task. In addition to using hierarchical linear regression as described by Goh et al. (2016) to analyze the data, we also include generalized additive mixed effects modelling (GAMMs; Wood, 2011) to control for participant random effects and, more importantly, investigate the potential for non-linear predictor effects. These models and their application to language data are described in detail in Baayen and Linke (2020) and Wieling (2018) and have been previously employed in various psycholinguistic studies (e.g., Kryuchkova et al., 2012; Lõo et al., 2018; Milin, Divjak, et al., 2017; Milin, Feldman, et al., 2017; Porretta et al., 2016; Wieling et al., 2016).

Goh et al. (2016) have significantly expanded our knowledge of semantic richness effects in comparison to previous works exploring spoken word recognition. However, the stimuli used by Goh et al. are restricted to 468 concrete nouns, where we note two important limitations—word-set size and homogeneity—that can be addressed through use of existing open archives. The MALD project (Tucker et al., 2019) provides access to both a massive stimuli bank of approximately 26,000 spoken words of different lexical and semantic characteristics and word types as well as a large number of participant responses to those stimuli in a lexical decision task. In Study 1 of the present work, we extract the

same item list described by Goh et al. in order to replicate their work. However, in contrast to the relatively small selection of stimuli described in Goh et al., in Study 2 we explore responses to a much larger set of 8,626 words belonging to the noun, verb, and adjective word classes. This expanded stimuli set allows us to address the issues related to the restrictions on semantic effects imposed by analyzing responses to only concrete nouns. That is to say, semantic richness effects seem to differ given information related to word class or other lexical characteristics (see Recchia & Jones, 2012; Sidhu et al., 2016; Tyler et al., 2002; Zdrazilova & Pexman, 2013). Through responses to a more diverse word-set we might observe whether or not the same patterns of semantic richness persist within the expanded word-set, and if there may be differences in semantic richness effects that vary as a function of word class or other word characteristics. Given the somewhat exploratory nature of our MALD data analysis, we refrain from making explicit hypotheses about the effects that semantic richness variables may exhibit.

We previously noted that models of spoken word recognition largely do not explicitly account for semantic effects in their architecture. And while some models do aim to account for semantic contributions, models that develop their mental lexicon as more than a list of unconnected items (e.g., Distributed Cohort Model, the Discriminative Lexicon approach, EARSHOT) still may have issues in accounting for certain semantic richness effects such as a word's valence, arousal, or danger. One goal of the present work is therefore to describe in more detail the effects of various semantic richness variables on isolated spoken word recognition. These findings can be used to inform and improve upon current models of spoken word recognition, specifically regarding the contributions of particular semantic variables. We also provide suggestions for how these effects may be incorporated within a computational model of (isolated) spoken word recognition.

The remainder of the paper is structured as follows. We first present the semantic richness variables used in our Studies 1 and 2 in more detail. These descriptions include brief definitions for each variable, as well as the sources from which they were obtained.

We then describe Study 1, which aims to replicate the findings of Goh et al. (2016). Additionally, we present our expanded data analysis for Study 1 which employs GAMMs to test for non-linear effects while controlling for random effects of subject and stimulus. We then proceed to Study 2, which uses GAMMs to analyze MALD data. Finally, the two studies are connected through a joint, general discussion. In this section we consider replicability of the semantic richness effects described by Goh et al. (2016), as well as their stability when a more diverse set of words is introduced (diverse in terms of word class and other word characteristics). We draw comparisons between the effects of semantic richness variables in the auditory versus visual modalities, and end the paper with suggestions for improving (computational) models of spoken word recognition.

Semantic richness variables used in the present work

Studies 1 and 2 both consider the same semantic richness variables analyzed by Goh et al. (2016), retrieving measures for analysis from the same sources. We present and describe these variables in two subcategories, classifying them as *rating-based measures* and *corpus-based measures* of semantic richness. The two categories draw attention to how each measure was obtained, that is, the methods applied to calculate each measure.

Rating-based measures

Rating-based measures of semantic richness are obtained as subjective ratings from a large number of raters. The first such measure we describe is number of semantic features; these values were compiled by McRae et al. (2005) and represent the number of distinct features assigned to a given word/concept. McRae et al. describe "features" in this context as denoting properties (physical, functional, taxonomic, etc.) that participants associate with each word/concept. For example, the word *knife* was often remarked to be "used for cutting" (25 participants), while some participants also noted that a knife "is shiny" (5) or "is a utensil" (9). Specifically, this measure refers to the number of features

assigned to each item by no fewer than 5 of the 30 participants in the McRae et al. study. Values for the number of semantic features range from 6 (e.g., *hut*) to 26 (e.g., *fawn*).

We obtained values for valence and arousal from Warriner et al. (2013), who collected them using the Amazon Mechanical Turk platform. Participants would rate each word or concept on a single dimension only, using a scale ranging from 1-9. Valence describes characteristics much like polarity or degrees of positivity, where mean values per word range from 1.26 (e.g., *torture* and *murder*) to 8.53 (e.g., *vacation* and *happiness*). The medium values are reserved for words with neutral valence (e.g., *cap* or *depend*). Arousal, on the other hand, refers to the extent to which a word/concept elicits an emotional or physiological reaction. Mean values per word for arousal within the data range from 1.60 (e.g., *grain*, *dull*, and *librarian*) to 7.79 (e.g., *insanity*, *sex*, and *gun*).

The final rating-based semantic richness variable we consider is concreteness, extracted from Brysbaert et al. (2014). Concreteness distinguishes concrete from abstract concepts—or, as described by Brysbaert et al., "the degree to which the concept denoted by a word refers to a perceptible entity" (p. 904). These ratings were also collected through Amazon Mechanical Turk, where participants were tasked with rating the degree of concreteness using a 5 point scale. Mean ratings per item range from 1.04 (e.g., *essentialness*) to 5 (e.g., *armchair* or *strawberry*).

Corpus-based measures

The corpus-based measures we consider were obtained as descriptions of the context in which words occur in large corpora documenting (written) use of English. The first such measure we explore is semantic neighborhood density (SND). Shaoul and Westbury (2010) express SND through the Average Radius of Co-occurrence (ARC) measure, which effectively captures the mean distance between a word and its semantic neighbours within a specified threshold (of distance). In other words, this measure aims to quantify semantic relatedness between a word or concept and its closely related "neighbors". Shaoul and

Westbury (2013) generated these estimates through the USENET corpus from years 2005 to 2009. The values of semantic neighborhood density range from 0.09 (e.g., *rakishly*) to 0.90. (*to* and *and* being good examples of words with high semantic neighborhood density).

We also used the semantic diversity data from Hoffman et al. (2013). Semantic diversity was calculated based on an extensive latent semantic analysis (Landauer & Dumais, 1997) of the British National Corpus (British National Corpus Consortium, 2007). This measure is meant to represent the variability in the contextual usage of a word. For example, certain words are only found in more limited semantic contexts (e.g., *garnish* with semantic diversity value of 0.08) while others are found in many different contexts (e.g., *other* or *also*, with the highest recorded semantic diversity being 2.41).

Study 1

The goal of Study 1 is to test whether or not the findings from Goh et al. (2016) can be replicated. We used stimuli from the MALD project that overlapped with the original Goh et al. (2016) stimuli to build a comparable, targeted experiment and test a new group of participants. The data were analyzed using the same statistical analysis employed in the original study (an earlier analysis of this data using hierarchical linear regression was already briefly presented in Tucker et al., 2019). In addition, we analyze the same data using GAMMs in order to control for participant random effects and explore potential non-linear predictor effects.

Method

Participants

The participant sample consisted of 24 native monolingual listeners of English (13 female and 11 male, age ranging from 18 to 31, $M = 20.88$, $SD = 2.85$). The participants were considered to be monolingual using the criterion from the MALD1 database: these participants were not exposed to any language besides English as they grew up in the first

five years of their life. All participants were students at the University of Alberta and received partial course credit for participation in the experiment. All participants provided informed consent prior to participating in the study. A brief hearing screening was performed using pure tone signals of 500, 1000, 2000, and 4000 Hz administered in turn at 20 dB. Ears were tested individually and participants were asked to acknowledge when and if they could hear each signal. One participant showed signs of limited hearing in the right ear at 500 Hz, but their experimental responses were on par with other participants and therefore retained.

Stimuli

Our stimuli involved 442 of the 468 nouns described in Goh et al. (2016); the remaining 26 items were not available in the MALD stimuli. Responses to an additional set of 442 randomly selected MALD pseudowords were also tested. The MALD stimuli set includes recordings of one 31 year old male, native speaker of western Canadian English. During recording words were presented to the speaker in standard English spelling, while pseudowords were presented using the International Phonetic Alphabet (the speaker was experienced and well-practiced with this alternative writing system). While reading, the speaker was instructed to produce words and pseudowords as naturally as possible; additional detail regarding stimuli and recording procedures is available in Tucker et al. (2019).

Procedure

Participants sat in a sound attenuated booth and stimuli were presented over MB Quart QP 805 DEMO headphones using E-Prime (Schneider et al., 2002). Participants used their dominant hand to respond “word”, and their non-dominant hand to respond “not a word” by pressing the appropriate button on a button-box. Each stimulus was preceded by a 500 ms fixation cross, and responses could be made during the stimulus presentation. If no response was made within 3000 ms, the experiment would automatically proceed to

the next fixation cross and stimulus. No stimuli were presented as practice and no feedback was provided during the experiment. Each session lasted approximately 25 minutes.

Results

Data analysis was performed in the R statistical environment (R Core Team, 2018). As in Goh et al. (2016), response latency measured from word onset served as the response variable in all models (note that response latency calculated from word offset correlated highly with response latency calculated from word onset, $r = .89$). Our initial analysis followed the data preparation procedure described by Goh et al. (2016), recreating the same item-level hierarchical linear regression. We excluded all items with an accuracy rate lower than 70%, all responses shorter than 200 ms, and all responses that were more than 2.5 SD away from each participant's mean response latency. We also excluded all items for which any of the predictor variable values were missing, and all responses to one noun which had an outlying (low) concreteness value. The exclusion process (when considering correct responses only) reduced the dataset from 10,608 responses by 22.19% to the final number of 8,254 retained responses. A similar percent of data was excluded by Goh et al. (2016), as they analyzed responses to 357 out of 468 words presented in their experiment. Although we were aware that responses to many words will have to be excluded because semantic richness variables were not available for them, we decided to use all of the words available in MALD that appeared in the Goh et al. (2016) study. This was done in order to follow their method as closely as possible and construct an experimental session that is similar in length and content. Lastly, we calculated the mean standardized RT per item, considering only correct responses.

In a second analysis, we modelled the same data using generalized additive mixed models in order to account for potential non-linear effects and random effects utilizing the *mgcv* (Wood, 2011) and *itsadug* (van Rij et al., 2020a) packages. Following the guidelines from Baayen and Milin (2010), we used the -1000/RT transform on the response time data.

This transform brings the residuals of the fitted models into a more normal distribution, as is needed in Gaussian-family regression models. Note that the negation of the values is needed so as to maintain the relationship between long RT values and high numbers, which was reversed when the reciprocal function was applied. Other data preparation procedures, barring standardizing and averaging response latency across participants, were kept identical.

Note that in the interest of space and clarity, we do not fully describe every step in the data preparation or analysis procedures, nor present every considered model; detailed comments and R code are available in an online archive (Nenadić et al., 2019).

Analysis 1 – hierarchical linear regression

Our exclusion criterion removed only three additional outlying responses. In total, we retained 8,254 responses to 363 different nouns – meaning 6 more items than in the original study. The difference is likely due to higher participant response accuracy, as nouns with an error rate of 30% or higher were excluded from the study. Overall error rate for words was 11% in the original study and only 3% in our replication experiment. However, the majority of excluded nouns were rejected because they did not appear in the databases of the various semantic richness variables used.

The first regression model included the following lexical control variables: word duration in milliseconds, logged word frequency as extracted from the SUBTLEX-US corpus (Brysbaert & New, 2009), phonological uniqueness point as extracted from the CMU dictionary (Weide, 2005), number of morphemes, and a structural principal component which combines measures of length and word-form similarity (number of phonemes, number of syllables, phonological Levenshtein distance, and negatively loaded phonological neighbourhood density). The structural principal component was calculated in the same manner as it was calculated in the analysis conducted by Goh et al. (2016). The model significantly predicted item-level standardized reaction times ($r^2 = .39$,

$F(5, 357) = 46.34$, $p < .001$). The effect of duration was inhibitory, the effects of frequency and the structural principal component were facilitative, and phonological uniqueness point and the number of morphemes did not significantly contribute to the model fit.

In the second step the model was expanded to include semantic richness variables (Table 1). The model significantly predicted item-level standardized reaction times ($r^2 = .44$, $F(11, 351) = 27.27$, $p < .001$) and was also significantly better than the first model ($F(6) = 7.29$, $p < .001$). The effects of lexical control predictors from model 1 remained the same. The effects of concreteness, number of semantic features, and valence were facilitative, whereas arousal, semantic neighbourhood density, and semantic diversity did not significantly improve model fit.

INSERT TABLE 1 APPROXIMATELY HERE

The third regression model introduced squared valence as a predictor and the fourth introduced interactions of arousal with valence and squared valence. Neither of these steps significantly improved model fit. The results of our hierarchical linear regression align almost completely with those described by Goh et al. (2016); we return to these results in the Study 1 Discussion section, as we compare them to the following analysis using generalized additive mixed effects models.

Analysis 2 – GAMM

The generalized additive mixed effects model included all predictor variables tested in step 4 of the hierarchical linear regression models described above while also introducing random effects for stimulus and participant, and random slopes of trial per participant. Note that squared valence was no longer required as GAMMs can deal well with non-linear effects, making this term redundant. The interaction between arousal and valence was included as a tinsor effect (since the main effects of arousal and valence are already present in the model, a tinsor produces just a tensor product interaction without these main effects).

The results of the initial, full model showed that phonological uniqueness point, number of morphemes, arousal, semantic neighborhood density, semantic diversity, and the tinsor (interaction) between valence and arousal were not significant predictors of response latency. These predictors were excluded in a backward-stepwise model fitting procedure, in order of least significance. After every exclusion, the `compareML()` function from the *itsadug* package was used to confirm that the new, simpler model is not worse than its more complex predecessor model. Qualitatively, the results did not change after the exclusions; significant effects remained significant, and nonsignificant effects remained nonsignificant until excluded.

The model including only the significant effects (presented in Table 2) is fully in line with our previously presented hierarchical linear regression models (specifically, step 2) and, by extension, with the results from Goh et al. (2016). The effects of all significant predictors, excluding trial number, were linear (as indicated by "edf" values which are 1 or close to 1), even though they were included as potential non-linear, smooth terms. Inspection of effects plots (Figure 1) indicated that, despite being non-linear, the effect of trial number was facilitative, especially for the first approximately 100 trials. The linear effect of duration was inhibitory, whereas the linear effects of frequency and the structural principal component were again related to shorter response latency. As in our previous analysis, concreteness, number of semantic features, and valence were found to be related to shorter response latency. Importantly, the effect of valence remained linear.

INSERT TABLE 2 APPROXIMATELY HERE

INSERT FIGURE 1 APPROXIMATELY HERE

However, as well as using significance and visual inspection of effects plots to determine whether a predictor should be retained in a model, we can also consider whether significant predictors actually improve model fit by comparing model fit of nested models (see van Rij et al., 2020b). We again excluded predictors one by one (in no particular order) and used the `compareML()` function to test whether the new model is significantly

worse than the preceding. In cases when there was no significant difference between two models, we selected the new, simpler model. Table 3 presents the model that includes only the predictors whose exclusion would significantly reduce model fit and these are duration, frequency, structural principal component, and concreteness. As can be seen in the effects plots in the additional material available in an online archive (Nenadić et al., 2019), the effects of these variables remain roughly the same as in the previous model (Figure 1), which included all significant effects.

INSERT TABLE 3 APPROXIMATELY HERE

Discussion

In this study we successfully replicate the findings from Goh et al. (2016), which indicate that higher values of semantic richness variables facilitate auditory word recognition. Specifically, effects related to the lexical control variables are entirely in line with those regularly obtained in auditory lexical decision studies see, e.g., Ferrand et al., 2018; Goh et al., 2016; Tucker et al., 2019: longer word duration is positively correlated, whereas higher frequency, trial number (recognized through the GAMM analysis), and the structural principal component are negatively correlated with response latency. Although it also comprises the number of phonemes and the number of syllables a word has, the effect of the structural principal component is facilitative because it represents the phonological distinctiveness of a word. Words with more phonemes and syllables have fewer similarly-sounding words and the structural principal component also includes phonological neighborhood density and phonological Levenshtein distance. Moreover, the effect of this predictor is facilitative primarily because the effect of duration is taken into account in the statistical models; of two words that have the same duration in ms, the participants will more quickly recognize the one with more phonemes/syllables as more segmental information has been passed in the same amount of time and the word with more phonemes/syllables likely has fewer neighbors. Excluding the contentious status of the

effect of valence on response latency, effects related to the remaining semantic richness variables tested within this work are also in line with previous research: we find that higher values of concreteness (see also Tyler et al., 2000) and number of semantic features (see also Sajin & Connine, 2014) facilitate processing in our data as well. Like Goh et al. (2016), we also find that arousal, semantic neighbourhood density, and semantic diversity do not reliably affect participant speed.

However, there are two differences in our results. First, unlike Goh et al. (2016), we did not find the number of morphemes comprising a word to be predictive of response latency in our data. We note that number of morphemes was not a significant predictor in the semantic categorization task also described in that work. This difference might suggest the influence of number of morphemes on response latency is fragile or at least less reliable. Moreover, given that nearly all of the nouns used were monomorphemic, we argue that further analyses of this variable's influence is required while incorporating a balanced set of monomorphemic and polymorphemic words. Second, we do not find a quadratic, non-linear effect of valence—even when using GAMMs. Instead, we find support for the linear effect of valence reported by Kuperman et al. (2014). Kousta et al. (2009) noted that asymmetrical (or linear) effects of valence (also recorded by, e.g., Estes & Verges, 2008) may be the consequence of poor control of other relevant lexical predictors, or otherwise an issue related to a biased sampling of word-stimuli. Yet our study used the same word stimuli as Goh et al. (2016) and we still only found that negatively-charged words are more often associated with longer response latency. One possibility is that the quadratic (non-linear) effect of valence may be volatile, much like the effect of number of morphemes. An alternate source for this discrepancy could lie in two often neglected aspects of a lexical decision experiment: its pseudowords and participants, as the pseudowords used and participants tested were the only important differences between our study and the lexical decision experiment reported by Goh et al. (2016).

Additionally, our (likely more strict) method of testing the influence of a predictor

upon response latency showed that considering trial number and, more importantly, valence and number of semantic features does not significantly improve model fit. In other words, the impact these variables have is small or unreliable enough to not warrant inclusion if model simplicity (parsimony) is favored over a marginal increase in predictiveness. This would leave only word concreteness as a semantic richness variable potent enough to be included alongside the larger effects attributed to frequency, combined measure of length and word-form similarity, and especially recording duration.

We refrain from a more in-depth discussion of the implications of our findings for two main reasons. First, we recognize that the results discussed above are based on experiments utilizing limited stimuli sets. The words that participants heard in our study and in Goh et al. (2016) were limited to nouns exclusively—which were further characterised by, e.g., high concreteness values, low arousal values, and being monomorphemic. We note that this type of work would likely benefit from expanded stimuli sets. Specifically, stimuli embodying a wider range of semantic richness and incorporating other parts of speech would be valuable, as limitation in the by-item range of semantic richness and the parts of speech explored may influence the effects registered through experimentation. Secondly, our study had a smaller number of participants relative to Goh et al. (2016), which may have impacted our experimental power (Brysbaert, 2019). Though we replicate many of the effects found by Goh et al. (2016), our study is based on a sub-set from a larger independent mega-study, and thus an absolutely faithful replication was not the original goal of our experiment. We therefore turn next to the full scale MALD project.

Study 2

In Study 1 we explored the roles of numerous semantic richness variables in spoken word recognition, providing additional support for their importance in auditory lexical decision tasks. However, work to date is typically restricted to investigation of only one or

two semantic richness variables at a time, explores relatively small stimuli sets, and/or investigates stimuli restricted to only one word class (e.g., Goh et al., 2016; Sajin & Connine, 2014; Tyler et al., 2000). Moreover, analyses in such works are generally limited to simple regression, disallowing the inclusion of random effects and/or non-linear effects. In Study 2 we aim to build upon previous works through exploration of the MALD data (Tucker et al., 2019).

The MALD project includes a database of responses as contributed by hundreds of participants to thousands of word and pseudoword stimuli. Stimuli are organized into 134 unique lists, each comprised of 800 items. These lists incorporate 1/67 MALD word-sets and 1/24 MALD pseudoword-sets, each set containing 400 stimuli-items. Each of the 67 word-sets was randomly paired (without replacement) with 2 of the pseudoword-sets, totalling the 134 unique lists to draw from. Participants could return for up to three sessions, each time completing a different list and experiencing stimuli in follow-up sessions that they have not encountered previously. In Study 2, we explore data from version 1.1 of the database[1] which holds 459,147 responses to 26,793 words and 9,592 pseudowords, as made by 440 participants in a total of 574 sessions. We subset the data and use generalized additive mixed models to analyse responses to word items for which there are available estimates of semantic richness variables.

Method

Participants

As in Study 1, participants in Study 2 were recruited through the University of Alberta's undergraduate participation pool. All participants provided informed consent prior to participating in the study. Hearing screenings were administered as in Study 1, where 41 of the 440 participants were excluded due to hearing issues (9%). 399 monolingual native listeners of English with typical hearing (76% female, 24% male; age 16

[1] The data is available for download at https://doi.org/10.7939/r3-kh0r-r116.

to 29, $M = 20.02$, $SD = 2.41$) provided data retained in the study in a total of 516 experimental sessions.

Stimuli

We explore responses to 8,626 word-items from the MALD database—more than a 20-fold increase in the number of stimuli sample size compared to Study 1. These items were selected for available estimates of concreteness, valence, arousal, semantic neighborhood density, semantic diversity, and number of morphemes. Although significant in Study 1, number of semantic features was not included as a predictor in Study 2 as the values for semantic richness collected by McRae et al. (2005) were limited to a set of 541 words. A more recent set was created by E. M. Buchanan et al. (2019) and includes values for 4,436 words. However, using these ratings would still reduce the number of words we could consider in Study 2 by approximately half, so we have chosen to exclude this variable in order to maximize the stimuli set.

This expanded stimuli set allows us to explore a more complete range of semantic richness values. The most notable improvement can be seen in the case of word concreteness (Figure 2), as the range of values expressing this feature is no longer restricted to words with high values of concreteness (cf. Goh et al., 2016). Improvement in the case of other semantic richness variables is less pronounced, where benefit comes primarily through the expanded number of words providing a more holistic representation of the scales, even in the extremes.

INSERT FIGURE 2 APPROXIMATELY HERE

The stimuli used in Study 2 incorporate multiple parts of speech as well (referred to as *word class* herein), so are not restricted to the "noun" class exclusively. The word class information used is available in the MALD dataset and is defined as the frequency-dominant part of speech in the SUBTLEX-US corpus (we will use 'part of speech' and 'word class' interchangeably in the remainder of the manuscript; Brysbaert

et al., 2012). The selected 8,626 words belong to the noun (5,338; 62% of the words), verb (1,144; 13% of the words), and adjective (2,144; 25% of the words) classes. Proper names, adverbs, function words, interjections, and numbers (59 items total) were excluded from our analyses despite available measures of semantic richness due to low representation of these classes in the stimuli set.

We do note that many of the selected words can simultaneously belong to more than one word class. For example, the word *salted* exists most frequently as an adjective, but can also function as a verb. The word *account*, depending on the context, can be both a noun (predominantly) or a verb. In the word stimuli explored in Study 2, 52% of words that are categorized as nouns, 48% of words that are categorized as verbs, and 64% of words that are categorized as adjectives also have at least one other part of speech marked in the SUBTLEX-US corpus. We decided to keep these items in the current analysis as limiting the stimuli pool to unambiguous words imposes limits to the generalizability of the findings; these words represent a significant part of the lexicon speakers and listeners use. We return to the issue of a word possibly belonging to multiple parts of speech in the General Discussion.

Having confirmed that distributions of semantic richness variables generally encompass a more holistic range of values than we see in previous works, we next explore the increased representation by part of speech in our stimuli—that is, the potential to consider semantic richness for nouns, verbs, and adjectives separately. Figure 3 shows that in most cases the distributions for each variable appear similar across word classes. In other words, we see substantial overlap in the distributions for different parts of speech for nearly all variables, with the clear exceptions of values associated with concreteness and, to a lesser extent, semantic diversity. Nouns in our data are predominantly associated with relatively high concreteness, whereas adjectives and verbs are typically associated with lower concreteness values. Nevertheless, in spite of skewed distributions observed for certain variables, we generally find a more complete representation of the entire scale of

potential semantic richness when compared to previous works.

INSERT FIGURE 3 APPROXIMATELY HERE

In summary, Study 2 builds on Study 1 (and previous works) through a 20-fold increase in the number of words tested, exploration of additional part of speech classes, and a more complete representation of the range of variability in semantic richness measures.

Procedure

Data collection procedures were identical to those described in Study 1 with the following exceptions. First, MALD sessions involved one of 134 MALD lists, each containing a total of 800 items; sessions in Study 1 included a total of 884 items (442 words and 442 pseudowords). Second, MALD participants could choose to return for up to three sessions total in Study 2, whereas in Study 1 participants experienced all available stimuli within a single session (precluding the possibility of follow-up sessions without overlapping stimuli).

Results

In Study 2 we use GAMMs to explore a total of 63,016 responses to 8,626 words, collected during 516 MALD experimental sessions from 399 subjects. The number of retained responses per word ranged from 2 to 14, but only 0.08% of the words had 5 or fewer responses recorded for them (Q1 = 6, Q2 = 7, Q3 = 8, M = 7.31, SD = 1.33). The data preparation procedure was identical to the one used in Study 1, and the models created were also very similar to those from Study 1. However, as mentioned above, we do not include number of listed features as a predictor. Random effects for word-stimuli were not included because computational time becomes unfeasible and because there were not enough responses per item to justify their inclusion, especially since it may lead to high concurvity/collinearity between random intercepts and fixed factors (Baayen & Linke, 2020). Additionally, we now take into account word class as a predictor in order to test whether semantic richness effects may vary for different parts of speech.

The initial, full model, including all parametric and non-linear, smooth terms is presented in Table 4. All included variables were significant predictors of participant response latency, though some semantic richness predictors were only significant for certain parts of speech.

INSERT TABLE 4 APPROXIMATELY HERE

We first briefly overview the effects of the standard predictors regardless of word class, presented in Figure 4. Again we register an effect of trial number that is generally facilitative, especially in the first 100 trials. The effects of word frequency and the structural principal component are also facilitative, although the effect is lost for medium to high values of the structural principal component. Duration remains the strongest predictor of response latency, with longer recordings being associated with longer response latency.

The effects of phonological uniqueness point and number of morphemes appear to be the smallest. While it would be expected that higher (i.e., later) phonological uniqueness points would be related to longer response latency, we see a relatively flat curve because the effect of recording duration has already been taken into account. The effect of number of morphemes is also relatively flat, with monomorphemic words and words consisting of four morphemes being responded to with a slightly longer response latency than words with two or three morphemes. We note, however, that despite the increase in the number of words in comparison to Study 1, our sample of words is still dominated by monomorphemic words (68%) and words with two morphemes (27%), so these results should be interpreted with caution.

INSERT FIGURE 4 APPROXIMATELY HERE

We continue our overview of the full model by focusing on part of speech, and the semantic richness effects that we examined separately for each part of speech. Figure 5 includes plots for all significant effects. Effects related to nouns are given in red, those related to verbs in blue, and effects related to adjectives are presented in green. We can see

in the plot shown in the first row of the second column that response latency was highest for nouns, followed by verbs, and response latency for adjectives was shortest. There is no statistically significant difference between nouns and verbs, but participants responded to adjectives more quickly than to nouns on a statistically significant level (as evidenced in Table 4, given that "noun" is the reference level). The difference between verbs and adjectives is statistically significant, tested by Wald post-hoc tests using the `wald_gam()` function ($\chi^2(1) = 5.46, p < .05$). Note that the figure showing the difference between parts of speech presents transformed response latency (-1000/RT) as the dependent variable on the x-axis.

INSERT FIGURE 5 APPROXIMATELY HERE

More importantly, Table 4 shows that semantic richness effects are predictive of response latency, but are not necessarily predictive for all three word classes. The effect of concreteness is significant in nouns and adjectives only; the plot in row one, column three of Figure 5 shows these effects. For adjectives, presented in green, there is a linear negative relationship as higher concreteness values (given on the x-axis) were associated with shorter response latency (presented as back-transformed response latency on y-axis). For nouns however, in red, the line is rather flat for concreteness values up to approximately 4, after which we see a steep decline marking a strong relationship between higher values of concreteness and shorter response latency. Therefore, in nouns the facilitative effect of concreteness only exists for its highest values (4 to 5). The effect for verbs is not plotted as it would simply present a mostly flat line.

The significant effects of concreteness in Table 4 and the concreteness effect plot show us that this effect is significant in both nouns and adjectives; however, they do not indicate whether or not the effect(s) of concreteness may be different for nouns and adjectives. In order to compare the effect concreteness has on nouns versus adjectives, we need to plot the difference curve based on model predictions by using the plot_diff function from the *itsadug* package (van Rij et al., 2020a). This difference is shown in the

plot in row one, column four of Figure 5. The difference in transformed response latency between effects of these two parts of speech is given on the y-axis ("N" stands for noun and "A" for adjective). The range of the concreteness distribution in which adjectives and nouns significantly differ is marked in purple on the x-axis, and limited by vertical purple dashed lines on the plot. We see that for values of concreteness up to 2.5 there is no difference between nouns and adjectives. After that, adjectives with medium or medium-high concreteness values (2.5 to 4) are responded to faster than nouns with the same concreteness values. For concreteness values above 4 this difference diminishes, until it becomes nonsignificant for the highest values (those close to 5).

The second row of Figure 5 is dedicated entirely to the effects of valence, as these effects registered for all of the word classes we examined. The first plot in the second row shows that the effects look rather similar for nouns, verbs, and adjectives: words with low valence values (negative valence words) and words with high valence values (positive valence words) are responded to faster than words with medium valence values (neutral words). In nouns, the facilitative effect of negative valence is somewhat larger than the facilitative effect of positive valence. In verbs and especially adjectives (where there is little difference between low and medium valence values) we find the opposite: high valence is related to shortest response latency. The remaining three plots in the second row of Figure 5 examine the differences between word classes. We see that the curves are pretty much the same when verbs are compared to nouns and adjectives ("V" stands for verb), as the significant portion of the difference (again, indicated in purple) is rather small. In other words, verbs find themselves somewhere between nouns and adjectives and are not significantly different from either. Nouns and adjectives, however, are different for values of valence 3 or higher. Longer response latencies are recorded for medium-valence nouns, and this difference further increases for words with the highest valence values. These higher valence values appear to facilitate adjective processing more than noun processing.

In the plots in the first two columns of the third row of Figure 5 we see the effects of

rated word arousal. As with concreteness, the effect of arousal is only significant for nouns and adjectives. The effect for nouns is mostly linear, as higher arousal is associated with shorter response latency. Note that the curve becomes more flat for low values of arousal; this effect is not as strong for low to medium arousal nouns. The effect for adjectives is less straightforward. The line looks mostly flat, except for arousal values of approximately 4, where longer response latency is registered. It is unclear why medium-low values of arousal should be more difficult to process, though it may be that a particular set of adjectives that have these arousal values are difficult for our listeners. We suggest that this result be interpreted with caution, especially since the significance level of this effect was the lowest of all significant effects observed ($p < .05$). The difference plot between nouns and adjectives shows that adjectives are associated with shorter response latency, despite the rise in response latency for adjectives with arousal ratings above approximately 4. This difference is lost only in the highest arousal values of 7 and 8.

The plot in row three, column three represents the interaction between valence and arousal in adjectives; this interaction was not significant for nouns or verbs. The plot can be read like a map marking elevation through color, where higher elevation regions—the brown hills—indicate longer response latency in comparison to lower elevation regions depicted as the green flat regions and the blue sea level. The numbers on the plot associated with the regions delineated by red lines represent transformed RT. The white portions of the plot represent combinations of arousal and valence with too few data points. We see that for adjectives the longest response latency is recorded for words that have low to medium valence and low to medium arousal. This is in line with the non-linear valence effect recorded for adjectives, as valence does not appear to be facilitating responses (certainly not as much as it does in nouns). However, by looking at the upper left portion of the plot, we see regions of blue and green. Low valence does not have to be inhibitory in adjectives: if an adjective has high arousal and low valence (being a negatively-charged word), we still record very short response latency. This facilitation is

not the case for medium (neutral) valence words, as having high arousal does not help in those cases. Lastly, high valence is associated with short response latencies in adjectives regardless of adjective arousal, as can be seen by observing the green and blue regions on the right hand side of the plot.

The fourth row of Figure 5 shows the effects of corpus-based measures of semantic richness. Semantic neighborhood density was a significant predictor of response latency in nouns and adjectives, but not in verbs. In both nouns and adjectives, higher values of semantic neighborhood density were associated with shorter response latency; the two lines are mostly parallel. In the difference plot, nouns simply always have a higher overall response latency than adjectives, regardless of the value of semantic neighborhood density (although, admittedly, this difference somewhat declines for the highest levels of semantic neighborhood density). Lastly, the effect of semantic diversity was significant in adjectives only. The effect is facilitative for low to medium values of semantic diversity, after which the effect is lost; the line becomes mostly flat, with some tendency to increase slightly toward the far end of extremely high semantic diversity values.

These results can also be summarized by word class. Response latency to MALD nouns is related to word-item concreteness, valence, arousal, and semantic neighborhood density. Response latency to adjectives is related to all five semantic richness variables considered in this study. In verbs, only the (roughly) quadratic effect of valence is recorded. Other considered measures of semantic richness were not predictive of response latency when a word that has "verb" as its frequency-dominant part of speech in the SUBTLEX-US corpus was presented to the listener.

The results above paint a detailed and complex picture of how different word characteristics, including their semantic characteristics, influence the process of isolated spoken word recognition for different word classes. However, some of the noted effects are clearly larger (e.g., frequency) than others (e.g., arousal). Thus, as in Study 1, we also test the contribution of all of our predictors to the overall model fit using the `compareML()`

function. Once again, we excluded predictors one by one, and assessed their contribution by comparing the model including a given predictor to the model excluding that predictor. In the case of semantic richness variables, we first removed the interaction with part of speech. If that exclusion was warranted, we then proceeded to test whether the semantic richness variable itself significantly contributes to model fit. Through this process, we excluded phonological uniqueness point, the interaction with part of speech for all semantic richness variables (but not part of speech itself as a parametric coefficient), and semantic diversity. The optimized model—the one including only those predictors whose exclusion would yield a model with significantly worse fit—is presented in Table 5.

INSERT TABLE 5 APPROXIMATELY HERE

As can be seen in Figure 6, we find no qualitative difference in comparison to the model that included all statistically significant predictors (i.e., the model presented in Table 4) where the retained control predictors are concerned.

Where the semantic richness predictors are concerned, the shape of their effects (given in Figure 6) for the most part mimics the shape of the same effect recorded for nouns in the model that retained all statistically significant effects. Specifically, we see that the effect of concreteness is mostly flat until high concreteness values are reached (4 or above), after which there is a negative linear relationship between concreteness and response latency. The case is similar with arousal, but the negative linear relationship between arousal and response latency starts earlier, at medium-low values of arousal (approximately 3.5 and above). The effect of valence could be roughly described as quadratic, as both low and high valence are associated with shorter response latency while words with medium (neutral) valence are associated with longer response latency. Finally, the effect of semantic neighborhood density is entirely linear, again with higher values of this semantic richness variable being related to shorter response latency.

The effect of word class is now included only as a parametric term. However, due to the exclusion of some other variables from the model, the differences between word classes

changed in comparison to the model including all statistically significant effects. Nouns are now responded to with a longer response latency than both adjectives and verbs, not adjectives only. At the same time, there is now no statistically significant difference between adjectives and verbs, as determined by using the `wald_gam()` function ($\chi^2(1) = 1.78, p = .18$).

INSERT FIGURE 6 APPROXIMATELY HERE

Discussion

In this study we have built upon Study 1, and other works exploring the role of semantics in auditory lexical decision, to contribute further support for the influence of semantic richness variables in auditory word recognition. We considered responses related to an expanded word-set with increased variability in both word class and the range of semantic richness for each variable, and provided evidence for the following. (1) We have found differences when comparing results across the two studies, largely supporting the importance of representing a full range of values for semantic richness variables and for considering as many varied items as possible during analyses. (2) We have shown that certain semantic richness effects (concreteness, valence) can exhibit non-linear effects, thereby supporting our use of GAMMs for data analysis. Lastly, (3) we have found differences indicating that the influence of a given semantic richness variable should not be assumed uniform across word classes. Importantly, the differences in certain variables recognized as predictive for a given word class appear to reflect characteristics of that word class.

As mentioned in Study 1 Results section, three different criteria are often employed in determining whether a certain predictor should be retained in a generalized additive mixed model. These are significance tests shown in the model summary, visual inspection of the effect plot, and model comparison procedures that compare model fit of nested models. Differences in semantic richness effects across word classes were recorded as

statistically significant and are visible on the effects plots, but model comparison procedures indicated that including them did not significantly improve model fit. Note that the criterion of significant improvement to model fit also excluded the effects of phonological uniqueness point and semantic diversity.

One reason why the interaction between semantic richness variables and word class did not improve model fit may be explained by noun predominance. Approximately two-thirds of our items were nouns. Additionally, we only found a significant effect of word valence for verbs. It could be that the verbs were under-represented even with 1,144 verbs, and that the sample size was still insufficient to capture potential trends. Whatever the reason, verbs did not alter the shape of the effect for most semantic richness variables (they attenuated it at best), while their number was small enough not to diminish the otherwise noticeable effect noted in nouns. In the model that did not take interactions between semantic richness predictors and word class into account, semantic richness effects closely resembled those recorded for nouns in the model that did consider this interaction. In other words, the distinction that was noticed between word classes is not sufficiently large to improve model fit (in a noun-predominant stimulus set), but could still be informative for important differences regarding how listeners conceptualize and access different parts of speech. The question then becomes whether we favor the simplest good model or whether we want a model to take into account all of the factors (variables, predictors) apparently playing a role in the spoken word recognition process, even though their contribution is relatively small in comparison to the complexity that they introduce to the model. The more conservative approach would be to use model comparison and exclude the interactions, but the more liberal and more informative models include these interactions as significant predictors (alongside the effects of phonological uniqueness point and semantic diversity). If our concern is not with model complexity, then the interactions should be retained, but their small effect size also warns that future studies may be needed to confirm their, somewhat subtle, effects.

The remainder of this discussion explores Study 2 findings in more detail, mainly by comparing and contrasting them to those of Study 1. Although we describe and discuss the recorded differences in semantic richness effects in different word classes, we presently still refrain from making strong theoretical claims about the need to distinguish between parts of speech as our results are not conclusive and merely call for future studies to scrutinize the issue further. The General Discussion to follow offers suggestions about how to continue investigating semantic richness effects in (spoken) word recognition, and situates our findings in theories of lexical access and retrieval.

Differences between Study 1 and 2 results

First, the clearest distinction between Studies 1 and 2 is in the increase of semantic richness variables that exert an influence on response latency. In Study 1, concreteness, valence, and number of semantic features were found to be significant predictors of our targeted replication experiment, while concreteness alone was found to also significantly improve model fit. In Study 2, with an increased number of word-items and responses, we find that all five semantic richness variables considered in the analysis (excluding number of semantic features) are significant predictors of response latency. In a model that only included terms that significantly improved model fit, semantic diversity was the only semantic richness variable excluded. Concreteness, valence, arousal, and semantic neighborhood density were all found to be important contributors to model fit.

Second, the shape of the effect that certain semantic richness variables had on response latency changed in Study 2. This is likely a consequence of an expanded word set—i.e., a more complete range of semantic richness attributable to the stimuli explored in Study 2—and is most strikingly the case for concreteness and valence. Recall that in Study 1 stimuli were restricted to nouns characterized by high levels of concreteness (as in Goh et al. (2016)), whereas the stimuli in Study 2 better represent the entire range of concreteness (and other semantic richness variables). Facilitative effects of concreteness

were observed for nouns in both Studies 1 and 2, though results from Study 2 suggest that this effect is restricted to high levels of concreteness only. That is, we have shown the effect of concreteness in nouns is not wholly linear, but instead is null until it becomes linear after reaching the required threshold of high rated concreteness. Similarly, the effect of valence in Study 1 was found to be linear, where instead we find a roughly quadratic effect when analyzing the expanded MALD word-set in Study 2. Through observing the effect patterns of these variables, one crucial conclusion emerges: It seems a full range of potential semantic richness values is often required to understand how they may influence lexical access. Even though certain descriptions may be reasonable in the context of restricted data (e.g., concreteness as described by Goh et al., 2016), the nuances of these semantic richness effects can be missed or misidentified through the methodologies.

Lastly, when considering the expanded range of semantic richness values alongside the diversity in word class available through Study 2, we also find a more complete assessment of semantic richness than was previously available (at least with regard to the semantic richness variables tested). Certain variables were recognized to influence multiple word classes in both similar and different ways, perhaps reflecting the way these characteristics are conceptualized within the mental lexicon. Differences observed across participant responses seem to reflect the relative salience of semantic attributes associated with items as weighted by word class. Put another way, certain semantic characteristics appear more likely to influence processing of a given item when those characteristics are more pertinent to that item's word class. Furthermore, such differences in processing support accounts suggesting that the semantic characteristics which broadly define certain item types (e.g., word classes) are stored in the mental lexicon as part of each individual entry, and that such properties are likely activated alongside a given lexical item during word recognition, thus aiding in selection (cf. Marslen-Wilson, 1987).

Effects of rating-based semantic richness variables in Study 2

Concreteness is recognized as a viable predictor for response latency to both the noun (Studies 1 and 2) and adjective classes (Study 2), though not for verbs (Study 2). This finding is not altogether surprising when considering the functions of these word classes: Nouns often represent tangible objects occupying space in the physical world and, as a class, adjectives serve to describe properties of those objects—many of which are experienced through our senses. Meanwhile, verbs typically represent abstract concepts and actions, which exist outside of a material context (that is to say, you cannot touch *running* or *amplify*). The differences in the distribution of concreteness values may be another cause for concreteness not having an effect in verbs, since very few verbs have high concreteness values. We also draw attention to this effect and the representation of values observed in adjectives, which spans the entire range of word concreteness, whereas the same effect is restricted to only high concreteness for nouns. For example, it takes more time to process a mid-concreteness adjective such as *disorderly* than a high-concreteness adjective such as *wet*, but it takes even more time to process a low-concreteness adjective such as *inexplicable*. In nouns, however, the participants are the fastest when responding to high-concreteness nouns such as *apple*, but are equally fast when processing mid-concreteness nouns such as *department* and other nouns in decreasing concreteness: *tension*, *insistence*, *privilege*, all the way to the lowest concreteness nouns such as *belief*. The finding that all non-concrete nouns are treated as equally abstract, while a fine gradient of concreteness is observed in adjectives remains to be verified and explored further.

The effect of valence in Study 1 was found to be linear, with higher values of valence being associated with shorter response latency. In Study 2, all three word classes showed similar significant effects of valence, where any such influence was strongest at the extreme values and diminished for the more neutral grades. For example, negative valence nouns *torture* and *disease* and positive valence nouns *enjoyment* and *joy* are both responded to

faster than neutral valence nouns such as *essay* or *plank*. The case is the same for verbs, as *hate* and *love* or *thank* are both responded to faster than neutral valence verbs like *indicate*, and adjectives as well, as *unhappy* and *fantastic* are responded to faster than *financial* or *canned*. Despite similarities, we do note that the facilitative effect of low valence is stronger in nouns, whereas the facilitative effect of high valence appears to be stronger in adjectives. Additionally, an interaction between valence and arousal was recorded in adjectives only. Facilitation attributed to low valence values occurred for high arousal adjectives, while the longest response latency was recorded in low valence adjectives with low arousal ratings (more detail given in the following paragraph).

Arousal is scaled similarly to concreteness through its unidirectional influence. The scale accounts for intermediary grades between null and strong, but there can be no negative arousal. With this fact in mind, we were somewhat surprised to find no significant effect of arousal for our verb items as arousal is itself a state of being, and therefore seems the product of something inherently related to verbs. One potential explanation for this finding is that perceiving arousing stimuli (nouns like *gun* or *snake*) and their characteristics (adjectives such as *exciting* or *erotic*) acts as a call to action, arguably leading to faster responses than when a listener is exposed to low arousal words such as *grain* or *quiet*. In turn, arousing actions themselves (verbs such as *scare* or *celebrate*), especially presented out of context, do not require or incite a reaction any more than *count* or *comb* would. The effect for adjectives also appears to be rather flat, with an unexpected, small increase in response latency registered for arousal values between roughly 3 and 4.5. Like the effect of valence, we argue that the effect of arousal for adjectives may be better interpreted through the interaction between valence and arousal. To repeat, higher values of arousal appear to make little difference for medium-to-high valence adjectives, but for low valence adjectives there is a recorded facilitative effect of higher arousal. Therefore, positive valence words are always processed and responded to relatively quickly. So are negative valence words with high arousal, such as *heartbroken* or *murderous*. Extremely

negative valence words with extremely low arousal are rare in our dataset, but the slowest response latencies are registered in words that have relatively low (i.e., negative) valence and arousal, such as words *nameless*, *sad*, *lifeless*, *unsuccessful*, *depressing* and similar.

As described by, i.a., Kuperman et al. (2014) and Goh et al. (2016), results regarding the effects of valence and arousal are mixed and dependent on a number of other stimuli, design, and analytical choices. Therefore, we primarily offer our findings as another example to be included as part of a set of studies considered in future focused theoretical and/or meta-analytical investigation of effects of valence and arousal. However, based on our methods and findings, we also wish to stress two more potentially relevant factors: (1) modality, as our stimuli were presented in the auditory modality which may yield different results in comparison to the visual modality and (2) word class, as we recorded differences between nouns, verbs, and adjectives that appear to merit further inspection.

Effects of corpus-based semantic richness variables in Study 2

Having discussed our rating-based semantic richness variables above, we next explore our corpus-based variables: semantic diversity and semantic neighbourhood density. Goh et al. (2016) and our Study 1 did not find an effect of these variables. Study 2 models, however, show that semantic neighbourhood density can influence responses to nouns and adjectives, and that semantic diversity influenced participant responses to adjectives only. These findings too suggest that using a larger and more varied set of stimuli may enable a detection of effects which would otherwise be missed. The effect of semantic neighborhood density is facilitative and similar for nouns and adjectives. The effect remains largely unchanged when the interaction between semantic neighborhood density and part of speech is excluded from the model. Simply put, it is easier to recognize words that situate their meaning relatively close to a larger number of other words (and their meaning). Nouns like *way* and *place* and adjectives like *great* and *little* are recognized faster than *jogger* or *exhilarated*. One might notice the differences in word frequency

among the examples given. Indeed, we did register a moderate correlation between logged frequency and semantic neighborhood density which was not included in our statistical models, so it is important to bear in mind the possible bias in these results.

We observed semantic diversity effects for adjectives only, which explains why such effects were missed by Goh et al. (2016) and in our Study 1. Indeed, when we remove the distinction by part of speech, this effect is lost in the noun-dominant full dataset. Adjectives can generally be applied in a broad assortment of contexts, although some adjectives more than others; there are relatively few restrictions which might prevent a particular object from being described as either *new* or *main*, in contrast to, for example *gastric* or *symphonic*. This apparent variety in freedom seems to parallel our measure of semantic diversity, which reflects the breadth of contexts in which an item might appropriately be applied. It, therefore, makes sense that responses to adjectives alone could be predicted using measures of semantic diversity, as items with the potential for more broad application were recognized more quickly (and perhaps resemble something closer to the prototypical, i.e., versatile adjective in this way). To be more precise, participant response latency is higher if an adjective is restricted to specific contexts, i.e., has low semantic diversity. However, this relationship between semantic diversity and response latency exists only for low to medium semantic diversity: a medium level of variety of contexts in which an adjective can be found seems sufficient to enable its unimpeded processing, as there is not much difference between responses to medium and high semantic diversity adjectives.

General Discussion

Whether or not we consider differences in the influence of semantic richness variables across word classes, the present study shows that semantic richness variables influence lexical access during spoken word recognition, even when words are presented in isolation (i.e., when lacking context). Although semantic richness effects are somewhat smaller in

comparison to well-established effects related to word length (e.g., duration in milliseconds, number of phones, number of syllables), frequency, and form (e.g., phonological neighborhood density or phonological Levenshtein distance), there seems to be little doubt that meaning-related word characteristics shape the process of spoken word recognition.

Importantly, expanding the stimuli set and increasing the number of responses lead to different results; that is, a more detailed understanding of these effects. In our view, the presently used semantic richness variables relate to processing speed in at least two distinct manners. The first is the manner of variety or clarity of contexts in which a single word-item may appear, expressed through the corpus measures of semantic neighborhood density and semantic diversity and, possibly, in some part by word concreteness. The second is the manner of urgency, as words carrying meaning that may be of immediate and higher importance to the listener, expressed through higher arousal and higher negative or positive valence, are responded to faster by listeners.

Given that the effects observed are clearly dependent on the words used in the experiment and on the analytical methods employed, we also draw attention to the necessity of collecting semantic richness information—as well as lexical decision or other task data—for as many words as possible. Most projects collecting measures of semantic richness considered in this paper do so, and so do many large-scale studies collecting participant responses such as MALD (Tucker et al., 2019) or the Auditory English Lexicon Project AELP (Goh et al., 2020). Besides the number of features that was excluded from our broader analysis in Study 2, other semantic richness variables that may be included in future analyses are danger and usefulness (Wurm & Vakoch, 2000, pertaining to word urgency) or the more recently developed information regarding embodied semantic richness variables such as body-object interaction or sensory experience ratings (Pexman et al., 2019).

However, the sets of words often do not fully overlap in these different projects, reducing the number of words that can be considered. In this regard, missing rating-based

measures of semantic richness are particularly difficult to replace, as they normally require a large sample of raters; collecting this data is costly in many ways. A potential replacement or an approximation for rating-based measures, such as the one calculated from co-occurrence statistics (Hofmann et al., 2018; Hollis & Westbury, 2016; Hollis et al., 2017), could be a good way to alleviate this issue and expand the set of words we could use, even in the present analysis (although we limit ourselves to the same set of variables used by Goh et al. (2016) to enable direct comparison).

We maintain that the above described results should be considered carefully. When analyzing behavioral data or when using computational models of spoken word recognition, each trial is most often treated as separate or isolated (hence isolated spoken word recognition). In reality, the participant is aware of preceding stimuli and has already activated their meaning, making these preceding stimuli act as potential—albeit weak—context for the current stimulus. In other words, an unknown amount of variance captured by both rating-based and corpus-based measures could be attributed to what Balota et al. (2012) refer to as overall list context effects (see also, e.g., Dorfman & Glanzer, 1988; Lupker et al., 1997; Perea & Carreiras, 2003; Van de Ven et al., 2011). We suspect that there should be a general tendency for words with higher semantic richness variables to be semantically connected to a larger number of other words, or that the effect of a semantic richness variable (e.g., high arousal in a particular word) might be attenuated if the item list includes many versus few high arousal words. Future analyses could operationalize the semantic relationship of the current word with a number of previously presented words to test whether this relationship predicts response latency and accounts for a portion of semantic richness effects.

The present study also allows us to assess how we might improve on the quality of studies investigating semantic richness and other effects in spoken word recognition, as well as theories of how these processes unfold. Suggestions for such improvement are the focus of the three subsections to follow. First, we discuss the issue of word class effects and their

representation in studies that use the auditory lexical decision task (or similar methods) to investigate semantic richness effects. We then discuss some ramifications of our findings on computational models of spoken word recognition, before finally drawing attention to an important aspect of semantic richness effects: namely, their time course.

The operationalization of word class

With regard to the extent that semantic richness variables may impact spoken word recognition differently for different word classes, we see this question as still open. In light of the evidence described above, the answer depends largely on whether one would take predictor significance or model fit improvement as criterion for predictor retention. Additional studies and analyses are needed to further investigate this issue. One important related decision exists in item categorization, i.e., how to best assign stimuli items with multiple meanings to a single word class. It was clear from our previously given examples that our study included words like *comb* or *count* which can belong to more than one part of speech. When describing our Study 2 stimuli, we noted that 52% of words we categorized as nouns, 48% of words we categorized as verbs, and 64% of words that we categorized as adjectives have at least one additional part of speech marked in the SUBTLEX-US corpus. Simply, less ambiguous stimuli will surely make for clearer findings.

In the auditory lexical decision task, stimuli are presented outside of the sentence, i.e., with little semantic context and within an unusual prosodic environment. Although we may speculate that these circumstances could potentially bias the listener towards perceiving nouns, we cannot know which meaning listeners identified when they were presented with a particular string of sounds and recognized them as a word of English. Some previous research suggests that listeners may be activating multiple meanings (all potential parts of speech) or at least show some degree of difficulty being associated with processing ambiguous words whose distinct meanings can belong to separate parts of speech (Mirman et al., 2010; Onifer & Swinney, 1981; Seidenberg et al., 1982). Other

accounts stress that the clearly biased (dominant) meaning of a word is accessed quickly even in neutral context (Duffy et al., 1989; Rayner & Duffy, 1987); it is important to note that most of these studies are concerned with the effect that disambiguating context has on lexical access of ambiguous words. To the best of our knowledge this question has not been conclusively resolved, especially with regard to isolated presentation, i.e., without a prime or sentence context.

The assumption about what happens when an ambiguous stimulus is presented, however, directly impacts how words are sampled for the analysis. We see three potential basic ways to address the fact that many English content words can belong to more than one word class, depending on context and/or prosody. The first option is to only use words with unambiguous word class in the corpus. As we mentioned before, this approach is problematic because it significantly reduces the amount of data analyzed and because it makes the findings generalizable to unambiguous words only. Moreover, this approach misrepresents the reality of the English language.

A second approach would be to observe the class-specific frequency of words that can belong to more than one class and select words that have a dominant word class. The threshold for dominance would still have to be decided on. In fact, our present study is a special case of this option—the threshold was just as low as possible, meaning that the frequency-dominant part of speech was selected regardless how small the difference in frequency between its occurrences representing different parts of speech may be. At the same time, an additional variable could be included to differentiate between words that belong to more than one class and those that do not. This differentiation could be made even more precise by keeping track of the exact number of classes a word belongs to or even the relative frequency of the dominant class for a particular word in comparison to frequency recorded for its other classes. These measures are available in SUBTLEX-US Brysbaert et al. (2012).

A third solution would be to categorize a word according to all classes it belongs to:

words that can be both, for example, a noun and an adjective (*tart*, *uniform*) would be classified into their own group, separately from words that can be both a noun and a verb (*bluff*, *jail*), as well as from words that are nouns only (*acre*, *translation*). This approach would be most fitting if we assume that all potential meanings (and hence word classes) of a word are activated simultaneously—and more in line with the standard approach to estimating word frequency in lexical decision and other similar tasks, which usually lumps together all word occurrences in a corpus regardless of the word class in which the target word appears. Such an approach to classification would still induce some small word loss. There are groups of words that belong to three or more classes and these groups do not have enough members to be represented in a statistical analysis. Even if we restrict our analysis to those words that belong to two classes only, the presently used set of words cannot represent some of the groups well. For example, there are only 312 words that belong to the adjective and the verb class in our dataset. There are more such words in the MALD1.1 database (1,632), but semantic richness variables are not available for all of them. Alternatively, one could create separate statistical models for each word class and include all words that are at least once tagged as belonging to that class. For example, the word *salted* would then be included in both the adjective and the verb statistical models. However, this may also artificially reduce the differences between different word classes, given large overlap in the items found in these models. All the same, any differences in the captured effects would still likely remain driven by the unambiguous-class words in the sample, as these items are the only difference between, for example, a noun and a verb model formed in this manner.

The representation of semantic richness effects in models of spoken word recognition

Our results indicate that the (resting) activation of words may depend on their meaning and relationship with other words. One concern for computational modelling of

these effects may be that they are somewhat small; it may appear more parsimonious to eschew them when generating model estimates and comparing them to behavioral data. Still, we should try to include these variables as predictors in our computational models and see whether or not the model fits participant data better. Models of spoken word recognition must be able to explain why certain semantic richness variables are predictive of response latency while others are not (for certain parts of speech). Moreover, we should not be limiting the breadth of our knowledge about the process of auditory lexical access to only large effects such as those driven by duration and frequency.

Based strictly on the results described by Goh et al. (2016), one would assume that representation of the mental lexicon in terms of co-occurrences is inadequate. Because semantic diversity and semantic neighborhood density did not predict response latency in that study, the connection of the target word to other words in the lexicon does not seem directly relevant to the process of lexical access. However, the results described in our Study 2, which used an expanded word set, show a significant effect of semantic neighborhood density, while the effect of semantic diversity was noted in adjectives only and lost once the interaction between this measure of semantic richness and word class was removed from the model. Models like the Distributed Cohort Model (Gaskell & Marslen-Wilson, 1997; Gaskell & Marslen–Wilson, 1999; Gaskell & Marslen-Wilson, 2002) and the Discriminative lexicon (Baayen et al., 2019) provide means to represent and take into account relationships between words (i.e., capture the effects of these corpus-based measures of semantic richness), and will therefore be useful as we move forward building more holistic models of this process.

Still, such a representation of the mental lexicon does not capture effects related to rating-based measures of semantic richness like concreteness, valence, or arousal. It seems that models that make use of top-down weights, such as TRACE (McClelland & Elman, 1986) or DIANA (Ten Bosch et al., 2022; ten Bosch et al., 2015) have a way of capturing these effects if the values of these variables are stored within the lexicon. It would be best

if we could represent these through separate top-down weights or resting activation increases in order to avoid conflating different factors under a single factor. Combining a more developed representation of the mental lexicon and top-down weights may be necessary to successfully simulate the effects of both coprus-based and rating-based measures of semantic richness (even if they are extrapolated from corpus-based measures; see Hollis et al., 2017).

Other conceptions of the mental lexicon that explicitly acknowledge connections between different entries could also provide useful direction for future research. Some of these conceptions stem from research in spoken word recognition, and some do not. Within network science-based approaches, the mental lexicon is conceptualized as a mathematical graph where nodes are connected to each other based on similarities such as phonological form (Luce & Pisoni, 1998; Vitevitch, 2021) or semantic closeness (Borge-Holthoefer & Arenas, 2010; Miller, 1995). A more modern approach utilizes multi-layer networks to represent more than one relation between items in the mental lexicon at the same time (Vitevitch, 2021). These types of network science approaches have been fruitful in modeling a variety of human behavior regarding language comprehension (Levy et al., 2021), acquisition (Stella et al., 2017), and difficulties (Castro et al., 2020). Similarly, it is a core tenet of constructionist approaches to language that an individual's total knowledge of a language is made up of a complex network between different constructions, i.e., a constructicon (Fillmore et al., 2012; Goldberg, 2003; Goldberg & Herbst, 2021). Finally, Libben (2022) provides a critical history of the lexicon-as-list/dictionary conceptualization and then goes on to express the mental lexicon as a dynamic system in which representations of words are hierarchies of actions that are based in part on the connections between items the language user knows. All of these advances could be used in studies simulating the process of spoken word recognition in order to enrich the representation of the mental lexicon, especially regarding its apparent semantic interconnections.

The time course of semantic richness effects

The notion of semantic richness, however, is not necessarily as simple as *if* semantic information plays a role in this context; one might also ask *when* semantic effects take place in the time course of word recognition for a more detailed discussion see, e.g., Carreiras et al., 2014; Kryuchkova et al., 2012; Pulvermüller et al., 2009; Rabovsky et al., 2012. Such time course analyses speak directly to theories addressing the organization and structure of the mental lexicon, and how units may be accessed within it see, e.g., Tyler et al., 2002, and references cited within. An extensive toolkit of statistical methods and experimental tasks have been employed to establish the absolute temporal onset and relative chronology of lexical effects in visual word recognition. For example, Staub et al. (2010) examined the onset of word frequency effects in eye-movement fixations using a non-parametric distributional analysis using the vincentile plotting technique (Ratcliff, 1979; Vincent, 1912). Further, a non-parametric survival analysis approach (Reingold & Sheridan, 2014) was applied to visual lexical decision latency and eye-movement fixation duration with the aim of assessing the time course of lexical effects in derived (Schmidtke et al., 2017) and compound word recognition (Schmidtke & Kuperman, 2019). These studies find that both the absolute onset—as well as the relative ordering of semantic effects–occur early within the time course of lexical processing. Finally, piece-wise exponential additive mixed models (PAMMs; Bender et al., 2018) have also proven useful in modelling lexical decision responses as time-to-event data, wherein the event is the binary "word" or "not a word" decision made by the participant. These models attempt to predict the probability of a response occurring at any given point in time beginning with word onset and given the values of relevant predictors. Using this technique Hendrix (2018) and Hendrix and Sun (2020) find evidence for early semantic neighbourhood density effects in visual lexical decision. These types of analyses have not yet been employed to explore data related to auditory lexical decision, but could be used to address the temporal dynamics of semantic processing during spoken word recognition. Our initial attempts have shown somewhat

volatile results, while also demanding weeks of computational time per model. We hope to explore these effects as they unfold in the auditory lexical decision task in the future.

One of the primary characteristics of spoken word recognition is that the signal unfolds in time, with early acoustic information going hand in hand with uncertainty about which item in the mental lexicon the signal refers to. Most theories and models of spoken word recognition assume that at this point a number of candidates are gaining activation, competing against one another. If we are truly looking at the time course of semantic richness effects (or other lexical effects, for that matter) it may be worthwhile to also assess whether and how lexical or semantic characteristics of assumed plausible competitors, and not just the target word, affect lexical access. In other words, we may want to explore ways for relevant lexical and semantic characteristics to be assessed based on the remaining set of plausible competitors, rather than on the actual word being presented only, given that the human listener cannot know which word is presented until the decision threshold has been met. Right now, we do that for formal characteristics of competitors, as we look at phonological neighborhood density or other ways to determine the number of plausible competitors, but we do not take into account characteristics which we find relevant for the word being presented – characteristics related to frequency, morphology, or, as this study shows, semantics. The assumptions about the exact shape of these effects would depend on the model (e.g., whether there is lateral inhibition between words or not), but it would be interesting and potentially important to test whether and how frequency or concreteness of a words' close competitors impacts lexical access for that target word.

Conclusion

We replicated the semantic richness effects as previously found in isolated spoken word recognition in a limited set of concrete nouns. Additionally, we explored these effects in a substantially larger sample of word-items from the Massive Auditory Lexical Decision project, while also considering potential non-linear effects and taking into account the

interaction between word class and semantic richness variables. We registered non-linear semantic richness effects in this expanded set of stimuli. We also found significant interactions between word class and the semantic richness predictors. These interaction effects had relatively small contributions to the model fit and were removed from subsequent statistical models that relied on model comparison as the criterion for predictor retention. However, we believe that the present results are encouraging and future investigation should consider effects of word class on semantic richness. These results also raise three important topics for future consideration: (1) the non-linearity of semantic richness effects, (2) the operationalization of word class in experiments investigating isolated spoken word recognition, and (3) the representation of semantic richness effects in (computational) models of spoken word recognition.

References

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, *4895481*.

Baayen, R. H., & Linke, M. (2020). Generalized additive mixed models. *A practical handbook of corpus linguistics* (pp. 563–591). Springer.

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

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods, orthography and phonology* (pp. 90–115). Hove, England: Psychology Press.

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47–53.

Bender, A., Groll, A., & Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, *18*(3-4), 299–321.

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic networks: Structure and dynamics. *Entropy*, *12*(5), 1264–1302.

British National Corpus Consortium. (2007). British National Corpus Version 3.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*.

Brysbaert, M., & New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word

frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991–997.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*(4), 1849–1863.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*(3), 531–544.

Carreiras, M., Armstrong, B. C., Perea, M., & Frost, R. (2014). The what, when, where, and how of visual word recognition. *Trends in Cognitive Sciences*, *18*(2), 90–98.

Castro, N., Stella, M., & Siew, C. S. (2020). Quantifying the interplay of semantics and phonology during failures of word retrieval by people with aphasia using a multiplex lexical network. *Cognitive Science*, *44*(9), e12881.

Devereux, B. J., Taylor, K. I., Randall, B., Geertzen, J., & Tyler, L. K. (2016). Feature statistics modulate the activation of meaning during spoken word processing. *Cognitive Science*, *40*(2), 325–350.

Dorfman, D., & Glanzer, M. (1988). List composition effects in lexical decision and recognition memory. *Journal of Memory and Language*, *27*(6), 633–648.

Duffy, S. A., Henderson, J. M., & Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 791.

Estes, Z., & Verges, M. (2008). Freeze or flee? negative stimuli elicit selective responding. *Cognition*, *108*(2), 557–565.

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., Dufau, S., Mathôt, S., & Grainger, J. (2018). MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, *50*(3), 1285–1307.

Fillmore, C. J., Lee-Goldman, R., & Rhodes, R. (2012). The framenet constructicon. *Sign-based construction grammar*, (193), 309–372.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*(5-6), 613–656.

Gaskell, M. G., & Marslen–Wilson, W. D. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science*, *23*(4), 439–462.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, *45*(2), 220–266.

Gianico-Relyea, J. L., & Altarriba, J. (2012). Word concreteness as a moderator of the tip-of-the-tongue effect. *The Psychological Record*, *62*(4), 763–776.

Goh, W. D., Yap, M. J., & Chee, Q. W. (2020). The auditory english lexicon project: A multi-talker, multi-region psycholinguistic database of 10,170 spoken words and nonwords. *Behavior Research Methods*, *52*(5), 2202–2231.

Goh, W. D., Yap, M. J., Lau, M. C., Ng, M. M., & Tan, L.-C. (2016). Semantic richness effects in spoken word recognition: A lexical decision and semantic categorization megastudy. *Frontiers in Psychology*, *7*, 976.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224.

Goldberg, A. E., & Herbst, T. (2021). The nice-of-you construction and its fragments. *Linguistics*, *59*(1), 285–318.

Hannagan, T., Magnuson, J. S., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, *4*, 563.

Hargreaves, I. S., Pexman, P. M., Johnson, J. S., & Zdrazilova, L. (2012). Richer concepts are better remembered: Number of features effects in free recall. *Frontiers in Human Neuroscience*, *6*, 73.

Hendrix, P., & Sun, C. (2020). A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/https://doi.org/10.1037/xlm0000819

Hendrix, P. (2018). A cross-linguistic investigation of response time distributions in lexical decision. *Proceedings of the 11th International Conference on the Mental Lexicon*, 16–17. https://mentallexicon2018.ca/wp-content/uploads/2018/09/MenLex2018_AbstractBooklet.pdf

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730.

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple co-occurrence statistics reproducibly predict association ratings. *Cognitive Science*, *42*(7), 2287–2312.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744–1756.

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology*, *70*(8), 1603–1619.

Hu, X., Tang, J., Gao, H., & Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. *Proceedings of the 22nd international conference on World Wide Web*, 607–618.

Kaushanskaya, M., & Rechtzigel, K. (2012). Concreteness effects in bilingual and monolingual word learning. *Psychonomic Bulletin & Review*, *19*(5), 935–941.

Kousta, S.-T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*(3), 473–481.

Kryuchkova, T., Tucker, B. V., Wurm, L. H., & Baayen, R. H. (2012). Danger and usefulness are detected early in auditory lexical processing: Evidence from electroencephalography. *Brain and Language*, *122*(2), 81–91.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*(3), 1065.

Ladefoged, P. (1990). Some reflections on the ipa. *Journal of Phonetics*, *18*(3), 335–346.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Levy, O., Kenett, Y. N., Oxenberg, O., Castro, N., De Deyne, S., Vitevitch, M. S., & Havlin, S. (2021). Unveiling the nature of interaction between semantics and phonology in lexical access based on multilayer networks. *Scientific Reports*, *11*(1), 1–14.

Libben, G. (2022). From lexicon to flexicon: The principles of morphological transcendence and lexical superstates in the characterization of words in the mind. *Frontiers in Artificial Intelligence*, *4*, 788430. https://doi.org/10.3389/frai.2021.788430

Lindblom, B. (1990). On the notion of "possible speech sound". *Journal of Phonetics*, *18*(2), 135–152.

Lõo, K., Järvikivi, J., Tomaschek, F., Tucker, B. V., & Baayen, R. H. (2018). Production of estonian case-inflected nouns shows whole-word frequency and paradigmatic effects. *Morphology*, *28*(1), 71–97.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics*, *39*(3), 155–158.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and parsyn. *Attention, Perception, & Psychophysics*, *62*(3), 615–625.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1.

Lupker, S. J., Brown, P., & Colombo, L. (1997). Strategic control in a naming task: Changing routes or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 570.

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabi, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., et al. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, *44*(4), e12823.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

Milin, P., Divjak, D., & Baayen, R. H. (2017). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1730.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS one*, *12*(2), e0171935.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, *3*(4), 235–244. https://doi.org/10.1093/ijl/3.4.235

Mirman, D., Strauss, T. J., Dixon, J. A., & Magnuson, J. S. (2010). Effect of representational distance between meanings on recognition of ambiguous spoken words. *Cognitive Science*, *34*(1), 161–173.

Nenadić, F., Podlubny, R. G., Kelley, M. C., Schmidtke, D., & Tucker, B. V. (2019). Semantic richness effects in isolated spoken word recognition: Evidence from Massive Auditory Lexical Decision [data and analysis scripts]. Massive Auditory Lexical Decision (MALD) Database, Education and Research Archive (ERA), University of Alberta, Edmonton, Canada. https://doi.org/10.7939/r3-7ejm-yf30

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Onifer, W., & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory & Cognition*, *9*(3), 225–236.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Osgood, C. E., & Suci, G. J. (1955). Factor analysis of meaning. *Journal of Experimental Psychology*, *50*(5), 325.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois press.

Perea, M., & Carreiras, M. (2003). Sequential effects in the lexical decision task: The role of the item frequency of the previous trial. *Quarterly Journal of Experimental Psychology*, *56*(3), 385–402.

Pexman, P. M., Hargreaves, I. S., Edwards, J. D., Henry, L. C., & Goodyear, B. G. (2007). The neural consequences of semantic richness. *Psychological Science*, *18*(5), 401–406.

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*(1), 161–167.

Pexman, P. M., Holyk, G. G., & Monfils, M.-H. (2003). Number-of-features effects and semantic processing. *Memory & Cognition*, *31*(6), 842–855.

Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review*, *9*(3), 542–549.

Pexman, P. M., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 english words. *Behavior Research Methods*, *51*(2), 453–466.

Pexman, P. M., Siakaluk, P. D., & Yap, M. J. (2013). Introduction to the research topic meaning in mind: Semantic richness effects in language processing. *Frontiers in human neuroscience*, *7*, 723.

Podlubny, R., Geeraert, K., & Tucker, B. V. (2015). It's all about, like, acoustics. *ICPhS*.

Porretta, V., Tucker, B. V., & Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics*, *58*, 1–21.

Pulvermüller, F., Shtyrov, Y., & Hauk, O. (2009). Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language*, *110*(2), 81–94.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rabovsky, M., Sommer, W., & Abdel Rahman, R. (2012). The time course of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience*, *6*, 11.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446.

Rayner, K., & Duffy, S. A. (1987). Eye movements and lexical ambiguity. *Eye movements from physiology to cognition* (pp. 521–529). Elsevier.

Rayner, K., Inhoff, A. W., Morrison, R. E., Slowiaczek, M. L., & Bertera, J. H. (1981). Masking of foveal and parafoveal vision during eye fixations in reading. *Journal of Experimental Psychology: Human perception and performance*, *7*(1), 167.

Recchia, G., & Jones, M. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, *6*, 315.

Reingold, E. M., & Sheridan, H. (2014). Estimating the divergence point: A novel distributional analysis procedure for determining the onset of the influence of experimental variables. *Frontiers in Psychology*, *5*, 1432.

Sajin, S. M., & Connine, C. M. (2014). Semantic richness: The role of semantic features in processing spoken words. *Journal of Memory and Language*, *70*, 13–35.

Sajin, S. M., & Connine, C. M. (2017). The influence of speech rate and accent on access and use of semantic information. *The Quarterly Journal of Experimental Psychology*, *70*(4), 619–636.

Schmidtke, D., & Kuperman, V. (2019). A paradox of apparent brainless behavior: The time-course of compound word recognition. *Cortex*, *116*, 250–267.

Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1793.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime reference guide*. Psychology Software Tools, Incorporated.

Schwanenflugel, P. J. (2013). Why are abstract concepts hard to understand? *The psychology of word meanings* (pp. 235–262). Psychology Press.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, *27*(5), 499.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 82.

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*(4), 489–537.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*(2), 393–413. http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.arcs.ncounts.html

Shaoul, C., & Westbury, C. (2013). A reduced redundancy USENET corpus (2005-2011). http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html

Sidhu, D. M., Heard, A., & Pexman, P. M. (2016). Is more always better for verbs? Semantic richness effects and verb meaning. *Frontiers in Psychology*, *7*, 798.

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1280.

Stella, M., Beckage, N. M., & Brede, M. (2017). Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports*, *7*(1), 1–10.

Taler, V., Kousaie, S., & Lopez Zunini, R. (2013). Erp measures of semantic richness: The case of multiple senses. *Frontiers in Human Neuroscience*, *7*, 5.

Ten Bosch, L., Boves, L., & Ernestus, M. (2022). Diana, a process-oriented model of human auditory word recognition. *Brain Sciences*, *12*(5), 681.

ten Bosch, L., Boves, L., & Ernestus, M. (2015). DIANA, an end-to-end computational model of human word comprehension. *The 18th International Congress of Phonetic Sciences (ICPhS 2015)*.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The massive auditory lexical decision (MALD) database. *Behavior Research Methods*, *51*(3), 1187–1204.

Tyler, L. K., Moss, H. E., Galpin, A., & Voice, J. K. (2002). Activating meaning in time: The role of imageability and form-class. *Language and Cognitive Processes*, *17*(5), 471–502.

Tyler, L. K., Voice, J. K., & Moss, H. E. (2000). The interaction of meaning and sound in spoken word recognition. *Psychonomic Bulletin & Review*, *7*(2), 320–326.

Van de Ven, M., Tucker, B. V., & Ernestus, M. (2011). Semantic context effects in the comprehension of reduced pronunciation variants. *Memory & Cognition*, *39*(7), 1301–1316.

Van Petten, C. (2014). Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence. *International Journal of Psychophysiology*, *94*(3), 407–419.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2020a). itsadug: Interpreting time series and autocorrelated data using gamms [R package version 2.4].

van Rij, J., Vaci, N., Wurm, L. H., & Feldman, L. B. (2020b). Alternative quantitative methods in psycholinguistics: Implications for theory and design. *Word Knowledge and Word Usage*, *83*.

Vincent, S. B. (1912). *The functions of the vibrissae in the behavior of the white rat* (Vol. 1). University of Chicago.

Vitevitch, M. S. (2021). What can network science tell us about phonology and language processing? *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12532 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12532

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, *45*(4), 1191–1207.

Weide, R. (2005). The Carnegie Mellon pronouncing dictionary [cmudict. 0.6]. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Westbury, C., & Moroschan, G. (2009). Imageability x phonology interactions during lexical access: Effects of modality, phonological neighbourhood, and phonological processing efficiency. *The Mental Lexicon*, *4*(1), 115–145.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of english. *Journal of Phonetics*, *70*, 86–116.

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S. N., & Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, *59*, 122–143.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36.

Wurm, L. H. (1996). Dimensions of speech perception: Semantic associations in the affective lexicon. *Cognition & Emotion*, *10*(4), 409–424.

Wurm, L. H., & Vakoch, D. A. (2000). The adaptive value of lexical connotation in speech perception. *Cognition & Emotion*, *14*(2), 177–191.

Wurm, L. H., Vakoch, D. A., & Seaman, S. R. (2004). Recognition of spoken words: Semantic effects in lexical access. *Language and Speech*, *47*(2), 175–204.

Wurm, L. H., Whitman, R. D., Seaman, S. R., Hill, L., & Ulstad, H. M. (2007). Semantic processing in auditory lexical decision: Ear-of-presentation and sex differences. *Cognition and Emotion*, *21*(7), 1470–1495.

Yap, M. J., Lim, G. Y., & Pexman, P. M. (2015). Semantic richness effects in lexical decision: The role of feedback. *Memory & Cognition*, *43*(8), 1148–1167.

Yap, M. J., Pexman, P. M., Wellsby, M., Hargreaves, I. S., & Huff, M. (2012). An abundance of riches: Cross-task comparisons of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience*, *6*, 72.

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*(4), 742–750.

You, H., & Magnuson, J. S. (2018). Tisk 1.0: An easy-to-use python implementation of the time-invariant string kernel model of spoken word recognition. *Behavior Research Methods*, *50*(3), 871–889.

Zdrazilova, L., & Pexman, P. M. (2013). Grasping the invisible: Semantic processing of abstract words. *Psychonomic Bulletin & Review*, *20*(6), 1312–1318.

Zhuang, J., Randall, B., Stamatakis, E. A., Marslen-Wilson, W. D., & Tyler, L. K. (2011). The interaction of lexical semantics and cohort competition in spoken word recognition: An fMRI study. *Journal of Cognitive Neuroscience*, *23*(12), 3778–3790.

Table 1

The best (second) of the four hierarchical linear regression models predicting item-level standardized reaction times in Study 1.

Variable	β	p-value
Duration	.67	$< .001$
Log subtitle frequency	-.17	$< .01$
Phonological uniqueness point	.11	.26
Structural principal component	-35	$< .01$
Number of morphemes	-.03	.41
Concreteness	-.16	$< .001$
Valence	-.09	$< .05$
Arousal	-.07	.11
Number of semantic features	-.15	$< .001$
Semantic neighborhood density	.06	.36
Semantic diversity	-.08	.09

Table 2

The generalized additive mixed effects model fit for response latency in Study 1. This model includes only the statistically significant predictors. Predictors that are marked with "s" represent smoothed predictors.

Parametric coefficients	Estimate	Std. Err.	t-value	Pr(>|t|)
(Intercept)	-1.28	0.02	-52.44	< .001
Smooth terms	**edf**	**Ref.df**	**F**	**p-value**
s(Trial, Subject)	117.28	215.00	11.47	< .001
s(Item)	238.32	356.00	2.01	< .001
s(Trial)	5.56	6.43	3.34	< .01
s(Duration)	1.00	1.00	205.37	< .001
s(Log subtitle frequency)	1.99	2.17	7.86	< .001
s(Structural principal component)	1.00	1.00	20.51	< .001
s(Concreteness)	1.00	1.00	12.80	< .001
s(Valence)	1.00	1.00	5.12	< .05
s(Number of semantic features)	1.00	1.00	12.50	< .001

Table 3

The generalized additive mixed effects model fit for response latency in Study 1. This model includes only the predictors whose exclusion significantly reduces model fit. Predictors that are marked with "s" represent smoothed predictors.

Parametric coefficients	Estimate	Std. Err.	t-value	Pr(>|t|)
(Intercept)	-1.28	0.02	-52.42	< .001
Smooth terms	**edf**	**Ref.df**	**F**	**p-value**
s(Trial, Subject)	126.70	215.00	11.79	< .001
s(Item)	245.20	358.00	2.14	< .001
s(Duration)	1.00	1.00	198.71	< .001
s(Log subtitle frequency)	1.00	1.00	27.40	< .001
s(Structural principal component)	1.00	1.00	24.35	< .001
s(Concreteness)	1.00	1.00	20.26	< .001

Table 4

The generalized additive mixed effects model fit for response latency in Study 2. Predictors that are marked with "s" represent smoothed predictors and those marked with "ti" represent a smooth for an interaction. Names of significant predictors are given in bold.

Parametric coefficients	Estimate	Std. Err.	z-value	Pr(|z|)
(Intercept)	-1.16	0.01	-208.99	< .001
PoS Adjective	-0.01	0.00	-4.94	< .001
PoS Verb	-0.00	0.00	-1.04	.303
Smooth terms	edf	Ref.df	Chi.sq	p-value
s(Trial, Subject)	1269.14	3590.00	5.94	< .001
s(Trial)	6.46	7.50	34.62	< .001
s(Duration)	5.11	6.23	769.01	< .001
s(Log subtitle frequency)	3.12	3.94	62.93	< .001
s(Phonological uniqueness point)	4.89	5.47	4.73	< .01
s(Structural principal component)	5.71	6.97	41.32	< .001
s(Number of morphemes)	1.88	1.99	14.75	< .001
s(Concreteness):PoS Noun	4.75	5.79	13.66	< .001
s(Concreteness):PoS Adjective	1.00	1.00	18.35	< .001
s(Concreteness):PoS Verb	1.00	1.00	0.33	.565
s(Valence):PoS Noun	4.76	5.85	12.84	< .001
s(Valence):PoS Adjective	3.32	4.16	7.09	< .001
s(Valence):PoS Verb	3.04	3.83	4.01	< .01
s(Arousal):PoS Noun	2.04	2.59	4.56	< .01
s(Arousal):PoS Adjective	3.15	3.96	3.28	< .05
s(Arousal):PoS Verb	1.00	1.00	0.89	.347
s(Semantic neighborhood density):PoS Noun	1.95	2.48	12.02	< .001
s(Semantic neighborhood density):PoS Adjective	1.00	1.00	7.38	< .01
s(Semantic neighborhood density):PoS Verb	1.00	1.00	0.32	.573
s(Semantic diversity):PoS Noun	1.00	1.00	0.22	.639
s(Semantic diversity):PoS Adjective	3.04	3.89	4.03	< .01
s(Semantic diversity):PoS Verb	1.00	1.00	0.00	.990
ti(Valence, Arousal):PoS Noun	1.00	1.00	0.34	.563
ti(Valence, Arousal):PoS Adjective	1.00	1.00	8.14	< .01
ti(Valence, Arousal):PoS Verb	1.23	1.42	0.87	.465

Table 5

The generalized additive mixed effects model fit for response latency in Study 2. This model includes only the predictors whose exclusion significantly reduces model fit. Predictors that are marked with "s" represent smoothed predictors and those marked with "ti" represent a smooth for an interaction.

Parametric coefficients	Estimate	Std. Err.	z-value	Pr(|z|)
(Intercept)	-1.16	0.01	-209.27	$< .001$
PoS Adjective	-0.01	0.00	-5.27	$< .001$
PoS Vers	-0.01	0.00	-2.86	$< .01$
Smooth terms	**edf**	**Ref.df**	**Chi.sq**	**p-value**
s(Trial, Subject)	1268.48	3590.00	5.94	$< .001$
s(Trial)	6.48	7.52	34.56	$< .001$
s(Duration)	5.05	6.17	777.95	$< .001$
s(Log subtitle frequency)	3.07	3.87	72.86	$< .001$
s(Structural principal component)	6.53	7.70	65.90	$< .001$
s(Number of morphemes)	1.90	1.99	15.44	$< .001$
s(Concreteness)	5.46	6.58	11.55	$< .001$
s(Valence)	4.84	5.92	18.32	$< .001$
s(Arousal)	2.75	3.49	6.10	$< .001$
s(Semantic neighborhood density)	1.23	1.42	16.18	$< .001$

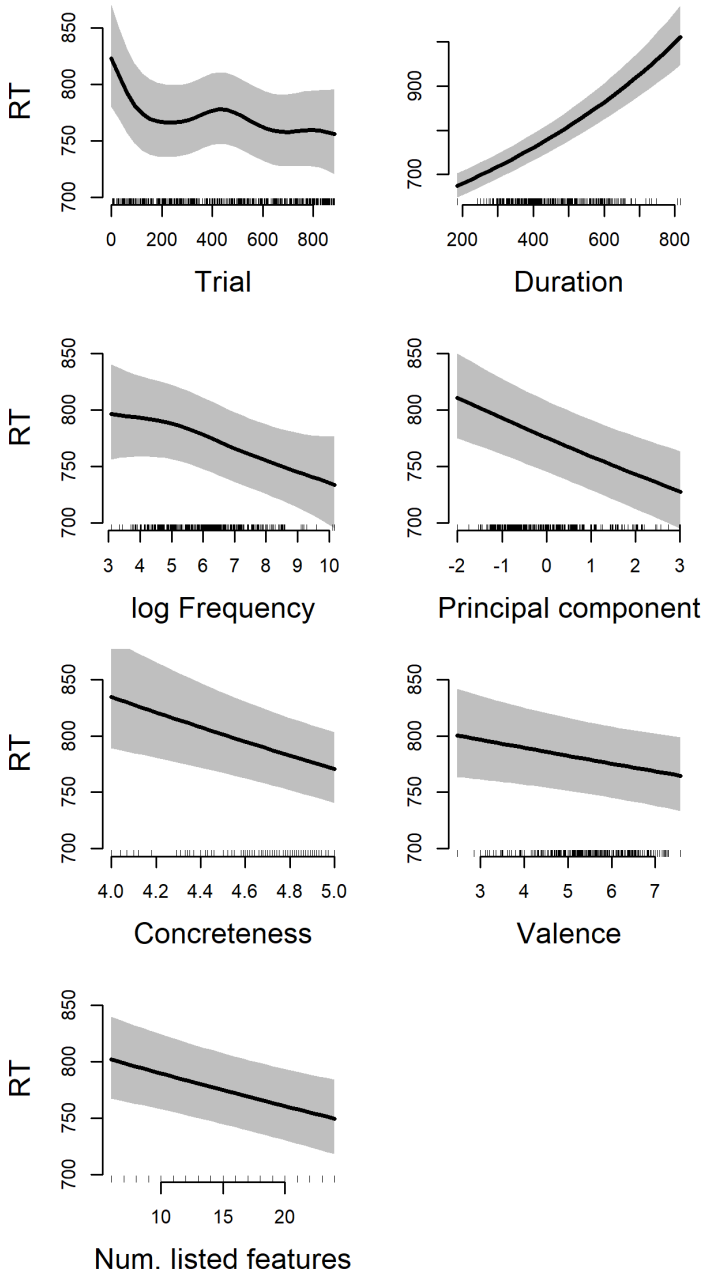

Figure 1

Effects plots for the GAM model that includes all the significant predictors in Study 1. Back-transformed response latency is on the y-axis, while different predictors are shown on the x-axis. Y-axes are all limited to the range between 700 and 870 ms to enable the comparison of effect size between predictors. The exception to this rule is duration, where the y-axis range used is much wider to encompass the large recorded effect of duration.

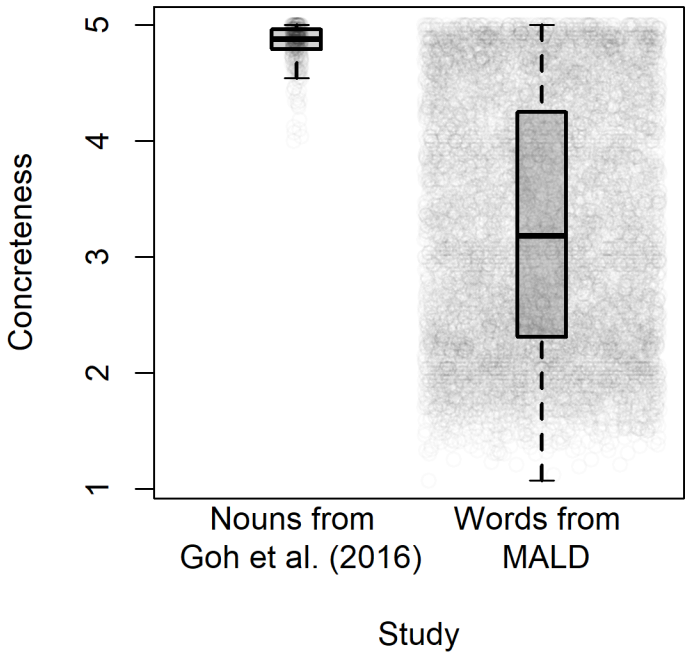

Figure 2

Distributions of word concreteness in the 442 concrete nouns from Goh et al. (2016), and in the 8,626 MALD words used in Study 2. Translucent circles represent separate observations with jitter introduced for better visibility.

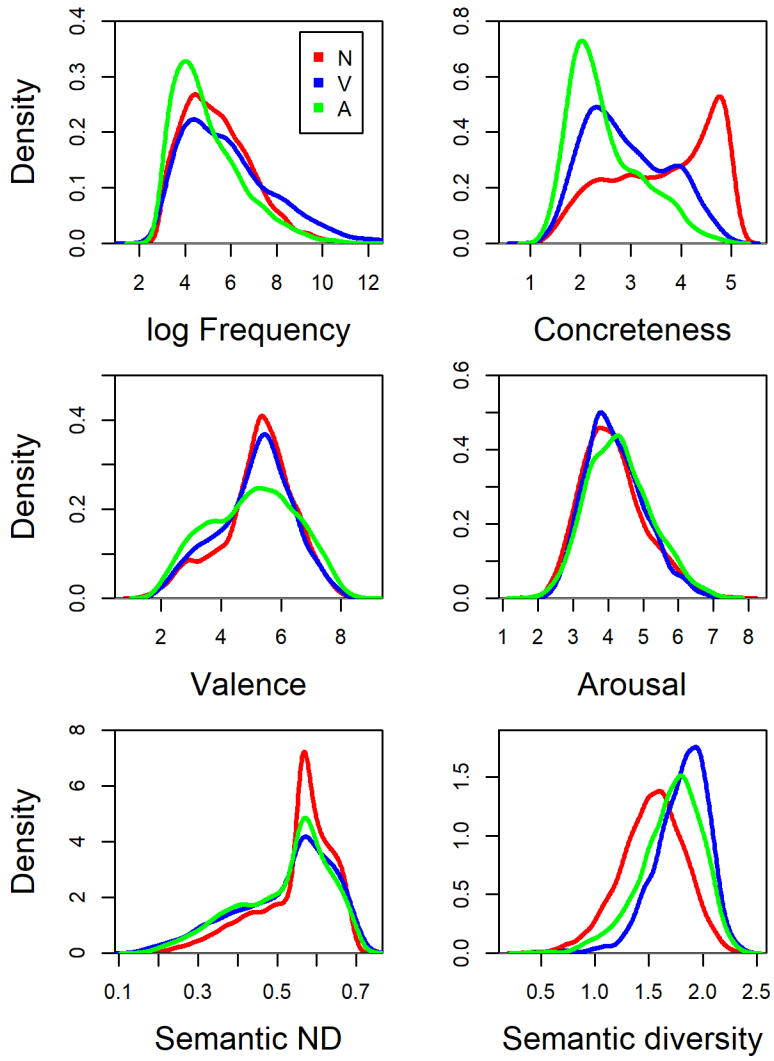

Figure 3

Density distributions for logged word frequency and the five semantic richness variables used in Study 2. Separate lines distinguish distributions by part of speech classes: nouns are in red, verbs are in blue, and adjectives are in green.

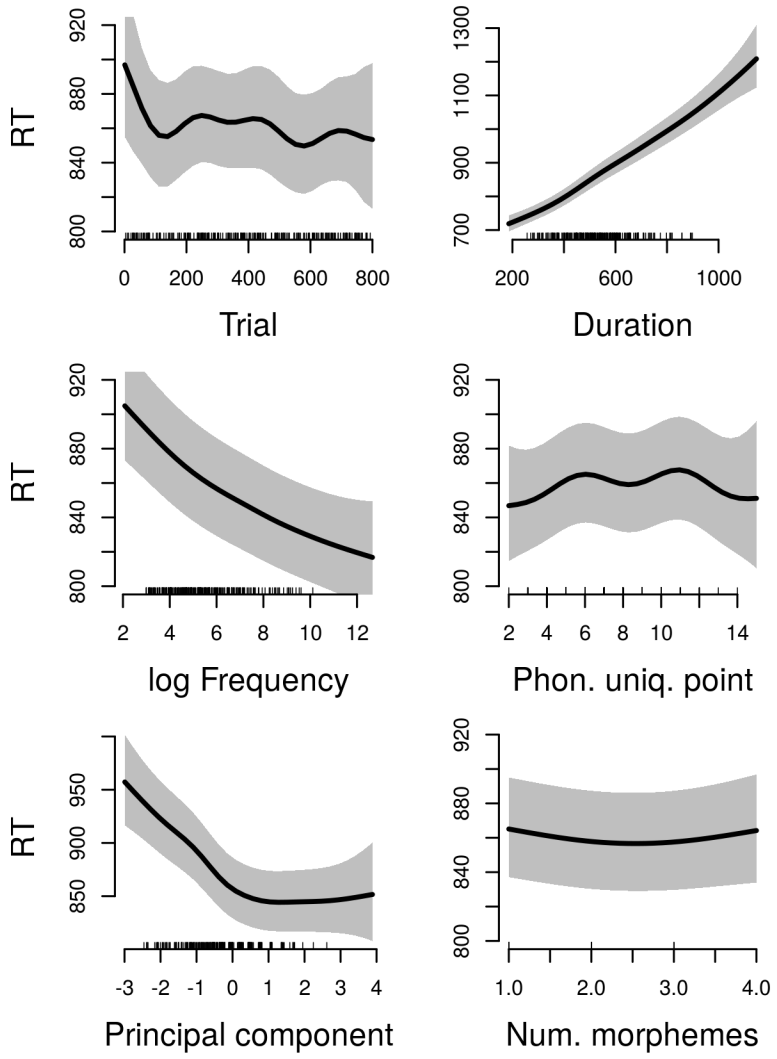

Figure 4

Effects plots of control variables for the GAM model that includes all the significant predictors in Study 1. Back-transformed response latency is on the y-axis, while different predictors are shown on the x-axis. Y-axes are all limited to the range between 800 and 920 ms to enable the comparison of effect size between predictors. The exceptions to this rule are duration and the structural principal component, where the y-axis range used is wider to encompass their effects.

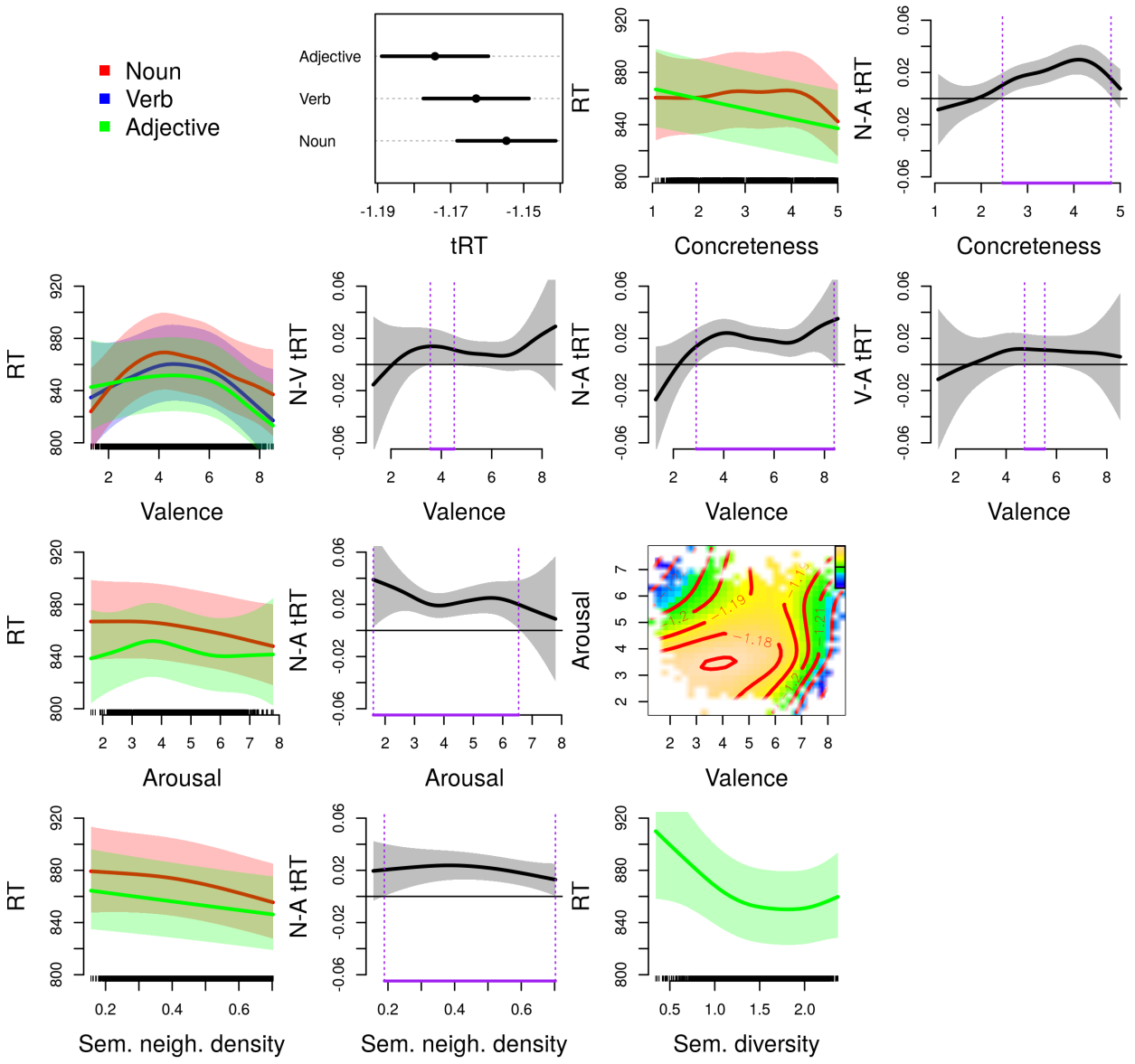

Figure 5

Effects plots for part of speech and semantic richness variables in the GAM model that includes all the significant predictors in Study 2. The plots are described in the text.

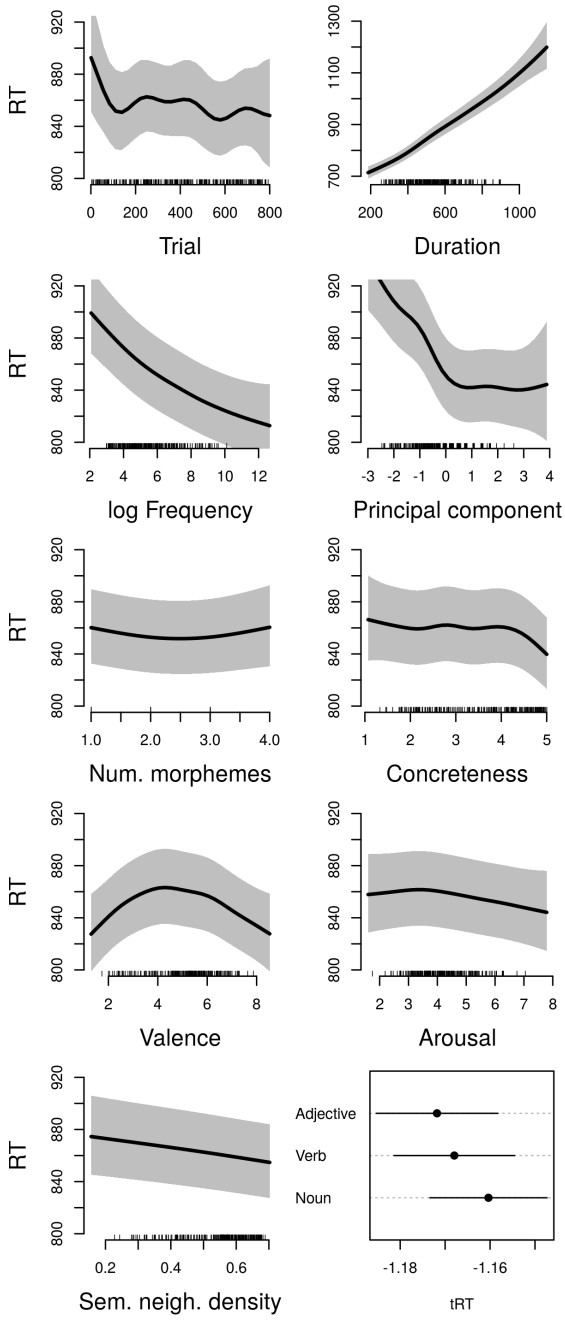

Figure 6

Effects plots for the GAM model that includes all the predictors whose exclusion would reduce model fit in Study 2. Back-transformed response latency is on the y-axis, while different predictors are shown on the x-axis. Y-axes are all limited to the range between 780 and 940 ms to enable the comparison of effect size between predictors. The exception to this rule is duration, where the y-axis range used is much wider to encompass the large recorded effect of duration.