**UNIVERSITY COLLEGE LONDON (UCL)**

# Computational explorations of semantic cognition

PHD THESIS

**Armand Stefan Rotaru**

**Supervisors:**

**Primary: Prof. Gabriella Vigliocco**

**Secondary: Prof. Lewis Griffin**

I, Armand Stefan Rotaru, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Name: Armand Stefan Rotaru Date: 25/07/2020

Signature:

# Abstract

Motivated by the widespread use of distributional models of semantics within the cognitive science community, we follow a computational modelling approach in order to better understand and expand the applicability of such models, as well as to test potential ways in which they can be improved and extended.

We review evidence in favour of the assumption that distributional models capture important aspects of semantic cognition. We look at the models' ability to account for behavioural data and fMRI patterns of brain activity, and investigate the structure of model-based, semantic networks.

We test whether introducing affective information, obtained from a neural network model designed to predict emojis from co-occurring text, can improve the performance of linguistic and linguistic-visual models of semantics, in accounting for similarity/relatedness ratings. We find that adding visual and affective representations improves performance, especially for concrete and abstract words, respectively.

We describe a processing model based on distributional semantics, in which activation spreads throughout a semantic network, as dictated by the patterns of semantic similarity between words. We show that the activation profile of the network, measured at various time points, can account for response time and accuracies in lexical and semantic decision tasks, as well as for concreteness/imageability and similarity/relatedness ratings.

We evaluate the differences between concrete and abstract words, in terms of the structure of the semantic networks derived from distributional models of semantics. We examine how the structure is related to a number of factors that have been argued to differ between concrete and abstract words, namely imageability, age of acquisition, hedonic valence, contextual diversity, and semantic diversity.

We use distributional models to explore factors that might be responsible for the poor linguistic performance of children suffering from Developmental Language Disorder. Based on the assumption that certain model parameters can be given a psychological interpretation, we start from "healthy" models, and generate "lesioned" models, by manipulating the parameters. This allows us to determine the importance of each factor, and their effects with respect to learning concrete vs abstract words.

# Impact statement

Distributional models of semantics have become essential tools in the study of semantic memory. In this context, in Chapter 1 we describe the original research included in our dissertation, namely examining the cognitive plausibility of distributional models, testing effective ways in which their performance in predicting behavioural data can be improved, looking at factors that can influence model-based semantic networks, and showing the potential of using distributional models to better understand semantic deficits.

In Chapter 2, we discuss some of the evidence in favour of the cognitive plausibility of distributional models. To the best of our knowledge, this is the first study that reviews the use of behavioural data and brain imaging in evaluating distributional models, from a psychological perspective. By highlighting limitations in the how the models were tested, this chapter can serve as a source of inspiration for further studies concerned with assessing the degree to which distributional models can provide a better understanding of semantic cognition.

In Chapter 3, we evaluate the effects of adding affective information to linguistic and linguistic-visual models of semantics, in order to improve their psychological plausibility. Our approach demonstrates the utility of employing distributional models of emotion to better capture semantic task performance, while also identifying contexts where including affective information can have detrimental effects. We also show that distributional models of emotion, trained on text-emoji co-occurrence patterns, can capture rich affective information for a very large number of words. Thus, our research suggests an attractive alternative to the use of traditional representations of emotion.

In Chapter 4, we examine the benefits of bringing together the connectionist and distributional approaches to studying semantic cognition. We advocate a more natural method of investigating semantics, by acknowledging the tight interplay between representations and processes. To do so, we test the effects of adding a spreading activation mechanism to structural, distributional models of semantics. The results show that considering the dynamics of semantic processing significantly increases the performance of the distributional models in predicting behavioural data,

across various tasks and datasets. The resulting dynamic models can be used to improve the performance of the structural models upon which they are based.

In Chapter 5, we investigate certain differences between concrete and abstract words, as reflected in the structure of semantic networks derived from distributional models. We take an in-depth look at some of the factors that might shape the structure of model-based semantic networks. This kind of analysis also allows us to test the predictions of several different theories of semantics.

In Chapter 6, we employ distributional models in order to explore potential causes behind a specific semantic impairment (i.e., Developmental Language Disorder). We show that it is possible to “lesion” certain distributional models, in a manner that is cognitively meaningful, by suggesting a psychological interpretation for model parameters. Such an approach has been employed in the past for connectionist models, but, as far as we are aware, not in the context of distributional models. Our approach serves as proof of concept.

# Acknowledgements

First and foremost I would like to thank my doctoral advisors, Professors Gabriella Vigliocco and Lewis Griffin, for the numerous insightful discussions that have shaped my current research, and for offering me a stimulating intellectual and emotional environment. Throughout the doctoral programme, their patience, goodwill and optimism have been invaluable to me. Furthermore, their firm belief in my potential as an aspiring scientist has helped me overcome all the setbacks and delays encountered along the way. Moreover, I am very grateful to Doctors Stefan Frank and Alessandro Lenci, for their fruitful collaboration and kind, constant help.

I would also like to thank the members of my doctoral committee, Doctor Maarten Speekenbrink and Professor Max Louwerse, for the time and effort invested in carefully reviewing my thesis, which resulted in both a very informative viva and a considerable number of helpful suggestions.

In addition, I owe a debt of gratitude to my colleagues at the Language and Cognition Laboratory, with whom I have had many pleasant conversations and to whom I have often turned for practical advice. Finally, I wish to thank my family and friends for their unwavering support and encouragement.

# Table of contents

# 1. Introduction

Over the past 25 years, progress in the study of semantic cognition has been accelerated by both the availability of extensive behavioural norms (Johns, Jamieson, & James, 2020), and the development of a wide range of new computational models, trained on very large text corpora. Comprehensive norms have been collected for semantic tasks such as free association (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019; Nelson, McEvoy, & Schreiber, 2004), similarity judgement (Bruni, Tran, & Baroni, 2014; Silberer & Lapata, 2014), feature generation (McRae, Cree, Seidenberg, & McNorgan, 2005; Vinson & Vigliocco, 2008), and semantic priming (Hutchison et al., 2013). In addition, extensive norms have been obtained for tasks that rely primarily on orthographic and phonological processing, but also include a semantic component, such as lexical decision (Balota et al., 2007; Keuleers, Lacey, Rastle, & Brysbaert, 2012), and speeded naming (Balota et al., 2007).

This wealth of data, together with advances in computational linguistics, has made it possible for researchers to develop and test increasingly more sophisticated distributional models of semantic memory. The theoretical foundation behind these models is the "distributional hypothesis", introduced by Harris (1954), which claims that "words that occur in similar contexts tend to have similar meanings" (Turney & Pantel, 2010). In other words, the collection of linguistic contexts in which a particular word occurs reflects important aspects of that word's meaning, such that commonalities in meaning between two words can be identified and even quantified by evaluating the overlap between the distributions of contexts associated with each word. For instance, the words "cat" and "dog" both frequently appear in linguistic contexts containing the words "animal", "pet", "furry", "house", and "vet", which suggests that they are similar in meaning; in contrast, the words "vacation" and "longbow" are usually encountered in very different linguistic contexts, which makes it likely that they are semantically dissimilar. Since the distributional hypothesis does not define context in a precise manner, certain models (i.e., "document-as-context" models; e.g, Latent Semantic Analysis; LSA; Landauer & Dumais, 1997; Topic; Griffiths, Steyvers, & Tenenbaum, 2007) assume that the context consists of the documents in which a given word occurs, whereas other models (i.e., "word-as-

context" models; e.g, Hyperspace Analogue to Language; HAL; Lund & Burgess, 1996; Skip-gram; Mikolov, Chen, Corrado, & Dean, 2013) consider that the words immediately following or preceding a given word make up the context for that word. Within the area of linguistic models, a number of studies have attempted a systematic exploration of how to best extract semantic information from linguistic contexts, by optimising the various parameters that influence the underlying semantic model (Bullinaria & Levy, 2007, 2012; Lapesa & Evert, 2014; Levy, Goldberg, & Dagan, 2015), such as the size of the linguistic corpus, the dimensionality of the semantic representations, the relative importance of each dimension, and the measure of semantic distance. Other studies have explored that benefits of including information about word order (Jones & Mewhort, 2007; Andrews & Vigliocco, 2010), syntactic dependencies (Padó & Lapata, 2007), and types of semantic relations (e.g., hypernymy; Baroni, Murphy, Barbu, & Poesio, 2010).

Moving beyond purely linguistic models, several researchers have looked at means of enriching semantic representations, by combining verbal and perceptual (i.e., visual) information. This approach is inspired by several embodied theories of cognition in which the semantic system is considered to rely on both modal and amodal representations (Barsalou, Santos, Simmons, & Wilson, 2008; Louwerse & Jeuniaux, 2008; Vigliocco, Meteyard, Andrews, & Kousta, 2009). Studies following this approach (for reviews, see Baroni, 2016; Bruni et al., 2014) have shown that integrating information from two modalities provides a better account of behavioural data than that offered by the individual modalities (see Chapter 2), across a wide range of models and integration methods, in tasks such as free association (Hill & Korhonen, 2014; Silberer & Lapata, 2012, 2014), similarity/relatedness judgement (Bruni et al., 2014; Hill, Reichart, & Korhonen, 2014; Kiela & Bottou, 2014; Silberer & Lapata, 2014), and categorization (Bruni, Tran, & Baroni, 2011; Silberer & Lapata, 2014). Moreover, it has been found that even for abstract concepts, such as "peace" and "freedom", the addition of (indirect) perceptual information improves the fit of the models to the human data (Bruni et al., 2014; Hill & Korhonen, 2014; Hill et al., 2014). The results are consistent with those of previous studies (Andrews, Vigliocco, & Vinson, 2009; Louwerse, 2011, 2018; Maki & Buchanan, 2008; Riordan & Jones, 2011; Sadeghi, McClelland, & Hoffman, 2015; Steyvers, 2010), indicating that language and

perception can be seen as complementary, yet highly redundant, sources of behaviourally relevant information.

The approaches presented so far examined semantics at the level of individual words, for instance, by looking at the contribution of linguistic and extralinguistic information, as well as at the processes thorough which semantic representations might be learned. In contrast, rather than focusing on individual representations, another prominent research trend investigated the patterns of semantic relations that link the individual representations within semantic memory. According to this perspective, the mental lexicon can be regarded as a network (Collins & Loftus, 1975), whose nodes are words, such that two nodes are connected by an edge whenever a certain type of semantic relation exists between the words associated with the nodes. As a result, network analyses of semantic networks have attracted an increasing amount of attention in recent years (for reviews on applications of network science in the study of language, see Borge-Holthoefer & Arenas, 2010; Choudhury & Mukherjee, 2009; Cong & Liu, 2014; Mehler, 2008; Solé, Corominas-Murtra, Valverde, & Steels, 2010; for general reviews of network-based analyses of cognition, see Baronchelli, Ferrer i Cancho, Pastor-Satorras, Chater, & Christiansen, 2013; Siew, Wulff, Beckage, & Kenett, 2019).

Some of those studies adopted a macroscopic view of semantic network structure, by taking a global perspective, for instance when testing whether a particular network has a number of specific structural properties (Ferrer i Cancho & Solé, 2001; Sigman & Cecchi, 2002; Steyvers & Tenenbaum, 2005; De Deyne & Storms, 2008b). This global perspective has shown that semantic networks exhibit a highly non-random, “small-world” structure (Watts & Strogatz, 1998), characterised by a combination of neighbourhood sparseness, high local clustering, and small average distance between nodes. The same approach has revealed structural commonalities across different types of semantic networks (e.g., those derived from free association vs word co-occurrence data; Steyvers & Tenenbaum, 2005; Griffiths, Steyvers, & Tenenbaum, 2007), and has shed some light on the developmental processes that shape the organization of semantic memory (e.g., preferential vs random attachment of new nodes; Gruenenfelder, Recchia, Rubin, & Jones, 2015; Utsumi, 2015).

In contrast, other studies opted for a microscopic view of semantic networks, which involves investigating the local characteristics of various semantic subnetworks.

These studies have linked network properties to observable task behaviour, by demonstrating that the size and interconnectivity of a word's local semantic neighbourhood are significant predictors of response times and accuracies in semantic tasks such as categorization and synonymy judgement, as well as in tasks where semantic processes play a secondary role, such as standard and go/no-go lexical decision, speeded naming, and artificial language learning (for reviews, see Jones, Johns, & Recchia, 2012; Mirman & Magnuson, 2008; Newcombe, Campbell, Siakaluk, & Pexman, 2012; Recchia & Jones, 2012; Yap, Pexman, Wellsby, Hargreaves, & Huff, 2012; Zdrazilova & Pexman, 2013). Moreover, other studies found support for the existence of representational differences between concrete and abstract words, whereby concrete concepts are richer than abstract ones when it comes to perceptual and motor elements, but poorer with respect to introspective and linguistic elements (for reviews, see Gee, Nelson, & Krawczyk, 1999; Hill, Korhonen, et al., 2014; Pecher, Boot, & Van Dantzig, 2011; Speed, Vinson, & Vigliocco, 2015; Vinson, Ponari, & Vigliocco, 2014; Wiemer-Hastings & Xu, 2005). These semantic richness effects have been shown to be both task-general and task-specific, to have both an early and a late impact on task behaviour (Hargreaves & Pexman, 2014), and to either facilitate or hinder task performance (Mirman & Magnuson, 2008).

In the context of current research on distributional models, the work presented in our dissertation covers three main topics. Firstly, we both review some of the most representative studies that that have directly or indirectly examined the cognitive plausibility of distributional models (Chapter 2), and demonstrate how the ability of such models to predict behavioural data can be improved by incorporating rich emotional information (Chapter 3). Secondly, we explore novel directions in the study of model-based semantic networks, by showing how automatic semantic processing can employ network structure during task performance (Chapter 4), and by investigating how several psychological factors such as word frequency and concreteness, relate to the previously mentioned structure (Chapter 5). Thirdly, we attempt to obtain a better understanding of the causes behind a prevalent semantic impairment (i.e., Developmental Language Disorder), by testing the effects of damaging various components of state-of-the-art models of semantics (Chapter 6). A more detailed description of the contents of the dissertation is given below.

In Chapter 2, we review evidence in favour of the assumption that distributional models, both unimodal and multimodal, capture important aspects of semantic cognition. We first look at the models' ability to account for behavioural data from a variety of semantic tasks, with a focus on the similarity/relatedness rating, free association, and semantic categorization tasks. We discuss both the main findings and certain limitations for each task. Then, we examine the degree to which networks derived from free association norms, and networks derived from distributional models, exhibit a "small-word" and "scale-free" structure, characteristic of many real-world networks. Finally, we investigate whether fMRI patterns of brain activity, observed during semantic tasks, can be predicted by employing distributional models. Overall, we find that distributional models have a relatively high degree of cognitive plausibility, but also that there are several factors which can influence model performance with respect to a given task, such as the nature of task, the properties of the stimuli (most notably, word concreteness), the architecture of the model (e.g., the definition of linguistic context that it employs), and the type of information (i.e., linguistic vs multimodal) to which the model has access, among other variables.

In Chapter 3, we start from the observation that emotion plays an important role in learning and representing abstract concepts, based on empirical work. However, distributional models integrating emotion along with linguistic and visual information are lacking. Therefore, we first improve on visual and affective representations, derived from state-of-the-art existing models, by choosing models that best fit available human semantic data and extending their coverage. Crucially then, we assess whether adding affective representations (obtained from a neural network model designed to predict emojis from co-occurring text) to purely linguistic or linguistic-visual models, enhances their ability to fit semantic similarity/relatedness judgements. We find that, given specific weights assigned to the models, adding both visual and affective representations contribute significantly to performance, with visual representations providing an improvement especially for more concrete words, and affective representations increasing the fit especially for more abstract words.

In Chapter 4, we bring together two relatively independent research directions in the computational modelling of semantics, one concentrating on semantic representations (i.e., the distributional approach), and the other concentrating on semantic processes (i.e., the connectionist approach). In order to fruitfully combine the

two approaches, we put forward a processing model based on distributional semantics in which activation spreads throughout a semantic network, as dictated by the patterns of semantic similarity between words. We then show that the activation profile of the network, measured at various time points, can successfully account for response times in lexical and semantic decision tasks, as well as for subjective concreteness and imageability ratings. We also demonstrate that the dynamics of the network is predictive of performance in relational semantic tasks, such as similarity/relatedness rating. Our results indicate that simulating the spread of activation within distributional semantic networks provides a good fit to both automatic lexical processing (as indexed by lexical and semantic decisions), as well as more deliberate processing (as indexed by ratings), above and beyond what has been reported for previous models that take into account only similarity resulting from network structure.

In Chapter 5, we evaluate some of the differences between concrete and abstract words, in terms of the structure of the semantic networks derived from distributional models of semantics. Our investigation is driven by the observation that our current picture of the network differences between concrete and abstract words, as well as of the behavioural consequences of these differences, is very unclear. For instance, interpretations based on imageability (Paivio, 1971, 1986) and contextual/semantic diversity (Hoffman, Lambon Ralph, & Rogers, 2013; Jones et al., 2012) predict mutually incompatible trends with respect to the number of neighbours for concrete and abstract words. Therefore, in order to better empirically test such theories of semantic representation, we investigate some of the relations between neighbourhood structure and a number of important factors, such as imageability, age of acquisition, emotional valence, contextual diversity, and semantic diversity, that differ between concrete and abstract words. We find that that concrete and abstract words are very similar in terms of number of neighbours, but the local neighbourhoods of concrete words are considerably more interconnected than those of abstract words. Moreover, the strongest predictors of local neighbourhood size and structure are contextual and semantic diversity, followed by concreteness and age of acquisition, while emotional valence has little effect.

In Chapter 6, we use distributional models to explore factors that might be responsible for the poor performance of children suffering from DLD, in linguistic tasks. In the clinical literature, such poor performance has been attributed to impairments in

working memory, statistical learning, and attention, among other factors. Our computational experiments are based on the assumption that certain model parameters can be given a clear psychological interpretation: for instance, the size of the sliding window is linked to the capacity of verbal working memory, whereas the learning rate is related to the efficiency and precision of the statistical learning mechanisms involved in the acquisition of word meanings from exposure to language. We start from a “healthy” model, using parameter values that have been shown to provide the best fit to behavioural data and/or are recommended by the authors of the models, and then generate several “lesioned” models, by factorially manipulating the parameters of the “healthy” model. By measuring the degree of impairment in terms of the dissimilarity between the representations produced by the “healthy” and “lesioned”, we find that certain factors (e.g., learning rate) have a large impact, while other factors (e.g., sliding window size) have a small or negligible impact. In addition, the effect of each factor is independent of word concreteness. This finding is consistent with the results of a recent behavioural study, conducted in our lab, showing that children with DLD have a similar level of impairment for both concrete and abstract words, which suggests that the role of linguistic input is comparable in the learning of abstract vs concrete words.

In Chapter 7, which concludes the dissertation, we provide a summary of the original research presented in our dissertation, and of what it tells us with respect to the nature of semantic cognition.

# 2. The psychological plausibility of distributional models

## 2.1. Introduction

In the past 30 years, distributional models of semantics have become very popular tools for developing semantic representations (for reviews, see Clark, 2015; Erk, 2012; Jones, Willits, & Dennis, 2015; Lenci, 2018; Turney & Pantel, 2010). Due to their versatility, such models have found applications in a number of fields, including psycholinguistics, clinical psychology, neuroscience, computational linguistics, and machine learning.

The use of distributional representations in the study of semantic cognition has a number of significant advantages over that of its main competitors, namely feature-based representations (McRae et al., 2005; Vinson & Vigliocco, 2008), and free association-based representations (De Deyne et al., 2019; Nelson et al., 2004). Firstly, distributional models can automatically learn semantic representations for an impressive number of words. For instance, the Global Vectors model (GloVe; Pennington, Socher, & Manning, 2014), trained over a corpus of 840 billion words, provides distributional representations for a total of over 2 million words. In contrast, the largest and most recent free association norms (De Deyne et al., 2019) and semantic feature norms (Buchanan, Valentine, & Maxwell, 2019) cover only slightly over 12,000 and 4,000 words, respectively, having been collected over a period of more than 7 years. Furthermore, since many large text corpora, pre-trained models, and well-documented model libraries are freely available, using distributional models requires a minimal investment of time and money.

Secondly, distributional models make use of texts that reflect ordinary linguistic experience (i.e., encyclopedia entries, news articles, classic literature, etc.), whereas free association responses and semantic features are elicited in non-naturalistic tasks: it is very unlikely that a typical person would routinely would often have to generate free associates (outside psychoanalytic therapy or brainstorming sessions) or to verbalize semantic features (unless describing an object unknown to someone).

Thirdly, having control over what and how the model learns allows researchers to easily test various hypotheses regarding semantic cognition. For instance, it is possible to compare semantic representations for two or more groups (e.g., children vs young adults vs older adults, monolinguals vs bilinguals, clinical vs healthy populations), by tailoring the corpora over which a model is trained (Johns, Jones, & Mewhort, 2019). Similarly, differences in semantic processing can be simulated by changing the values for the hyperparameters[1] of a model.

However, the potential advantages of using distributional models in psychological studies depends crucially on the cognitive plausibility of the models. In this chapter, we review a large number of representative studies, which tackle this issue in a direct or indirect manner. When choosing the studies described in the rest of the chapter, we followed a number of selection criteria, namely:

- besides studies where the models were tested with respect to their ability to account for behavioural data, which form the bulk of the chapter, we decided to also include both studies that analysed the properties of model-based semantic networks, and studies that examined the models' ability to predict fMRI activation patterns
- we focused only on tasks which rely primarily on semantic processing (i.e., similarity/relatedness rating, free association, and semantic categorization), as opposed to tasks where semantics plays a secondary or incidental role (e.g., lexical decision, naming; Recchia & Jones, 2012)
- where possible, we picked studies which used relatively large datasets (i.e., consisting of hundreds to thousands of individual words or word pairs), covering a diversity of parts of speech and concreteness levels
- we selected studies employing a wide variety of model architectures
- whenever possible, we included studies focusing on both unimodal models (i.e., linguistic, visual), and multimodal models (i.e., linguistic-visual)
- for each collective of researchers, we typically included only one or two studies, in order to reduce redundancy and provide a balanced picture

[1] A hyperparameter is a structural parameter of a model (e.g., window size, vocabulary size), which is not learned/optimized during the training process.

In the following subchapters, we review empirical evidence for (and against) the cognitive plausibility of distributional models, based on their (a) degree of success in fitting behavioural results from the similarity/relatedness rating, free association, and semantic categorization tasks, (b) ability of capture the structural properties of free association-based networks, and (c) performance in predicting fMRI activation patterns.

## 2.2. Comparing models using behavioural data

Perhaps the simplest and most direct method of evaluating the psychological plausibility of distributional models is to test whether they can successful account for behavioural data. A vast array of semantic tasks have be employed to this end (for a comprehensive review, see Bakarov, 2018), some of the best known being similarity/relatedness rating (Bruni et al., 2014; Gerz, Vulić, Hill, Reichart, & Korhonen, 2016), psycholinguistic property rating (e.g., valence, arousal, dominance, concreteness; Hollis, Westbury, & Lefsrud, 2017), semantic decision (Pexman, Heard, Lloyd, & Yap, 2017), semantic categorization (Riordan & Jones, 2011), free association (Cattle & Ma, 2017), semantic feature prediction (Johns & Jones, 2012), semantic priming (Mandera, Keuleers, & Brysbaert, 2017), synonymy judgement (Landauer & Dumais, 1997), and identification of lexical relations (Vylomova, Rimell, Cohn, & Baldwin, 2016).

In the rest of this chapter we extensively discuss only three tasks, namely similarity/relatedness rating, free association, and semantic categorization. We chose these particular tasks because each of them has been modeled in an extensive number of computational studies, employing several comprehensive datasets and a multitude of distributional models. Moreover, the cognitive representations and processes involved in performing the tasks are relatively well-understood (i.e., entire volumes have been dedicated to each task; e.g., Cramer, 1968; Hahn & Ramscar, 2001). Before we look at each individual task, we find it useful to review two very comprehensive and informative studies, which investigated task performance across

multiple tasks, datasets, and model architectures. Their main focus was to compare two prominent classes of distributional models, namely “count” and “predict” models. The former class consists of models that explicitly count co-occurences between a given word and the words that make up its linguistic contexts, and then apply certain transformations to these counts, in order to highlight particularly informative co-occurrences. In contrast, the latter class includes models which are trained to predict the linguistic contexts in which a given word appears (i.e., they extract co-occurrence data in an explicit, task-driven manner, typically using artificial neural networks).

The first study is that by Baroni and collaborators (2014). For the “count” models, the authors started from a representative model (Bullinaria & Levy, 2007, 2012), and created 36 versions of it, by varying the size of the sliding window, the dimensionality of the representations, the co-occurrence weighting function, and the dimensionality reduction technique (if any). For the “predict” models, the authors started from the Continuous Bag-of-Words model (CBOW; Mikolov, Chen, et al., 2013), and generated 48 versions of it, by changing the size of the sliding window, the dimensionality of the representations, the optimization techniques (i.e., hierarchical softmax vs negative sampling, with different numbers of negative samples), and the degree of subsampling for frequent words. The results showed the superiority of “predict” models, over “count” models, in all the tasks employed, namely similarity rating, synonymy judgement, semantic categorization, selectional preference rating, and analogy solving. This finding was shown to be remarkably robust, applying to setups which compared the best “predict” models and the best “count” models, on each task and across the tasks, and the worst such models, across the tasks. Moreover, the performance of the top “predict” models was comparable to, or better than, that of the state-of-the-art models, for each task.

The second study is that by Schnabel, Labutov, Mimno, and Joachims (2015). It used the same tasks, but focused on different types of “predict” models (i.e., CBOW; GloVe; C&W; Collobert & Weston, 2008), and “count” models (i.e., Hellinger Principal Component Analysis; Lebret & Collobert, 2014; Two Step Canonical Correlation Analysis; TSCCA; Dhillon, Rodu, Foster, & Ungar, 2012; Sparse Random Projections; Li, Hastie, & Church, 2006). Overall, the findings supported the conclusion of the previous study, such that, on average, the “predict” models strongly outperformed the “count” models, for all the tasks. However, within each class, the results indicated high

variability in performance, between the models: for the “predict” class, the CBOW model had a small advantage over the GloVe model, but a large advantage over the C&W model; for the “count” class, the TSCCA model outperformed the Hellinger PCA and Sparse Random Projections models, by a large margin, while the last two models had comparable levels of performance.

We now move on to describing more specific studies, by looking at the main findings in each study, after having introduced the relevant model architectures, training corpora, and behavioural datasets.

**Semantic similarity/relatedness rating:**

*General overview:*

Some of the main technical details regarding the studies reviewed in this subchapter are shown in Table 1.

Table 1. Computational studies focused on the similarity/relatedness rating task, covered in this subchapter. The entry for each study describes the linguistic and/or visual models employed (i.e., their type and the text/image corpora on which they are trained), as well as the behavioural datasets on which the models are tested.

| Study | Linguistic models | | Visual models | | Behavioural datasets |
|---|---|---|---|---|---|
| | Types | Corpora | Types | Corpora | |
| Hill, Reichart, & Korhonen (2015) | C&W<br>HAL<br>HSMN<br>Skip-gram | Various | - | - | MEN<br>SimLex-999<br>WordSim-353 |
| Gerz et al. (2016) | HAL<br>Skip-gram | ukWaC<br>Wikipedia | - | - | SimLex-999<br>SimVerb-3500 |
| Bruni, Boleda, Baroni, & Tran (2012) | DM<br>HAL<br>Topic | ukWaC<br>Wikipedia | SIFT | ESP Game<br>Pascal VOC | MEN<br>WordSim-353 |
| Bruni et al. (2014) | HAL | ukWaC<br>Wikipedia | SIFT | ESP Game | MEN<br>WordSim-353 |
| Kiela & Bottou (2014) | Skip-gram | BNC<br>Text8 | SIFT<br>AlexNet | ImageNet<br>ESP Game | MEN<br>WordSim-353 |
| Silberer & Lapata (2014) | Strudel | Wikipedia | Visual attributes | ImageNet | SL |

*Datasets:*

The norms of semantic similarity/relatedness data were collected by typically asking participants to rate the degree to which two words (e.g., “doctor”-“lawyer”) are similar/related in meaning. The datasets covered in our discussion are WordSim-353 (Finkelstein et al., 2001), SimLex-999 (Hill et al., 2015), SimVerb-3500 (Gerz et al., 2016), MEN (Bruni et al., 2014), and SL (Silberer & Lapata, 2014).

WordSim-353 consists of 350 noun pairs, rated from 0 (completely unrelated words) to 10 (very strongly related words). In a further study (Agirre et al., 2009), the dataset was divided into a similarity subset (i.e., 203 word pairs, classified as synonyms, antonyms, identical, or hyponym-hyperonym, as well as all the pairs with a rating below 5), and a relatedness subset (i.e., 250 word pairs, classified as meronym-holonym as well as all the pairs with a rating below 5) with some overlap between the two subsets[2].

SimLex-999 includes 666 noun pairs, 222 verb pairs, and 111 adjective pairs. Similarly to the case for WordSim-353, the pairs are rated on a scale from 0 to 10, but with the important difference that the participants were explicitly asked to evaluate only similarity, and not relatedness (i.e., the task instructs the raters to “consider how close the words are (or are not) to being synonymous”). Also, unlike other datasets, such as MEN and SL, which cover only concrete words, SimLex-999 is made up of pairs that span the whole spectrum of word concreteness.

SimVerb-3500 consists of 3,500 verb pairs, ranging from very concrete to very abstract, and including verbs belonging to more than 100 semantic classes (Kipper, Korhonen, Ryant, & Palmer, 2008). The methodology was very close to that employed in collecting the SimLex-999 norms, using a scale from 0 to 10, and again making sure that the answers given by participants reflect genuine similarity, and not relatedness.

MEN includes 3,000 noun, verb, and adjective pairs, rated on a scale from 0 to 50. In contrast to SimLex-999 and SimVerb-3500, most word pairs were concrete (i.e., they occurred more than 50 times as tags in two large collections of images), and the

[2] Two words form a hyponym-hyperonym pair, or a meronym-holonym pair, if the referents of the first word are included in the referents of the second word (i.e., a specific-general relation, such as “desk”-“furniture”, ‘knife”-“cutlery”), or denote parts of the referents of the second word (i.e., a part-whole relation, such as “door”-“house”, “mouth”-“head”), respectively.

authors did not distinguish between semantic similarity and relatedness. Furthermore, rather than asking for absolute ratings (e.g., “how related are the words *pie-kitchen*?”), the researchers instructed the participants to select the most related pair, from sets of two alternatives (e.g., “which two words are more related, *pie-kitchen* or *arrow-cup*?”), which encourages a more natural decision process, the authors argue.

Finally, SL consists of 7,576 noun pairs, rated on a scale from 1 to 5. The words are all very concrete, covering only the normed words from the feature norms by McRae and collaborators (2005). Like in the case of MEN, the authors did not instruct the participants to focus only on similarity or relatedness.

*Results:*

In the study by Hill and collaborators (2015), the authors tested the performance of several distributional models, in accounting for similarity ratings from the WordSim-353, MEN, and SimLex-999 norms. The linguistic models consisted of Skip-gram, two other neural network models (C&W; Collobert & Weston, 2008; HSMN; Huang, Socher, Manning, & Ng, 2012), and two HAL-based models (Kiela & Clark, 2014). The results indicated that the Skip-gram model strongly outperformed the other two neural network models, for the MEN and SimLex-999 datasets, whereas all three models had comparable levels of performance, for the WordSim-353 dataset. In addition, the Skip-gram model performed similarly to the two HAL-based models, for all the datasets. When comparing the results across the three norms, the authors found that the best model (i.e., Skip-gram) performed near the ceiling, for the WordSim-353 and MEN norms, but rather modestly, for the SimLex-999 norms. Another finding was that, for the best models, tested on the SimLex-999 norms, the highest performance was obtained for adjectives, followed by nouns, followed by verbs, with large differences in performance between the three parts of speech. In conclusion, the Skip-gram model outperformed other neural network models, when tested on a large corpus (i.e., around 1 billion words), but was comparable to typical “count” models, when tested on a smaller corpus (i.e., around 150 million words). However, even for a state-of-the-art model such as Skip-gram, the goodness of fit for the SimLex-999 dataset was notably poorer than for the WordSim-353 and MEN datasets. This suggests that the SimLex-

999 dataset, in virtue of its relative difficulty, might offer a more accurate index of model performance, than other datasets.

In a related, further study (Gerz et al., 2016), the authors used two versions of the Skip-gram model (i.e., with or without dependency-based information; Levy & Goldberg, 2014a), as well as a HAL-based model (Baroni, Dinu, & Kruszewski, 2014), and tested them on both the SimVerb-3500 norms, and the verb pairs from the SimLex-999 norms. Unlike in the previous study, where most of the words were nouns, the researchers found that the Skip-gram models clearly outperformed the HAL-based model, for both verb datasets. However, the levels of performance were relatively low and very close to those reported in the previous study.

Various other studies examined the potential advantages of integrating visual and linguistic representations, over employing only linguistic representations. In the study by Bruni and collaborators (2012), the authors tested several linguistic and multimodal models, using the WordSim-353 and MEN datasets. The linguistic models consisted of a HAL-based model, Topic, and Distributional Memory (DM; Baroni & Lenci, 2010), while the visual model was based on Scale-Invariant Feature Transform representations (SIFT; Lowe, 2004)[3]. With respect to the linguistic models, the HAL-based model had a slightly better performance than the Topic model, and both models considerably outperformed the DM model. In the case of the multimodal models, the combined models had a better fit than the purely linguistic models, for the MEN dataset, whereas the opposite pattern was found for the WordSim-353 dataset. These results are likely to be due to the fact that the models were tuned on part of the MEN dataset, as well as that the word pairs in the MEN dataset are, on average, more imageable (i.e., concrete) than the word pairs in the WordSim-353 dataset. In conclusion, adding visual information is not always beneficial, since it is likely to improve the representation of concrete words, but degrade that of abstract words.

In the study by Bruni and collaborators (2014), the authors used a HAL-based linguistic model, and a visual model consisting of SIFT representations. In order to perform a more thorough analysis of the space of possible multimodal models, they factorially manipulated the size of the multimodal representations, the weights assigned to the linguistic and visual components, as well as the method employed in

[3] A brief description of the SIFT model can be found in Appendix A.

combining the two modalities (i.e., either at the feature level, via weighted concatenation, or at the scoring level, as a weighted sum of independently calculated, text-based and image-based similarities). The models were tested on the WordSim-353 and MEN datasets, and the results showed a small advantage in favour of the multimodal models, as compared to the linguistic models. In addition, the multimodal models were particularly good at capturing the visual similarity between objects from the same category (e.g., “cheetah”-“lion”, “stream”-“waterfall”), while the linguistic models excelled at capturing more abstract relations, such as that between an object and its properties (e.g., “skyscraper”-“tall”, “cat”-”feline”). The authors also found that, when dividing the MEN norms into a concrete subset and an abstract subset (based on word concreteness), the addition of visual information improves the performance of the linguistic model for the concrete subset, but has no effect for the abstract subset. Overall, the results are consistent with those of the previous study, and suggest that the effectiveness of integrating visual and linguistic information depends on the concreteness of the word pairs.

In the study by Kiela and Bottou (2014), the linguistic model was Skip-gram, while the visual models consisted of SIFT and an AlexNet[4]-like, convolutional neural network (Oquab, Bottou, Laptev, & Sivic, 2014). The models were tested on the WordSim-353 and MEN datasets, and the results indicated that the multimodal models incorporating the AlexNet-like model substantially outperformed the purely linguistic model. In contrast, the performance of the multimodal models based on SIFT features did not differ significantly from that of the purely linguistic model. The results also showed that the increase in performance, when adding visual information, was comparable across the ImageNet (Deng et al., 2009) and ESP Game (Von Ahn & Dabbish, 2004) image datasets. In conclusion, the potentially positive effect of integrating visual and linguistic information seems to strongly depend on the quality of the visual model, such that newer, neural network models should offer a notably larger gain than older, bag-of-visual-words models (i.e., models which represent images as numerical distributions of visual salient patterns, regardless of their spatial position; see the previous footnote). Furthermore, the characteristics of the image corpus, such as image quality (i.e., high, for ImageNet, vs low, for ESP Game) and coverage (i.e.,

[4] For more details on the AlexNet model, see Appendix A.

low, for ImageNet, vs high, for ESP Game), appear to have little effect on the performance of the visual models.

In the study by Silberer and Lapata (2014), the models consisted of stacked autoencoders (i.e., deep neural networks; e.g., Bengio, 2009), trained over Strudel representations (Baroni et al., 2010), for the linguistic model, and over visual attribute representations (Silberer, Ferrari, & Lapata, 2013), for the visual model. The models were tested using the SL norms. The researchers evaluated multiple methods of integrating the two modalities, and found that projecting the unimodal representations to a bimodal representation (via the hidden layers of a stacked autoencoder), as well as using kernelized canonical correlation analysis (Hardoon, Szedmak, & Shawe-Taylor, 2004), led to a small increase in model performance, whereas using singular value decomposition (Bruni et al., 2014) produced a small decrease in model performance. In conclusion, whether combining visual and linguistic information provides a boost in performance over using only linguistic information, seems to depend on the method employed in bringing the two modalities together.

Despite their widespread use, semantic similarity ratings are known to suffer from a number of shortcomings (Batchkarov, Kober, Reffin, Weeds, & Weir, 2016; Faruqui, Tsvetkov, Rastogi, & Dyer, 2016; Gladkova & Drozd, 2016). Firstly, instead of measuring only similarity, certain sets of ratings also capture relatedness. For instance, in the MEN norms (Bruni et al., 2014), very high ratings are given to both pairs of similar words, such as “ocean”-“sea”, and pairs of related (but not similar) words, such as “burn”-“flame”. One solution to this problem, implemented for the WordSim-353 norms, is to split the data into a relatedness subset and a similarity subset (Agirre et al., 2009), although it is not always possible to tease the two apart (e.g., “chair” and “table” are similar, in terms of appearance and functionality, but they are also related, since tables usually come with chairs). Another solution is to make the task instructions very specific, in order to exclude words which are related, but not similar. This approach was adopted for the SimLex-999 (Hill et al., 2015) and SimVerb-3500 (Gerz et al. 2016) norms.

Secondly, the level of agreement between raters is relatively low, as compared to that in other semantic tasks. For instance, by binning the ratings and using Cohen’s κ as a measure of inter rater agreement for the WordSim-353 and MEN norms, Batcharov and collaborators (2016) found values of $\kappa = .21 - .61$, depending on the

number of bins chosen. In contrast, considerably higher values have been reported in other tasks, such as κ = .70 - .91, for sentiment classification (Gamon, Aue, Corston-Oliver, & Ringger, 2005; Kim & Hovy, 2004; Wilson, Wiebe, & Hoffmann, 2005), or κ = .73 - .94, for semantic feature classification (Lenci, Baroni, Cazzolli, & Marotta, 2013; Montefinese, Ambrosini, Fairfield, & Mammarella, 2013; Recchia & Jones, 2012).

Thirdly, the task implicitly assumes that words have a single meaning. However, most words are either polysemous, having distinct, but related meanings (e.g., “paper” can denote both a writing material and a document), or homonyms, having distinct, but unrelated meanings (e.g., “file” can denote both a folder and a tool). Since the words are not presented within an informative context, it is not clear which particular meaning is assigned to an ambiguous word, by participants. For example, the participant might focus on the most frequent or recently encountered meaning of the word (e.g., “bank” would represent an institution), or might use the other word in the pair as a disambiguating context (e.g., “apple”, within the pair “apple”-“download”, would represent a company), or might select a meaning in a random fashion. In contrast to these possibilities, with few exceptions (e.g., Huang et al., 2012; Reisinger & Mooney, 2010), the vectors produced by distributional models typically conflate the various meanings of a given word, where the contribution of each meaning is roughly proportional to its frequency within the corpus.

Finally, the ability of distributional models to account for similarity ratings, is a poor predictor of task performance in a number of related tasks. For instance, in the study by Schnabel and collaborators (2015), described previously, six popular distributional models were tested in their ability to predict similarity ratings from five similarity datasets. The results revealed large differences in performance between the models, with the relative performance of the models being consistent across the datasets. However, when the same models were tested on a noun phrase chunking task and a sentiment classification task, the differences in performance between the models became very small, and the hierarchy changed from that corresponding to the similarity ratings. Similarly, a study by Tsvetkov, Faruqui, Ling, Lample, and Dye (2015), which employed eleven distributional models and three similarity datasets, found that task performance in predicting similarity ratings correlates rather poorly with

task performance in a number of other tasks (i.e., document categorization, sentiment analysis, and metaphor detection), with most correlations falling in the .4 - .6 range.

A number of conclusions can be draw from the studies reviewed so far, as well as other similar studies, namely:

- with respect to linguistic models, “predict” models, especially CBOW (but also Skip-gram and GloVe), consistently outperform “count” models, such as LSA and HAL (Baroni et al., 2014; Mandera et al., 2017; Pereira, Gershman, Ritter, & Botvinick, 2016; Schnabel et al., 2015), where model performance is measured by the ability to fit behavioural data (i.e., ratings, probabilities, and semantic categories).

- adding visual information increases the performance of linguistic models, for concrete words, but has a null or negative effect on performance, for abstract words (Bruni, Boleda, et al., 2012; Bruni et al., 2014; Kiela, Hill, Korhonen, & Clark, 2014). The most likely explanation for the very modest contribution of visual models, with respect to concrete word pairs, is the fact that the performance of the linguistic models is close to the ceiling: for instance, in the case of the MEN and SL norms, the best linguistic models have correlations of .73 and .65 with the norm data, with an estimated inter-rater agreement of .84 and .76, respectively.
- multimodal models which employ convolutional neural networks usually have better performance than those which make use of bag-of-visual-words representations (Kiela & Bottou, 2014; Lazaridou, Pham, & Baroni, 2015; Silberer, Ferrari, & Lapata, 2017). Given that convolutional neural networks obtain state-of-the-art results in multiple computer vision tasks (e.g., Rawat & Wang, 2017), and are heavily informed by findings from visual neuroscience (e.g., LeCun, Bengio, & Hinton, 2015), it is not surprising that they maintain their dominance over older models (e.g., SIFT), when combined with linguistic models.
- the nature of the image corpus used in training the visual models has a relatively small effect on performance (Kiela & Bottou, 2014). In a more recent study, by Kiela, Verő, and Clark (2016), the authors showed that various convolutional neural networks can produce good performance even with as little as 20 images

per word, using images found by web search engines such as Google, Bing, and Flicker, instead of relying on large datasets of hand-labelled images, such as the ImageNet and ESP Game corpora.

- the mechanism through which the linguistic and visual representations are combined, plays an important role in determining the size of the resulting increase (or decrease) in performance. More specifically, integration methods that make use of information shared by the two modalities, by projecting linguistic and visual representations onto a common, multimodal space (e.g., via singular value decomposition or autoencoders; Bruni et al., 2014; Bruni, Uijlings, Baroni, & Sebe, 2012; Silberer & Lapata, 2014), lead to better model performance than integration methods which keep the two modalities separate (e.g., via concatenation).
- model performance, as measured by the Spearman correlation between model-derived similarity estimates and subjective ratings, is considerably poorer for datasets that focus strictly on word similarity (e.g., SimLex-999, SimVerb-3500), than for datasets that do not clearly distinguish between similarity and relatedness (e.g., WordSim-353, MEN, SL). For instance, in the study by Hill, Reichart, and collaborators (2015), the Skip-gram model achieved a correlation of .80, for the MEN dataset, as opposed to .41, for the SimLex-999 dataset. As noted by the authors, this effect is most likely due to the fact that pairs of words which are related (e.g., “bacon”-“pan”), and pairs of words which are similar (e.g., “bacon”-“pastrami”), can frequently occur in the same kinds of context (e.g., in discussions related to cooking), which makes it difficult for distributional models to separate the two types of relations. In support of this hypothesis, the authors found that, when considering only the one third most related pairs in the SimLex-999 dataset (based on free association data; Nelson et al., 2004), the performance of the Skip-gram model dropped from .41 to .26.
- model performance is markedly higher for pairs of nouns, than for pairs of verbs. This finding holds for subjective ratings, such as those from the SimLex-999 norms, as well as for WordNet-based proxies of subjective ratings (Hill, Reichart, et al., 2014). This finding might be explained by the fact that, in comparison to nouns, verbs are more polysemous and context-sensitive (e.g., Gentner, 2006). In support of this hypothesis, the authors of the first study

showed that adding syntactic information about the contexts in which words occur, results in a greater increase in model performance for verbs, than for nouns.

- although similarity ratings are a "gold standard" for evaluating the psychological plausibility of distributional models, they suffer from several shortcomings (Batcharov et al., 2016; Faruqui et al, 2016; Gladkova & Drozd, 2016), such as not distinguishing between similarity and relatedness, having low inter-rater agreement, not dealing with polysemy and homonymy, and over-representing concrete nouns. Part of these problems have been addressed when creating the most recent norms, such as SimLex-999 and SimVerb-3500, but the vast majority of studies still rely (at least partially) on relatively old norms, such as WordSim-353. Furthermore, model performance in accounting for similarity ratings does not automatically translate into similar levels of performance for a variety of semantic tasks (Schnabel et al., 2015; Tsvetkov et al., 2015).

**Free association:**

*General overview:*

Some of the main technical details regarding the studies reviewed in this subchapter are shown in Table 2.

Table 2. Computational studies focused on the free association task, covered in this subchapter. The entry for each study describes the linguistic and/or visual models employed (i.e., their type and the text/image corpora on which they are trained), as well as the behavioural datasets on which the models are tested.

| Study | Linguistic models | | Visual models | | Behavioural datasets |
|---|---|---|---|---|---|
| | Types | Corpora | Types | Corpora | |
| Cattle & Ma (2017) | GloVe<br>Skip-gram<br>Topic<br>Word2Gauss | Various | - | - | USF<br>EAT |
| Thawani, Srivastava, & Singh (2019) | FastText<br>GloVe<br>HAL<br>Skip-gram | Various | - | - | SWoW |
| Feng & Lapata (2010) | Topic | BBC News | SIFT | BBC News | USF |
| Hill & Korhonen (2014) | Skip-gram | Text8 | Image tags | ESP Game | USF |

*Datasets:*

The datasets of free association data were obtained by typically asking participants to read a cue word (e.g, “bread) and then record the first word that comes to mind when thinking about the cue (e.g., “butter”). Then, for each cue, the researchers calculated the probability that a given word would be produced in response to a cue (e.g., if 4 out of 20 participants generated “butter” as an associate of “bread”, then the probability of the pair “bread”-“butter” would be equal to 4/20 = 0.2). The norms from the studies included in our analysis are the University of South Florida dataset (USF; Nelson, McEvoy, & Schreiber, 2004), the Edinburgh Associative Thesaurus dataset (EAT; Kiss, Armstrong, Milroy, & Piper, 1973), and the Small World of Words dataset (SWoW; De Deyne et al., 2019)[5].

USF consists of more than 72,000 cue-associate pairs, produced in response to slightly over 5,000 cues. With respect to part of speech, nouns are by far the most frequent cues (76%), followed by adjectives (13%), verbs (7%), and other parts of speech.

EAT includes nearly 325,000 cue-associate pairs, for 8,400 cue words.

SWoW consists of nearly 1,390,000 cue-associate pairs, covering more than 12,000 cues. What distinguishes the SWoW norms from the USF and EAT norms, besides the greater number of cues, is the fact that the participants were asked to generate three associates in response to each cue, as opposed to only a single associate. Unless otherwise stated, the studies described in the next section employed (parts of) the USF norms.

*Results:*

In the study by Cattle and Ma (2017), the authors compared the ability of four distributional models, to predict free association probabilities from the USF and EAT datasets. The linguistic models included Skip-gram, GloVe, Topic, and Word2Gauss

[5] The large differences between the three datasets, in terms of average number of associates generated per cue, are largely a result of the fact that the EAT and SWoW norms list idiosyncratic responses (i.e., cue-associate pairs produced only once), whereas such responses are filtered out from the USF norms.

(Vilnis & McCallum, 2014). In order to predict the association between two words, the authors employed both the cosine similarity between the two vectors (which is the measure of choice in the literature), as well as the difference of the two vectors. For the USF norms, the Skip-gram and GloVe models, using cosine similarity, produced similar levels of fit, and outperformed all the other models (using either cosine similarity or vector difference) by a large margin. In contrast, for the EAT norms, the vector difference led to higher levels of performance than the cosine similarity, with the Skip-gram and GloVe models once again being the winners and having comparable performance. Overall, the results corroborate with those of other studies (e.g., Pennington et al., 2014), in showing that the Skip-gram and GloVe models are some of the best performing models currently available. Furthermore, the findings suggest that certain similarity measures, based on vector difference or the Euclidean distance between vectors (e.g., Pereira et al., 2016), for instance, might be a better alternative to cosine similarity, at least for some tasks.

In the study by Thawani and collaborators (2019), the authors tested the overlap between the associates of the words in the SWoW dataset, and the closest neighbours of the same words, but based on the representations produced by several distributional models. The distributional models consisted of Skip-gram, GloVe, FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017), and a HAL-based model (Baroni et al., 2014). The best model, as judged by the highest degree of overlap, was FastText, followed by Skip-gram and GloVe (with no significant difference in performance between the two), and followed by the HAL-based model. The results are consistent with those reported in the previous literature, suggesting that “predict” models are better than “count” models (e.g., Baroni et al., 2014), and that the use of morphological information improves model performance (Bojanowski et al., 2017).

Several other studies examined the issue of whether adding visual information to linguistic, distributional models, improves their performance in predicting free association data, obtained from the USF norms. In the study by Feng and Lapata (2010), the linguistic model was Topic, and the visual model was based on SIFT features. The authors used three measures for the strength of association between words, namely the Kullback-Leibler and Jensen-Shannon divergence, as well as the conditional probability of generating one word, given the other. The results showed a significant benefit of combining visual and linguistic information, over using only

linguistic information. With respect to the association measures, the conditional probability was slightly better than the Jensen-Shannon divergence, and both measures considerably outperformed the Kullback-Leibler divergence. Similarly to the study by Cattle and Ma (2017), this study highlights that the choice of association measure can have a large effect on model performance.

In the study by Hill and Korhonen (2014), the linguistic model was Skip-gram, while the visual model consisted of bag-of-visual-feature representations. As expected, for concrete nouns, the multimodal model outperformed the purely linguistic one. Moreover, the results indicated that the manner in which the linguistic and visual models were combined had an effect on model performance, such that the best fit was obtained when a modified Skip-gram model was trained over a corpus that mixed words and visual features of images associated with those words. In contrast, more traditional methods of integrating the two modalities (i.e., concatenation; canonical correlation analysis; Hardoon et al., 2004; singular value decomposition; Bruni et al., 2014) produced lower levels of performance.

However, the findings were more complex when considering pairs of abstract nouns, concrete verbs, and abstract verbs. Since the image corpus did not cover these words, their visual representations had to be inferred from those of the items in the corpus (i.e., for the concrete nouns), as well as from the linguistic representations. This propagation of visual information, via the linguistic representations, was implemented through the method proposed by Johns and Jones (2012), or via ridge regression, or simply by employing the multimodal representations generated by the authors' modified Skip-gram model. The results showed that either the modified Skip-gram model was better than the ridge regression model, which in turn was better than the Johns and Jones model (i.e., for all nouns, concrete verbs, and all verbs), or instead that all models had comparable levels of performance (i.e., for abstract nouns and abstract verbs). Moreover, for each of the three multimodal models, whether that model outperformed the purely linguistic model depended on the type of word pairs over which it was tested. Another finding was that increasing the contribution of the visual component usually improved performance for the concrete pairs, but had the opposite effect for the abstract pairs.

Taken together, the results of this study seem to indicate that adding visual information does not always improve the performance of a linguistic model. Instead,

the direction of the effect, as well as its size, depends on the properties of the free association pairs (e.g., concrete/abstract, noun/verb), on how the visual component is combined with the linguistic component (e.g., it is better to integrate the two streams of information during learning, than after it), and on whether the visual information is extracted directly or indirectly from the image corpus, among other factors. However, it is not clear how robust the results are, given that the subsets of the USF norms used in the study included only 295 abstract noun pairs, 66 concrete verb pairs, and 127 abstract verb pairs, as opposed to 1,418 concrete noun pairs.

Despite the popularity of employing free association norms for testing and comparing distributional models, there are a couple of issues with respect to the free association task. One such issue was examined in detail by Nelson, Dyrdal, and Goodmon (2005). In their study, the authors found that forward strength (i.e., the probability of generating the target, given the cue) has a moderate, negative correlation with cue competitor strength (i.e., the probability of generating words related to the cue, but unrelated to the target, given the cue), and small, positive correlations with backward strength (i.e., the probability of generating the cue, given the target), as well as target competitor strength (i.e., the probability of generating words related to the target, but unrelated to the cue, given the target) and target activation strength (i.e., the probability of generating either an associate of the target, or the target, given another associate of the target). These results, although being just correlational in nature, seem to suggest that forward strength does not depend only on the direct association between the cue and the target, but also on a subset of both direct and indirect associative relations between the cue, target and their associates. The potential role of the cue competitor strength is likely to reflect a competition for retrieval between the target and other words (e.g., Nelson, McKinney, Gee, & Janczura, 1998; Raaijmakers & Shiffrin, 1981), whereas the potential influence of backward strength, target competitor strength and target activation strength, could be accounted for by spreading activation theories of semantics (e.g., Anderson, 1983; Collins & Loftus, 1975). This means that, unlike similarity ratings, free association probabilities are unlikely to provide an absolute index of semantic association between a cue and a target.

Another issue is that performance in modelling free association data does not always correlate highly with performance in other semantic tasks. In the study by

Thawani and collaborators (2019), mentioned previously, the authors also looked at model performance with respect to tasks such as sentiment analysis, chunking, natural language inference, named entity recognition, and part-of-speech tagging. Their findings indicated that the relative goodness of fit for the models was strongly task dependent, and did not always follow the trend from the free association results: for instance, all the models performed equally well in the named entity recognition task; moreover, the FastText model was no longer the clear winner in the sentiment analysis, chunking, and part-of-speech tagging tasks; also, the Skip-gram and GloVe models did not outperform the HAL-based model in the natural language inference task. In addition to these findings, the study showed that the manner in which the vector representations were employed had a strong impact on performance, such that the differences in performance between the FastText, Skip-gram and GloVe models effectively vanished when the vector representations were fine-tuned for each task, separately, instead of being kept constant.

Several conclusions can be drawn from the studies presented so far, as well as other similar studies, namely:

- as is the case for other tasks, “predict” models, such as Skip-gram, FastText, and GloVe, usually outperform older, “count” models, such as LSA, HAL, and Topic (e.g., Cattle & Ma, 2017; Thawani et al., 2019).
- for concrete words, adding visual information to linguistic models typically improves their performance. This result holds regardless of whether the visual representations are derived from image corpora (Feng & Lapata, 2010; Hill & Korhonen, 2014), feature norms (Andrews et al., 2009; Silberer & Lapata, 2012), or combinations of the two (Silberer et al., 2013). In contrast, for abstract words, including visual information seems to degrade the performance of linguistic models (Hill & Korhonen, 2014; Kiela et al., 2014).
- it can often be difficult to make comparisons between studies, given that a wide variety of association measures are used in accounting for free association data. These measures include cosine similarity, Euclidean distance (Pereira et al., 2016), vector difference (Cattle & Ma, 2017), Kullback-Leibler divergence, Jensen-Shannon divergence, and conditional probability (Feng & Lapata, 2010). Moreover, some of the results obtained by comparing different association measures are very counter-intuitive. For instance, since free

association probabilities are asymmetrical, while similarity ratings are symmetrical, it would seem very plausible that the former should be better captured by asymmetrical measures (e.g., Kullback-Leibler divergence), whereas the latter should instead be best predicted by symmetrical measures (e.g., Jensen-Shannon divergence). Instead, Feng and Lapata (2010) found the exact opposite pattern. A related problem is that the number of word pairs used in each study varies widely, such as roughly 900 (Kiela et al., 2014), 2,000 (Hill & Korhonen, 2014; Hill, Reichart, et al., 2014), 5,000 (Hill et al., 2013), and more that 20,000 (Feng & Lapata, 2010; Silberer & Lapata, 2012).

- free association probabilities provide a relative measure of semantic association between words, which is partially dependent on a combination of automatic and voluntary retrieval processes, such as spreading activation and competition for retrieval. In contrast, word similarity measures derived from distributional models most likely estimate the amount of information shared by the vector representations, without taking into consideration the effects of semantic processes operating over the representations. While it is possible to add such processes to distributional models, for instance by emplying Luce's choice rule (Luce, 1959) in order to simulate probabilistic, competition-based memory retrieval (Nelson et al., 1998; Raaijmakers & Shiffrin, 1981), none of the studies described in this subchapter have done so, with the notable exception of that by Mandera and collaborators (2017), which might explain the relatively poor performance of both the unimodal and multimodal models (e.g., the largest correlations with the free association probabilities fall in the .2 - .4 range; Cattle & Ma, 2017; Hill & Korhonen, 2014; Silberer & Lapata, 2012, 2014). In addition, as in the case of similarity ratings, performance in predicting free association probabilities is not always a reliable indicator of performance in other semantic tasks, such as sentiment analysis, natural language inference, and named entity recognition (Thawani et al., 2019).

**Semantic categorization:**

*General overview:*

Some of the main technical details regarding the studies reviewed in this subchapter are shown in Table 3.

Table 3. Computational studies focused on the semantic categorization task, covered in this subchapter. The entry for each study describes the linguistic and/or visual models employed (i.e., their type and the text/image corpora on which they are trained), as well as the behavioural datasets on which the models are tested.

| Study | Linguistic models | | Visual models | | Behavioural datasets |
|---|---|---|---|---|---|
| | Types | Corpora | Types | Corpora | |
| Bruni et al. (2011) | DM | BNC<br>Web<br>Wikipedia | SIFT | ESP Game | Almuhareb-Poesio<br>Battig |
| Bruni et al. (2014) | HAL | ukWaC<br>Wikipedia | SIFT | ESP Game | Almuhareb-Poesio<br>Battig |
| Riordan & Jones (2011) | BEAGLE<br>COALS<br>CS-LL<br>CS-LO<br>HAL<br>HiDEx<br>LSA<br>PPMI<br>Topic | TASA | - | - | HJC<br>MCSM<br>VV |
| Silberer & Lapata (2014) | Strudel | Wikipedia | Visual attributes | ImageNet | Fountain-Lapata |

*Datasets:*

The datasets of semantic categorization were obtained by selecting a set of typically concrete nouns, and then dividing the set into several categories (e.g., animals, vegetables, tools, etc.). The categorization data included in our discussion consists of the HJC norms (Howell, Jankowicz, & Becker, 2005), MCSM norms (McRae et al., 2005), VV norms (Vinson & Vigliocco, 2008), Almuhareb-Poesio norms (Almuhareb & Poesio, 2005), Battig norms (Baroni, Murphy, Barbu, & Poesio, 2010), and Fountain-Lapata norms (Fountain & Lapata, 2010).

The HJC norms include 442 words typically learned at an early age, namely 352 nouns, divided into 11 categories, and 90 verbs, divided into 9 categories. The MCSM norms consist of 541 concrete nouns, and are a compilation of stimuli employed in a variety of studies on semantic memory. The VV norms include 456 words, namely 240 nouns, divided into 21 categories and 216 verbs, divided into 33 categories.

The Almuhareb-Poesio norms consist of 402 nouns, mostly concrete, divided into 21 categories. When designing the dataset, the authors focused on balancing the items with respect to three factors, namely representativeness (i.e., each category covers one of the 21 primary classes in the WordNet hierarchy; Fellbaum, 1998), frequency (i.e., high, medium, and low frequency words each constitute a third of the items), and polysemy (i.e., based on WordNet, words with 4 senses, 2-3 senses, and a single sense each make up a third of the items).

The Battig norms include 83 concrete nouns, divided into 10 categories. The categories where chosen from the study by Van Overschelde, Rawson, and Dunlosky (2010), such that the words are very typical for each each category, and they form part of the stimuli MCSM norms.

The Fountain-Lapata norms are made up of 541 concrete nouns, divided into 41 categories. The words consist of all the stimuli covered by the MCSM norms. In contrast to all the previous norms, the category labels and members where not decided by the authors. Instead, they were derived from the answers provided by a number of participants, who were shown a selection of items and asked to name the categories to which the items belong. With the exception of the Fountain-Lapata norms, all the other norms derive their categories from WordNet, and/or the verb classification by Levin (1993), and/or the Arthur Communicative Development Inventory (MCDI; Fenson et al., 2000).

*Results:*

By far the most comprehensive and informative study is that by Riordan and Jones (2011), which focused on comparing several distributional and feature-based models of semantics. The linguistic models consisted of HAL, High Dimensional

Explorer (HiDEx; Shaoul & Westbury, 2006), Correlated Occurrence Analogue to Lexical Semantics (COALS; Rohde, Gonnerman, & Plaut, 2006), Contextual Similarity - Log-Likelihood (CS-LL; McDonald, 2000), Contextual Similarity - Log Odss (CS-LO; Lowe & McDonald, 2000), PPMI (Bullinaria & Levy, 2007), LSA, Topic, and Bound Encoding of the Aggregate Language Environment (BEAGLE; Jones & Mewhort, 2007). The feature-based models were constructed using the features corresponding to the words in the HJC, MCSM, and VV datasets.

For the dataset provided in the study by McRae and collaborators (2005), consisting only of concrete nouns, the results indicated that the "word-as-context" models (with the exception of BEAGLE) outperformed both the feature-based model, and the "document-as-context" models (i.e., LSA and Topic). The authors also found differences between the feature-based model (i.e., MCSM) and the best linguistic model (i.e., COALS), in terms of cluster structure: for the animals class, the MCSM model produced two clusters (i.e., flying vs non-flying), whereas the COALS model produced a single cluster; for the artifacts class, the MCSM model generated two clusters, based on material (i.e., wood vs metal), while the COALS model generated three clusters, based on the type of action associated (i.e., wearing vs riding vs holding); for the foods class, both models produced a single cluster. A more detailed analysis revealed that the semantic similarity structure of the MCSM representations depended on different types of features being associated with different semantic classes (e.g., behaviours, for the animals class, vs materials, for the artifacts class), while the semantic similarity structure of the COALS representations relied on frequent action verbs (e.g., "eat", "fly").

For the dataset described in the study by Vinson and Vigliocco (2008), consisting of object nouns, action verbs, and event nouns, both concrete and abstract, the authors ran two analyses, one for the object nouns, and the other for the action verbs. The event nouns were left out, for reasons of compatibility with the previous analysis. With respect to the object nouns, the results indicated that the COALS and VV models outperformed all the others, but did not differ significantly. In the case of the action verbs, the VV model outperformed all the linguistic models, among which the best performance was obtained for the COALS model.

For the dataset from the study by Howell and collaborators (2005), consisting of nouns and verbs known to young children, the authors once again ran two analyses,

one for the nouns, and the other for the verbs. With respect to the nouns, the HJC model outperformed the COALS model, which in turn performed better than all the other models. In the case of the verbs, the CS-LL and HJC models outperformed all the others, but did not differ significantly.

In conclusion, with regard to their ability to categorize words based on meaning, distributional models are comparable to feature-based models. This suggests that the perceptual, motor, and affective properties of a word's referent, as measured by feature norms, are reflected in the linguistic contexts in which that word is encountered, as captured by distributional models. However, as the authors note, the redundancy between feature-based and distributional representations is only partial: for instance, when performing semantic categorization over objects, feature-based models relied mostly on internal/external properties and materials, while distributional models were guided especially by actions, functions, and situations. Another finding was that distributional models performed comparably to feature-based models, not just for concrete nouns, but also for concrete and abstract verbs. Overall performance for verbs was poorer than that for nouns, but this difference is likely to be due to the fact that verbs have a relatively shallow taxonomic structure, as compared to nouns (Fellbaum, 1998), and the fact that the classifications for verbs contain more and small categories than those for nouns. Within the family of distributional models, "word-as-context" models (especially COALS) appear to be slightly better "document-as-context" models, but the differences are very small and not always significant.

More recent studies focused on testing the hypothesis that multimodal models, which combine visual and linguistic representations, have a better performance than purely linguistic models. In the study by Bruni, Tran, and Baroni (2011), the linguistic model was DM, and the visual model was based on SIFT features. The effect of adding a visual representation was found to depend on the dataset: the multimodal model outperformed the linguistic model on part of the Battig dataset, but the two models did not differ significantly for the Almuhareb-Poesio dataset. The authors attributed this difference to the fact that the Battig dataset consists only of concrete words, whereas the Almuhareb-Poesio dataset includes both concrete and abstract words, under the assumption that visual information is particularly important for representing concrete words, but not abstract ones. In a similar, further study (Bruni et al., 2014), the authors used a HAL-based linguistic model, instead of the DM model, and employed several

versions of the multimodal model, by tuning them on the Battig dataset, and then testing them on the Almuhareb-Poesio dataset. Their results indicated that, while certain multimodal models had a slightly better performance than the purely linguistic model, the differences in performance did not reach statistical significance, mirroring the results from the previous study. In another study, by Silberer and Lapata (2014), described earlier in this chapter, the authors found that combining visual and linguistic information by using an autoencoder provided a moderate boost in performance, while employing kernelized canonical correlation analysis and singular value decomposition resulted in little or no improvement in performance.

A few conclusions can be derived from the studies presented so far, namely:

- distributional models are remarkably good when employed for semantic categorization, performing comparably to models based on subjective features (Riodan & Jones, 2011). Their level of performance is especially notable, given that the models from the study by Riordan and Jones (2011) were trained on a small corpus (i.e., containing less than 11 million words; Zeno, Ivens, Millard, & Duvvuri, 1995), and tested on several datasets (Howell et al., 2005; McRae et al., 2005; Vinson & Vigliocco, 2008), consisting of both nouns and verbs. However, although being similar in performance, the distributional models focused more on information about how word referents related to the outside world, whereas the feature-based models emphasized the perceptual characteristics of the word referents[6].
- within the class of "count" models, there is some evidence that "word-as-context" models slightly outperform "document-as-context" models (Riordan & Jones, 2011). One potential explanation for this result might be that documents (typically consisting of hundreds of words) lump together words with various level of semantic association to a given word, whereas narrow text windows (typically consisting of 4-10 words) include mostly words that are strongly

[6] Distributional models can have indirect access to a limited amount of perceptual and motor information, by virtue of it being reflected in language use (Louwerse, 2011, 2018; Riordan & Jones, 2011; Sadeghi et al., 2015). To give an example, if two objects are visually similar (e.g., "table" and "chair"), their verbal descriptions are very likely to contain similar words. For a comprehensive examination of the types of information that can or cannot be easily extracted from distributional representations, see the study by Utsumi (2020).

associated with a given word. As a result, “document-as-context” models might capture a rougher measure of semantic similarity.

- in contrast to the results for the semantic similarity/relatedness and free association tasks, adding visual information (derived from image corpora) has little or no positive effect on model performance (Bruni et al., 2011, 2014; Silberer & Lapata, 2014). This is likely due to the fact that, as shown in several studies (Baroni et al., 2014; Bruni et al., 2011, 2014; Riordan & Jones, 2011), the linguistic models perform near ceiling, leaving little room for improvement.

## 2.3. Comparing distributional models, by using networks derived from free association norms

Free association norms (e.g., Nelson et al., 2004) are widely assumed to offer a quantitative description of the associative structure of semantic memory (Cramer, 1968; Deese, 1965; Palermo & Jenkins, 1964). One natural representation of the cues and the associates that they elicit, is in the form of a network (or graph), where the vertices are words, and the edges are semantic associations linking the words. A number of previous studies, most notably that of Steyvers and Tenenbaum (2005), showed that semantic networks based on free association exhibit certain properties that clearly distinguish them from networks generated in a random manner (Erdős & Rényi, 1960).

More specifically, semantic networks have a "small-world" structure (Watts & Strogatz, 1998), characterized by highly clustered neighbourhoods (i.e., for any given word, many of its associates are also associated to one another) and small shortest paths between nodes (i.e., for any two words, a short chain of associations can be found, linking the two words). Words with high clustering coefficients have been found to have an advantage in recognition and recall tasks (Nelson et al., 1998), and to generate stronger priming effects (Nelson & Goodmon, 2002). These results can be explained using spreading activation theories of semantics (e.g., Anderson, 1983; Collins & Loftus, 1975): cues with highly interconnected associates are able to retain more activation within their immediate neighbourhood, some of which can then feed back into the cue, thus facilitating its processing.

In addition to being highly interconnected, free association networks also display a "scale-free" structure, where node degrees (i.e., the number of associates for each word), denoted by *k*, follow a either a power-law distribution (Steyvers & Tenenbaum, 2005), such that $\boldsymbol{P(k) \sim k^{-\gamma}}$, or a truncated power-law distribution (Morais, Olsson, & Schooler, 2013), such that $\boldsymbol{P(k) \sim k^{-\gamma} e^{-\delta k}}$, where γ and δ are positive constants. In comparison to random networks, "scale-free" networks, such as free association networks, are particularly resilient in the face of (random) node deletions (Albert, Jeong, & Barabási, 2000), given that such networks contain a significant number of hubs (i.e., nodes connected to a very large number of other nodes). Also, short path lengths, which follow from the "scale-free" property, are likely

to support efficient memory retrieval, since path lengths in the semantic networks of typically developing children are shorter than those of children with delays in language development (e.g., “late talkers”; Beckage, Smith, & Hills, 2011; and children with cohlear implants; Kenett et al., 2013).

**General overview:**

Some of the main technical details regarding the studies reviewed in this subchapter are shown in Table 4.

Table 4. Computational studies focused on semantic network structure, covered in this subchapter. The entry for each study describes the linguistic models employed (i.e., their type and the text corpora on which they are trained), as well as the behavioural datasets on which the models are tested.

| Study | Linguistic models | | Behavioural datasets |
|---|---|---|---|
| | Types | Corpora | |
| Griffiths, Steyvers, & Tenenbaum (2007) | LSA<br>Topic | TASA | USF |
| Gruenenfelder et al. (2015) | BEAGLE<br>LSA<br>POC<br>Topic | TASA | USF |
| Jones, Gruenenfelder, & Recchia (2018) | BEAGLE<br>Jaccard Index<br>LSA | TASA | USF |
| Kajić & Eliasmith (2018) | GloVe<br>Skip-gram | Common Crawl<br>Google News | USF |
| Utsumi (2015) | HAL<br>LSA | BNC | USF |

**Datasets:**

All the studies included in our analysis make use of the USF dataset. As described in the previous subchapter, the USF norms cover more than 72,000 cue-associate pairs, covering to over 5,000 cues, representing various parts of speech.

**Results:**

Semantic networks derived from free association responses have been used in evaluating the cognitive plausibility of various types of distributional models (Griffiths, Steyvers, & Tenenbaum, 2007; Gruenenfelder et al., 2015; Jones et al., 2018; Kajić & Eliasmith, 2018; Utsumi, 2015). In the study by Griffiths, Steyvers and Tenenbaum (2007), the authors compared the networks generated from the LSA and Topic models, in terms of their relation to the networks extracted from the USF norms. One of the comparisons involved looking at the relation between word frequency and neighbourhood size. In free association networks, there is a strong positive correlation between the two measures (Steyvers & Tenenbaum, 2005), which can be explained by the fact that both word frequency and neighbourhood size are an index of a word's utility. More specifically, words with many neighbours are high in frequency since they are retrieved in response to a large number of cues (i.e., they are relevant in numerous contexts, making them especially useful). The Topic model produced a strong correlation between the two measures which was very similar to that observed in the free association network, whereas the correlation calculated from the LSA model was noticeably weaker. Furthermore, the Topic model performed better than the LSA model in matching the slope of the best fitting power-law distribution and the mean clustering coefficient, and vice-versa for the mean shortest path length. With respect to the LSA model, the authors also tested whether performance would improve when using inner product, instead of cosine, for computing word associations. The found that the fit increased when considering the correlation between neighbourhood size and word frequency, as well the mean shortest path length, but also that the fit decreased for the slope of the best fitting power-law distribution, as well as the mean clustering coefficient. In conclusion, the Topic model appears to have an advantage over the LSA model, given that its network properties more closely mirror those of the USF networks.

In the study by Gruenefelder and collaborators (2015), the authors compared the properties of semantic networks derived from the USF norms, with those of networks obtained from several distributional models, namely LSA, BEAGLE, Topic, and Proportion Of Co-occurrence (POC). The measures being examined consisted of the shape and slope of the degree distribution, the mean clustering coefficient, and the mean shortest path length. With respect to the mean clustering coefficient, the

performance of the four models was comparable, such that LSA, BEAGLE and Topic models overpredicted the amount of clustering in the free association norms, while the POC model underpredicted the same quantity. The difference between the two trends is most likely due to the manner in which word associations are computed in each of the models: in contextual models (i.e., LSA, BEAGLE, and Topic), words become associated if they occur in similar contexts, regardless of whether they co-occur in a document or not; in contrast, associative models (i.e., POC) require words to co-occur within a document, in order for them to be associated, which means that contextual similarity does not always translate into word association, for those models. Since the POC model is more restrictive in generating associations, it is not surprising for it to produce less interconnected networks. With respect to the mean shortest path length and the slope of the best fitting power-law distribution, the Topic and POC models proved considerably better than the LSA and BEAGLE models, at matching the free association data. Moreover, the authors also found that the Topic and POC models better fitted a power-law distribution, than an exponential distribution, while the opposite was true for the LSA and BEAGLE models.

Gruenenfelder and collaborators (2015) concluded that the Topic and POC models seemed to be more cognitively plausible than the LSA and BEAGLE models, based on their superior performance, even though none of the four models provided a tight fit to the properties of the free association network. They also remarked that contextual and associative models tap into different sources of information, and that retrieval from semantic memory might rely on both types of information. This hypothesis is consistent with behavioural data from the category verification task, where participants must decide whether an exemplar (e.g., “apple”) belongs to a certain category (e.g., “fruit”). Participants are faster to reject a false sentence when the wrong category is unrelated to the correct one (e.g., “occupation” vs “fruit”), than when they are coordinate categories (e.g., “vegetable” vs “fruit”; McCloskey & Glucksberg, 1979). Since the strength of the association between the two types of wrong categories is very low and comparable, it is also insufficient for determining the correct answer. Instead, it is likely that participants make a comparison between the featural representations of the exemplar and category provided, representations which might be captured by contextual models. In contrast, there is no difference in response times between cases when the wrong category is a coordinate of the exemplar (e.g.,

“An apple is a banana”), and cases when the wrong category and the exemplar are unrelated (e.g., “An apple is a vehicle”; Gruenenfelder, 1986). Here, the decision does not seem to depend on the overlap between the featural representations of the two words, and instead appears to be better accounted for in terms of their associative strength, which might be captured by associative models.

The involvement of both types of representations (i.e., contextual and associative) might also explain findings from a study by Recchia and Jones (2012), which used naming and lexical decision response times, as a means of comparing the semantic processing of concrete and abstract words. The researchers found a double dissociation between concrete and abstract words: concrete words benefited from having more contextual neighbours (measured by the overlap between featural representations), whereas no effect was detected for the number of associative neighbours (measured by the probability of co-occurrence with a text corpus); the reverse pattern became evident for abstract words, with a positive impact for having more associative neighbours, but no impact for the number of contextual neighbours. Given the behavioural evidence for dual representations, Gruenenfelder and collaborators (2015) argued that hybrid models, which combine a contextual model (i.e., LSA, BEAGLE, and Topic) and an associative one (i.e., POC), should outperform individual models from either class. The authors tested this prediction and discovered that, indeed, the fit of all the contextual models improved considerably, when adding the associative model, with the Topic-POC hybrid fitting all the properties of the free association network almost perfectly. All in all, their results showed that the Topic and POC models, taken either individually or in combination, best fit the free association data, which suggests that they are more cognitively plausible than the LSA and BEAGLE models.

In the study by Utsumi (2015), the author tested the properties of semantic networks derived from a wide variety of “count” models, by factorially manipulating a number of important hyperparameters, namely: the size of word neighbourhoods (i.e., by matching the neighbourhood sizes with those from the free association norms, or by using a cumulative similarity ratio); the definition of context (i.e., the documents in which a word occurs, like in the LSA model, or the words preceding and following a word, like in the HAL model); the weighting function for the elements of the context matrix (i.e., TF-IDF, PPMI, or none); and dimensionality reduction (i.e., present or

absent). The results indicated that all the model-based networks had a “small-world” structure (i.e., large clustering coefficients and small shortest path lengths). In addition, the neighbourhood sizes in all the networks followed either a power-law distribution or a truncated power-law distribution, as opposed to an exponential distribution. This finding is inconsistent with the results obtained by Gruenenfelder and collaborators (2015), where the neighbourhood sizes from the LSA model did not fit a power-law distribution better than an exponential one. The discrepancy is most likely caused by significant differences in methodology between the two studies, especially in terms of how the node degrees for the model-based networks were matched to those from the free association network, and how the fit to a power-law distribution was calculated.

Another finding from the study by Utsumi (2015) was that “word-as-context” models (including HAL) outperformed “document-as-context” models (including LSA), as measured by the similarity between of the degree distributions of the model-based and free association-based networks. Also, employing a weighting function or reducing the dimensionality of the context matrix usually improved the fit of the “word-as-context” models, but had null or detrimental effects for the fit of the “document-as-context” models. In summary, the match between the properties of model-based networks and those of free association networks, which can be used as an estimate of the cognitive plausibility of the models, strongly depends on the choice of hyperparameters, as well as on how the linguistic context is defined, such that, on average, “word-as-context” models provide a superior fit to behavioural data, as compared to “document-as-context” models.

In the study by Kajić & Eliasmith (2018), the authors examined the “small-world” and “scale-free” properties of networks derived from the Skip-gram and GloVe models, comparing them to those of networks obtained from the USF norms. Directed and undirected versions of networks were constructed, and word neighbourhoods were determined using the two methods described in the studies by Utsumi (2015) and Griffiths, Steyvers and Tenenbaum (2007). The researchers found that the networks derived from both models had “small-world” properties, namely large clustering coefficients and small shortest path lengths, with negligible differences between the models. With respect to the “scale-free” property, the neighbourhood sizes for all the networks followed a truncated power-law distribution (but not a pure one), while for some of the networks, also the lognormal and/or exponential distributions were

plausible candidates. In conclusion, networks based on the Skip-gram and GloVe models are structurally similar to free association networks, although the power-law nature of the degree distributions is not always evident.

Some studies (Jones et al., 2018; Griffiths, Steyvers, & Tenenbaum, 2007) have taken a more indirect approach to comparing model-based networks and free association networks, by investigating the symmetry of associations (related to the degree to which directed networks can be approximated by, or derived from, undirected networks), as well their transitivity (related to the clustering or interconnectivity of the networks). The two studies previously mentioned start by making a distinction between spatial and probabilistic models. In spatial models (e.g., LSA), words are represented as points in a multidimensional space, and associations are based on a measure of the distance between words. Importantly, the distances need to satisfy four axioms, for any points $x$, $y$, and $z$, namely:

- non-negativity (i.e., distances are always be positive: $d(x,y) \geq 0$)
- identity of indiscernibles (i.e., two points are at a distance of 0 if and only if they coincide: $d(x,y) = 0 \Leftrightarrow x = y$)
- symmetry: (i.e., distances are symmetrical: $d(x,y) = d(y,x)$)
- subadditivity (i.e., distances follow the triangle inequality: $d(x,z) \leq d(x,y) + d(y,z)$)

Free association probabilities violate the symmetry and subadditivity axioms (Griffiths, Steyvers, & Tenenbaum, 2007). Asymmetric associations (e.g., “snake” is frequently produced in response to “cobra”, but not vice-versa) can be largely explained in terms of differences in word frequency, such that participants are likely to generate high frequency targets, regardless of the cue, while associations that do not follow the triangle inequality (e.g., a strong association exists between “asteroid” and “belt”, as well as between “belt” and “buckle”, but not between “asteroid” and “buckle”) might be accounted for by differences in the semantic context related to each association. In contrast to spatial models, in probabilistic models (e.g., Topic) words are represented as probability distributions, which means that associations depend on the probability of one word, given the other. Associations derived from probabilistic models are not subject to symmetry and subadditivity constraints, which suggests that probabilistic models are more psychologically plausible than spatial models.

As a response to this critique of spatial models, Jones and collaborators (2018) noted that the associates produced by participants in the free association task are the result of a semantic process, operating over semantic representations. They argued that distance measures, in spatial models, and conditional probabilities, in probabilistic models, only reflect the similarity and/or relatedness of word representations, without providing any information about the mechanisms that employ such representations, in order to generate cue-target pairs. The fact that distributional models do not embed task-specific semantic processes is supported by the finding that state-of-the-art distributional models, such as CBOW, Skip-gram, and GloVe, perform very well on certain tasks, such as similarity/relatedness rating, semantic categorization and analogy solving (Baroni et al., 2014), but rather poorly on other tasks, such as free association (Cattle & Ma, 2017), and semantic priming (Ettinger & Linzen, 2016; Auguste, Rey, & Favre, 2017).

Jones and collaborators (2018) suggested that task behaviour in the free association task, which appears not to be consistent with the metric axioms, might be at least partially explained by employing Luce's (1959) choice rule. This rule provides a means of computing the conditional probability of generating the target $T_j$, in response to cue $C_i$, based on a (metric) similarity space, as follows:

$$\boldsymbol{P(T_j|C_i) = \frac{\beta_j \eta_{i,j}}{\sum_{k \in V} \beta_k \eta_{i,k}}}$$

where $\boldsymbol{\beta_j}$ is the response bias for word *j*, *V* is the vocabulary of the model, and $\boldsymbol{\eta_{i,j}}$ is the similarity between words *i* and *j*. The authors tested a simplified version of the choice rule, where all the response biases are equal to 1, as follows:

$$\boldsymbol{P(T_j|C_i) = \frac{\cos(C_i, T_j)}{\sum_{k=1}^{\tau} \cos(C_i, T_k)}}$$

where $\boldsymbol{\tau}$ is a minimum similarity threshold hyperparameter (i.e., the sum in the denominator is computed only over the cosines with values greater than $\boldsymbol{\tau}$). In order to determine whether the application of the choice rule results in violations of the metric axioms, similar to those observed in free association responses, they used three distributional models, namely the Jaccard Index (Jaccard, 1912), LSA, and BEAGLE, trained over the corpus described in the study by Griffiths, Steyvers, and Tenenbaum (2007).

To evaluate the asymmetry of associations, the authors first selected all the highly asymmetric cue-target pairs (i.e., pairs for which $\boldsymbol{P}(\boldsymbol{C_i}|\boldsymbol{T_j})$ and $\boldsymbol{P}(\boldsymbol{T_j}|\boldsymbol{C_i})$ differ by at least an order of magnitude) from the USF norms, and then tested the degree to which the three aforementioned models could predict the direction of the asymmetry. They found that the Jaccard and BEAGLE models performed as well as the Topic model, unlike the LSA model, which performed only slightly better than chance. Consequently, these results prove that, while certain spatial models (e.g., LSA) might be less psychologically plausible than certain probabilistic models (e.g., Topic), based on how well they capture the asymmetry of free association probabilities, this conclusion cannot be extended to the entire class of spatial models.

To examine departures from the triangle inequality, Jones and collaborators (2015) started by attempting to reproduce the quantitative results from Griffiths, Steyvers, and Tenenbaum (2007). In that study, the authors looked at the relationship between the association probabilities $\boldsymbol{P}(\boldsymbol{W_2}|\boldsymbol{W_1})$, $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_2})$, and $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_1})$, where $W_1$, $W_2$ and $W_3$ are words from the USF norms. Under the natural assumption that association probabilities are inversely related to distances within a mental similarity space, the triangle inequality implies that if $\boldsymbol{P}(\boldsymbol{W_2}|\boldsymbol{W_1})$ and $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_2})$ are large (i.e., $W_1$ is close to $W_2$, and $W_2$ is close to $W_3$), then $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_1})$ should be large as well (i.e., $W_1$ should be close to $W_3$, by transitivity). In contrast, the authors found that for the vast majority of the cases where $\boldsymbol{P}(\boldsymbol{W_2}|\boldsymbol{W_1})$ and $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_2})$ are large, it is the case that $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_1})$ is either small or equal to 0, which translates into divergence from the triangle inequality. However, Jones and collaborators (2015) noted that it is problematic to use associations where $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_1}) = 0$, given that the inferences derived from unobserved events might be inexact: for instance, collecting responses from a larger number of participants might produce a value of $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_1})$ greater than 0; also, having two associations with a probability of 0 does not necessarily imply that they are equal, but rather that they are both too weak to reliably detect, when collecting responses from a relatively small number of participants. Therefore, the authors removed associations with $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_1}) = 0$, before re-running the analyses. Surprisingly, they obtained a reasonably strong, positive correlation between $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_1})$ and $\min(\boldsymbol{P}(\boldsymbol{W_2}|\boldsymbol{W_1}), \boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_2}))$, meaning that free association probabilities do follow the triangle inequality, at least to a certain extent: when $\boldsymbol{P}(\boldsymbol{W_2}|\boldsymbol{W_1})$ and $\boldsymbol{P}(\boldsymbol{W_3}|\boldsymbol{W_2})$ are large (i.e, the $W_1$ and $W_3$ are close to $W_2$), then the

positive correlation typically implies that $\boldsymbol{P(W_3|W_1)}$ is large as well (i.e., $W_1$ and $W_3$ are close to each other).

Thus, given that apparently small methodological differences (e.g., including vs excluding associations with a probability of 0) can dramatically change the results of the analyses, the authors concluded that more research is needed to determine whether and how models of semantics can be compared, in terms of psychological plausibility, based on their adherence (or lack of) to the triangle inequality.

Several conclusions can be drawn from the findings described in this subchapter, namely:

- although the LSA model is arguably the most widely known and used distributional model, it is consistently outperformed by both the Topic model (Griffiths, Steyvers, and Tenenbaum, 2007; Gruenenfelder et al., 2015; Jones et al., 2018), and by HAL-based models (Utsumi, 2015), which define context as short text windows, as opposed to large fragments (i.e., documents). The results argue against employing LSA as the default distributional model, given its apparent lack of cognitive plausibility. This conclusion is also supported by a number of comparative, modelling studies (Baroni et al., 2014; Pereira et al., 2016; Schnabel et al., 2015), which have shown that neural network models (e.g., CBOW, Skip-gram, GloVe) are almost always better than traditional models (e.g., LSA, HAL), when tested on data obtained from tasks such as semantic similarity rating, semantic categorization, and semantic priming.
- it is very difficult to estimate the generality of the results, due to considerable methodological differences between the studies. Some of these differences are related to how the free association network is defined, which depends on choices regarding the directedness of the network (i.e., having directed vs undirected edges), the criterion for including an edge (i.e., the minimum absolute frequency of the association corresponding to that edge), and the form of the power-law distribution being evaluated for the model (i.e., truncated vs not truncated). Other differences stem from how the model networks are defined, such as the method used in deriving word neighbourhoods (i.e., using the *k*-nn method, the *cs*-method, or the $\varepsilon$-method; Utsumi, 2015), and the transformation applied to model associations (i.e., no transformation vs applying the choice rule; Jones et al., 2018). Moreover, some studies compare

single models (e.g., the LSA and Topic models; Griffiths, Steyvers, & Tenenbaum, 2007), while other studies contrast entire classes of related models (e.g., “word-as-context” and “document-as-context” models; Utsumi, 2015). Also, all the studies test their models only on the USF norms.

- it might be the case that responses in the free association task use multiple types of linguistic information (e.g., associative and contextual; Gruenenfelder et al., 2015), as suggested by the remarkable performance of the Topic-POC hybrid model, in comparison to that of single models. The distinction between associative and contextual representations can be reframed in terms of the difference between syntagmatic (thematic) and paradigmatic (taxonomic) relations (Kacmajor & Kelleher, 2019; Lapesa, Evert, & Schulte im Walde, 2014; Mirman, Landrigan, & Britt, 2017; Utsumi, 2015). Two words are syntagmatically related if they frequently co-occur in a temporal, spatial and/or linguistic context (e.g., “dog”-“bone”, “sky”-“blue”, “eat”-“fork”), but share few features, if any. In contrast, two words are paradigmatically related if they have many common features (e.g., “duck”-“goose”, “open”-“close”, “green”-“blue”), but do not necessarily co-occur within a spatio-temporal or linguistic context. In this context, the argument put forward by Gruenenfelder and collaborators (2015) rests on the assumption that “a distributional model accumulated from co-occurrence information contains syntagmatic relations between words, while a distributional model accumulated from information about shared neighbors contains paradigmatic relations between words” (Sahlgren, 2008). Therefore, contextual models should be particularly good at predicting semantic similarity ratings (i.e., capturing paradigmatic relations), and considerably less successful at accounting for semantic relatedness ratings (i.e., capturing syntagmatic relations). However, this is the exact opposite of what is reported in the literature (Gerz et al., 2016; Hill et al., 2015). Another related assumption is that similarity between word representations in contextual models is highly correlated with the similarity between featural representations (McRae et al., 2005; Vinson & Vigliocco, 2008), produced by feature generation tasks. This assumption, like the previous one, is incompatible with certain behavioural findings. For instance, by using multidimensional scaling, Maki and Buchanan (2008) showed that the similarity spaces associated with featural

representations are very different from those associated with contextual models. Furthermore, Sadeghi and collaborators (2015) found only a relatively weak correlation between the two similarity spaces. Thus, the psychological justification for the unusual performance of hybrid, associative-contextual models, such as the Topic-POC model, remains elusive.

## 2.4. Predicting brain-imaging data

Functional resonance magnetic imagining has rapidly become one of the most popular and influential techniques for investigating the neural correlates of cognitive processes (Huettel, Song, & McCarthy, 2014). In its most common version, this brain imaging tool measures the metabolic changes triggered by task-related modifications in neuronal processing (i.e., the hemodynamic response), in the form of a blood-oxygen-level dependent (BOLD) signal. The reasons behind its success as a neuroimaging technique are the multiple advantages that it provides. fMRI is non-invasive, highly unlikely to pose a health risk for the overwhelming majority of participants in experiments and has very good spatial resolution. Finally, fMRI allows researchers to evaluate the functional connectivity between different brain regions (Van Den Heuvel & Pol, 2010), either directly (e.g., by observing the structure of white matter tracts, using motion contrast), or indirectly (e.g., by looking at the coactivation of distinct neuronal populations, as indexed through correlation between BOLD signals).

Therefore, fMRI designs have found a variety of application within cognitive science, with respect to the study of several types of aspects: cognitive (e.g., linking individual differences in perception to the topology of the neocortex; Schwarzkopf, Song, & Rees, 2011), clinical (e.g., tracing functional reorganization after brain damage; Rossini et al., 1998; detecting compensatory activation before the onset of neurodegenerative diseases; Bookheimer et al., 2000), developmental (e.g., mapping the maturation of specific functional regions and their interactions; Bunge, Dudukovic, Thomason, Vaidya, & Gabrieli, 2002), and social (e.g., studying the role of the mirror

neuron system in social interactions; Schippers, Roebroeck, Renken, Nanetti, & Keysers, 2010), to name but a few.

## General overview:

Some of the main technical details regarding the studies reviewed in this subchapter are shown in Table 5.

Table 5. Computational studies focused on brain imaging data, covered in this subchapter. The entry for each study describes the linguistic and/or visual models employed (i.e., their type and the text/image corpora on which they are trained), as well as the fMRI datasets on which the models are tested.

| Study | Linguistic models | | Visual models | | Behavioural datasets |
|---|---|---|---|---|---|
| | Types | Corpora | Types | Corpora | |
| Abnar, Ahmed, Mijnheer, & Zuidema (2018) | FastText<br>GloVe<br>LexVec<br>Skip-gram | Various | - | - | Mitchell et al. (2008) |
| Anderson, Bruni, Bordignon, Poesio, & Baroni (2013) | HAL | ukWaC<br>Wikipedia | SIFT | ImageNet | Mitchell et al. (2008) |
| Anderson, Bruni, Lopopolo, Poesio, & Baroni (2015) | HAL | ukWaC<br>Wikipedia | SIFT | ImageNet | Mitchell et al. (2008) |
| Anderson, Kiela, Clark, & Poesio (2017) | Skip-gram | Wikipedia | AlexNet | Google Images | Anderson et al. (2017) |
| Bulat, Clark, & Shutova (2017) | HAL<br>Skip-gram | Wikipedia | AlexNet | Google Images | Mitchell et al. (2008) |
| Murphy, Talukdar, & Mitchell (2012) | HAL<br>LSA | Web-based | - | - | Mitchell et al. (2008) |

*Datasets:*

The fMRI datasets were obtained by scanning participants while performing semantic tasks. The datasets from the studies included in our analysis are those collected by Mitchell and collaborators (2008), and by Anderson and collaborators (2017). The dataset by Mitchell and collaborators (2008) covers 60 concrete nouns, divided into 12 categories. In order to record the fMRI patterns, the participants were shown sets of word-drawing pairs, where the drawing associated with each word depicted that word's referent, and were asked to think about the properties of each

word. The dataset by Anderson and collaborators (2017) consists of 35 concrete nouns, divided into 7 categories, and 35 abstract nouns, divided into the same 7 categories. The fMRI patterns were collected by presenting the participants with a list of words, and asking them to think of a situation they associate with each word. Unless otherwise stated, the studies presented in the next section employed the dataset by gathered by Mitchell and collaborators (2008).

*Results:*

Starting with the influential work by Mitchell and collaborators (2008), a number of studies have tried to evaluate the cognitive plausibility of distributional models of semantics (Abnar et al., 2018; Anderson et al., 2013, 2015, 2017; Bulat et al., 2017; Murphy et al., 2012). The study by Murphy and collaborators (2012) compared various linguistic models, in terms of predicting fMRI patterns. The distributional models included an LSA-like model, five HAL-based models, and a dependency-based model. The researchers found that the dependency-based model had the best performance, followed by the HAL-based models, followed by the LSA-like model. However, the differences in goodness of fit between the models were relatively small. In conclusion, "word-as-context" models appear to slightly outperform "document-as-context" models, and the performance of "word-as-context" models can be improved by including dependency relations.

The study by Abnar and collaborators (2018) focused on comparing a number of linguistic models, based on their ability to predict neural activation patterns. In this regard, it can be viewed as an extension/update of the study by Murphy and collaborators (2012). The linguistic models consisted of GloVe, Skip-gram, FastText, a dependency-based, Skip-gram model (Levy & Goldberg, 2014a), and Lexical Vectors (LexVec; Salle, Idiart, & Villavicencio, 2016). The best performance was obtained for the dependency-based model, whereas the worst performance was obtained by the Skip-gram model. Also, no significant differences in goodness of fit were found between "count" models (i.e., LexVec) and "predict" models (i.e., GloVe, as well as Skip-gram and its variants). To sum up, adding dependency and morphological information improves the performance of the Skip-gram model.

Several studies have also compared the performance of unimodal models (i.e., linguistic or visual) and multimodal models (i.e., linguistic-visual). The study by

Anderson and collaborators (2013) examined the similarities between fMRI brain activation patterns, and representations obtained from linguistic and/or visual (distributional) models. The distributional models included a HAL-based linguistic model (i.e., Window2), SIFT visual models depicting either objects (i.e., Object), or the physical context in which the objects occur (i.e., Context), or objects in context (i.e., Object&Context), as well as combinations between the linguistic model and the three visual models. The authors tested whether different models best correlated with brain activation patterns in areas associated with the same modalities as the models. More specifically, their hypotheses was that the best performing models would be the following, per brain region: (1) for the occipital lobe, the Object model (since the main function of the lobe is to process low-level and mid-level features, such as oriented edges; Bruce, Green, & Georgeson, 2003); (2) for the temporal lobe, the Object model (since the fusiform gyrus is associated with object categorisation; Chao, Haxby, & Martin, 1999; Kanwisher & Yovel, 2006; Peelen & Downing, 2005), the Context model (since the parahippocampus is associated with scene processing; Epstein, 2008), as well as the Window2&Object&Context model (since the medial temporal gyrus, inferior temporal gyrus, and ventral temporal lobe are associated with multimodal integration and concept retrieval; Binder, Desai, Graves, & Conant, 2009); (3) for the parietal lobe, the Context model (since the lobe is associated with spatial cognition; Sack, 2009), and the Window2&Context model (since the angular gyrus is associated with multimodal integration and information retrieval; Binder et al., 2009); (4) for the frontal lobe, the Window2 model (since the lobe is associated with the processing of abstract information; Miller, Freedman, & Wallis, 2002).

With respect to the unimodal models, the results indicated large and statistically significant correlations between the model representations and the fMRI patterns in the occipital, parietal and temporal lobes. The same kind of correlations, but for the frontal lobe, were smaller and did not reach statistical significance. In contrast to the researchers' expectations, the results of the analyses revealed no statistically significant interactions between model type and lobe. With respect to the multimodal models, the researchers found that the linguistic-visual representations were almost always better than their unimodal components, as reflected in stronger correlations with the fMRI patterns. Moreover, for the visual models, the performance of the Object&Context model was usually better than that of the Object and Context models.

In conclusion, both unimodal and multimodal models of semantics can reliably predict brain activity patterns, with a small advantage for the multimodal models. However, the relative performance of the models did not vary as a function of the anatomical region over which the models were tested.

The study by Anderson and collaborators (2015) focused on testing embodied theories of semantics, according to which perceptual information is automatically accessed and employed during word processing. This study is similar to that by Anderson and collaborators (2013), but with the notable difference that it uses a more fine-grained anatomical parcellation. The linguistic model consisted of HAL-based representations, while the visual model was based on SIFT representations. The researchers found that the visual model was better than the linguistic model in predicting activation patterns in the left middle occipital gyrus and the left ventral temporal cortex, two areas associated with mental imagery (Reddy, Tsuchiya, & Serre, 2010; Stokes, Thompson, Cusack, & Duncan, 2009). In contrast, the linguistic model outperformed the visual model in the left middle temporal gyrus, the left inferior frontal gyrus, and the left posterior intraparietal area, three regions typically activated during language-oriented tasks (Devereux, Clarke, Marouchos, & Tyler, 2013; Fairhall & Caramazza, 2013). Interestingly, the results also showed no difference in performance between the models with respect to the left inferior temporal gyrus, an area linked to multimodal integration (Carlson, Simmons, Kriegeskorte, & Slevc, 2014; Devereux, Clarke, Marouchos, & Tyler, 2013), and an advantage of the visual model, over the linguistic one, for the left supramarginal gyrus and left dorsomedial frontal cortex, two regions found to be involved in semantic processing (Binder and Desai, 2011; Binder et al., 2009). To sum up, linguistic models outperform visual models in areas associated with word processing, whereas the opposite finding applies to areas associated with mental imagery.

The study by Bulat and collaborators (2017) compared the performance of linguistic, visual and linguistic-visual (distributional) models, in accounting for neural activation patterns. The linguistic models consisted of two HAL-based models, with or without dimensionality reduction (i.e., DISTRIB and SVD300, respectively), two dependency-based models, with or without dimensionality reduction (i.e., DEPS and DEPS-SVD300, respectively), the Skip-gram model (i.e., EMBED-BOW), and a dependency-based, Skip-gram model (i.e., EMBED-DEPS; Levy & Goldberg, 2014a).

The visual model was AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), while the multimodal models were obtained by combining the visual model with each of the linguistic models. In order to test the models, the authors used both a regression-based approach (Mitchell et al., 2008) and two similarity-based approaches (Anderson, Zinszer, & Raizada, 2016). With respect to the linguistic models, the regression-based approach revealed that the sparse models (i.e., DISTRIB and DEPS) typically performed better than the dense models (i.e., SVD300, DEPS-SVD300, EMBED-BOW, and EMBED-DEPS), whereas the similarity-based approaches found no significant difference between the models. Also, for all three approaches, the visual and multimodal models outperformed the linguistic models, with no statistically significant difference in performance either between the visual and any of the multimodal models, or between the various linguistic-visual models. In conclusion, the authors found strong evidence for the superiority of visual and multimodal models over purely linguistic models, as well as limited evidence of enhanced performance for high-dimensional vs low-dimensional linguistic models.

The study by Anderson and collaborators (2017) focused on testing the Dual Coding theory of semantics (Paivio, 1971, 1986), according to which the mental representation of concrete words consists of both linguistic and visual information, whereas the representation of abstract words is predominantly linguistic. Therefore, linguistic and visual models should have comparable levels of performance when predicting fMRI patterns for concrete words, while for abstract words, there should be a clear advantage for linguistic models. The fMRI data was collected by the authors, for this particular study. The linguistic model was Skip-gram, the visual model was a convolutional neural network very similar to AlexNet, and the multimodal model was a combination of the linguistic and visual models. In line with their expectations, for the concrete words, the authors found no statistically significant difference in performance between the linguistic and visual models. Also, the results indicated that the multimodal model significantly outperformed the visual model, but not the linguistic one. In contrast, for the abstract words, a small advantage was observed for the linguistic model, over the visual model. All in all, based on the ability of distributional models to account for fMRI data, the authors found an interaction between model type

(i.e., linguistic vs visual) and word class (i.e., concrete vs abstract), in the direction predicted by the Dual Coding hypothesis.

The studies we have described provide converging evidence in favour of the cognitive plausibility of distributional models, either unimodal (i.e., linguistic or visual), or multimodal (i.e., linguistic-visual). Several conclusions can be drawn, namely:

- the generality of the findings is hard to estimate, mainly due to the fact that all the studies, with the notable exception of that by Anderson and collaborators (2017), used the fMRI data collected by Mitchell and collaborators (2008). The dataset has several shortcomings, as it is small (i.e., it contains only 60 words) and fails to capture the diversity of spoken/written language (i.e., it consists only of concrete nouns). This is in stark contrast to sets of behavioural data obtained from semantic tasks, such as similarity/relatedness rating (Bruni et al., 2014; Gerz et al., 2016) and free association (De Deyne et al., 2019; Nelson et al., 2004), which include thousands of words, covering various parts of speech and levels of concreteness.
- whether or not models can reliably predict fMRI patterns does not seem to depend on the particular method used for matching the two types of representations. Some studies employed a multiple linear regression approach (Abnar et al., 2018; Anderson et al., 2013; Bulat et al., 2017; Murphy et al., 2012), whereas other studies employed either representational similarity analysis (Anderson et al., 2015), or related approaches (Anderson et al., 2017; Bulat et al., 2017). Regardless of the approach, distributional models were successful in accounting for brain activation patterns.
- the relative performance of the various classes of models (i.e., linguistic, visual, and linguistic-visual) seems to be fairly consistent across participants, even though the results reveal substantial inter-subject variability (Anderson et al., 2017; Bulat et al., 2017). Most of this variability is likely due to inter-individual differences in functional localization (Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000), but other factors may also play a role: for instance, in the study by Mitchell and collaborators (2008), model performance had a strong, negative correlation with the amount of estimated head motion.

- the results depend on the how the fMRI data was partitioned into anatomical regions. For instance, the study by Anderson and collaborators (2013) found that, when using lobes as anatomical regions of interest, the relative performance of the models did not change as a function of brain region. However, in a subsequent study, Anderson and collaborators (2015) employed a more fine-grained division of brain areas, and found that the best performing models differed from region to region. For each region, the modality of winning model matched the dominant modality associated with that region, as inferred from the neuroscientific literature. In contrast, the rest of the studies focused on whole-brain analyses, which did not allow testing specific functional predictions.
- with respect to differences in performance between linguistic models, the studies provide a complex and sometimes surprising picture. For instance, models that rely on dependencies outperformed models that do not, in some cases (Abnar et al., 2018; Murphy et al., 2012), but not in others (Bulat et al., 2017). When comparing the results of the studies with those obtained from fitting behavioural data (e.g., Bruni, Boleda, et al., 2012; Bullinaria & Levy, 2012; Riordan & Jones, 2011), the qualitative trends were similar in some cases (e.g., an advantage for "word-as-context" models over "document-as-context" models; Murphy et al., 2012), while reversed in others (e.g., a decrease in performance after applying dimensionality reduction; Bulat et al., 2017).
- when comparing unimodal models to multimodal ones, the performance of the linguistic-visual models was usually at least as good as that of their linguistic and visual components (Anderson et al., 2013, 2017; Bulat et al., 2017). These results mirror those obtained in studies where distributional models were used to predict behavioural data (Bruni et al., 2014; Feng & Lapata, 2010; Silberer & Lapata, 2014), and provide evidence for the role of visual information in semantic processing. This role was also likely amplified by the fact that the objects were presented to the participants using both a graphical sketch and a verbal label, thus increasing the saliency of perceptual information. However, the relative performance of the linguistic and visual models was found to depend on the concreteness of the stimuli (Anderson et al., 2017), as well as on the anatomical region from which the fMRI data was recorded.

## 2.5. Conclusions

In this chapter we have presented results from a wide range of computational studies, which directly address the question of whether distributional models of semantics are cognitively plausible or not. The models were tested with respect to their ability to fit behavioural data (i.e., ratings, probabilities, and categories), to produce networks with structural properties similar to that of subjective (i.e., free association-based) semantic networks, and to account for fMRI activation patterns elicited by semantic tasks. We believe that the findings included in our review argue strongly in favour of the claim that distributional models of semantics, both unimodal and multimodal, provide a psychologically plausible account of semantic cognition, in terms of both representations and process (for a broader and more theoretical discussion on this issue, see Günther, Rinaldi, & Marelli, 2019). Therefore, although their origins and most frequent applications are in the field of artificial intelligence, this does not mean that distributional models cannot be useful in psychological studies of various aspects of semantic memory. Some of the most general findings from the studies described previously are as follows.

In general, “predict” models outperform “count” models in a wide variety of tasks, such as similarity rating (Pereira et al., 2016), free association (Thawani et al., 2019), semantic priming in lexical decision and naming (Mandera et al., 2017), semantic categorization, and analogy solving (Baroni et al., 2014). Some researchers (Hollis, 2017; Mandera et al., 2017) have argued that the superiority of “predict” models can be at least partially explained by the increased cognitive plausibility of their learning objectives. However, Levy and Goldberg (2014b) have shown that one of most popular “predict” models, namely Skip-gram, learns its representations by implicitly factorizing a (transformed) co-occurrence matrix, which makes it very similar to traditional “count” models, such as LSA and HAL. In addiction, “predict” models have more hyperparameters than “count” models (Levy et al., 2015). The extra hyperparameters (e.g., the number of negative samples) provide more flexibility to “predict” models, and can have a significant effect on performance.

Within the class of “count” models, there is some evidence that “word-as-context” models are better that “document-as-context” models, with respect to their

performance in the semantic categorization task (Riordan & Jones, 2011), as well as their ability to reproduce the degree distribution of free association networks (Utsumi, 2015), and to predict fMRI activation patterns (Murphy et al., 2012). This finding might arise from differences in the psychological interpretation of the learning mechanisms corresponding to the different classes of models. In "word-as-context" models, learning is likely to be driven predominantly by working memory, such that words which are active at the same time (i.e., they appear in the same, narrow text window) increase their association. In contrast, in "document-as-context" models, learning is almost entirely dependent on episodic memory, such that words which are encountered in the same episode (i.e., they appear in the same document) strengthen their association. If working memory plays a larger role than episodic memory, in word learning, than this might explain the relative performance of the two types of models.

Combining linguistic and visual models often increases model performance, as compared to purely linguistic and purely visual models. However, this effect is not universal, and appears to be influenced by a number of factors. Firstly, the addition of visual information typically improves performance for concrete words, but not for abstract words, when predicting similarity/relatedness ratings (Bruni, Boleda, et al., & Tran, 2012; Bruni et al., 2014; Kiela et al., 2014), free association probabilities (Hill & Korhonen, 2014; Kiela et al., 2014), and semantic categories (Bruni et al., 2011, 2014). These results are consistent with the Dual Coding theory of semantics, which states that concrete words are represented using both linguistic and visual information, while abstract words are associated with very little visual information. Secondly, the performance of the purely linguistic models matters: in certain tasks, such as relatedness rating (Bruni et al., 2014; Silberer & Lapata, 2014) and semantic categorization (Baroni et al., 2014; Bruni et al., 2011, 2014; Riordan & Jones, 2011), the linguistic models reaches ceiling performance, leaving little room for improvement. This does not mean that visual information does not play any significant role in those tasks, but instead it seems likely that the task-relevant visual information is already represented in the linguistic input (Louwerse, 2011, 2018; Louwerse & Jeuniaux, 2008). In contrast, in tasks such as free association (Feng & Lapata, 2010; Hill & Korhonen, 2014), where the linguistic models underperform, the addition of visual information can boost performance by between 50% and 200%. Thirdly, at least in the case of predicting relatedness ratings, the method through which the linguistic and

visual representations are combined has an impact on performance (Bruni et al., 2014; Bruni, Uijlings, et al., 2012; Silberer & Lapata, 2014). More specifically, methods which exploit the redundancy between the two modalities, by combining the linguistic and visual dimensions into multimodal dimensions, are more effective than methods which keep the two modalities separate, by concatenating the representations, such that each dimension is either purely linguistic, or purely visual.

The performance of distributional models appears to be strongly task-dependent: it can be either low, in free association (i.e., correlations in the .1 - .3 range; e.g., Hill & Korhonen, 2014), medium, in similarity rating (i.e., correlations in the .1 - .5 range; e.g., Hill et al., 2015), high, in relatedness rating (i.e., correlations in the .4 - .7 range; e.g., Bruni, Boleda, et al., 2012), or very high, in semantic categorization (i.e., purities in the 60% - 100% range; e.g., Baroni, Dinu, & Kruszewski, 2014).

There are at least three reasons why distributional models show relatively poor performance in certain tasks. Firstly, it might be the case that the information employed by the participants in carrying out the task is present in the linguistic input, but the distributional models are not sensitive enough to it. For instance, including syntactic information, in the form of dependencies, significantly improves model performance with respect to predicting similarity ratings (i.e., from the SimLex-999 and SimVerb-3500 norms), by allowing the models to better discriminate between semantic similarity and relatedness (Hill et al., 2015; Levy & Goldberg, 2014a). Similarly, introducing dependency data increases performance for predicting fMRI patterns (Abnar et al., 2018; Murphy et al., 2012).

Secondly, non-linguistic information, which is only partially captured by linguistic models, might play an important role in task performance. As discussed earlier in this chapter, adding visual information improves model performance for concrete words, in line with current theories of word representation (e.g., Dual Coding theory). With respect to abstract words, three recent studies (De Deyne et al., 2018; Rotaru & Vigliocco, 2019, 2020; see Chapter 3) found that including emotional information has a beneficial effect on performance, regardless of whether the emotional representations are derived from subjective ratings or word-emoji co-occurrences.

Thirdly, for certain tasks, the processes involved in generating a response might go beyond just performing a feature-by-feature comparison, as is the case when

computing the cosine similarity or Euclidean distance between two representations. More specifically, three studies (De Deyne, Verheyen, et al., 2016; Rotaru, Vigliocco, & Frank, 2016, 2018; see chapter 4) have shown that, by implementing a spreading activation mechanism (Collins & Loftus, 1975), operating over semantic representations, it is possible to increase model performance for eight different similarity and/or relatedness datasets, including SimLex-999, SimVerb-3500, MEN, and SL. Also, by simulating a probabilistic retrieval process, based on Luce's choice rule, Jones and collaborators (2018) improved the ability of model-based networks to match the network properties of free association-based semantic networks.

From a structural point of view, the properties of semantic networks derived from distributional models are similar to those obtained from free association norms. In particular, both types of networks typically have a small world structure (i.e., local neighbourhoods are strongly interconnected, and it is possible to connect any two words with very short paths), and are "scale-free" (i.e., they contain an unusually large number of hubs, namely words which are directly connected to many other words). However, it is worth noting that generating semantic networks from distributional models and free association norms is by no means a straightforward process, since it involves making numerous decisions, referring to the directedness of the networks (i.e., directed vs undirected), the criterion for including an edge in the networks, the transformation that is applied to the weights of the edges (e.g., no transformation vs normalization), and the type of power-law distribution that is matched against the network structure (i.e., pure vs truncated). The problem here is that, at the current moment, we do not know what the optimal (i.e., most cognitively plausible) set of choices actually is. Therefore, if we obtain a poor match between model-based and free association-based networks, it could be the case that either we have a poor model, or instead that the model is good, but we made a wrong set of choices in building the networks.

A related issue is that evaluating model performance becomes problematic. For instance, in the case of predicting relatedness/similarity ratings, goodness of fit is neatly captured by a single metric (i.e., Spearman/Pearson correlation between the vector cosines and the subjective ratings). In contrast, when matching network structures, goodness of fit can be assessed with respect to a variety of measures (e.g., Gruenenfelder et al., 2015; Utsumi, 2015), such as number of edges,

number/proportion of nodes in the largest strongly connected component, average node degree, maximum node degree, diameter, average shortest path length, and average clustering coefficient. Of all these factors, some are likely to have a stronger impact on task behaviour than others, and, therefore, should be considered more important (for a review, see Siew et al., 2019). From the literature on semantic richness, discussed in Chapter 5, we know that one such factor is node degree (i.e., number of neighbours), since it is a significant predictor of task performance in certain tasks. Nevertheless, since neighbourhood-based measures correlate significantly with at least 10 semantic and non-semantic factors (Yap et al., 2012; also see Chapter 4), ranging from imageability to number of letters, it is very difficult to attribute behavioural effects to network measures.

Distributional representations can reliably predict fMRI-based activation patterns, by learning a mapping between model-based and fMRI-based representations. This finding holds regardless of whether the mapping is performed using first-order, featural spaces (e.g., via linear regression; e.g., Abnar et al., 2018; Murphy et al., 2012), or second-order, similarity spaces (e.g., via Representational Similarity Analysis; e.g., Anderson et al., 2015).

In line with previous findings, multimodal models typically outperform unimodal models, and this effect depends on word concreteness. However, the spatial level of analysis plays a crucial role: multimodal models usually provide the best results only when predicting activation patterns for entire lobes or the whole brain. In contrast, when more specific anatomical regions are used, the winning models are the ones that match the modality associated with a given region (e.g., visual models outperform the other models when tested on predominantly visual areas).

Unfortunately, the use of fMRI data has at least two limitations. A general shortcoming arises from the fact that model performance depends on the interaction of three factors (Bullinaria & Levy, 2013), namely (1) the signal-to-noise ratio of the fMRI recordings, (2) the quality of the model representations, and (3) the complexity of the mapping between the two types of representations. When employing a multiple linear regression mapping, as is typically the case, the authors found that the amount of noise present in the fMRI recordings is by far the largest limiting factor with respect to model performance. Furthermore, reducing the noise level appears to be a non-trivial task: for instance, the authors showed that increasing the number of

presentations per stimulus seems to have little effect on performance. A more particular shortcoming is related to relying almost exclusively on the dataset collected by Mitchell and collaborators (2008). The dataset is relatively small (i.e., it contains only 60 words), and not very representative from a linguistic point of view (i.e., it consists only of concrete nouns). An additional problem is that the task on which the models are tested is relatively easy, leading to accuracy scores of over 70% (Abnar et al., 2018; Bulat et al., 2017; Murphy et al., 2012). Taken together, these two limitations suggest that the current model testing setup might not be sensitive enough for comparing state-of-the-art distributional models.

# 3. Constructing semantic models from words, images, and emojis[7]

## 3.1. Introduction

Despite the success of distributional, linguistic models in accounting for behavioural effects in a variety of semantic tasks, all these models suffer from the "symbol grounding problem" (Harnad, 1990). This problem refers to the fact that the meaning of symbols (i.e., words) is computed based on other symbols (i.e., other words), but without connecting those symbols to their real-world referents. For instance, distributional representations might tell us that "hammer" and "nail" are semantically linked, but they do not capture the sensory and motor information obtained by interacting with the two objects (e.g., the shape of the hammer, the texture of its handle, the motion used when hitting the nail, or the sound produced when the nail is struck). As a solution to this problem, embodied theories of semantics (e.g., Glenberg, Graesser, & de Vega, 2008) have argued that the sensory-motor representations generated by our experiences with the world play an important role in determining word meaning. A large number of behavioural studies provide support for embodied theories. For instance, Solomon and Barsalou (2004) used a property verification task (e.g., "is *face* a property of *gorilla*?") to show that variables associated with perceptual effort (e.g., the size and position of the "face", relative to a "gorilla") were significant predictors of response times and error rates, which suggests that perceptual simulations were involved. Neuroimaging studies offer additional evidence for embodied theories. These studies indicate that the brain regions responsible for perception, action, and emotion (partially) overlap with, or are anatomically close to, the brain regions used in processing conceptual knowledge related to perception and action. For instance, in an fMRI study by Hauk, Johnsrude, and Pulvermuller (2004), the researchers found that reading words referring to actions (e.g., "lick", "pick", "kick")

[7] Adapted from (Rotaru & Vigliocco, 2019, 2020).

produced an increased activation of the premotor cortex areas involved in actually performing those actions.

Recent computational models of semantics reconcile distributional and embodied theories, by combining linguistic and perceptual (i.e., visual) representations. The underlying assumption is that the two classes of representations capture complementary aspects of meaning. For example, using a semantic categorization task, Riordan and Jones (2011) showed that distributional models focus mostly on information about actions, functions, and situations, but not on the perceptual properties of objects. Instead, such properties are better captured by featural models, which are rich in perceptual information. The fact that language and vision provide complementary sources of information is best illustrated by the finding that multimodal, textual-visual models outperform both purely linguistic and purely visual models, in a wide range of tasks (see Chapter 2). Moreover, the superiority of multimodal models over unimodal ones holds when tested over a variety of model architectures, training datasets, and multimodal integration methods.

However, empirical work has shown that semantic representations are not only grounded in sensory-motor experience but also in emotion. A vast literature supports the finding that emotion plays a significant and pervasive role in human cognition (for a review, see Dolan, 2002). Given their behavioural relevance, it is perhaps not surprising that emotional stimuli capture attention more than non-emotional ones. For instance, using a visual search paradigm, Öhman, Flykt, and Esteves (2001) showed that, within an array of images, fear-relevant pictures are found more quickly than fear-irrelevant ones. Moreover, since the processing of fear-relevant stimuli is not affected by their position in the display, or by the number of distractors, it appears that such stimuli capture attention in an automatic, bottom-up manner. Emotional stimuli can also have an effect on perception, particularly on spatial perception. One study (Riener, Stefanucci, Proffitt, & Clore, 2011) found that participants judge a hill to be steeper while in a sad mood, as opposed to a happy mood. Another study (Gasper & Clore, 2002) showed that sad participants tend to focus more on the details of a visual display (i.e., local perception), while happy participants perceive the image in a more holistic manner (i.e. global perception). Emotion is an important factor in memory (Blaney, 1986; Eich, Macaulay, & Ryan, 1994), such that participants are better at recalling information for which the emotional content and/or mood during encoding,

are compatible with the participants' mood during testing. Finally, the effects of emotion have also been observed with respect to a number of higher cognitive functions, such as reasoning, decision-making and problem-solving (Forgas 1995; Isen, Daubman, & Nowicki, 1987; Schwarz & Clore 1996).

It is also known that emotion is an important factor in processing words (e.g., Kousta, Vinson, & Vigliocco, 2009). Kousta, Vigliocco, Vinson, Andrews, and Del Campo (2011) found that a much larger number of abstract than concrete concepts are valenced (have positive or negative emotional associations) and by virtue of being valenced, they are processed faster than neutral matched words. Vigliocco et al. (2014) further showed that because of their greater affective associations, abstract word processing engages the limbic emotional system and Ponari, Norbury, and Vigliocco (2018) showed that emotionally valenced words are learnt earlier and better recognized by children up to 9 years of age. Within a general embodiment framework, the hypothesis is that semantic representations do not only embed sensorimotor properties but also emotional properties. Emotional properties may be especially important for abstract concepts (e.g., "religion", "society", "idea"), however, emotional associations are not limited to abstract words and therefore, we argue, they play a general role in semantic representation.

While many models have integrated linguistic and visual information, only one previous study has considered emotional information along with visual and linguistic information (De Deyne, Navarro, Collell, & Perfors, 2018). The authors examined the change in performance for distributional models of semantics, when adding visual and emotional information. They tested the assumption that external language models (i.e., distributional models, trained on word corpora) are relatively poor at representing visual and affective information, in comparison to internal language models (i.e., models based on free association norms). They found that adding visual and emotional information led to little or no improvement for internal language models, but a moderate positive effect for external language models. Here, we develop a quite different multimodal model of semantics that incorporates linguistic, visual and emotional information from corpora of text, images and emojis, and test the multimodal model against existing datasets of ratings of semantic similarity/relatedness of words. We use a state-of-the-art emotion model (i.e., DeepMoji; Felbo, Mislove, Søgaard, Rahwan, & Lehmann, 2017) and we improve the coverage of the visual model we use.

While state-of-the art distributional language models (Pereira et al. 2016) have large coverage of words and have been widely tested for their ability to fit human semantic similarity/relatedness data, this is not the case for visual models. Thus, before being able to develop models that embed linguistic, visual and emotional information, we extend the coverage of existing visual models and carry out their evaluation in order to decide which one to use for our multimodal models. We expect that the integrated model will outperform a purely linguistic, as well as models that combine linguistic-visual and linguistic-emotional information. In addition, we expect that adding visual or emotional representations will especially be beneficial for more concrete concepts whereas emotional information will especially be beneficial for more abstract concepts, in line with the empirical evidence reviewed above (and with initial findings from De Deyne et al., 2018).

## 3.2. Methods

**Datasets of behavioural data:**

We use four datasets of similarity/relatedness ratings to carry out evaluation of the models. The datasets are SimLex-999, SimVerb-3500, MEN, and SL[8] (see Chapter 2). We chose these norms mainly because they are some of the largest datasets currently available, but also because the word pairs they contain cover are very diverse in terms of concreteness and valence, as well as parts of speech. With respect to word pair concreteness, SimLex-999 (*M* = 3.62, *SD* = 1.07) and SimVerb-3500 (*M* = 3.1, *SD* = 0.7) cover a broad range of values, whereas MEN (*M* = 4.4, *SD* = 0.49) and SL (*M* = 4.83, *SD* = 0.14) consist predominantly of concrete words.

[8] The SL norms contain both semantic and visual similarity ratings. To make the analyses comparable across the different datasets, we employ only the semantic similarity data.

**Linguistic model:**

Our linguistic model of choice is GloVe (Pennington et al., 2014), trained on a corpus of 6 billion words, using 300-dimensional representations. This model was proposed as a solution to certain (potential) shortcomings of two classes of popular distributional models, namely global matrix factorization models, such as LSA (Landauer & Dumais, 1997), and local context window models, such as CBOW and Skip-gram (Mikolov, Chen, et al., 2013). According to the authors, the first class of models perform poorly on word analogy tasks, denoting a sub-optimal vector space structure, while the second class of models do not exploit global co-occurrence information. GloVe has been shown to have a performance better than, or equal to, several state-of-the-art distributional models, such as vLBL and ivLBL (Mnih & Kavukcuoglu, 2013), HPCA (Lebret & Collobert, 2014), as well as CBOW and Skip-gram, in tasks that involve solving analogies, predicting similarity ratings, and recognizing named entities. This makes GloVe one of the best linguistic models available.

**Visual model:**

To select the best model for our study, we compared five of the most popular visual models for object recognition, based on their performance in predicting subjective similarity/relatedness ratings. For a technical description of each model, see Appendix A[9]. The first model (K&B) is the convolutional model employed by Kiela and Bottou (2014; 6144 dimensions), trained on the ESP Game dataset (Von Ahn & Dabbish, 2004), using the mean of the feature vectors per each word. The second, third, and fourth models are AlexNet (Krizhevsky et al., 2012; 4,096 dimensions), GoogLeNet (Szegedy et al., 2015; 1,024 dimensions), and VGG-19 (Simonyan & Zisserman, 2014; 4,096 dimensions), trained on images obtained from Google Image Search, following the approach employed by Kiela and collaborators (2016). The fifth

[9] Since the K&B model is extremely similar to the AlexNet model, we only describe the latter.

model uses SIFT descriptors (Lowe, 2004), computed over the NUS-WIDE dataset (Chua et al., 2009; 500 dimensions)[10]. The models were tested on similarity/relatedness ratings for 7,611 word pairs, covered by all models and obtained by merging the four sets of ratings. Before merging, the scores in each set were linearly rescaled to fall in the interval [0,1], to make them comparable across datasets. The performance of the models was evaluated using the Spearman correlation between the cosine similarity of the model representations, and the similarity/relatedness ratings from the norms. The results are shown in Figure 1.

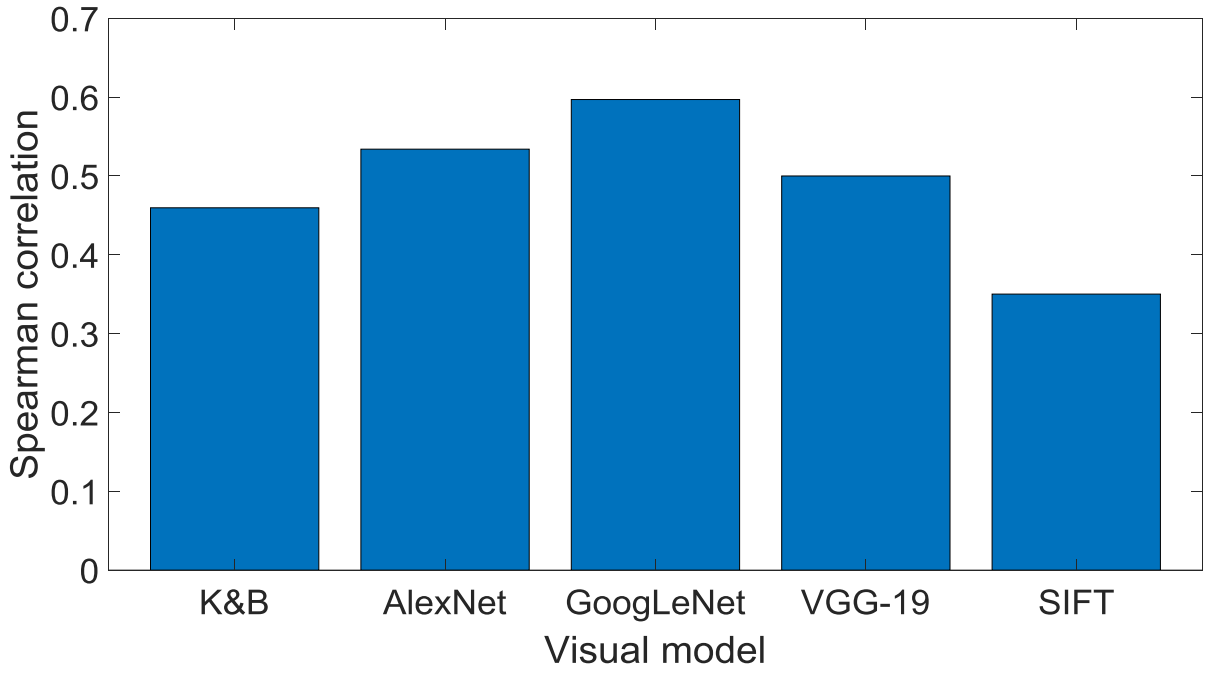


Figure 1. Spearman correlations between cosine similarities, from the visual models, and subjective similarity/relatedness ratings, from the SimLex-999, SimVerb-3500, MEN, and SL datasets.

All the correlations are significant ($p < .001$), suggesting that model-based similarities are reliable predictors of subjective similarity/relatedness ratings. Since we want to find the best model, we apply the Fisher Z-Transformation and then run two-tailed Z-tests for all the 10 possible pairings of models. All the differences are

[10] For all the CNN models, we used the hyperparameter values from their original papers. We trained the models for 10 epochs, with a batch size of 32. For the SIFT model, we used the representations from (Chua et al., 2009).

significant ($p < .004$), and they reveal that GoogLeNet has the highest performance, followed by AlexNet, VGG-19, K&B, and SIFT. Thus, we use GoogLeNet.

**Emotional model:**

Selecting a quality emotional model is not nearly as straightforward as choosing the linguistic and visual models. Most modern models of emotion typically fall in one of two classes, namely psychological models and computational models. Psychological models are based on behavioural data, typically in the form of affective word ratings. Such models are usually informed either by discrete (or categorical) theories of emotion, or by dimensional theories of emotion (for a review, see Hamann, 2012). Discrete theories assume the existence of a set of basic emotions (e.g., “anger”, “disgust”, “fear”, “happiness”, “sadness”, and “surprise”; Ekman, 1992; “acceptance”, “anger”, “anticipation”, “disgust”, “joy”, “fear”, “sadness”, “surprise”; Plutchik, 1980), which share a number of essential characteristics, such as the fact that they are innate, contribute to the survival of the individual and the species, exist in all cultures, and can be reliably identified on the basis of facial and vocal expressions, as well as autonomic responses (Ortony & Turner, 1990). Also, their basic status means that they serve as components for more complex emotions. In contrast, according to dimensional theories, emotions can be seen as regions in a multidimensional space (Rubin & Talarico, 2009). Each dimension forms a continuum, and the most frequently used dimensions are valence (e.g., the pleasantness of the emotion), arousal (i.e., the intensity of the emotion), and dominance (i.e., the amount of control over the cause of the emotion) (Russell & Mehrabian, 1977). Following either of the two classes of theories, a variety of affective norms have been collected, such as SenticNet 3.0 (i.e., 30,000 words, rated for polarity/valence; Cambria, Olsher, & Rajagopal, 2014), EmoSenticNet (i.e., 13,741 words, associated with “joy”, “disgust”, “sadness”, “surprise”, and/or “fear”; Poria, Gelbukh, Cambria, Hussain, & Huang, 2014), WordNet-Affect (i.e., 4,787 words, labelled as emotions, moods, situations eliciting emotions, or emotional responses; Strapparava & Valitutti, 2004), and

SentiWordNet 3.0 (i.e., 114,759 word senses, rated for positivity, neutrality, and negativity; Baccianella, Esuli & Sebastiani, 2010).

In contrast to psychological models, computational models derive affective word representations in an indirect manner, by predicting affective labels/classes associated with the documents in a corpus. Two broad classes of computational models can be distinguished, namely bag-of-words models and neural network models. Bag-of-words models are typically trained to perform sentiment classification (i.e., using positive vs negative ratings; Maas et al., 2011), emotion classification (i.e., using basic emotions; Mohammad & Kiritchenko, 2015), or mood classification (Leshed & Kaye, 2006; Mishne, 2005). One important shortcoming of such models is that they do not take syntax into consideration: for instance, given that the words “lack” and “flaws” have a negative connotation, the expression “a lack of flaws” would be classified as being (strongly) negative, rather than positive. Another problem is that the dimensionality of the representations is relatively low (i.e., typically less than 50), being limited by either the nature of the nature of the affective information, namely the number of classes, or by the computational resources needed to train the models. Modern models are usually based on recurrent neural networks (Abdul-Mageed & Ungar, 2017; Tai, Socher, & Manning, 2015), as well as feedforward neural networks (Tang et al., 2014) and convolutional neural networks (Kim, 2014). They are trained on the same tasks as the traditional models, as well as on the emoji classification task (i.e., determining the type of emoji present in a document), which exploits a rich source of affective information. Unlike traditional models, recurrent neural networks have the advantage of being sensitive to syntax, given that they process text in a sequential manner and have an internal representation (or memory) of the words processed before the current word. An additional benefit of using neural network models is the fact that the dimensionality of the representations is high (typically greater than 100), since it is no longer tied to the number of classes.

In order to select an appropriate model for our study, we compared four of the most recent and well-performing models of emotion. For a technical description of the general architectures upon which the models are based, see Appendix A. The first

model (CNN)[11] is the convolutional model from (Coman, Nechaev, & Zara, 2018; 300 dimensions). The second and the third model are recurrent neural networks, namely the gated recurrent unit (GRU) model and the long short-term memory (LSTM) model from (Çöltekin & Rama, 2018; 128 dimensions)[12]. The fourth model is the stacked LSTM model (DeepMoji) from (Felbo et al., 2017; 256 dimensions)[13]. All the models were trained on the Twemoji dataset (Cappallo, Svetlichnaya, Garrigues, Mensink, & Snoek, 2019), consisting of 15 million tweets, each containing one or more emojis. From the full corpus, we kept only the tweets associated with emojis of facial expressions, as they are reliable and unambiguous indicators of emotion. This resulted in a subset of almost 10 million tweets. The task of the models was to predict the emoji(s) co-occurring with each tweet, based on the text contained by the tweet. After training, we tested the models' ability to account for similarity/relatedness ratings, for 12,659 word pairs covered by all the models and generated by combining the four sets of ratings, and then linearly scaling the values to the range [0,1]. Model performance was measured using the Spearman correlation between the cosine similarity of the model representations, and the ratings from the norms. The results are shown in Figure 2.

[11] During training we used the hyperparameter values from Table 2 in (Coman, Zara, Nechaev, Barlacchi, & Moschitti, 2018), except for the dimensionality of the representations (i.e., denoted as "Output dimension" in the table), which we set to 300, the number of words in the vocabulary (i.e., denoted as "Input dimension" in the table), which we set to 100,000, and the maximum sequence length, which we set to 50. We made these changes in order to replicate the model from (Coman, Nechaev, et al., 2018). The model was trained for 10 epochs, with a batch size of 512.

[12] During training we used the hyperparameter values from Section 2.2 of (Çöltekin & Rama, 2018). The models were trained for 10 epochs, with a batch size of 512.

[13] During training we used the hyperparameter values from Section 3.2 of (Felbo et al., 2017). The model was trained for 10 epochs, with a batch size of 256.

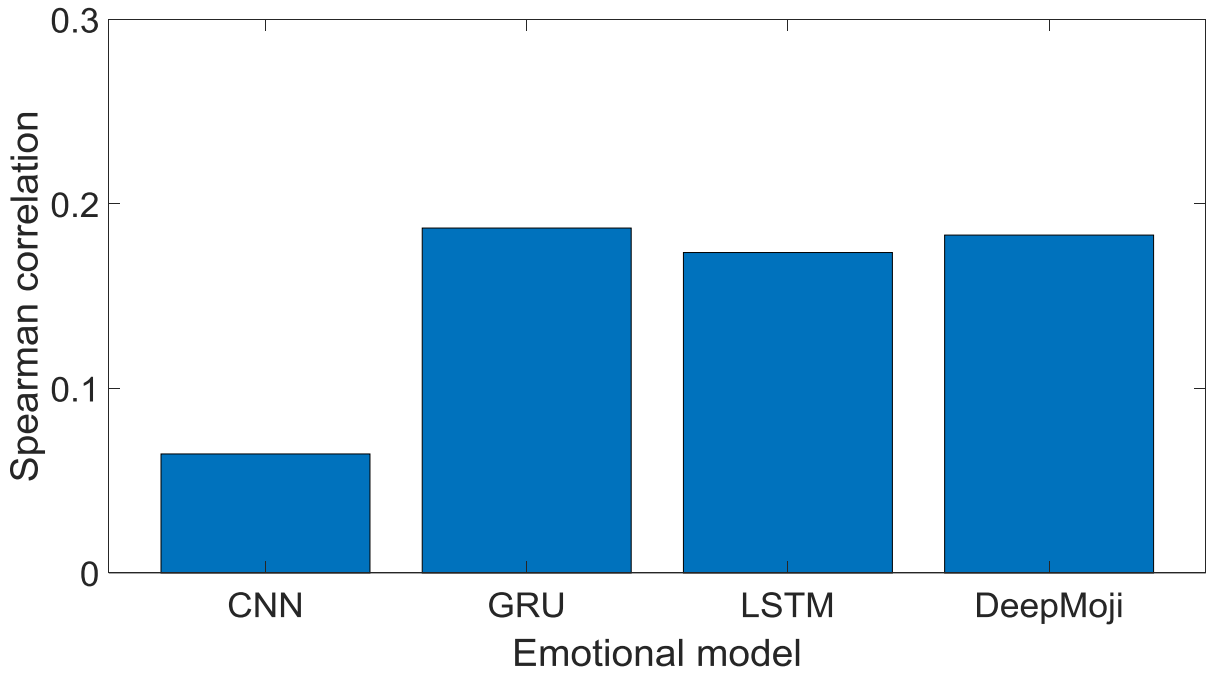


Figure 2. Spearman correlations between cosine similarities, from the emotional models, and subjective similarity/relatedness ratings, from the SimLex-999, SimVerb-3500, MEN, and SL datasets.

All the correlations are significant ($p < .0001$)[14], which indicates that the models capture some of the affective information reflected in the subjective similarity/relatedness ratings. Given that we want to find the best model, we apply the Fisher Z-Transformation and then run two-tailed Z-tests for all the 6 possible pairings of models. Only the differences between CNN and all the other models are significant ($p < .0001$), and they reveal that GRU, LSTM, and DeepMoji are comparable in performance, while being better than CNN.

As a result, our model of choice is DeepMoji, but trained over a large corpus of 1.2 billion tweets, as made publicly available by the authors of the model. This version of the model has been shown to obtain state-of-the-art performance in tasks involving emotion and sentiment analysis, as well as sarcasm detection. The model is very different from the one by De Deyne and collaborators (2018), which was constructed by concatenating valence, arousal, and potency ratings, for men and women separately (i.e., 6 dimensions), from the study by Warriner, Kuperman, and Brysbaert (2013), with valence, arousal, and dominance ratings, from the study by Mohammad

[14] The Bonferroni correction was applied when assessing the statistical significance of all the results presented in this study.

(2018). DeepMoji provides better representations for our purposes than ratings because firstly, a model trained over a corpus of tweets, rather than subjective ratings, makes the emotion model more comparable to the linguistic and visual models, both trained over corpora. Secondly, DeepMoji covers 50,000 words, whereas the combined affective norms cover less than 14,000 words. Finally, the model operates with 256-dimensional vector representations, and is trained to predict the occurrence of 64 types of emojis, and thus should be able to represent complex patterns of word similarity, driven by richer emotional information than that captured by subjective norms. The emojis employed by the model, as well as their frequency, are shown in Figure 3.

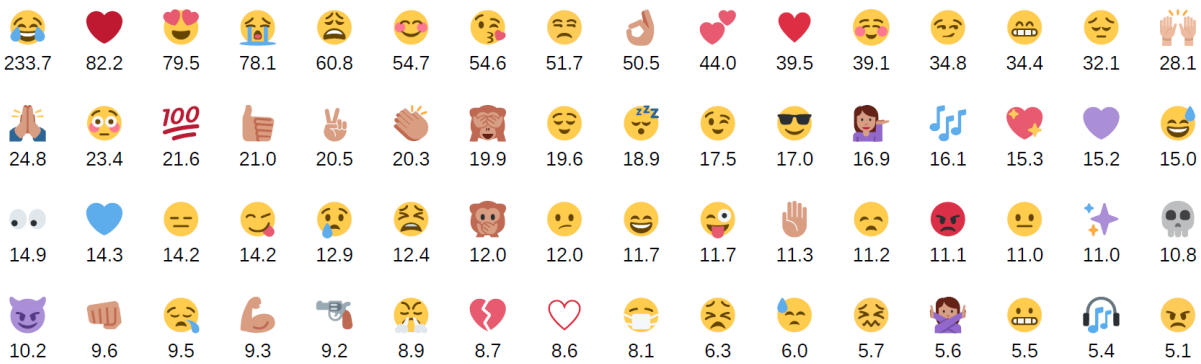


Figure 3. Emojis covered by the DeepMoji model, together with their frequency, in millions. From “Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm”, by B. Felbo and collaborators. (2017). Reprinted with permission.

In order to obtain a more detailed understanding of the emotional information captured by the DeepMoji model, we used PCA and extracted the first 10 principal components from the model representations. Then, we computed the (absolute) Spearman correlations between each component and the affective norms collected by Mohammad (2018), covering subjective ratings of word valence, arousal, and dominance. Our analysis included the 13,678 words common to both the model and the norms. The results are shown in Figure 4.

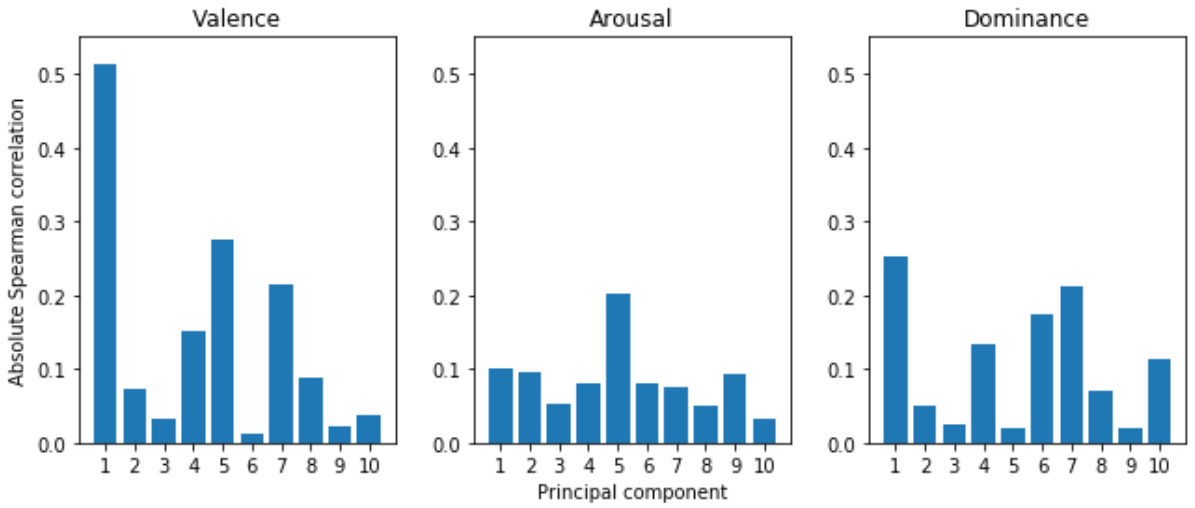


Figure 4. Absolute Spearman correlations between DeepMoji-based principal components and valence, arousal, and dominance ratings.

All the 10 components correlate significantly with valence (except for P6, where $p = .19$), arousal, and dominance. This provides additional evidence that the DeepMoji model is sensitive to affective dimensions of word meaning. Since PC1, PC5, and PC7 seem to be most strongly correlated with the subjective ratings, we also decided to find the words corresponding to the most extreme values for each of the three components. The words and their mean affective ratings are shown in Table 6.

Table 6. Words with the lowest and highest values for PC1, PC5, and PC7, as well as the words' mean affective ratings. The words with the lowest values are written in boldface, for better visibility.

| | **PC1 (low)** | PC1 (high) | **PC5 (low)** | PC5 (high) | **PC7 (low)** | PC7 (high) |
|---|---|---|---|---|---|---|
| Words | **headache** | follow | **bitch** | happiest | **naked** | important |
| | **sad** | birthday | **beg** | hope | **stalk** | appreciate |
| | **sucks** | happy | **hood** | smile | **imagine** | great |
| | **stressed** | direction | **hoe** | gift | **secret** | deserve |
| | **tired** | amazing | **thug** | dream | **sunshine** | follow |
| | **fail** | proud | **savage** | happiness | **addicted** | merry |
| | **crying** | happiest | **fuck** | tour | **nervous** | respect |
| | **impossible** | pizza | **hell** | notice | **drunk** | proud |
| | **migraine** | cute | **pussy** | happy | **sos** | inspiration |
| | **pain** | sexy | **ass** | reading | **beg** | supporting |
| Mean val. | **0.14** | 0.88 | **0.25** | 0.86 | **0.47** | 0.85 |
| Mean aro. | **0.62** | 0.65 | **0.75** | 0.49 | **0.56** | 0.55 |
| Mean dom. | **0.29** | 0.68 | **0.44** | 0.6 | **0.46** | 0.75 |

When comparing the ratings for two sets of extreme words, the results indicate significant differences in valence (all $|t| > 4.29$, $p < .001$) and dominance (all $|t| > 2.28$, $p < .035$), for each of the three components. However, the two sets of extreme words differ significantly, in terms of arousal ($t = 2.83$, $p = .011$), only for PC5. These results are consistent with those from the correlation analysis, and suggest that the DeepMoji model is most sensitive to valence, followed by dominance, followed by arousal. In addition, PC1, PC5, and PC7 seem to capture different aspects of emotion. This is most evident when looking at the words from the lower extreme (i.e., the ones displayed in boldface): for PC1, they seem to be related to bodily causes and effects of emotion (e.g., “headache”, “migraine”, “pain”, “crying”, “tired”), and belong to polite language; for PC5, they refer mostly to individuals (e.g., “bitch”, “hoe”, “thug”, “savage”), and belong to vulgar language; for PC7, they include actions (e.g., “stalk”, “beg”), physical states (e.g., “addicted”, “drunk”), and pieces of information (e.g., “SOS”, “secret”), all belonging to polite language.

## 3.3. Results

We tested whether linguistic-visual and linguistic-emotional models are indeed better than a purely linguistic one, as well as whether it is the case that linguistic-visual-emotional models are better than linguistic-visual, linguistic-emotional and purely linguistic ones. We also examined whether the models behave differently for concrete and abstract word pairs.

**Linguistic-visual and linguistic-emotional models vs purely linguistic model:**

To evaluate the change in goodness of fit associated with adding a visual component to the purely linguistic model, we began by normalizing the linguistic and the visual representations to unit length. Next, we concatenated the linguistic representations with the visual ones, assigning a weight of 1 to the linguistic components, and weights from 0.2 to 2, in steps of 0.2, to the visual components. Both here and in our further analyses, we tested various weights, since it was not clear

which weight would produce optimal results. Finally, for each of the four similarity/relatedness datasets, we compared the 10 resulting linguistic-visual models with the purely linguistic model, by normalizing the correlations and using two-tailed Z-tests. The same type of analyses were run for the linguistic-emotional models. The results are shown in Figure 5, Figure 6, Figure 7, and Figure 8.

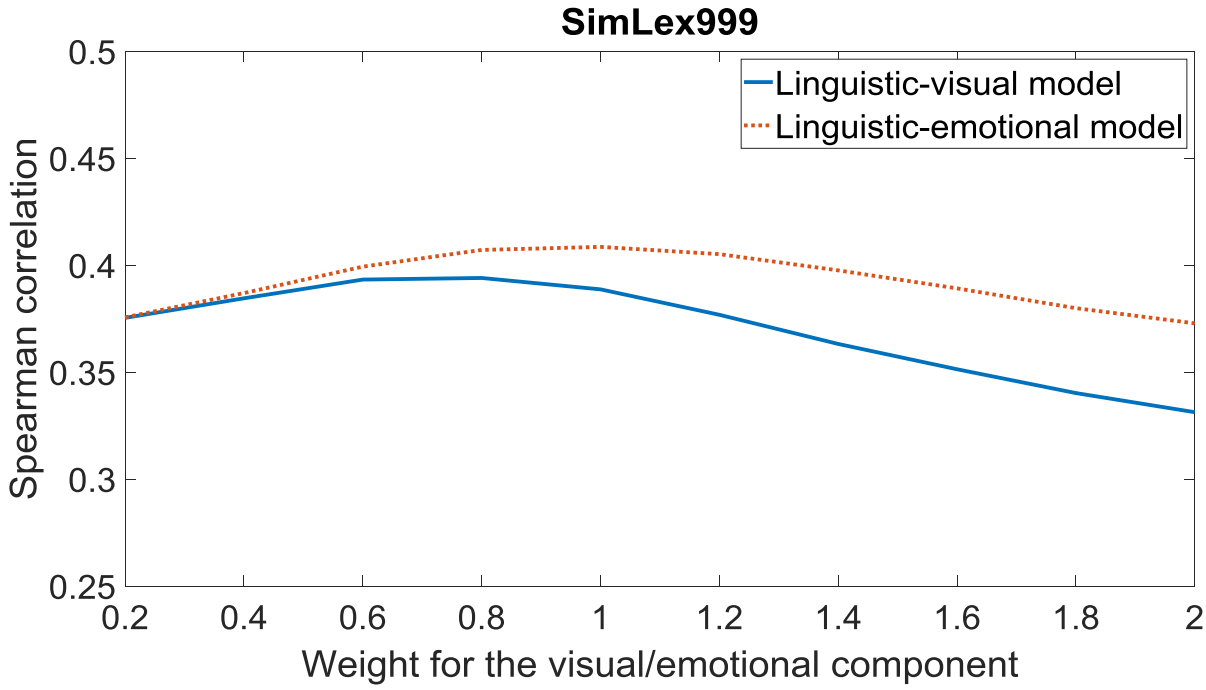


Figure 5. Model performance for the linguistic-visual and linguistic-emotional models, in predicting similarity/relatedness ratings from the SimLex-999 dataset. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2.

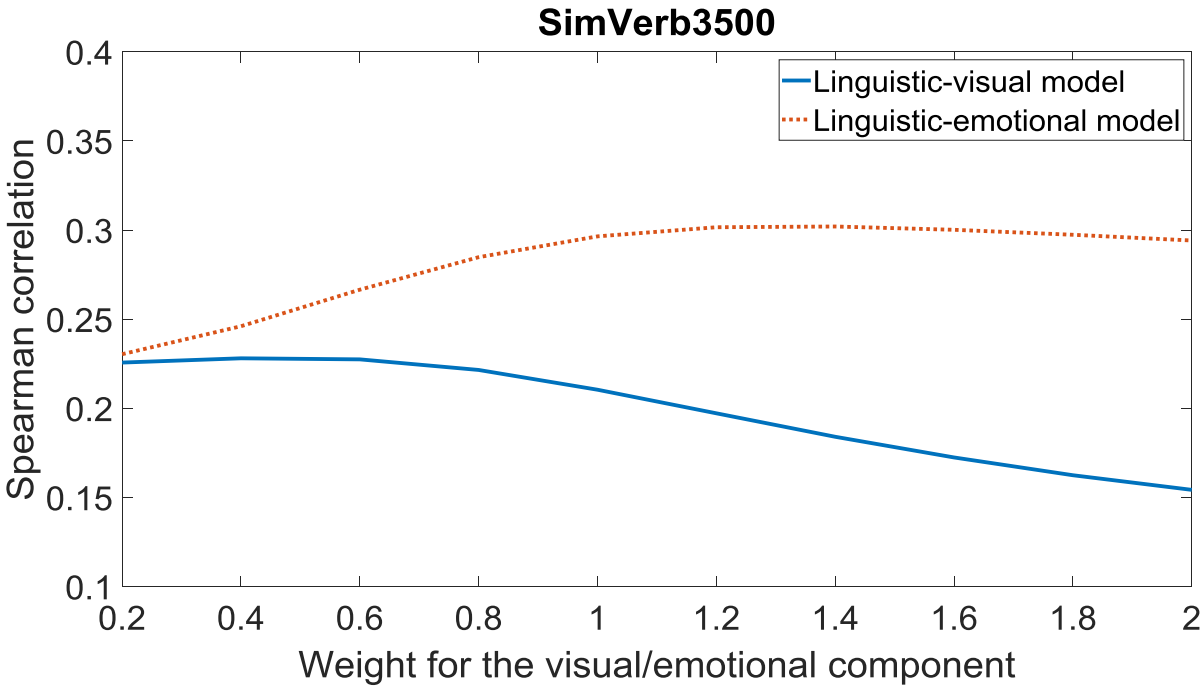


Figure 6. Model performance for the linguistic-visual and linguistic-emotional models, in predicting similarity/relatedness ratings from the SimVerb-3500 dataset. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2.

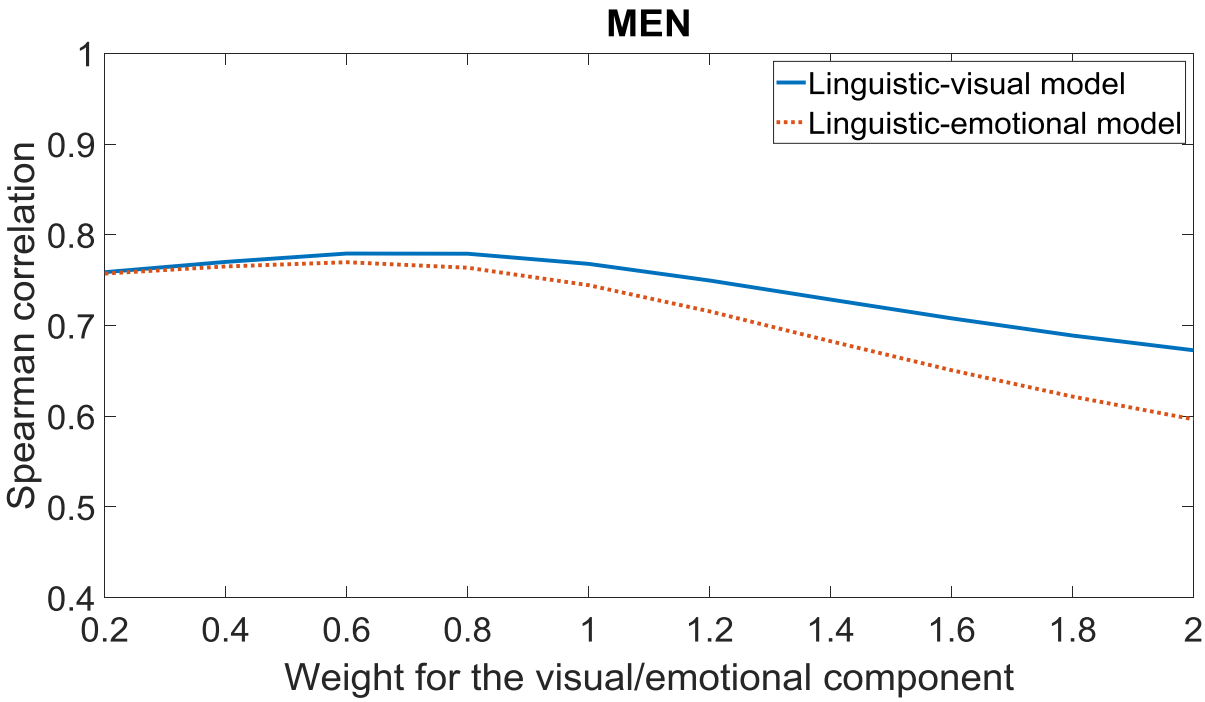


Figure 7. Model performance for the linguistic-visual and linguistic-emotional models, in predicting similarity/relatedness ratings from the MEN dataset. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2.

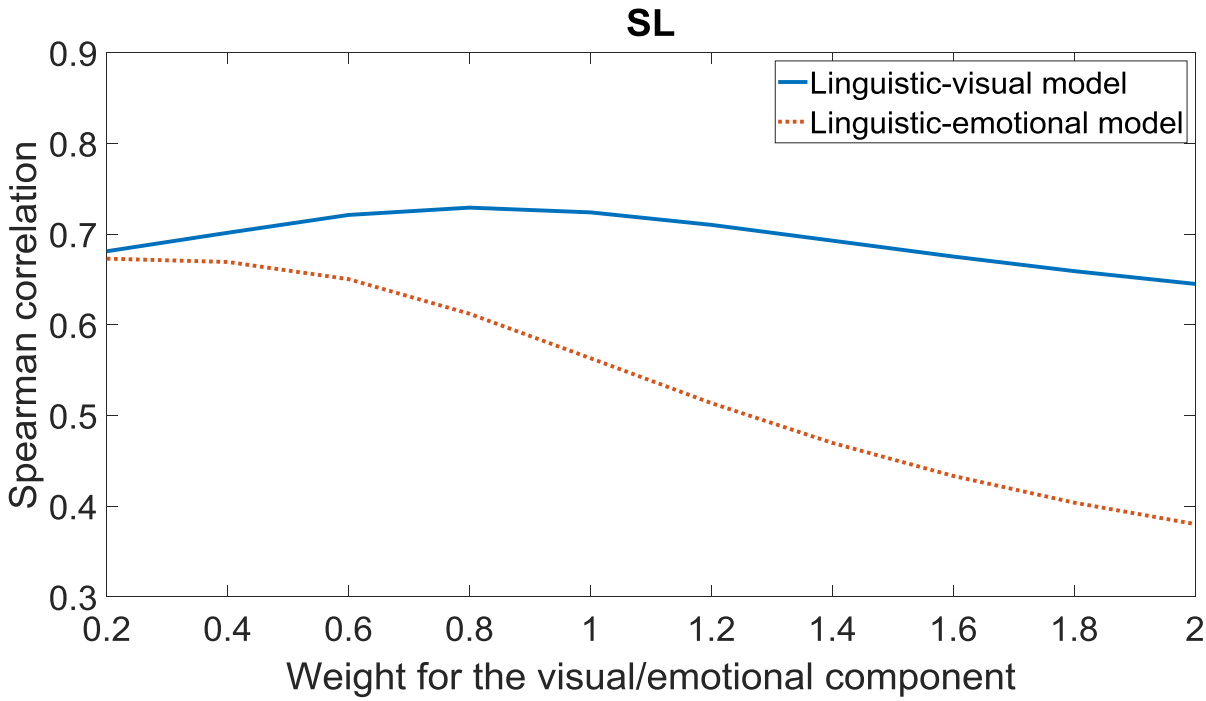


Figure 8. Model performance for the linguistic-visual and linguistic-emotional models, in predicting similarity/relatedness ratings from the SL dataset. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2.

The tests indicate that adding visual information has a significant positive effect only for the SL dataset ($p < .001$), for weights ranging from 0.6 to 1.2, and a significant negative effect for the MEN dataset ($p < .001$), for weights between 1.6 and 2. These results seem to be at odds with previous studies showing that linguistic-visual models always perform slightly better than purely linguistic ones. However, firstly, in almost all the other studies, the authors either weigh the linguistic and visual representations equally, by default (e.g., Kiela et al., 2014; Silberer et al., 2013), or they only employ the weight that gives the best results for the integration (e.g., Bruni et al., 2014; Bruni, Uijlings, et al., 2012), which leaves room for null or detrimental results of linguistic-visual integration, when employing sub-optimal weights. Secondly, we use a linguistic model that is trained over a corpus of 6 billion words, whereas other studies (e.g., Hill & Korhonen, 2014; Kiela & Bottou, 2014; Silberer & Lapata, 2012) typically employ considerably smaller corpora (i.e., containing between 80 and 800 million words). Since smaller corpora lead to a poorer performance of the linguistic model, this leaves more room for a beneficial effect of adding visual information in the other studies, as compared to our study.

Adding emotional information is significantly beneficial only for the SimVerb-3500 dataset ($p < .00125$), for weights ranging from 1.2 to 1.6, while it is significantly detrimental for the MEN dataset ($p < .001$), for weights between 1.4 and 2, and for the SL dataset ($p < .001$), for weights between 0.6 and 2. The SimVerb-3500 dataset is different from all the others in that it is the only one including only verbs (which are not highly represented in any other dataset). As verbs (words referring to events) are considered to be more abstract, this finding is in line with the view that emotional information is especially important for abstract words (Kousta et al., 2011).

**Linguistic-visual-emotional model vs linguistic-visual, linguistic-emotional, and purely linguistic models:**

In order to compare the trimodal model with the bimodal and unimodal ones, we again start by normalizing the linguistic, visual, and emotional representations, to unit length. We then construct trimodal models by assigning a weight of 1 to the linguistic components, and weights from 0.2 to 2, in steps of 0.2, to the visual and emotional components, in all pairwise combinations for the last two components. Next, for each dataset, we select the best five and worst five trimodal models, in terms of performance, and compare them to their corresponding linguistic-visual models (i.e., obtained by removing the emotional component), linguistic-emotional models (i.e., obtained by removing the visual component), and purely linguistic model (i.e., obtained by removing both the visual and emotional components). The results are shown in Figure 9, Table 7, and Table 8.

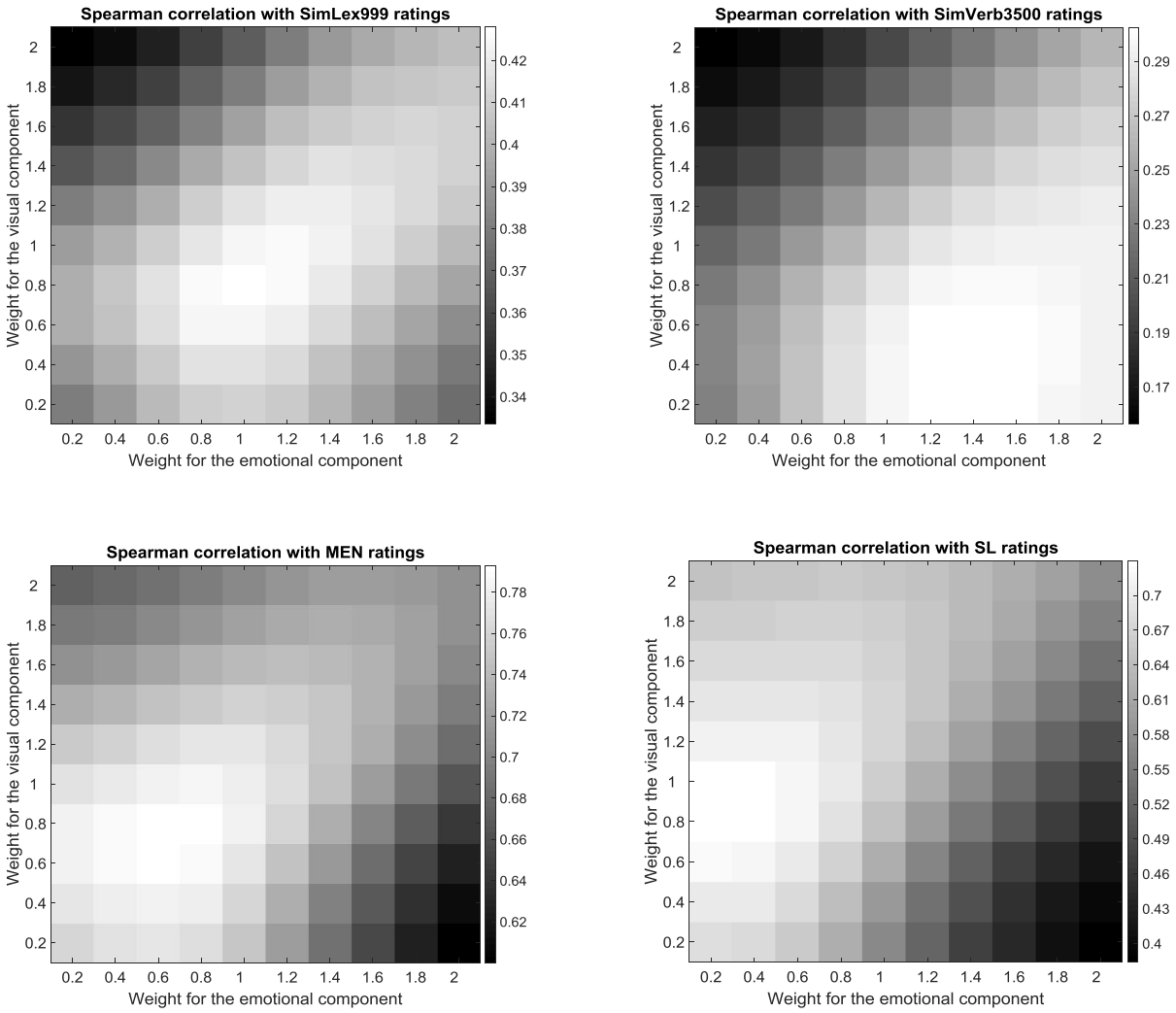


Figure 9. Model performance for the linguistic-visual-emotional models. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2.

Table 7. *p* values for comparing the best and worst linguistic-visual-emotional (LVE) models, for the SimLex-999 and SimVerb-3500 datesets, to their corresponding linguistic-visual (LV), linguistic-emotional (LE), and purely linguistic (L) models, respectively. The values in bold and italic correspond to cases where the performance of the trimodal models is significantly better or worse than that of their restricted counterparts, respectively.

| Vis. weight | Emo. weight | LVE vs LV | LVE vs LE | LVE vs L | Vis. weight | Emo. weight | LVE vs LV | LVE vs LE | LVE vs L |
|---|---|---|---|---|---|---|---|---|---|
| SimLex-999 – Best five models | | | | | SimLex-999 – Worst five models | | | | |
| 0.8 | 1.0 | .367 | .604 | .138 | 2.0 | 0.2 | .956 | .288 | .339 |
| 1.0 | 1.2 | .322 | .576 | .151 | 2.0 | 0.4 | .851 | .221 | .410 |
| 0.8 | 1.2 | .401 | .584 | .154 | 1.8 | 0.2 | .961 | .397 | .459 |
| 0.8 | 0.8 | .409 | .631 | .159 | 2.0 | 0.6 | .701 | .177 | .531 |
| 1.0 | 1.0 | .340 | .664 | .162 | 1.8 | 0.4 | .838 | .325 | .559 |
| SimVerb-3500 – Best five models | | | | | SimVerb-3500 – Worst five models | | | | |
| 0.4 | 1.4 | **.001** | .981 | **.001** | 2.0 | 0.2 | .933 | .002 | .005 |
| 0.2 | 1.4 | **.001** | .990 | **.001** | 2.0 | 0.4 | .746 | < *.001* | .010 |
| 0.4 | 1.2 | **.002** | .987 | **.001** | 1.8 | 0.2 | .926 | .007 | .013 |
| 0.6 | 1.4 | **.001** | .998 | **.001** | 1.8 | 0.4 | .705 | .002 | .028 |
| 0.2 | 1.2 | **.001** | .988 | **.001** | 2.0 | 0.6 | .475 | < *.001* | .028 |

Table 8. *p* values for comparing the best and worst linguistic-visual-emotional (LVE) models, for the MEN and SL datesets, to their corresponding linguistic-visual (LV), linguistic-emotional (LE), and purely linguistic (L) models, respectively. The values in bold and italic correspond to cases where the performance of the trimodal models is significantly better or worse than that of their restricted counterparts, respectively.

| Vis. weight | Emo. weight | LVE vs LV | LVE vs LE | LVE vs L | Vis. weight | Emo. weight | LVE vs LV | LVE vs LE | LVE vs L |
|---|---|---|---|---|---|---|---|---|---|
| MEN – Best five models | | | | | MEN – Worst five models | | | | |
| 0.8 | 0.6 | .179 | .027 | **< .001** | 0.2 | 2.0 | < *.001* | .834 | < *.001* |
| 0.6 | 0.6 | .201 | .030 | **< .001** | 0.4 | 2.0 | < *.001* | .420 | < *.001* |
| 0.8 | 0.8 | .217 | .008 | **< .001** | 0.2 | 1.8 | < *.001* | .816 | < *.001* |
| 0.6 | 0.8 | .374 | .020 | **.001** | 0.6 | 2.0 | < *.001* | .084 | < *.001* |
| 0.6 | 0.4 | .387 | .030 | **.002** | 0.4 | 1.8 | < *.001* | .353 | < *.001* |
| SL – Best five models | | | | | SL – Worst five models | | | | |
| 0.8 | 0.2 | .963 | **< .001** | **< .001** | 0.2 | 2.0 | < *.001* | .809 | < *.001* |
| 0.8 | 0.4 | .847 | **< .001** | **< .001** | 0.4 | 2.0 | < *.001* | .335 | < *.001* |
| 1.0 | 0.2 | .938 | **< .001** | **< .001** | 0.2 | 1.8 | < *.001* | .773 | < *.001* |
| 1.0 | 0.4 | .968 | **< .001** | **< .001** | 0.6 | 2.0 | < *.001* | .034 | < *.001* |
| 0.6 | 0.2 | .978 | **< .001** | **< .001** | 0.4 | 1.8 | < *.001* | .254 | < *.001* |

When comparing the performance of the trimodal models to that of their corresponding linguistic-visual models, the addition of an emotional component has a significant positive effect for the best models on the SimVerb-3500 dataset ($p < .0016$), and a significant negative effect for the worst models on the MEN and SL datasets ($p < .001$). These results are very similar to those found when comparing the linguistic-emotional models to the purely linguistic one, and might be explained by the fact that verbs, such as those that make up the SimVerb-3500 norms, are relatively abstract. In contrast, for concrete nouns, which form the majority of pairs from the MEN and SL norms, emotion should not have a positive effect (the finding of a detrimental effect is unexpected but potentially interesting as it may indicate that adding affective information may reduce the separation between different types of words).

The comparison between the trimodal models and their corresponding linguistic-emotional models reveals that including a visual component is significantly beneficial for the best models on the SL dataset ($p < .001$), but significantly detrimental for two of the worst models on the SimVerb-3500 datasets ($p < .001$). Again, SL consists only of concrete nouns, for which visual information is very salient, while SimVerb-3500 consists only of verbs, the semantics of which is likely not to be properly captured in a few tens of images per word, due to its complexity.

Finally, contrasting the trimodal models with the purely linguistic one, we find that bringing in both visual and emotional information significantly increases performance for the best models on the SimVerb-3500, MEN, and SL datasets ($p < .0016$), while it significantly decreases performance for the worst models on the MEN and SL datasets ($p < .001$). These results are a combination of the partial results regarding the effects of appending visual and emotional components to the purely linguistic and bimodal models, which indicates little overlap between vision and emotional representation.

**Comparing the models for concrete and abstract words:**

In order to test whether visual content is more important for more concrete words, while emotional content for more abstract words, we first combined the SimLex-999 and SimVerb-3500 datasets, as they cover a broader range of concreteness

ratings than MEN and SL. Then, we divided the merged dataset into a low and a high concreteness subset. More specifically, we selected the bottom 25% and the top 25% of pairs, based on the mean concreteness of each word pair covered by the concreteness norms of Brysbaert, Warriner, and Kuperman (2014). We then tested the performance of the emotional and visual models, the two bimodal models, and the trimodal models, setting all the weights to 1. The results are displayed in Figure 10.

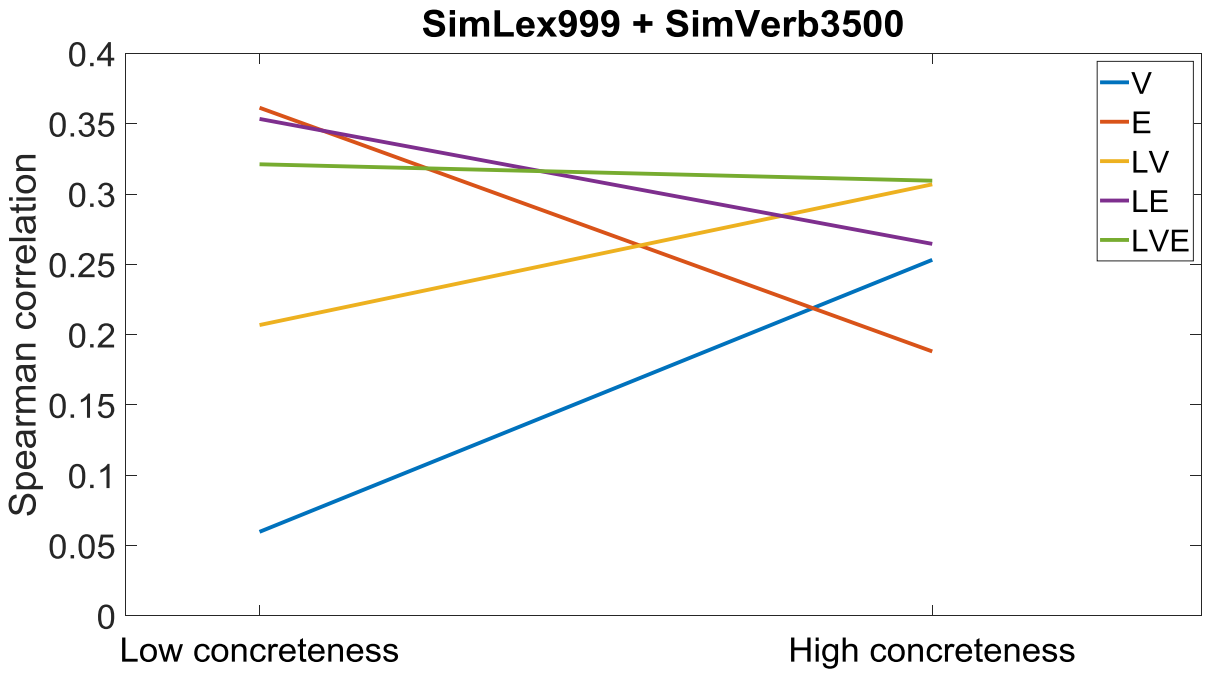


Figure 10. Model performance for low and high concreteness word pairs. V = visual model; E = emotional model; LV = linguistic-visual model; LE = linguistic-emotional model; LVE = linguistic-visual-emotional model.

Using one-tailed *Z*-tests, after normalizing the correlations, we found that the performance of the visual model is higher for more concrete pairs, in comparison to the less concrete ones, for the visual ($p < .001$) and linguistic-visual ($p < .01$) models. Also, the emotional model has a better performance for the more abstract pairs, as opposed to the less abstract ones. Non-significant results were obtained for the linguistic-emotional and trimodal models. These results seem to suggest that the positive effect of adding visual information should be greatest for datasets consisting mainly of more concrete words, such as MEN and SL, while the beneficial effect of

including emotional information should be largest for datasets made up mainly of more abstract words, such as SimLex-999 and SimVerb-3500.

## 3.4. Conclusions

A first goal of this chapter was to present an evaluation of visual and emotional models, in order to identify the model(s) better fitting behavioural semantic data. In the case of the emotional models, models based on recurrent neural networks (i.e.,GRU, LSTM, DeepMoji) perform comparably, while being better than a convolutional neural network model (perhaps unsurprisingly, knowing that CNNs are not designed for operating with linguistic representations). For the visual models, we found that convolutional neural networks models (i.e., K&B, AlexNet, GoogLeNet, VGG-19) have a better performance than a classical, bag-of-visual-words model (i.e., SIFT), when tested over a large dataset of similarity/relatedness ratings. Among the convolutional models, GoogLeNet gave the best results, followed by AlexNet, VGG-19, and K&B.

The second and main goal was to develop models that integrate linguistic, visual and emotional information and to assess their performance against purely linguistic models and models that only include either visual or emotional features. We chose the DeepMoji model for a number of reasons, namely: its state-of-the-art performance in a number of emotional tasks; its distributional nature, since it predicts the occurrence of an emoji based on its immediate linguistic context; its capacity to use rich emotional information, as it is trained on tweets containing 64 types of emojis; its high dimensionality, which allows it to encode complex patterns of emotion-based word similarity. Moreover, with respect to the DeepMoji representations, we found that the first 10 principal components are significantly correlated with subjective valence, arousal, and dominance ratings, which provides additional support to the hypothesis that the DeepMoji model captures affective information.

In order to better understand the relative importance of each visual and emotional component, we carried out comparisons in which we parametrically varied the weight of visual and/or emotional information. In this manner, we could see when adding this information leads to better or worse performance. We found that adding

visual information had a positive effect in 4/40 cases, no significant effect in 33/40 cases, and a negative effect in 3/40 cases. When including emotional information, there was a positive effect in 3/40 cases, no significant effect in 26/40 cases, and a negative effect in 11/40 cases. Finally, when introducing both visual and emotional information, for the best models, the analyses revealed a positive effect in 15/20 cases, no effect in 5/20, and a negative effect in 0/20 cases; in contrast, for the worst models, the results indicated a positive effect in 0/20 cases, no significant effect in 10/20 cases, and a negative effect in 10/20 cases. In general, we found that whether the addition of non-linguistic increases or decreases model performance, or instead has no effect, is determined by the weights attributed to the different types of information, which may have practical value for future modeling.

In addition, it appears that this impact depends on whether the dataset includes predominantly concrete or abstract words. As expected on the basis of previous literature (e.g., Kousta et al., 2011), we found that including visual information is particularly beneficial to more concrete concepts, whereas including emotional information is particularly beneficial to more abstract concepts. This is clearly visible when we assess model performance separately for more concrete and abstract words (see Figure 10). It is also clear from the comparison between MEN (only concrete words) and SimVerb-3500 (only verbs, hence more abstract): Across comparisons, we see that indeed visual information brings more benefit to the former, whereas emotional information brings more benefit to the latter.

As mentioned in the introduction, a previous study (De Deyne et al., 2018) also examined the change in performance for distributional models of semantics, when adding experiential (i.e., visual and emotional) information. They found that including experiential information led to little or no improvement for internal language models, but had a moderate positive effect for external language models. Moreover, they also found that adding visual information had the greatest effect for concrete words while introducing affective information had the largest impact for abstract words. This finding mirrors our own, when comparing the linguistic-visual and linguistic-emotional models to the purely linguistic model.

However, there are a number of key differences between their approach and ours. First, we avoided the potentially controversial distinction between external and internal language models. In the study by De Deyne and collaborators (2018), external

language models derive semantic representations from corpora of language, whereas internal language models derive semantic representations from free associations. Thus, the models differ in whether they use objective or subjective data (based on a metacognitive task), but both might be argued to tap into the same construct. We focus on an objective corpus-based approach, to avoid such potential criticisms. Second, in a similar vein, we decided to use an emotional model that learns affective information indirectly, by predicting the co-occurrence of emojis and text in a corpus, rather than using emotional representations derived directly from valence, arousal, and dominance norms (Mohammad, 2018; Warriner et al., 2013). This also increases the coverage of our model. Finally, since the resulting representations in our model are high-dimensional, they might provide more fine-grained information than representations with only a few dimensions.

# 4. Modelling the structure and dynamics of semantic processing[15]

## 4.1. Introduction

In the past 40 years, the connectionist approach has been one of most influential paradigms in the computational modelling of cognition, in general (e.g., Houghton, 2005; McClelland, Rumelhart, & the PDP Research Group; 1986), and of semantics, in particular (e.g., Rogers & McClelland, 2004). Connectionist models consist of artificial neural networks, built from layers of simple units (i.e., "neurons"), and weighted interconnections between the layers (i.e., "synapses"). Learning in such models involves activating the representation of a stimulus in the input layer (i.e., simulating the "neuronal firing pattern" associated with the stimulus), then allowing activation to propagate through a number of intermediate layers, until, finally, the output layer becomes activated (i.e., simulating the behavioural response to the stimulus). The activation pattern in the output layer is compared with the expected activation pattern, and the difference between the two patterns, which represents the prediction error, is used in order to modify the strength of the individual connections between the units, thus simulating a biological learning process.

Despite their apparent simplicity, connectionist models have been successfully used to study cognitive processes and representations in a variety of contexts (for a review, see Thomas & McClelland, 2008), perhaps most prominently in the fields of developmental psychology (Elman et al., 1996) and clinical psychology (Aakerlund & Hemmingsen, 1998). With respect to semantics, such models have been able to account for certain aspects of task behaviour in semantic tasks, such as feature verification and semantic priming (McRae, De Sa, & Seidenberg, 1997), as well as some of the behavioural effects of semantic impairments, such as deep dyslexia (Hinton & Shallice, 1991), Alzheimer's disease (Devlin, Gonnerman, Andersen, &

[15] Adapted from (Rotaru et al., 2016, 2018).

Seidenberg, 1998), and semantic dementia (Rogers et al., 2004). Connectionist models have also been employed in capturing lexical development during childhood (Horst, McMurray, & Samuelson, 2006; Li, Zhao, & MacWhinney, 2007), for both typical and atypical populations (Thomas & Karmiloff-Smith, 2003).

However, especially in the last decade, distributional models of semantics have become considerably more popular than connectionist models. As discussed in Chapter 2, their popularity is due to at least two factors, namely the ability to automatically extract statistical patterns from huge text corpora (e.g., in the order of hundreds of billions of words), for nearly all the words in a language, and their remarkable ability to predict behavioural and neuroimaging data (e.g., in tasks such as synonymy judgement, state-of-the-art models produce perfect performance; Bullinaria & Levy, 2012).

Interestingly, connectionist and distributional models have complementary strengths (and weaknesses). Connectionist models usually focus on the semantic processes involved in learning representations and performing various tasks. However, most of these models do not include realistic representations but, rather, simplified ones, either for computational feasibility, or because the models are meant as a proof of concept. For instance, in the study of study by Devlin and collaborators (1998), only 60 words were employed, and each word was represented as a binary feature vector, thus discarding any quantitative information regarding relative feature importance. In contrast, distributional models usually operate with real-valued vectors, containing hundreds of dimensions, and having a vocabulary of tens or hundreds of thousands of words. However, the processes that operate over the distributional representations when performing a semantic task are typically very simple, and almost always involve little more than computing a measure of vector (dis)similarity between words, such as cosine similarity or Euclidean distance (Bullinaria & Levy, 2007).

In this context, our aim is to test whether we can obtain the best of both worlds, by bringing together the two classes of models and retaining their advantages, while eliminating most of their shortcomings. In order to achieve this goal, we combine distributional models of semantic structure and processing models of lexical activation. Firstly, we model both the structural properties of semantic networks, as well as their dynamic aspects, by considering the flow of semantic activation (Anderson, 1983; Collins & Loftus, 1975) generated by the automatic processing of individual words. An

important consequence of looking at both structure and dynamics is that it allows us to assess the effects of direct, as well as indirect, mediated semantic relations between words (e.g., “coin”-“round”-“moon”, which links “coin” and “moon”), rather than limiting our analysis to strong, direct semantic links (e.g., “coin”-“round”, “moon”-“round”). Previous research using models of semantics based on free association (De Deyne, Navarro, & Storms, 2013; Steyvers, Shiffrin, & Nelson, 2005) shows that indirect associations provide a complementary source of semantic information, in tasks including lexical decision, semantic similarity rating, and extralist cued recall. However, to the best of our knowledge, there are very few studies that investigate the explanatory power of indirect semantic relations in text-based models of semantics, as well as their temporal dynamics (for an exception, see De Deyne, Verheyen, & Storms, 2016). Starting from standard distributional models of semantics, we allow activation to spread throughout the semantic network, as dictated by the patterns of semantic similarity between words, and record the activation of each word, as a function of time. We then study how the activation pattern at each time point relates to task performance in a number of tasks, as a means of linking dynamics to observable task behaviour.

Secondly, we assume that both strong and weak semantic relations between words, as indexed by standard measures of semantic similarity (e.g., vector cosine), contribute to performance in semantic tasks (Chen & Mirman, 2012; Mirman & Magnuson, 2008), rather than focusing only on the strong relations, as is traditionally done when performing network analyses (Buchanan, Westbury, & Burgess, 2001; Griffiths, Steyvers, et al., 2007; Gruenenfelder et al. 2015; Utsumi, 2015). The significant influence of distant neighbours is likely to be a direct result of the fact that words have considerably more distant neighbours than close ones, given that semantic similarity based on the cosine measure follows a power law distribution (Griffiths, Steyvers, et al., 2007). Therefore, we keep both classes of neighbours in our models, and we do not make any a priori assumptions about any privileged role that close neighbours might have over distant ones (or vice-versa), in the course of semantic processing.

Within our dynamic models, semantic activation flows from an initial concept to its neighbours, then to the neighbours of its neighbours, and so on, until the system reaches a global “attractor” state. However, unlike many other connectionist models

(Chen & Mirman, 2012; Hoffman & Woollams, 2015; Rogers & McClelland, 2004), they have a large number of nodes and feedforward/feedback/recurrent connections, making them more realistic models of human lexico-semantic knowledge. As a result, it is expected they should provide better insight into the distinct contribution of structural and task-related aspects of semantic behaviour. Our models can also be seen as probabilistic, such that at each step, they make use of their underlying discrete-time Markov chain, in order to perform multi-step inferences. Thus, our approach lies at the intersection of connectionist (McClelland et al., 2010) and probabilistic (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010) modelling. In contrast to other probabilistic models, such as Topic (Griffiths, Steyvers, et al., 2007), our models are non-hierarchical and do not undergo any form of dimensionality reduction, which means that the inferences are easier to interpret and that less semantic information is lost.

## 4.2. Model development

**Distributional semantic models:**

Previous studies have shown that "word-as-context" models (e.g., HAL, Skip-gram, CBOW, GloVe), provide a better fit to behavioural data, as compared to "document-as-context" models, as well as that, within the class of "word-as-context" models, the CBOW and GloVe models have a clear advantage over their competitors (see Chapter 2). Given that these models have shown their superiority in a number of tasks, we adopt them as our models of choice. We include both CBOW and GloVe to test whether our findings generalize beyond a specific architecture. Moreover, to further assess if our results truly support a role for the dynamics of semantic activation beyond the structural assumptions, we also include the LSA model in our analyses. For our computational experiments, we use the *gensim* tool (Řehůřek & Sojka, 2010),

for the CBOW and LSA models, and the GloVe implementation provided by the authors of the model[16].

We derive our semantic representations by training the models on the written part of the British National Corpus (BNC; Leech, Garside, & Bryant, 1994), containing approximately 87 million words. The BNC consists of contemporary texts from a variety of sources (e.g., newspapers, journals, books, letters, essays), providing a comprehensive corpus of modern British English. In order to improve the quality of the resulting representations, we first pre-process the corpus by converting all the words to lowercase, eliminating punctuation marks and removing words whose absolute frequencies are less than five. We then construct 300-dimensional vector representations for the words in our corpus. For reasons of computational efficiency, we do not employ all the words covered by our models, but instead keep only the 28,592 words that are also part of the 30,000 most frequent nouns, verbs, and adjectives, according the SUBTLEX-UK frequency norms for British English (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014).

**Structure and dynamics:**

Since we are interested in obtaining semantic networks that reflect the semantic associations between words, we compute a representational similarity matrix *SM* (i.e., the structural model) from the vectors produced by each of the three distributional models, using vector cosine as a measure of similarity between the word representations. For each model, the matrix *SM* contains the structure of our semantic network, such that any value *SM*(*i,j*) can be interpreted as the strength of the (symmetric) semantic association between words $w_i$ and $w_j$. Within *SM*, large values (i.e., close to 1) indicate pairs of words that are close semantic neighbours, whereas small values (i.e., close to 0) correspond to pairs of words that are only weakly related. Given that negative cosine values are likely to provide very little or no useful semantic

[16] For the CBOW and LSA models, we used the default hyperparameter values from the *gensim* software library, as described at https://radimrehurek.com/gensim/models/word2vec.htm and https://radimrehurek.com/gensim/models/lsimodel.html. For the GloVe model, we employed the hyperparameter values from the paper by Pennington et al. (2014). The software package we used for training the GloVe model is available for download at https://github.com/stanfordnlp/GloVe.

information, word pairs with negative cosine similarity receive a zero value in *SM*, as a means of reducing the amount of noise present.

The matrices *SM* represent our structural models. In order to obtain our dynamic models, we assume that semantic activation spreads throughout the networks, such that the activation propagated from the source word $w_i$ to the target word $w_j$ is proportional to both the current activation level of $w_i$, and the value of $SM(i,j)$, following the principle that the more similar two words are, the more activation flows between them. We also impose that the total amount of activation present in the networks should remain constant. Thus, we set to zero all the diagonal elements (we deal with these recurrent connections separately; see below) and normalize the rows of the resulting matrices $SM_{NORM}$, such that each row sums to one[17], meaning that the total activation provided by $w_i$ to its semantic neighbours is exactly equal to its current level of activation. However, since it is very plausible that the source word $w_i$ also retains some of its activation, we employ the weighted average of $SM_{NORM}$, which indexes feedforward/feedback connections, and the identity matrix, with indexes recurrent connections, rather than $SM_{NORM}$ itself. The weight (i.e., 2/3 for $SM_{NORM}$ and 1/3 for the identity matrix) is chosen heuristically (see the study by De Deyne et al., 2016, for a similar approach).

We model the spreading of activation within the semantic network as occurring in discrete time steps, rather than being a continuous process, which allows us to express our models as a discrete-time Markov chain, denoted as *MC*. In this way we can further assess whether the initial steps better capture tasks that only implicitly tap into semantic knowledge (such as the lexical decision task) whereas tasks that explicitly require semantic activation (such as semantic decisions, but also ratings of concreteness and imageability) correspond to later steps of the chain. The probability matrix underlying *MC* is represented by *DM* (i.e., the dynamic model), such that $DM = (2 * SM_{NORM} + I_N) / 3$. In our analyses we will focus only on the first five time steps in the evolution of the chain, given that the subsequent time steps provide little new

[17] Each row can be seen as a probability distribution over the semantic neighbours of the word corresponding to that row, very similar to the distribution of association strengths for a given cue, in the free association task.

information[18]. An illustration of the structural and dynamic models is given in Figure 11 and Figure 12.

[18] It might seem strange and arbitrary that we do not look beyond the first five time steps. However, as verified by our calculations, the particular Markov chains employed in this chapter (i.e., derived from the CBOW, GloVe, and LSA models, trained over the BNC) have the property of being ergodic (e.g., Serfozo, 2009), which means that, as time goes to infinity, they converge to a fixed distribution, known as a steady-state/stationary distribution. As a result, after the first few steps, the patterns of activation become almost indistinguishable from one another, as they closely approximate the steady-state distribution. Therefore, since the information associated with these later time points is largely redundant, we discard them from the analysis. As a final note, the ergodicity of our models and their rapid convergence are not accidental, but instead result from the rich interconnectivity of model-based semantic networks (i.e., their "small-world" and "scale-free" structure, see Chapter 2).

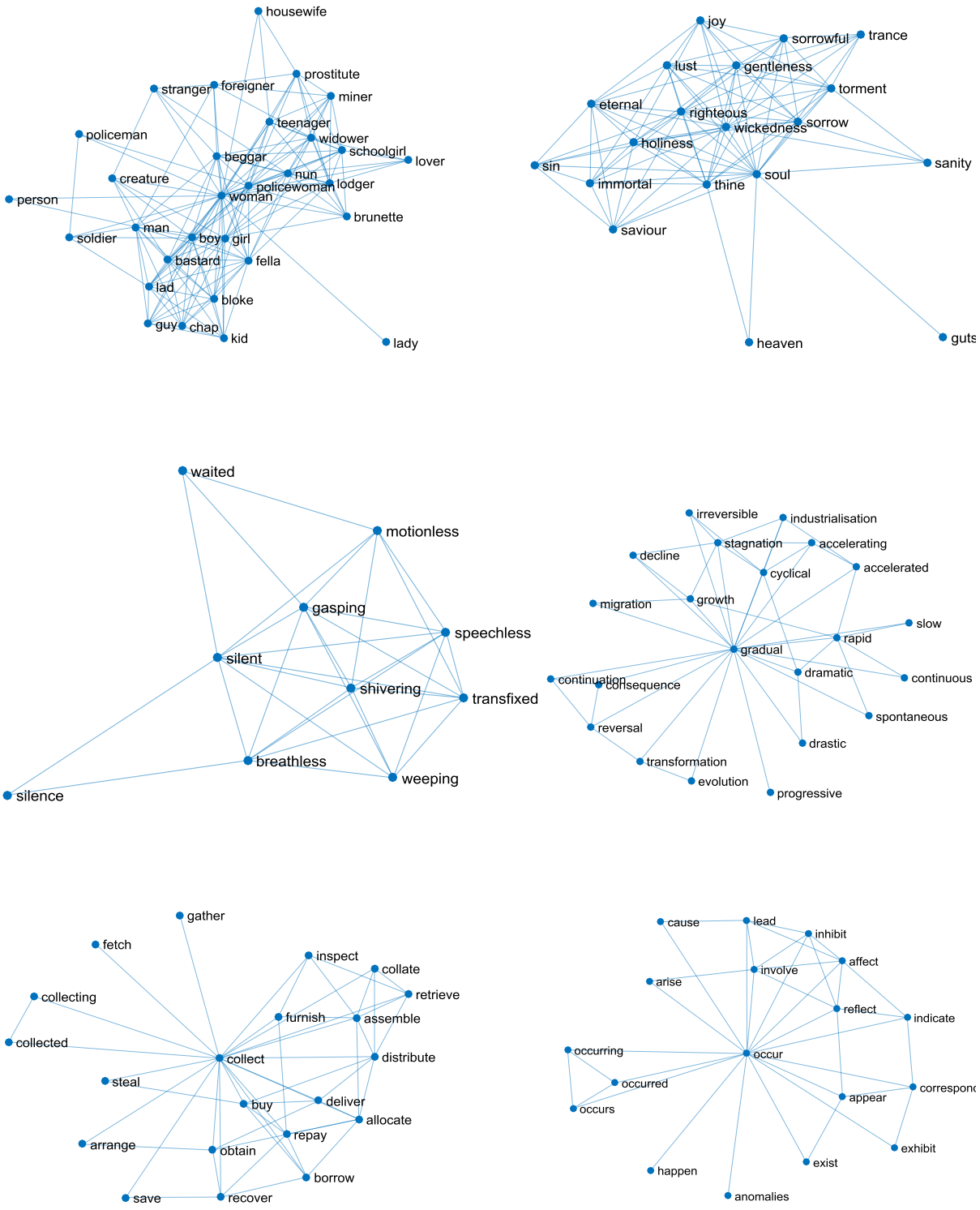


Figure 11. Local semantic neighbourhoods for three concrete words (i.e., “woman”, “silent”, “collect”; on the left) and three abstract words (i.e., “soul”, “gradual”, “occur”; on the right), covered by the CBOW model. We include only very strong neighbours for each word (i.e., pairs of words with cosine similarity greater than 0.425).


Figure 12. Toy example of the spreading of activation in our dynamic model. The network consists of four words and eight directional semantic associations between the words. The levels of activation are represented by the intensity of the colours for each word. Initially, only "Cioran" is activated; during step 1, "Borges" receives activation from "Cioran"; during step 2, "Paz" receives activation from "Borges"; during step 3, "Paz" and "Borges" exchange part of their activation, while "Cioran" and "Calvino" receive activation from "Paz"; during the remaining steps, the network reaches a state of equilibrium, such that the level of activation corresponding to each word remains almost constant.

Let $S_k(MC)$ denote the state of *MC* at time step *k*. In most of our empirical validations, we are interested in the number of neighbours of word $w_i$, at various distances from $w_i$. More specifically, we partition $S_k(MC)$ into 10 deciles, such that each word $w_j$ falls into one decile depending on its activation/probability, given by $S_k(MC)$. Each decile corresponds to a type of neighbourhood, ranging from very distant neighbourhoods (i.e., the first few deciles, characterized by weak associations between the word $w_i$ and its neighbours), to very close neighbourhoods (i.e., the last few deciles, characterized by strong associations between the word $w_i$ and its neighbours). The number of neighbours at each step *k* and in each decile *d,* denoted as $numNeigh_{k,d}$, forms the predictor for reaction times, response accuracies, and concreteness and imageability ratings. For modelling similarity/relatedness judgments (i.e., "how similar/related are $w_i$ and $w_j$?"), we take the level of activation corresponding to $w_j$, from the Markov chain starting with $w_i$ (i.e., the result of evaluating the pair by focusing first on $w_i$, and then on $w_j$), as well as the level of activation corresponding to $w_i$, from the Markov chain starting with $w_j$ (i.e., the result of evaluating the pair by focusing first on $w_j$, and then on $w_i$). We look at both forward and backward probabilities/activations because, whereas activation spreads in our network in an asymmetrical manner, we believe that similarity/relatedness judgements are largely symmetrical, although the issue of symmetry in (episodic and semantic) memory associations is still under debate (Kahana, 2002; Tversky, 1977).

Deriving the structural and dynamic models involves the following steps:

- use the CBOW/GloVe/LSA model in order to obtain 300-dimensional vector representations for all the words in a given set of size *N*, representations which we denote by *Vecs*. The matrix *Vecs* is of size *N* × 300, such that each row corresponds to the vector associated with a given word.
- compute a similarity matrix *M*, of size *N* × *N*, from said vectors, using vector cosine as a measure of similarity between vectors, such that $M = (Vecs \,/\, ||Vecs||) * (Vecs \,/\, ||Vecs||)^T$, where *T* denotes the matrix transpose, ||·|| denotes the Euclidian norm (computed for each row), and / denotes element-wise division.

- set to zero all the negative values in the cosine matrix, meaning that $SM(i,j) = M(i,j)$, if $M(i,j) > 0$, and $SM(i,j) = 0$, otherwise.
- employ the matrix $SM$ as a structural model.
- set to zero the diagonal elements of the matrix $SM$, then normalize its rows, such that each row sums to one. This means that $SM_{NORM}(i,j) = 0$, if $i = j$, and $SM_{NORM}(i,j) = SM(i,j) / \sum (\{SM(i,col) \mid 1 \leq col \leq N \text{ and } col \neq i\})$, otherwise.
- employ the matrix $DM = (2 * SM_{NORM} + I_N) / 3$ as the probability matrix for the Markov chain $MC$ representing our dynamic model, where $I_N$ is the identity matrix of size $N$.

Let $S_k(MC)$ denote the state of $MC$ at step $k$. This state can be computed by raising DM to the power of $k$, meaning that $S_k(MC) = DM^k$. Thus, for any row $i$ and column $j$, the value $S_k(MC)(i,j)$ represents the probability that $MC$ is in state $j$, at time step $k$, given that it started in state $i$. This probability gives us the amount of activation associated with word $w_j$, at time $k$, following the initial presentation of word $w_i$.

When modelling non-relational tasks (e.g., lexical or semantic decision, imageability or concreteness rating), for any word $w_i$ and time step $k$ between 1 and 5, we compute $numNeigh_{k,d}(i)$ as the number of elements on row $i$ of $S_k(MC)$ that have activations (i.e., probabilities) falling into the $d^{th}$ decile of all the activations in $S_k(MC)$. In other words, for $d$ = 1 and d = $10$, we count the weakest and the strongest neighbours of $w_i$, respectively, while for any $d$ between 2 and 9 we calculate how many of the neighbours have intermediate levels of activation. More formally, $numNeigh_{k,d}(i)$ is equal to the number of elements in the set $\{S_k(MC)(i,col) \mid \text{quantile}(S_k(MC), 10*(d-1)) < S_k(MC)(i,col) \leq \text{quantile}(S_k(MC), 10*d)$, for $1 \leq col \leq N\}$, for $1 \leq k \leq 5$ and $1 \leq d \leq 10$. For consistency, we also perform an analogous count for the cosine similarity values in the matrix $SM$, resulting in a total of (5 + 1) * 10 = 60 predictors for each of the CBOW, GloVe, and LSA models.

When modelling relational tasks (e.g., similarity/relatedness rating), for any two words $w_i$ and $w_j$, and time step $k$ between 1 and 5, we use the values $S_k(DM)(i,j)$ and $S_k(DM)(j,i)$ to represent the strength of the association between $w_i$ and $w_j$, and that between $w_j$ and $w_i$, respectively. We obtain a total of 5 * 2 = 10 predictors for each of the CBOW, GloVe, and LSA models.

## 4.3. Model testing

**Behavioural measures:**

We tested our models on a number of behavioural measures taken from existing sources. These are: (1) lexical decision response time and accuracy, for a subset of 2,328 words taken from (Keuleers et al., 2012); (2) semantic decision response time and accuracy, for a subset of 2,639 words from (Pexman et al., 2017) in which participants were asked to classify a word as either concrete or abstract; (3) concreteness ratings and (4) imageability ratings for the same words as (1) taken from (Keuleers et al., 2012); (5) semantic similarity/relatedness ratings taken from SL, MEN, SimVerb-3500, and SimLex-999 datasets (see Chapter 2). For all these tasks, we selected all the words covered by our models and norms.

**Baseline models:**

In order to assess the role of structural relationships among words and dynamic flow of activation, we first compared our models to a baseline model that included as many as possible of the other variables which are known to affect lexical and semantic decisions, as well as concreteness and imageability ratings. In order to evaluate our models conservatively, we crucially included a number of semantic and non-semantic variables to assess whether our structural measures provide a fit above and beyond the other semantic predictors. The choice of the specific variables to include in the baseline model for each task is dictated by the availability of relevant norms as well as considerations regarding the specific task used. Then, we compared a combination of the baseline model, the ten neighbourhood sizes from the structural models, and the ten neighbourhood sizes from the individual steps of the dynamic models, with a combination of the baseline model and the structural models.

For the analysis of the lexical decision respone time and accuracy, we used a baseline model including age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), familiarity (Gilhooly & Logie, 1980; Stadthagen-Gonzalez & Davis, 2006), log frequency, log contextual diversity (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014), semantic diversity (Hoffman et al., 2013), squared hedonic valence

(Warriner et al., 2013), number of letters, Coltheart's N (i.e., the number of words that can be produced by substituting one letter of a given word, for any other, such that the result is a valid word; Coltheart, Davelaar, Jonasson, & Besner, 1977), orthographic Levenshtein distance (OLD20; the average orthographic editing distance between a word and its twenty closest neighbours in the lexicon; Yarkoni, Balota, & Yap, 2008), and phonological Levenshtein distance (PLD20; the average phonological distance between a word and its twenty closest neighbours in the lexicon; Suárez, Tan, Yap, & Goh, 2011). For the analysis of semantic decision response time and accuracy, the baseline model included log frequency, semantic diversity, number of letters and orthographic Levenshtein distance, in order to attempt to replicate the findings from (Pexman et al., 2017).

For the analysis of concreteness and imageability rating tasks, the baseline model included age of acquisition, familiarity, (log) frequency, log contextual diversity, semantic diversity, squared hedonic valence, number of letters, Coltheart's N, OLD20, and PLD20. Finally, for the analysis of semantic similarity/relatedness ratings, we omitted a baseline model, given that performance in these tasks has been shown to be very well captured by the information provided by distributional models (see Chapter 2).

**Results:**

In order to test whether a purely structural model can fit the data better than a baseline model, and then, crucially, whether further including spreading of activation (across five consecutive steps) provides any further improvement of the fit, we used multiple linear regression models, taking in turn each behavioural measure (i.e., response times, accuracies, and ratings) as the dependent variable. The independent variables were (a subset of) those in the baseline measures (defined in this subchapter), structural measures and dynamical measures (defined in the previous subchapter). In order to deal with the problem of multiple comparisons, we employed the Bonferroni correction when reporting the statistical significance of each result.

*Lexical decision:*

The results for the lexical decision task are shown in Figure 13 and Table 9. For log response time, the fit was improved by the addition of the structural models (CBOW, GloVe, and LSA), as well as by the inclusion of the first step (CBOW, GloVe), in the case of the dynamic models. For accuracy, a significantly better fit was obtained when adding the structural models (CBOW, GloVe), as well as the first step (CBOW) and second step (GloVe), of the dynamic models. These results suggest that the dynamics of the semantic network, as captured by our models, provide a complementary source of information regarding semantic processing in the lexical decision task.

An additional interesting question is whether the models behave similarly for concrete and abstract words. In order to assess this, we divided our words into two classes, based on concreteness ratings, and ran separate analyses for each subset of words. Overall, it appears that the behavior of the models is largely comparable across the two word classes (see detailed results in Appendix B).

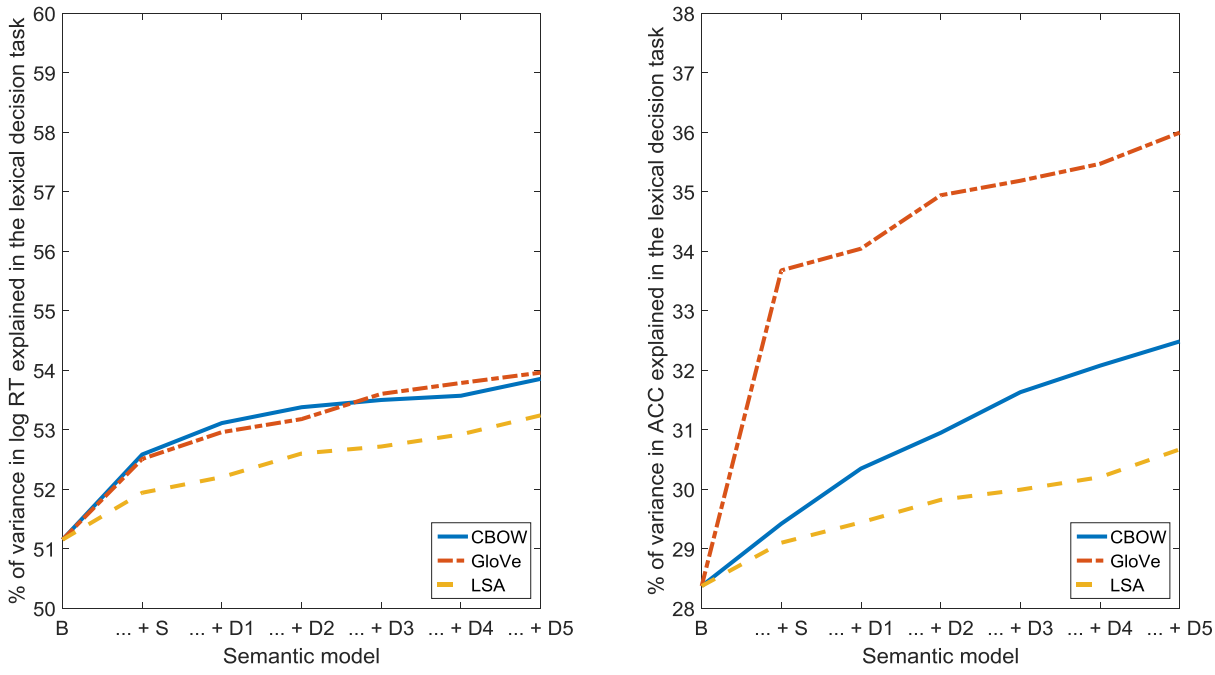


Figure 13. Percentage of variance in log response time (RT) and accuracy (ACC) in the lexical decision task, accounted for by the baseline model (B), the combination of the baseline model and the structural model (… + S), and the combination of the baseline model, the structural model, and consecutive steps of the dynamic model (… + D1 through … + D5).

Table 9. Results of model comparisons for predicting log response time and accuracy in the lexical decision task. B = baseline model; S = structural model; $D_{1...k}$ = first *k* individual steps of the dynamic model.

| | | Enhanced vs simple model | | | | | |
|---|---|---|---|---|---|---|---|
| Model | Statistic | B vs B + S | B + S vs B + S + $D_1$ | B + S + $D_1$ vs B + S + $D_{1...2}$ | B + S + $D_{1...2}$ vs B + S + $D_{1...3}$ | B + S + $D_{1...3}$ vs B + S + $D_{1...4}$ | B + S + $D_{1...4}$ vs B + S + $D_{1...5}$ |
| | | Log response time (lexical decision) | | | | | |
| CBOW | F | 6.96 | 2.60 | 1.47 | 1.00 | 0.51 | 1.21 |
| | (p) | (< .0001) | (.004) | (.15) | (.44) | (.89) | (.28) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| GloVe | F | 6.59 | 2.59 | 1.06 | 2.12 | 1.08 | 0.81 |
| | (p) | ( < .0001) | (.004) | (.39) | (.02) | (.38) | (.62) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| LSA | F | 3.79 | 1.17 | 2.24 | 0.33 | 0.97 | 1.63 |
| | (p) | (< .0001) | (.31) | (.01) | (.97) | (.46) | (.09) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| | | Accuracy (lexical decision) | | | | | |
| CBOW | F | 3.42 | 3.08 | 1.98 | 2.35 | 1.53 | 1.20 |
| | (p) | (.0002) | (.0007) | (.03) | (.01) | (.12) | (.29) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| GloVe | F | 18.45 | 1.38 | 3.21 | 0.88 | 0.91 | 1.76 |
| | (p) | (< .0001) | (.18) | (.0004) | (.55) | (.52) | (.06) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| LSA | F | 2.37 | 1.07 | 1.13 | 0.52 | 0.63 | 1.77 |
| | (p) | (.01) | (.38) | (.33) | (.87) | (.79) | (.06) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |

*Semantic decision:*

The results for the semantic decision task are shown in Figure 14 and Table 10. For log response time, the addition of the structural models significantly improved the fit in two out of three cases (CBOW, LSA). In the case of the dynamic models, the fit was ameliorated by the inclusion of step one (CBOW, GloVe, LSA) and step five (CBOW). For accuracy, however, only the addition of one of the structural models (LSA), and of step three (CBOW), improved the fit. It is important to note that our findings for log response time are in contradiction with the results of several previous studies (Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008; Yap et al., 2011, 2012; Zdrazilova & Pexman, 2012), where no effects of neighbourhood size and connectivity were detected (see Chapter 5). This discrepancy may come about because we perform a relatively fine-grained analysis of neigbourhood size, as a function of semantic distance, resulting in ten neighbourhoods per word, while all the other studies only focus on (very) close neighbourhoods, yielding one neighbourhood

per word. Also, we include both the structure and the dynamics of our semantic network, whereas the other approaches investigate only structural aspects.

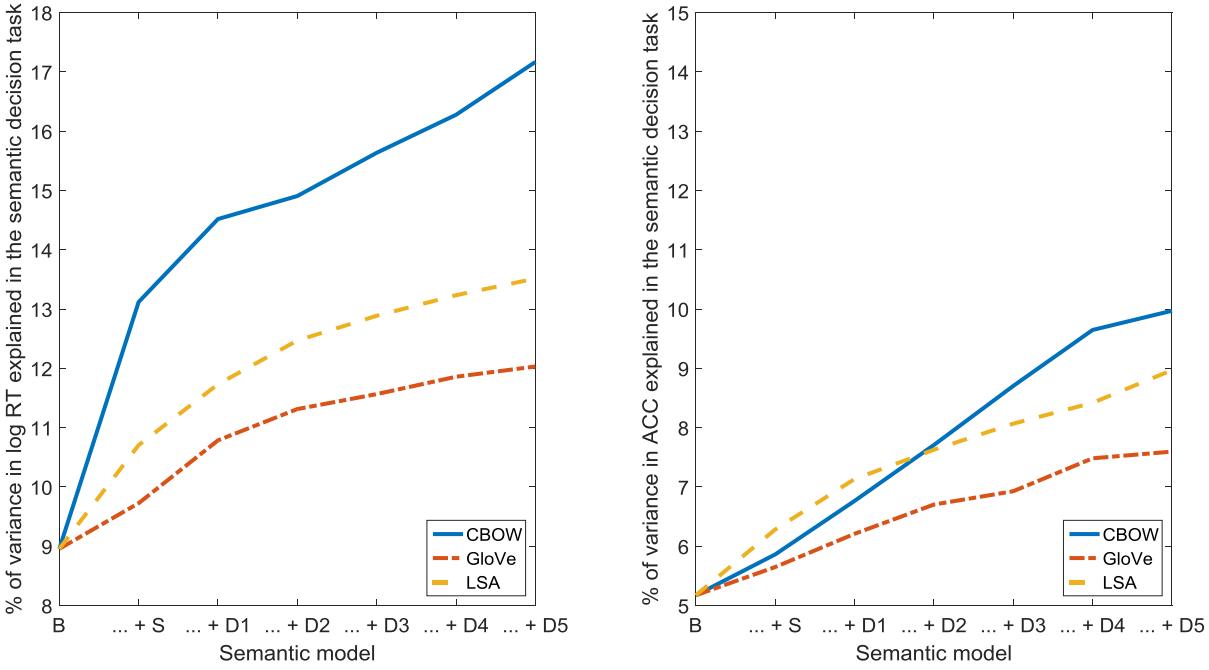


Figure 14. Percentage of variance in log response time (RT) and accuracy (ACC) in the semantic decision task, accounted for by the baseline model (B), the combination of the baseline model and the structural model (… + S), and the combination of the baseline model, the structural model, and consecutive steps of the dynamic model (… + D1 through … + D5).

Table 10. Results of model comparisons for predicting log response time and accuracy in the semantic decision task. B = baseline model; S = structural model; $D_{1...k}$ = first *k* individual steps of the dynamic model.

| | | Enhanced vs simple model | | | | | |
|---|---|---|---|---|---|---|---|
| Model | Statistic | B vs B + S | B + S vs B + S + $D_1$ | B + S + $D_1$ vs B + S + $D_{1...2}$ | B + S + $D_{1...2}$ vs B + S + $D_{1...3}$ | B + S + $D_{1...3}$ vs B + S + $D_{1...4}$ | B + S + $D_{1...4}$ vs B + S + $D_{1...5}$ |
| | | Log response time (semantic decision) | | | | | |
| CBOW | F | 12.57 | 4.28 | 1.22 | 2.2 | 1.84 | 2.78 |
| | (p) | (< .0001) | (< .0001) | (.27) | (.02) | (.05) | (.002) |
| | df | 10, 2624 | 10, 2614 | 10, 2604 | 10, 2594 | 10, 2584 | 10, 2574 |
| GloVe | F | 2.24 | 3.11 | 1.56 | 0.65 | 0.95 | 0.49 |
| | (p) | (.01) | (.0006) | (.11) | (.77) | (.48) | (.90) |
| | df | 10, 2624 | 10, 2614 | 10, 2604 | 10, 2594 | 10, 2584 | 10, 2574 |
| LSA | F | 5.14 | 3.04 | 2.20 | 1.11 | 1.10 | 0.82 |
| | (p) | (< .0001) | (.0008) | (.02) | (.35) | (.36) | (.61) |
| | df | 10, 2624 | 10, 2614 | 10, 2604 | 10, 2594 | 10, 2584 | 10, 2574 |
| | | Accuracy (semantic decision) | | | | | |
| CBOW | F | 1.92 | 2.53 | 2.51 | 2.79 | 2.30 | 0.85 |
| | (p) | (.04) | (.005) | (.005) | (.002) | (.01) | (.58) |
| | df | 10, 2624 | 10, 2614 | 10, 2604 | 10, 2594 | 10, 2584 | 10, 2574 |
| GloVe | F | 1.32 | 1.56 | 1.43 | 0.69 | 1.52 | 0.31 |
| | (p) | (.21) | (.11) | (.16) | (.73) | (.13) | (.98) |
| | df | 10, 2624 | 10, 2614 | 10, 2604 | 10, 2594 | 10, 2584 | 10, 2574 |
| LSA | F | 3.10 | 2.40 | 1.44 | 1.22 | 1.03 | 1.53 |
| | (p) | (.0006) | (.008) | (.16) | (.27) | (.42) | (.12) |
| | df | 10, 2624 | 10, 2614 | 10, 2604 | 10, 2594 | 10, 2584 | 10, 2574 |

*Concreteness and imageability rating:*

For the concreteness and imageability ratings, the results are the following (see Figure 15 and Table 11). With respect to concreteness, all the structural and dynamic models improved the fit, except for step one (GloVe, LSA), and step four (LSA), in the dynamic models. Similarily, with respect to imageability, all the structural and dynamic models improved the fit, except for step one (GloVe, LSA), and step four (GloVe, LSA), in the dynamic models. Our findings clearly indicate that concreteness and imageability are reflected in both the structure and dynamics of the semantic network.

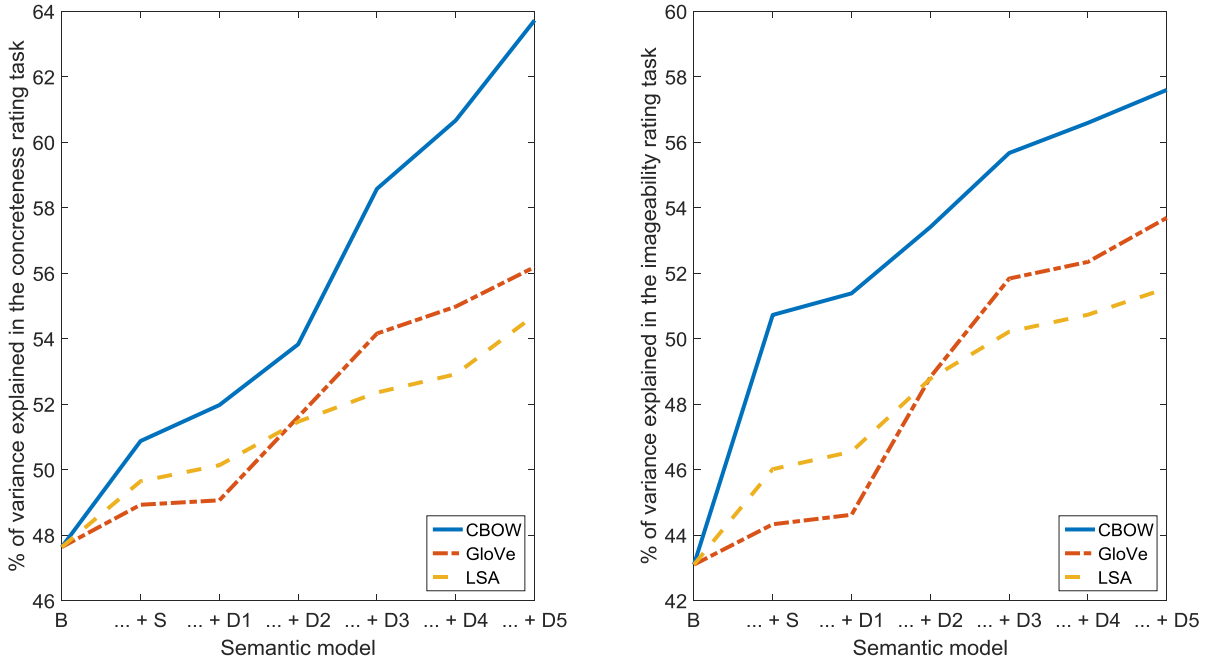


Figure 15. Percentage of variance in concreteness and imageability ratings, accounted for by the baseline model (B), the combination of the baseline model and the structural model (… + S), and the combination of the baseline model, the structural model, and consecutive steps of the dynamic model (… + D1 through … + D5).

Table 11. Results of model comparisons for predicting concreteness and imageability ratings. B = baseline model; S = structural model; $D_{1...k}$ = first *k* individual steps of the dynamic model.

| | | Enhanced vs simple model | | | | | |
|---|---|---|---|---|---|---|---|
| Model | Statistic | B vs B + S | B + S vs B + S + $D_1$ | B + S + $D_1$ vs B + S + $D_{1...2}$ | B + S + $D_{1...2}$ vs B + S + $D_{1...3}$ | B + S + $D_{1...3}$ vs B + S + $D_{1...4}$ | B + S + $D_{1...4}$ vs B + S + $D_{1...5}$ |
| | | Concreteness rating | | | | | |
| CBOW | F | 15.21 | 5.25 | 9.05 | 26.06 | 12.09 | 18.75 |
| | (p) | (< .0001) | (< .0001) | (< .0001) | (< .0001) | (< .0001) | (< .0001) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| GloVe | F | 5.84 | 0.74 | 12.02 | 12.69 | 4.18 | 6.17 |
| | (p) | (< .0001) | (.69) | (< .0001) | (< .0001) | (< .0001) | (< .0001) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| LSA | F | 9.21 | 2.22 | 6.32 | 4.21 | 2.64 | 8.70 |
| | (p) | (< .0001) | (.01) | (< .0001) | (< .0001) | (.003) | (< .0001) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| | | Imageability rating | | | | | |
| CBOW | F | 35.71 | 3.09 | 9.64 | 11.72 | 4.96 | 5.11 |
| | (p) | (< .0001) | (.0006) | (< .0001) | (< .0001) | (< .0001) | (< .0001) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| GloVe | F | 5.11 | 1.21 | 18.84 | 14.34 | 2.29 | 6.64 |
| | (p) | (< .0001) | (.28) | (< .0001) | (< .0001) | (.01) | (< .0001) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |
| LSA | F | 12.46 | 2.35 | 9.76 | 6.61 | 2.26 | 3.75 |
| | (p) | (< .0001) | (.01) | (< .0001) | (< .0001) | (.01) | (< .0001) |
| | df | 10, 2307 | 10, 2297 | 10, 2287 | 10, 2277 | 10, 2267 | 10, 2257 |

*Semantic similarity/relatedness ratings:*

For the semantic similarity/relatedness ratings (see Figure 16, Table 12, and Table 13), the addition of any of the steps in the dynamic models (CBOW, GloVe, LSA), with the exception of steps four (LSA) and five (CBOW, LSA), improved the fit to the SL dataset. For the MEN dataset, the fit was increased by the addition of steps one and two (CBOW, GloVe, LSA), step three (GloVe, LSA), as steps four and five (CBOW, LSA). Also, the addition of steps one (CBOW, GloVe), two (GloVe), three (GloVe, LSA), and four (CBOW) of the dynamic models, ameliorated the fit to the SimVerb-3500 dataset. In the case of the SimLex-999 dataset, the inclusion of steps two (GloVe), four (CBOW, GloVe) and five (LSA) in the dynamic models significantly contributed to the model fit. These results seem to suggest that similarity/relatedness judgements correlate strongly with both the structure and dynamics of the semantic network underlying our models. Our findings hold across datasets covering a wide range of word frequencies, semantic relations, and parts of speech.

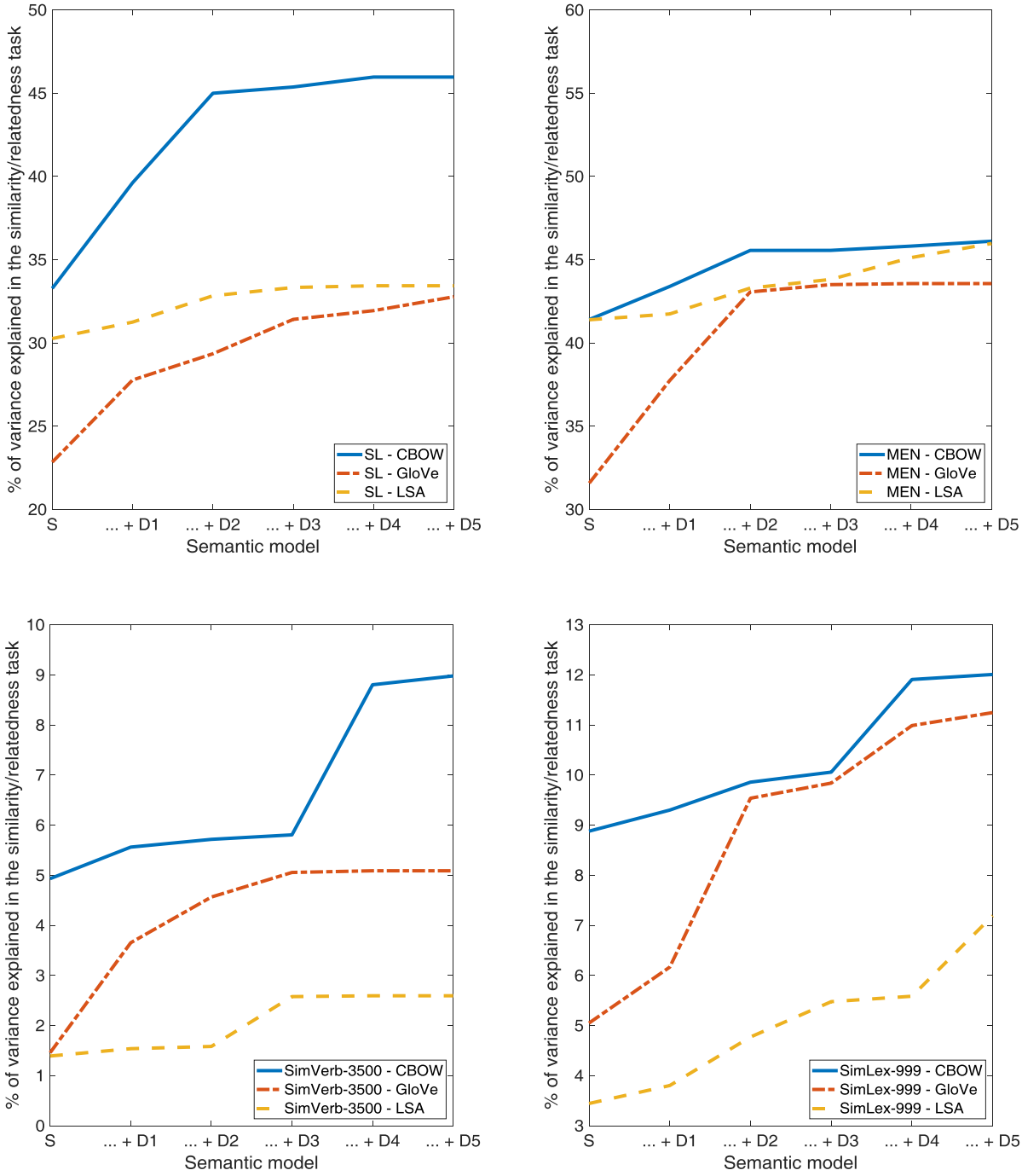


Figure 16. Percentage of variance in similarity and/or relatedness ratings (from the SL, MEN, SimVerb-3500, and SimLex-999 datasets), accounted for the structural model (S), and a combination of the structural model and consecutive steps of the dynamic model (… + D1 through … + D5).

Table 12. Results of model comparisons for predicting SL and MEN similarity/relatedness ratings. S = structural model; $D_{1...k}$ = first *k* individual steps of the dynamic model.

| | | Enhanced vs simple model | | | | |
|---|---|---|---|---|---|---|
| Model | Statistic | S vs S + $D_1$ | S + $D_1$ vs S + $D_{1...2}$ | S + $D_{1...2}$ vs S + $D_{1...3}$ | S + $D_{1...3}$ vs S + $D_{1...4}$ | S + $D_{1...4}$ vs S + $D_{1...5}$ |
| | | Similarity / relatedness rating (SL dataset) | | | | |
| CBOW | F | 315.66 | 293.50 | 20.59 | 33.17 | 0.00 |
| | (p) | (< .0001) | (< .0001) | (< .0001) | (< .0001) | (> .99) |
| | df | 2, 6007 | 2, 6005 | 2, 6003 | 2, 6001 | 2, 5999 |
| GloVe | F | 205.17 | 66.98 | 90.36 | 23.15 | 37.20 |
| | (p) | (< .0001) | (< .0001) | (< .0001) | (< .0001) | (< .0001) |
| | df | 2, 6007 | 2, 6005 | 2, 6003 | 2, 6001 | 2, 5999 |
| LSA | F | 43.22 | 70.73 | 22.46 | 4.61 | 0.00 |
| | (p) | (< .0001) | (< .0001) | (< .0001) | (.01) | (> .99) |
| | df | 2, 6007 | 2, 6005 | 2, 6003 | 2, 6001 | 2, 5999 |
| | | Similarity / relatedness rating (MEN dataset) | | | | |
| CBOW | F | 49.67 | 58.19 | 0.10 | 6.64 | 7.74 |
| | (p) | (< .0001) | (< .0001) | ( .90) | (.001) | (.0004) |
| | df | 2, 2831 | 2, 2829 | 2, 2827 | 2, 2825 | 2, 2823 |
| GloVe | F | 140.14 | 132.13 | 10.97 | 1.59 | 0.00 |
| | (p) | (< .0001) | (< .0001) | (< .0001) | (.20) | (> .99) |
| | df | 2, 2831 | 2, 2829 | 2, 2827 | 2, 2825 | 2, 2823 |
| LSA | F | 8.48 | 38.77 | 13.07 | 33.75 | 22.15 |
| | (p) | (.0002) | (< .0001) | (< .0001) | (< .0001) | (< .0001) |
| | df | 2, 2831 | 2, 2829 | 2, 2827 | 2, 2825 | 2, 2823 |

Table 13. Results of model comparisons for predicting SimVerb-3500 and SimLex-999 similarity/relatedness ratings. S = structural model; $D_{1...k}$ = first *k* individual steps of the dynamic model.

| | | Enhanced vs simple model | | | | |
|---|---|---|---|---|---|---|
| Model | Statistic | S vs S + $D_1$ | S + $D_1$ vs S + $D_{1...2}$ | S + $D_{1...2}$ vs S + $D_{1...3}$ | S + $D_{1...3}$ vs S + $D_{1...4}$ | S + $D_{1...4}$ vs S + $D_{1...5}$ |
| | | Similarity / relatedness rating (SimVerb-3500 dataset) | | | | |
| CBOW | F | 11.08 | 2.72 | 1.58 | 54.48 | 3.21 |
| | (p) | (< .0001) | (.07) | (.21) | (< .0001) | (.04) |
| | df | 2, 3322 | 2, 3320 | 2, 3318 | 2, 3316 | 2, 3314 |
| GloVe | F | 37.85 | 15.87 | 8.58 | 0.58 | 0.00 |
| | (p) | (< .0001) | (< .0001) | (.0002) | (.56) | (> .99) |
| | df | 2, 3322 | 2, 3320 | 2, 3318 | 2, 3316 | 2, 3314 |
| LSA | F | 2.47 | 0.75 | 16.95 | 0.28 | 0.00 |
| | (p) | (.08) | (.47) | (< .0001) | (.75) | (> .99) |
| | df | 2, 3322 | 2, 3320 | 2, 3318 | 2, 3316 | 2, 3314 |
| | | Similarity / relatedness rating (SimLex-999 dataset) | | | | |
| CBOW | F | 2.18 | 2.90 | 1.04 | 9.82 | 0.53 |
| | (p) | (.11) | (.06) | (.35) | (< .0001) | (.59) |
| | df | 2, 941 | 2, 939 | 2, 937 | 2, 935 | 2, 933 |
| GloVe | F | 5.60 | 17.50 | 1.57 | 6.04 | 1.36 |
| | (p) | (.004) | (< .0001) | (.21) | (.002) | (.26) |
| | df | 2, 941 | 2, 939 | 2, 937 | 2, 935 | 2, 933 |
| LSA | F | 1.75 | 4.78 | 3.50 | 0.54 | 8.08 |
| | (p) | (.17) | (.009) | (.03) | (.58) | (.0003) |
| | df | 2, 941 | 2, 939 | 2, 937 | 2, 935 | 2, 933 |

**Discussion:**

We described here three models that take into account the structural properties of semantic networks, as well as their dynamic aspects, namely the flow of semantic activation generated by the automatic processing of individual words. By embedding both structure and dynamics, we could assess the effects of both direct and indirect (mediated) semantic relations between words, rather than limiting our analysis to strong, direct links. We found that our dynamic models predict results in all tasks we have considered above and beyond what is predicted by a model that takes into account not only a large number of lexical and sub-lexical variables, but also semantic variables such as semantic diversity (Hoffman et al., 2013). Semantic diversity quantifies the similarity of the linguistic contexts in which a given word appears, has been found to account for a significant amount of variance in the lexical decision task

(Hoffman & Woollams, 2015), and has been argued to capture important differences in semantic processing, especially between concrete and abstract words.

Of the three dynamic models, the ones based on CBOW and GloVe generated better results than the one based on LSA, in almost all the tasks (with the exception of the semantic decision task), in line with the finding that “word-as-context” models typically yield a higher performance than “document-as-context” models, and that “predict” models are usually superior to “count” models (see Chapter 2). Importantly, however, even for the poorest performing model, namely LSA, adding the spreading activation mechanism improved the model fit in all tasks (except for lexical decision, where the other models also did not fare very well). Thus, we have reason to believe that the advantages of considering the spread of activation are not tied to a particular type of distributional model. However, this does not mean that the choice of model is irrelevant: better structural models are likely to produce better dynamic models, given that the flow of semantic activation employs information encoded in the structure of the semantic network.

We have shown that our models predict word processing in different tasks: both offline (untimed), semantic tasks such as providing ratings for concreteness and imageability, or for similarity/relatedness, but also online (timed) tasks that require more (semantic decision) or less (lexical decision) semantic information, both of which are assumed to recruit automatic spreading of activation across the semantic network (Dell, 1986; Roelofs, 1992). It is important to note here that although our models significantly predicted response time and accuracy in the lexical decision task, they are considerably more successful at predicting results from semantic rating and semantic decision tasks. A simple account for this difference is semantic decision and the other tasks tap into semantic processing to a greater extent than lexical decision. Importantly, however, the improvement in the fit of the models due to the dynamic steps was not limited to offline semantic tasks, but was found also in online tasks (semantic decision and, to a lesser extent, lexical decision). This result indicates that the mechanism we have described here can be thought of in terms of automatic spreading of activation across the network.

Overall, our results show the usefulness and plausibility of joining distributional probabilistic modelling of semantics with dynamic processes. There are however limitations that we need to take into account. First, we make a number of simplifying

assumptions in the models. For example, we assumed that all the words receive the same amount of initial activation, however, it is very likely that some words might benefit from a stronger initial activation than others, for a variety of reasons (e.g., due to increased imageability, valence, arousal, semantic and/or contextual diversity). We opted for this simplification because we simply do not know how much more activation particular words would receive. Another issue refers to the fact that, for the same reasons, we impose that the total amount of activation in our network remains constant, while it seems more cognitively realistic that activation first increases (i.e., semantic representations are accessed gradually), then reaches a plateau, and finally decreases (i.e., semantic representations are affected by competition for retrieval and time-dependent decay, among other factors). Since modelling this type of dynamics requires the addition of several theoretical assumptions and model parameters, we do not tackle this issue here, for reasons of simplicity.

Finally, our dynamic models rely on a process of spreading activation in order to access higher-order semantic relationships between words. Spreading of activation has long been considered as a psychologically plausible dynamic mechanism (e.g., Collins & Loftus, 1975; Dell, 1986). Our hypothesis is that, during word retrieval, the spreading activation mechanism accesses useful, but implicit information stored in the semantic representations. However, there is also the possibility that, by employing a considerably larger corpus and/or a more sophisticated model architecture (e.g., in the case of CBOW, by adding extra hidden layers, or combining word and context vectors), the implicit information might be made more explicit during learning, thus simulating the effects of retrieval-based spreading activation. Future work is needed to assess these alternative possibilities.

## 4.4. Similarities and differences with other models

Our dynamic models of semantic processing are similar to a number of other formal approaches to semantics, especially those put forward by Anderson (1983), De Deyne and collaborators (2016), and Steyvers and collaborators (2005). Moreover,

there are a number of other approaches to semantic cognition which share our interest in exploring the role of weak and indirect semantic relations between words, and in analysing the dynamics of semantic processing. These approaches examine task performance in tasks such as intralist and extralist cued recall (Bruza, Kitto, Nelson, & McEvoy, 2009; Nelson, Kitto, Galea, McEvoy, & Bruza, 2013), discrete free association and synonym generation (Howard, Shankar, & Jagadisan, 2011), continuous free association (De Deyne & Storms, 2008a, b), verbal fluency (Hills, Jones, & Todd, 2012; Hills, Todd, & Jones, 2015), as well as lexical decision and similarity rating (De Deyne et al., 2013). Given the large methodological differences between these studies and ours, we do not discuss them here in more detail.

De Deyne and collaborators (2016) investigated, among others, some of the differences that exist between two popular types of semantic representations, namely those based on discrete and continued word association, and those based on word co-occurrence in text corpora. The study also looked at the explanatory power of weak and/or indirect semantic relations, obtained using a spreading activation mechanism very similar to that employed by Anderson (1983). However, in contrast to our approach, the authors focused on the semantic categorisation task and semantic similarity ratings, whereas we examine lexical and semantic decision, as well as concreteness, imageability and similarity/relatedness ratings. Another difference between their linguistic model and ours is the manner in which activation spreads: we assume that the global distributional overlap between a source word and a target word (i.e., their cosine similarity) determines the amount of activation transmitted, whereas De Deyne and collaborators considered that this quantity is computed from the local probability of the source and target word directly co-occurring in text (i.e., their pointwise mutual information). Also, in their dynamic model, the authors examined only the equilibrium state, as opposed to our approach, where we look at both the initial steps in the spreading of activation, and the activation profile corresponding to the equilibrium state.

Steyvers and collaborators (2005) examined the role of direct and mediated semantic associations in a number of episodic memory tasks, involving the evaluation of similarity between novel and studied items in a recognition-based paradigm, the recollection of studied items in the extralist cued recall task, and the production of intrusions in the free recall task. Although the tasks rely primarily on episodic memory,

the authors did not include any episodic component within their model, focusing instead on the semantic similarity between the words presented during the tasks. The associative structure of semantic memory was obtained from the USF norms, which were first symmetrized, by combining cue-target and target-cue association probabilities, and then subjected to one of three treatments: (1) singular value decomposition for one-step associations; (2) singular value decomposition for both one-step and two-step associations; (3) multidimensional scaling for associative chains involving one or more steps. By employing dimensionality reduction techniques and multi-step associations, the resulting semantic network indexed both direct and indirect semantic relations between words, which is a defining feature of both their model and ours. Nevertheless, since our semantic representations are constructed automatically from large text corpora, we are not limited with respect to the number of words that we can include in our model, and we can make use of richer, more fine-grained information than that which can be gleaned from free association norms, given that the latter usually collect only between 100 and 200 associations per normed word. Another difference between the models is that we look beyond one-step and two-step associations, by taking into account the effects of associative chains of lengths from one to five. Admittedly, Steyvers and collaborators also explored the contribution of long associative chains, in the third version of their model, but they considered only the shortest chain between two words, whereas we employ all the chains between the same two words, regardless of length. A final difference is that we do not assume that semantic associations are symmetric (Tversky, 1977), especially given the strong asymmetry that is characteristic of free association probabilities (Nelson et al., 2005).

Anderson (1983) offered a unified account of various long-term memory phenomena, with an emphasis on memory retrieval. Similar to our models, human memory was represented as a network of associations between meaningful units (e.g., words or sentences), such that the retrieval of task-relevant units strongly depended on the spreading of activation between the elements of the network. However, there are at least two key differences between Anderson's model and ours. Firstly, although Anderson mentioned that the spreading activation mechanism was inspired by research related to semantic priming, his model did not have a particular focus on semantic memory, given that the tasks to which the model were applied are mainly episodic. The author provided a detailed description of a number of aspects that are

typically studied in the context of episodic memory, such as the occurrence of proactive and retroactive interference in the paired-associate paradigm, the improvement of memory performance with practice, and the levels-of-processing effect. Moreover, the author indicated how to compute the strength of the associations formed between items that are presented in the same episodic context, but he did not offer a means of quantifying the semantic associations formed between items that are related in meaning. As a result, since our interest lies exclusively with semantic memory, many important aspects of Anderson's model (e.g., the nature and structure of the memory representations, as well as the encoding, maintenance, and forgetting mechanisms associated with them) are not present in our models. Secondly, the semantic associations between words are computed very differently between the models, since the quantities involved in computing the associations for the Anderson model depend on an a non-relational variable (i.e., the "strength" of each word, based on the number and spacing of repetitions for that word), whereas the associations in our structural models are derived from a relational variable (i.e., the distributional similarity between pairs of words, based on the history of their co-occurrence with other words).

Thus, overall, our dynamic models are similar to the three other models described above, in that they allow for indirect, mediated semantic relations between words to contribute to task performance, in a variety of semantic tasks. However, the models also differ significantly in a number of respects. Firstly, given that most of the research on the dynamics of semantic activation has relied on free association norms (De Deyne et al., 2013; Nelson et al., 1998), it is not surprising that two of the three related models used semantic representations derived from free association data. In contrast, our models operate with text-based, distributional representations, which have the advantage of covering a considerably larger set of words, and of capturing a multitude of weak, but reliable semantic associations between words (De Deyne, Navarro, Perfors, & Storms, 2012), which are largely absent from free association norms. Also, since free association norms are task-based, whereas text corpora are task-independent, we believe that the semantic information accessed by our models is more general than that provided by free association norms. Secondly, the emphasis of our models is on the semantic process that extracts implicit information from the semantic representations, and on the additional data revealed at each step of the

process. The related models did not examine the individual steps in the evolution of the semantic networks, but instead collapsed all the available information into a new, enhanced representation (e.g., in order to reduce the sparsity of the representations; De Deyne et al., 2016). Finally, we look at the individual semantic neighbourhoods associated with a large number of words, whereas the other approaches either investigated global neighbourhoods (De Deyne et al., 2016), or were not directly concerned with network properties (Anderson, 1983; Steyvers et al., 2005).

The majority of the models presented here are based on distributional semantic models, and are in line with the mainstream approach of using co-occurrences of words in text as the only data source from which to learn semantic representations and their neighbourhood structure. It is the case, however, that a number of models have also been proposed that are not limited to linguistic information derived from texts, but also employ multimodal information, corresponding to sensory-motor and emotional properties of words as data from which semantic representations are learnt (e.g., Andrews et al., 2009; Bruni et al., 2014). These grounded (or embodied) models have been shown to provide better fit to behavioural data than models based solely on linguistic data. For example, Andrews and collaborators (2009) found that a Topic model (see Griffiths, Steyvers, et al., 2007) trained on both text and speaker-generated features (covering perceptual, motor and affective properties of referents) was better at predicting semantic effects in speech error data (specifically semantic errors among slips of the tongue), as well as in semantic priming experiments and in word association norms. One might wonder therefore if the structure of the neighbourhoods and the effect of spreading activation would be different in models of this type. We leave this question for future studies.

## 4.5. Conclusions

We have shown here that by supplementing state-of-the-art text-based models of semantic structure with relatively standard processing assumptions, these models can provide a much better fit to behavioural data from word processing tasks that

require the use of semantic information (ratings of concreteness/imageability, semantic similarity/relatedness, semantic decision), but also for tasks such as lexical decision, for which semantic information plays a secondary role. The improvement from structural models alone is especially important given the large number of lexical and semantic variables we had already included in most of our baseline comparison models. Thus, our work demonstrates that by bringing together large scale probabilistic models of semantic representations and processing models we can better account for a variety of behavioural results. Moreover, the distributional models we chose cover a representative selection of some of the most frequently used model architectures (e.g., “count” vs “predict; “word-as-context” vs “document-as-context”; see Chapter 2), suggesting that the gains of adding processing assumptions are not tied to a particular model or task. Our results extend those obtained by De Deyne and collaborators (2016), who used a similar methodology, but focused only on one type of linguistic model and two semantic tasks.

An important implication resulting from our findings is that dynamics are important and useful when modelling semantic behaviour. As a result, network analyses of semantics can be easily improved by combining structural and processing assumptions, either in a direct manner (e.g., via spreading activation, in neural network models, or multi-step inference, in probabilistic models), or in an indirect way (e.g., by examining shortest path, flow and random process based centrality measures; De Deyne et al., 2016; Griffiths, Steyvers, & Firl, 2007; Steyvers et al., 2005; for a technical introduction, see Koschützki et al., 2005).

# 5. Concreteness and semantic network structure

## 5.1. Introduction

A number of recent studies have investigated the patterns of semantic relations that link the representations of words within network models of semantic memory. Whether the network structure is taken to provide a model of semantic memory (Collins & Loftus, 1975) or whether, instead, it is seen as a convenient way to operationalise relations among words without assumptions concerning the structure of semantic memory, network representations afford detailed quantitative analyses. Thus, network analyses of semantic networks have become a relatively popular research topic (for reviews, see Baronchelli et al., 2013; Borge-Holthoefer & Arenas, 2010; Mehler, 2008; Siew et al., 2019).

Some of studies have looked at the topology of semantic networks obtained from distributional models and free association data (De Deyne, Kenett, Anaki, Faust, & Navarro, 2016; Gruenenfelder et al., 2015; Steyvers & Tenenbaum, 2005; Utsumi, 2015), and have found that the two classes of networks are structurally similar, having "small-world" and "scale-free" properties (see Chapter 2). Other studies have examined the effects of including semantic processes that operate over semantic networks, such as spreading activation (De Deyne et al., 2016; Rotaru, et al., 2016, 2018), and have shown that considering such processes increases model performance (see Chapter 4). Yet other studies (De Deyne et al., 2019; Rotaru et al., 2016, 2018; see Chapter 4) have illustrated the ability of network-based measures to predict behavioural data, from a variety of tasks. Also, the differences between semantic networks associated with typical and atypical populations have been investigated (e.g., highly creative persons have more strongly interconnected networks; Kenett, Anaki, & Faust, 2014).

Here, we explore the differences between concrete and abstract words, in terms of the structure of the semantic networks derived from distributional models of semantics. We believe that that our current picture of the network differences between concrete and abstract words, as well as of the behavioural consequences of these

differences, is rather incomplete. For instance, imageability-based interpretations of concreteness (Paivio, 1971, 1986) predict that concrete words should have more and stronger neighbours than abstract ones, since concrete words have a richer perceptual content. On the other hand, theories that emphasize the role of contextual/semantic diversity (Hoffman et al., 2013; Jones et al., 2012) postulate that abstract words should have larger semantic neighbourhoods, given that they occur in a broader variety of linguistic contexts, as compared to concrete ones.

Our approach is very different from that of typical studies on semantic networks, described earlier, which focus either on general structural properties of networks (e.g., the distribution of neighbourhood sizes; Steyvers & Tenenbaum, 2005), or the behavioural effects of local neighbourhood structure (e.g., the relation between the neighbourhood size of individual words and measures of task performance for those words; Recchia & Jones, 2012). More specifically, we provide a comprehensive description of neighbourhood structure taking into account some of the factors, such as imageability, age of acquisition, squared hedonic valence, contextual diversity, and semantic diversity, which have been argued to differ between concrete and abstract words. This also allows us to test some predictions of existing accounts of representational differences between concrete and abstract words.

## 5.2. Differences in semantic richness between concrete and abstract words

Differences in representational richness have been argued to underscore a variety of behavioural effects, as well as the distinction between concrete and abstract words. These differences are usually assumed to be only quantitative, such that concrete words contain more perceptual and motor information, but less introspective and linguistic information, than abstract words (Barsalou & Wiemer-Hastings, 2005; Danguecan & Buchanan, 2016; Newcombe et al., 2012; Kousta et al, 2011; Paivio, 1971, 1986; Wiemer-Hastings & Xu, 2005). In contrast to concrete words, abstract words (e.g., “justice”, “theorem”) are not directly linked to physical objects, which makes it necessary to rely mainly on introspective (e.g., emotional) and linguistic

elements. While this does not exclude any contribution from perception and action (e.g., “justice” is typically associated with a “courtroom”, where a “defendant” is being “tried”), their role is arguably not essential in learning and representing abstract words. As a result, in some cases it is argued that concrete words are semantically richer (Paivio, 1971, 1986), whereas in other cases the opposite argument is made (Rechia & Jones, 2012).

**Are concrete words richer than abstract ones?**

Richness can and has been operationalised in a variety of ways. Three of the most commonly examined richness measures are the number of features (based on feature generation norms), the number of associates (based on free association norms), and the number of semantic neighbours (based on distributional models of semantics). With respect to the number of features, concrete words have an advantage over abstract ones (Recchia & Jones, 2012). An examination of the types of features characteristic to each word class indicates that concrete words have more entity, concrete context and taxonomic features than abstract words, whereas the opposite is true when it comes to introspective features. For concrete words, the number of features facilitates performance in lexical decision, naming and semantic decision tasks (Pexman et al., 2008; Recchia & Jones, 2012; Yap et al., 2011, 2012). However, it not yet clear whether the same effect becomes apparent for abstract words. To the best of our knowledge, the only study to investigate this hypothesis is that by Recchia and Jones (2012), which did not find any effect of number of features, on lexical decision and naming response times. One possibility is that the effect is indeed present, but the scarcity of features corresponding to abstract words makes it challenging to detect, especially when certain confounding variables (e.g., imageability and familiarity) are not controlled for. Another explanation might be that the relevance of features is different for concrete vs abstract words, such that they describe essential characteristics of a word’s representation, in the former case, but have a more peripheral role, being similar to verbal associations, in the latter case. A related alternative is that contextual information is more difficult to access for abstract than for concrete words (Schwanenflugel & Shoben, 1983), meaning that the features of

abstract words might be activated too late in order to have an impact on performance, especially for time-constrained tasks, such as lexical decision and naming.

In relation to the number of associates, concrete words are slightly poorer than abstract ones (Hill, Korhonen, et al., 2014). Unlike for number of features, evidence for the fact that the number of associates influences task performance is mixed. For concrete words, certain studies (Buchanan et al., 2001, exp. 2; Duñabeitia, Avilés, & Carreiras, 2008; Mirman & Magnuson, 2008; Pexman et al., 2008) showed that the numbers of associates correlates positively with performance in lexical decision, naming, semantic decision and progressive demasking tasks. In contrast, no such effect was found in other studies (Buchanan et al., 2001, exp. 1; Recchia & Jones, 2012; Yap et al., 2011, 2012). On the other hand, abstract words do not seem to benefit from an increased number of associates (Recchia & Jones, 2012, exp. 2; Zdrazilova & Pexman, 2013). There are several potential reasons for the lack of an effect of number of associates. For instance, the associates of a particular word might become activated as a result of relatively high level cognitive processing (Buchanan et al, 2001), which would explain why their effect is not reliably detected in lexical decision and naming. It might also be case that the number of associates offers a rather incomplete picture of a word's associative network. More specifically, one study (De Deyne et al., 2013) found that the number of incoming associations (i.e., the number of other words that produce a particular word as an associate), as well as the interconnectivity of the associative network, are vastly more predictive of lexical decision performance, than the number of outgoing associations (i.e., the number of associates generated by a particular word).

Finally, concrete words have fewer semantic neighbours than abstract ones (Recchia & Jones, 2012; Hargreaves & Pexman, 2014). For concrete words, the size of the semantic neighbourhood appears to improve performance in the lexical decision and progressive demasking tasks, but to have no effect in naming and semantic decision (Danguecan & Buchanan, 2016; Mirman & Magnuson, 2008; Moffat, Siakaluk, Sidhu, & Pexman, 2015; Pexman et al., 2008; Recchia & Jones, 2012, exp. 1; Yap et al., 2011, 2012). The same pattern of results is observed for abstract words (Danguecan & Buchanan, 2016; Newcombe et al., 2012; Recchia & Jones, 2012; Moffat et al., 2015). While the null effect reported for naming is not surprising, given that semantic factors generally have a very weak influence on task behaviour, not

finding any effect for semantic decision is more difficult to account for. One possibility is that semantic neighbourhood measures inadvertently conflate close and distant neighbours (Mirman & Magnuson, 2008). Since close neighbours are likely to have an inhibitory effect (by competing with the stimulus for retrieval), while distant neighbours should have a facilitatory effect (by increasing the familiarity of the stimulus), it could be the case that the two effects cancel each other out. Alternatively, as mentioned in our discussion of associative networks, the number of semantic neighbours might provide an impoverished perspective on the important characteristics of the semantic neighbourhood. Adding to this potential problem is the fact that, whereas the number of features and the number of associates each have the same definition across studies, and are obtained from the same set of norms, the number of semantic neighbours is computed in a variety of manners, either directly or indirectly. Depending on the study, semantic neighbourhood size is captured by measures such as semantic density (Buchanan et al., 2001), average radius of co-occurrence (Shaoul & Westbury, 2010), and inverse of neighbour count (Newcombe et al., 2012), with no clear indication of the reasons why a particular measure is preferable to any of its many alternatives (e.g., on grounds of its cognitive plausibility).

At least two different accounts on the source of richness effects have been put forward (Balota, 1990; Hino & Lupker, 1996). Firstly, words with rich representations should elicit more activation within the semantic system, as compared to representationally poor words. In addition, the activation is thought to be automatic, regardless of whether the task explicitly includes a semantic component. This mechanism is likely to play a part in the lexical decision task, where the familiarity of a stimulus is a good indication of the stimulus being a valid word. Secondly, semantically wealthy words should automatically generate a stronger feedback from the semantic level to the orthographic and phonological layers. This mechanism ought to play a role in tasks such as lexical decision, naming, and progressive demasking, which rely heavily on non-semantic processing. Both mechanisms should also contribute to task performance in semantically-oriented tasks, such as semantic decision and sentence reading, although the effect should be considerably weaker (or even absent), given that these tasks are largely dependent on strategic semantic

retrieval, and that the decision stage involves examining the particular contents of a word's semantic representation.

Besides investigating which dimensions of semantic richness improve performance in a number of tasks, as well as how concrete and abstract concepts differ along those dimensions, a number of recent studies have focused on the time course of richness effects. From a theoretical perspective, embodied accounts of semantic cognition, such as the Language and Situated Simulation theory (Barsalou et al., 2008) and the Symbol Interdependency Hypothesis (Louwerse & Jeuniaux, 2008), assume that linguistic processing is faster than perceptual and motor processing, even though both components become active at roughly the same time. In order to test this prediction, Hargreaves and Pexman (2014) employed the signal-to-respond paradigm (for a comprehensive review, see Ratcliff, 2006), applying it to the lexical decision and semantic decision tasks. For lexical decision, the researchers failed to find an early effect of the number of semantic neighbours (which is a language-based measure), but detected a relatively late effect of the number of features (which is mostly a perception/action-based measure). In contrast, for semantic decision, both richness measures were significant predictors of task performance, such that the influence of the number of semantic neighbours became noticeable before that of the number of features. These findings are largely consistent with the assumption that, in light of their higher complexity, perceptual and motor simulations are slower than language-driven processes.

**Five factors and their impact on network structure:**

In our analyses, first we compute measures of neighbourhood size and interconnectivity (clustering coefficient) separately for 2,328 concrete and abstract words, classified based on a median split of available concreteness ratings (Brysbaert et al., 2014). For each measure, we then assess, using regression analysis, whether semantic factors such as imageability (Gilhooly & Logie, 1980; Stadthagen-Gonzalez & Davis, 2006), age of acquisition (Kuperman et al., 2012), squared hedonic valence (Warriner et al., 2013), but also log contextual diversity (Van Heuven et al., 2014) and semantic diversity (Hoffman et al., 2013), differentially affect the network structure for concrete and abstract words. We chose to include these five factors because each of

them has been hypothesized to play an important role with respect to semantic representations and processes.

Imageability is central to the Dual Coding theory (Paivio, 1971, 1986), which assumes that abstract words are represented in a predominantly verbal code, while concrete words benefit from having a representation that encompasses both a linguistic and a perceptual (i.e., mainly visual) component. According to this view, concrete words should have more semantic neighbours, under the assumption that the visually mediated, semantic relations between words are also reflected in language. All else being equal, visual relations should strengthen distributional similarity, based on the degree of overlap between the visual world and the linguistic discourse.

Age of acquisition is considered to reflect the centrality of a concept within semantic memory, such that very important and general words are learned earlier than less crucial and more specific words. Word learning appears to depend heavily on the process of differentiation, by which the meaning of newly encountered words is derived by gradually modifying the meaning of words that are already known to the learner. As a result, the newly acquired word and the previously learned word to which it relates become semantic neighbours within the semantic network, and also, some the old word's neighbours become neighbours of the new word as well. Since early words are candidates for differentiation more often that late ones, given the temporal asymmetry of the process, they should have a larger number of semantic neighbours (Steyvers & Tenenbaum, 2005). Additionally, words with many neighbours are assumed to be favoured as targets for differentiation, through the process of preferential attachment (Barabási & Albert, 1999), further increasing the advantage of words with a lower age of acquisition.

Emotional valence is related to age of acquisition, in that highly valenced words are acquired earlier than neutral ones. It has been hypothesized (Kousta et al., 2011) that emotional factors are especially relevant in the learning of abstract words, since they provide a form of grounding that is based on internal, object-independent, introspective states. This kind of grounding is different from that associated with concrete words, which draw upon external, object-dependent, perceptual information, and might support the transition from the learning of almost exclusively concrete words, in the early stages of childhood, to a more balanced learning of both abstract

and concrete words, later on. With respect to neighbourhood size and interconnectivity, it is not clear whether emotional valence has any effect, and, if so, what the specific nature of the effect is. In order to gain some initial insight into this problem, one might be tempted to look into the related literature on the number of associates (for a review, see Chapter 1 from Cramer, 1968), where the effects of emotional content have been investigated at length. Unfortunately, however, the results are difficult to interpret and to generalize beyond the free association paradigm: valenced words have an advantage over neutral ones in discrete association (i.e., a single response collected per stimulus), but the advantage is reversed in the case of continued association (i.e., multiple responses collected per stimulus).

Contextual diversity is another measure of semantic centrality, strongly related to word frequency (Adelman, Brown, & Quesada, 2006). However, whereas frequency is equal to the total number of times a given word is encountered, contextual diversity measures the number of different (linguistic) contexts in which that word appears. If a word is more contextually diverse than another, it can be inferred that the first word is more general and/or more polysemous than the second, since it can be employed in a wider variety of situations. Since each new context enhances the semantic richness of a word, it seems plausible that contextual diversity increases neighbourhood size.

Unlike contextual diversity, which is agnostic with respect to the meaning of words, semantic diversity measures the semantic variation in the contexts where a given word is encountered. It is closely related to word polysemy or ambiguity (Hoffman et al., 2013), based on the finding that abstract words appear in more diverse linguistic contexts, than concrete words. Indeed, it seems plausible that the referents of concrete words can typically be found in only a relatively small number of real situations (e.g., “banana” is associated almost exclusively with scenarios involving “eating” and “desserts”), whereas those of abstract words are significantly more general (e.g., “peace” is related to scenarios involving any kind of dispute between two parties: “the two countries signed a peace treaty”, “the police seems to have made peace with the gangsters”, “she’s finally found inner peace”, “the beach is perfect for peace and relaxation”, etc.). Therefore, overall, abstract words should have more semantic neighbours that concrete ones. Moreover, according to the Context Availability hypothesis (Schwanenflugen & Shoben, 1983), it is easier to retrieve linguistic contexts for concrete, than for abstract words, which seems to suggest that

abstract words have fewer close (but not distant) semantic neighbours than concrete words.

**Methods:**

In order to study the relation between various semantic factors and the structure of semantic memory, we use the CBOW model, as implemented within the *gensim* tool (Řehůřek & Sojka, 2010)[19]. In deriving our semantic representations, we follow all the steps described in Chapter 4 (i.e., we train the CBOW model on the pre-processed, written part of the BNC, thus generating 300-dimensional representations for 28,592 of the most frequent words in the SUBTLEX-UK norms). We then compute the matrix *S*, consisting of the cosine similarity values for all the word pairs covered by the model. As a means of reducing the level of noise, we set to zero all the negative values in *S*. Also, in order to remove outliers, we set to zero all the values above the 99.9$^{th}$ percentile of the strictly positive values.

Having created the model-based semantic network, we can now look at the relationships between the two measures of semantic richness (i.e., neighbourhood size and clustering coefficient), and the five factors we selected, namely imageability, age of acquisition, squared hedonic valence, log contextual diversity, and semantic diversity. In order to obtain a comprehensive picture, for each word we consider a range of neighbourhoods, from very close to very distant. We also distinguish between concrete and abstract words.

With respect to the number of neighbours, we first define 11 neighbourhood strength thresholds, such that $thr_d = \max(S) * (d\text{-}1) / 10$, for *i* between 1 and 11. Based on these thresholds, we then create 10 neighbourhoods bands, such that a neighbour $w_j$ of word $w_i$ is included in the $k^{th}$ neighbourhood band if and only if $thr_k < S(i,j) \leq thr_{k+1}$, for *k* between 1 and 10. In other words, when moving from the first band (1) to the last band (10), we are transitioning from very distant neghbours to very close neighbours. For each neighbourhood band *k*, with *k* between 1 and 10, we run a separate

[19] We used the hyperparameter values from (Rotaru, Vigliocco, & Frank, 2016).

regression, where the dependent variable is the number of neighbours in that neighbourhood band, for each word (i.e., *numNeigh*$_k$), and the independent variables are the five factors. More concisely, the models can be expressed as follows:

$$numNeigh_k \sim imag + aoa + val + contDiv + semDiv$$

where *imag*, *aoa*, *val*, *contDiv*, and *semDiv* represent the normed values for imageability, age of acquisition, squared hedonic valence, log contextual diversity, and semantic diversity, respectively, for each word.

With respect to the clustering coefficient, we start from the neighbourhood strength thresholds defined for the previous analyses. However, we find it more natural to employ a cumulative definition of neighbourhood bands, such that a neighbour $w_j$ of word $w_i$ is included in the $k$th neighbourhood band if and only if $thr_k < S(i,j)$, for $k$ between 1 and 10. Put differently, when moving from the first band (1) to the last band (10), we are transitioning from all the neghbours (i.e., a very liberal definition of neighbourhood) to only very close neighbours (i.e., a very conservative definition of neigbourhood). For each neighbourhood band $k$ and word $w_i$, we compute the clustering coefficient associated with the neighbours in that band, denoted by $clCoef_{k,i}$, in the following manner:

$$clCoef_{k,i} = \frac{2 \cdot numNeighConn_{k,i}}{numNeigh_{k,i} \cdot (numNeigh_{k,i} - 1)}$$

where $numNeigh_{k,i}$ denotes the number of neighbours that word $w_i$ has in the $k^{th}$ neighbourhood band, calculated as:

$$numNeigh_{k,i} = |\{j \mid thr_k < S(i,j), i \neq j\}|$$

and $numNeighConn_{k,i}$ denotes the number of distinct pairs of neighbours that word $w_i$ has in the $k^{th}$ neighbourhood band, such that the neighbours in the pair are also neigbours to one another (i.e., they are connected), calculated as:

$$numNeighConn_{k,i} = |\{(j_1, j_2) \mid thr_k < S(i,j_1), thr_k < S(i,j_2), thr_k < S(j_1,j_2), i \neq j_1, i \neq j_2, j_1 < j_2\}|$$

In other words, $clCoef_{k,i}$ consists of the probability that two randomly chosen neighbours of word $w_i$, in the $k^{th}$ neighourhood band, are also neighbours between themselves. Therefore, the clustering coefficient measures how interconnected a particular neighbourhood is. For each neighbourhood band $k$, with $k$ between 1 and 10, we again run a separate regression, where the dependent variable is the clustering coefficient for the neighbours in that neighbourhood band, for each word (i.e., $clCoef_k$),

and the independent variables are the five factors. More concisely, the models can be expressed as follows:

$$clCoef_k \sim imag + aoa + val + contDiv + semDiv$$

where *imag*, *aoa*, *val*, *contDiv*, and *semDiv* once again represent the normed values for imageability, age of acquisition, squared hedonic valence, log contextual diversity, and semantic diversity, respectively, for each word.

**Results and discussion:**

The size of the semantic neighbourhoods for concrete and abstract words, as a function of neighbour strength, is shown in Figure 17. Overall, the neighbours of concrete words are both more numerous ($t(2,326) = 14.54$, $p < .001$, two-tailed) and closer (i.e., they have a higher average cosine similarity; $t(40,540,780) = 605.37$, $p < .001$, two-tailed) than those of abstract words[20]. This slight advantage holds true for almost all neighbourhood strengths, with the exception of very weak neighbours, for which the pattern is reversed. Our result is in contrast to those of Recchia and Jones (2012), as well as Hargreaves and Pexman (2014), who found that abstract words are richer than concrete ones. However, it is worth pointing out that the aforementioned studies used radically different models and text corpora: the former employed pointwise mutual information, computed over the TASA corpus (17 million words; Zeno et al., 1995), in order to estimate the degree of association between words, whereas the latter relied on the High Dimensional Explorer model (Shaoul & Westbury, 2010), trained over a USENET-based corpus (>300 million words), for the same purpose. Also, the magnitude of the difference between the two classes of words varies considerably between the two studies: when comparing the average size of the neighbourhoods for concrete and abstract words, the first study obtained the values 167 and 201, respectively, while the second study produced the values 1,380 and 3,561, respectively. Our result seems to suggest that the advantage of abstract words over concrete ones, when it comes to semantic neighbourhood size, should not be

[20] In this global analysis, we employ all the neighbours of a given word, irrespective of their neighbourhood bands, and we use vector cosine as a measure of how close two words are.

taken for granted. We believe that our approach, using a state-of-the-art distributional model (i.e., the CBOW model) and a representative collection of texts (i.e., the BNC), might provide a slightly more accurate answer to this question of which class of words is richer in semantic neighbours, but we admit that our claim is largely speculative.

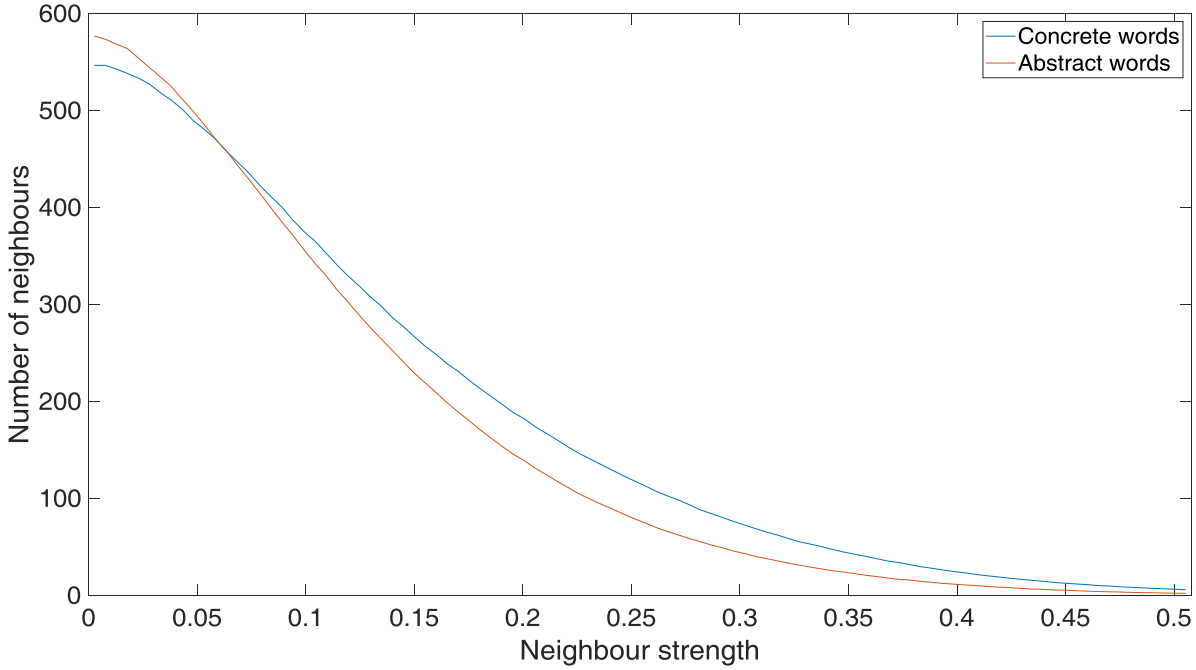


Figure 17. Average number of neighbours, as a function of semantic association strength between a word and its neighbours, measured using the vector cosine.

We hypothesize that the neighbours of concrete words might generally be richer and stronger due to the fact that concrete words are more imageable than abstract ones, while being less polysemous. In order to test this idea, we looked at how well neighbourhood sizes can be predicted by imageability ratings and corpus-based operationalizations of polysemy (i.e., semantic diversity and log transformed contextual diversity), as well as by other factors, such as age of acquisition and squared hedonic valence. As mentioned in the Methods section, we ran a regression analysis each for our concrete and abstract words, using these five independent variables, and number of neighbours, belonging to various strength intervals, as a dependent variable. The results are shown in Figure 18.

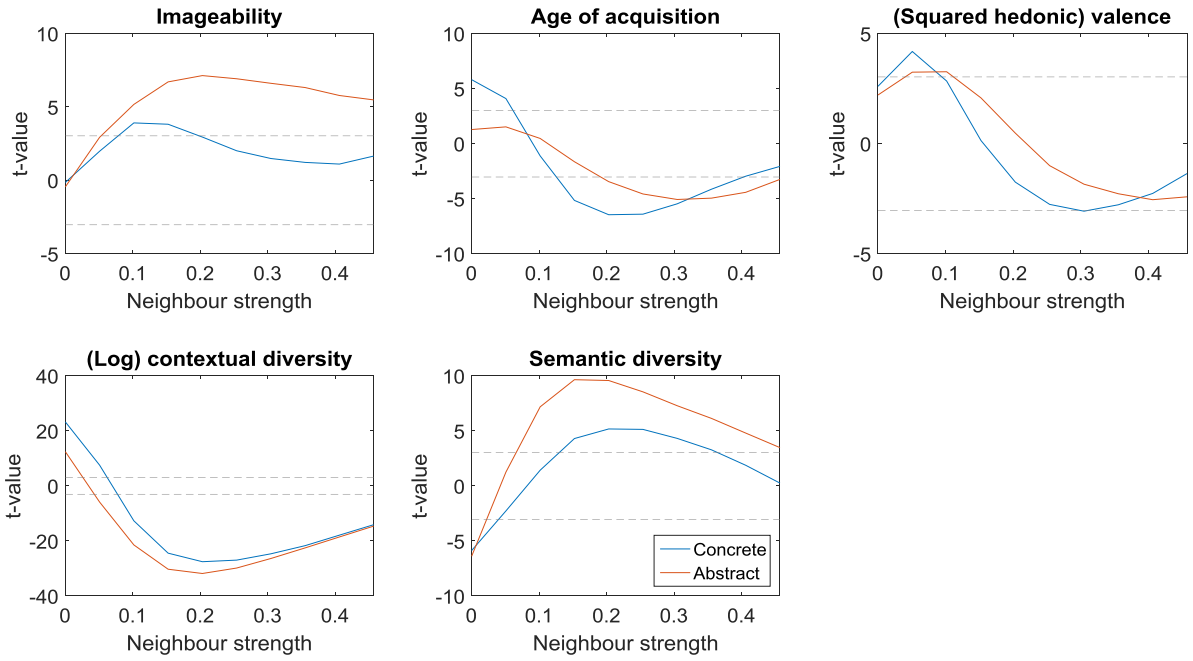


Figure 18. t-values for imageability, age of acquisition, squared hedonic valence, log contextual diversity, and semantic diversity, when entered simultaneously as linear predictors for the number of neighbours, for both concrete and abstract words. The dashed, grey lines correspond to Bonferroni-corrected, one-tailed significance thresholds.

Our results suggest that imageable words have more semantic neighbours than non-imageable ones, which might explain why concrete words have an advantage over abstract words. With respect to imageability, it seems very plausible that having material referents should strengthen the semantic association between concrete words, but not abstract ones, since they are not perceptually grounded. Our assertion is based on two hypotheses: firstly, that the semantics of concrete words can be accessed via both linguistic and perceptual channels, whereas the semantics of abstract words are accessed almost exclusively via linguistic channels; secondly, that the physical similarity, relatedness and/or co-occurrence of different objects is also reflected in the linguistic co-occurrence of the words that denote them. The former assumption is supported, among others, by the experimental evidence accumulated in favour of the Dual Coding theory (Paivio, 1971, 1986). The latter assumption is derived from two multimodal extensions of the "distributional hypothesis" (Harris, 1954), namely the "Appearance Hypothesis" (Griffin, Wahab, & Newell, 2013), according to which "words that occur in similar contexts tend to have referents with similar appearance", and its converse, the "Illustrated Distributional Hypothesis" (Bruni, Uijlings, et al., 2012), which states that "semantically similar objects will tend

to occur in similar environments in images". A recent study by Silberer and Lapata (2014) offers indirect evidence for all three hypotheses, by shedding some light on the significant amount of overlap that exists between visual and linguistic representations. The authors asked participants to rate 7,576 pairs of concrete nouns for both visual and semantic similarity, in order to allow for a more comprehensive evaluation of multimodal models of word meaning. According to our own calculations, the correlation between the two sets of ratings is very high ($r$ = .82, $n$ = 7,567, $p$ < .001), which is consistent with a strong degree of redundancy between visual and linguistic representations, especially given that purely verbal models of semantics (e.g., Baroni et al., 2014) are remarkably successful in accounting for subjective relatednes ratings. Two other studies (Griffin et al., 2013; Lazaridou, Bruni, & Baroni, 2014) bring more evidence in favour of the aforementioned hypotheses, by proving that the considerable match between visual and verbal semantics could serve as a powerful, "zero-shot learning" mechanism for children, allowing them to associate verbal labels (i.e., names) to newly encountered objects, with an accuracy rate well above chance level. Also, in a study by Sadeghi and collaborators (2015), it is shown that word similarity, derived from a distributional model of semantics (i.e., LSA), correlates positively with object co-occurrence probabilities ($r$ = .30), derived from a large set of natural scenes, and with feature-based similarity, derived from semantic feature norms ($r$ = .23).

With respect to age of acquisition, it seem rather strange that we find any effect at all, given that we control for contextual diversity, which is very strongly correlated with frequency. We believe that a possible explanation for our finding lies in the fact that concrete words are usually acquired earlier than abstract ones. As described in network-oriented studies of word learning, the semantics of newly acquired words is likely to be derived from that of already known words, such that the new word "inherits" some of the semantic associates of a known word, and also becomes associated with the word itself. Given that the first words learned by an infant are mostly concrete (Schwanenflugel, 1991; also, the average concreteness of the first 10% of words in our norms, in terms of age of acquisition, is higher than the average concreteness of all the words in our norms, i.e., 4.24 > 3.69, $t$(2,566) = 8.16, $p$ < .001), it is likely that these early words have more (and stronger) semantic associations between themselves, regardless of contextual diversity.

Squared hedonic valence also has a weak positive effect on the number of very distant neighbours. Our explanation for the effect of valence is very similar to that for imageability: the semantics of valenced words can be accessed through both linguistic and emotional channels, as opposed to that of neutral words, which lack emotional content. This should strengthen the semantic association between words of comparable valence. One study (Van Rensbergen, Storms & De Deyne, 2015) found that, in the case of semantic networks derived from free association norms, there are positive correlations between cue and target words, in the case of affective dimensions such as valence ($r^2$ = 0.31) and arousal ($r^2$ = 0.19). Although the semantic networks obtained from distributional models differ quite markedly from those produced by free association studies (e.g., Maki & Buchanan, 2008), we have no reason to believe that the slight tendency of valenced words to be associated to each other should not hold in the current situation, as well.

The two measures of polysemy, namely log contextual diversity and semantic diversity, have considerable positive or negative effects, depending on the strength of the neighbours (i.e., whether the neighbourhood is near or distant). Log contextual diversity has a very strong negative effect for near and distant neighbours, and a strong positive effect for very distant neighbours. Since we control for semantic diversity, contextual diversity becomes a (more cognitively meaningful) proxy for frequency, rather than a measure of polysemy. Within a text corpus, words with high contextual diversity have a larger probability of co-occurring with other words purely by chance, rather than due to semantic association. Therefore, in our interpretation, contextual diversity roughly indicates the amount of “semantic noise” that is present in the representation of a given word. This might explain why near and distant neighbours, which index meaningful semantic relations, are at a disadvantage, whereas very distant neighbours benefit from contextual diversity.

In contrast, semantic diversity has a strong positive effect for near and distant neighbours, and a strong negative effect for very distant neighbours. Given that we control for contextual diversity, semantic diversity can be seen as roughly quantifying the semantic complexity of a word (i.e, the more heterogeneous the linguistic contexts in which a word appears, the more multifaceted its meaning). Consequently, increasing semantic diversity has two opposing effects: on the one hand, more different contexts translates into more neighbours for a given word; on the other hand,

more neighbours means that the associations between a word and its neighbours become weaker. Whether the influence of semantic diversity on neighbourhood size is positive or negative depends upon the relative strength of the two tendencies: for near and distant neighbours, the first effect dominates the other (i.e., the contribution is positive), while the situation is reversed (i.e., the contribution is negative) for very distant neighbours.

In order to have a more exact picture of the factors that influence neighbourhood size, we repeated the analysis presented in Figure 18, but this time we divided each semantic neighbourhood into a neighbourhood containing only concrete words, and another containing only abstract ones. The results are shown in Figure 19.

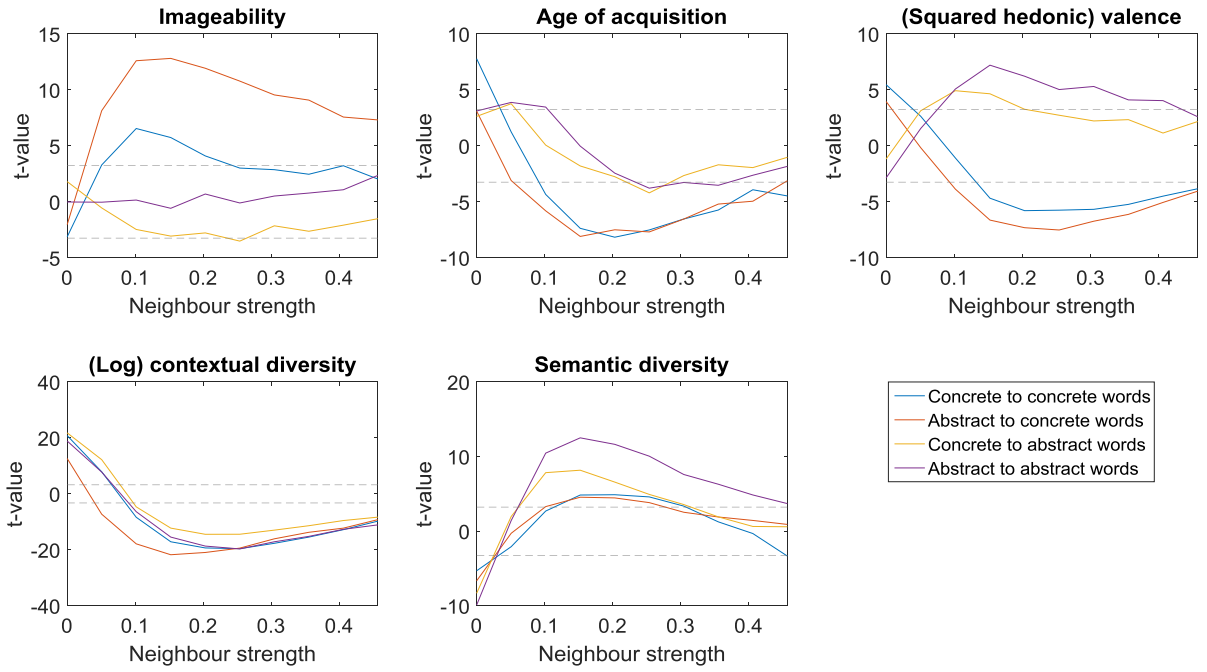


Figure 19. t-values for imageability, age of acquisition, squared hedonic valence, log contextual diversity, and semantic diversity, when entered simultaneously as linear predictors for the number of neighbours, for both concrete and abstract words. Both the initial words and their semantic associates are divided into concrete and abstract words. The dashed, grey lines correspond to Bonferroni-corrected, one-tailed significance thresholds.

With respect to imageability, we find once more that imageable words have more neighbours than non-imageable ones. However, this richness refers almost exclusively to concrete, rather than abstract neighbours, which suggests that semantic associations are strengthened only when both a word and its neighbour are highly imageable (i.e., when physical similarity/relatedness/co-occurrence can play a role).

For age of acquisition, it seems that the slight advantage for concrete words can be largely attributed to semantic associations between concrete words, as predicted by the fact that the early lexicon consists predominantly of concrete words. Put differently, when other relevant factors (e.g., contextual diversity) are controlled for, there is no reason to assume that the age at which a word is learned has different effects for concrete vs abstract words. The results for squared hedonic valence indicate that its influence is quite subtle, in that it manifests itself in the composition of the neighbourhoods, rather than in their size. More precisely, squared hedonic valence correlates negatively with the number of concrete neighbours, and positively with the number of concrete ones. Since abstract words are more valenced than concrete words (e.g., for the 2,328 words covered by our norms, the correlation between squared hedonic valence and concreteness is $r = -0.19$, $p < 0.001$), this finding lends additional support to our hypothesis that semantic associations are facilitated only between words which deviate from emotional neutrality (i.e., when shared affective content can come into play). Finally, the results for log contextual diversity and semantic diversity are very similar to those from our previous analysis, and we do not discuss them further here.

Having investigated some of the differences between the neighbourhoods of concrete and abstract words, as well as a few factors that might account for these differences, we now turn our attention to the interconnectivity of said neighbourhoods. The average clustering coefficient of the semantic neighbourhoods for concrete and abstract words, as a function of the minimum threshold for neighbourhood inclusion, is shown in Figure 20. Overall, the neighbours of concrete words are more clustered (i.e., they have a higher average clustering coefficient; $t(2,326) = 20.00$ , $p < .001$, two-tailed) than those of abstract words[21].

[21] We consider any two words with a strictly positive vector cosine to be neighbours (i.e., we have a single neighbourhood band), and we compute the clustering coefficient for each word, as decribed in the Methods section.

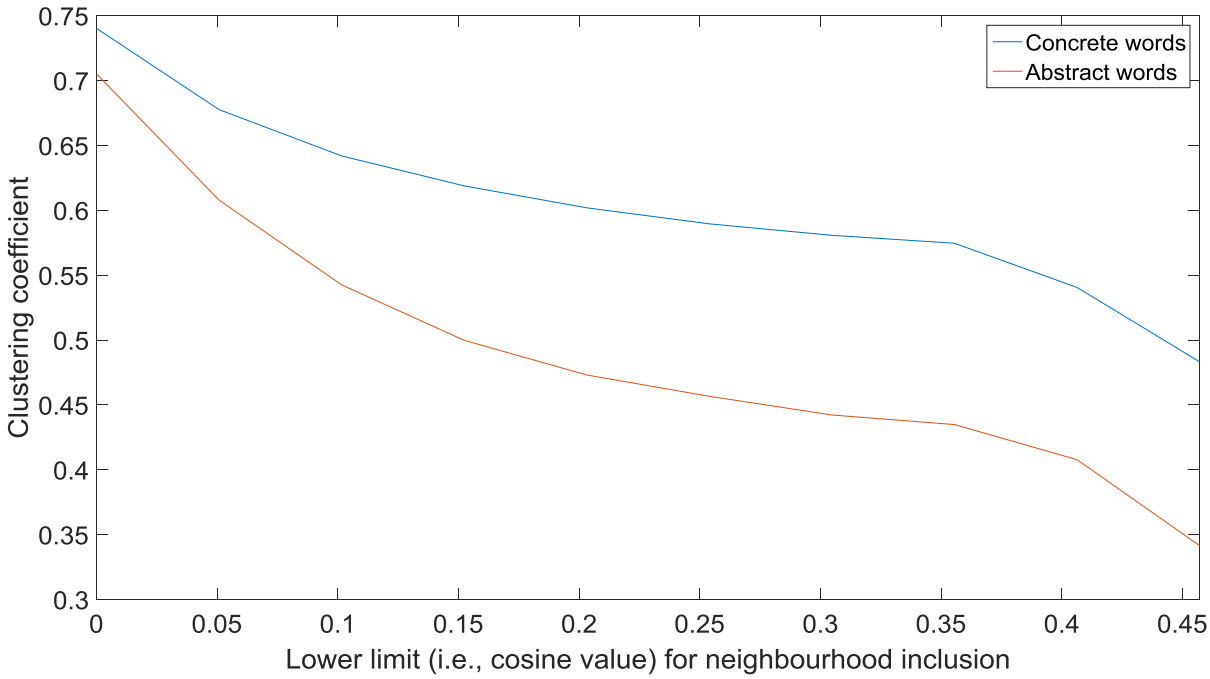


Figure 20. Average clustering coefficient, as a function of minimum semantic association strength between a word and its neighbours, measured using the vector cosine.

Like in the case of neighbourhood size, we hypothesize that the neighbourhoods of concrete words might be more interconnected due to the higher imageability, and lower polysemy, as compared to those of abstract words. To test our hypothesis, we looked at how well clustering coefficients can be predicted by imageability ratings and corpus-based operationalizations of polysemy (i.e, contextual and semantic diversity), as well as by age of acquisition and valence. As previously, we ran a regression analysis each for our concrete and abstract words, using these five independent variable, and clustering coefficient, for the various neighbourhood bands, as a dependent variable. The results are shown in Figure 21.

Imageable words have denser neighbourhoods than non-imageable ones, for both concrete and abstract words. We believe that this effect originates in the specific nature of the perceptual mechanisms that facilitate semantic associations: since physical similarity/relatedness/co-occurrence are transitive relations, up to a certain degree, this transitivity makes it likely that the vision-based neighbours of a given word are also vision-based neighbours of one another. This transitivity should be relatively weak, given that the three aforementioned perceptual relations between a word and its neighbours are likely to hold all at once, whereas only a part of them can be assumed to be characteristic of neighbours themselves (e.g., chairs are visually similar

to tables/sofas, are usually placed underneath tables or next to sofas, and are often found in rooms where tables and/or sofas are present; however, tables and sofas share few visually salient features, are not perceptually related in any obvious way, but do tend to physically co-occur in living rooms and kitchens). Therefore, we would expect that the effects of imageability should be most notable for the liberal definitions of a semantic neighbourhood, which is indeed the case.

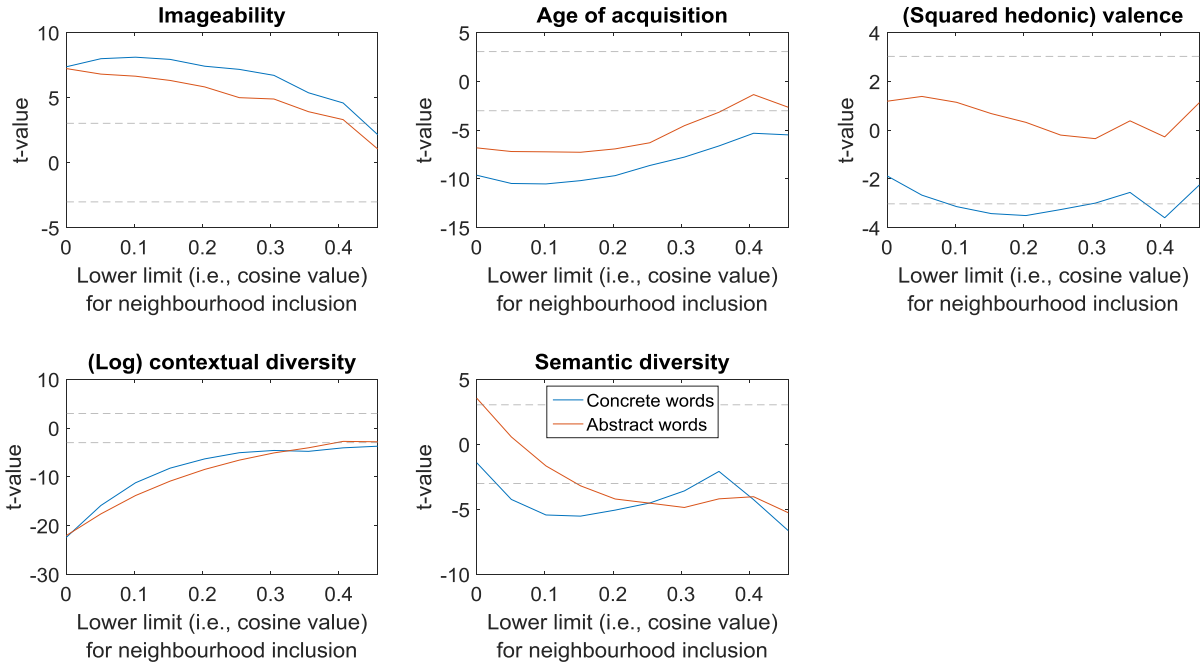


Figure 21. t-values for imageability, age of acquisition, squared hedonic valence, log contextual diversity, and semantic diversity, when entered simultaneously as linear predictors for the clustering coefficient, for both concrete and abstract words. The dashed, grey lines correspond to Bonferroni-corrected, one-tailed significance thresholds.

The same unusual relation between neighbourhood size and clustering coefficient can be observed for age of acquisition. As mentioned before, the early lexicon consists mostly of highly imageable words, which means that semantic associations between words are strengthened by the transitivity of perceptually mediated relations between word referents. Moreover, transitivity is facilitated by the reduced diversity of the physical contexts experienced by infants, and by the substantial physical similarity of any two exemplars belonging to the same category.

Squared hedonic valence has a very weak impact on neighbourhood interconnectivity, mirroring the finding for neighbourhood size. In contrast, log

contextual diversity has a strong negative effect on clustering coefficient, consistent with the idea that the semantic associations between (very) weak neighbours are negligible overall. The last of the factors, namely semantic diversity, has a relatively weak negative effect, compatible with the fact that the neighbours of polysemous words tend to organize themselves into a number of subclusters, corresponding to a different sense of the word. The elements of each subclusters are densely interconnected, whereas the subclusters themselves are weakly associated to one another, since they index partially incompatible interpretations of the same word.

Finally, we extended our previous analysis by examining the interconnectivity of concrete-concrete, concrete-abstract, abstract-concrete, and abstract-abstract word pairs, using the same five predictors. The results are displayed in Figure 22.

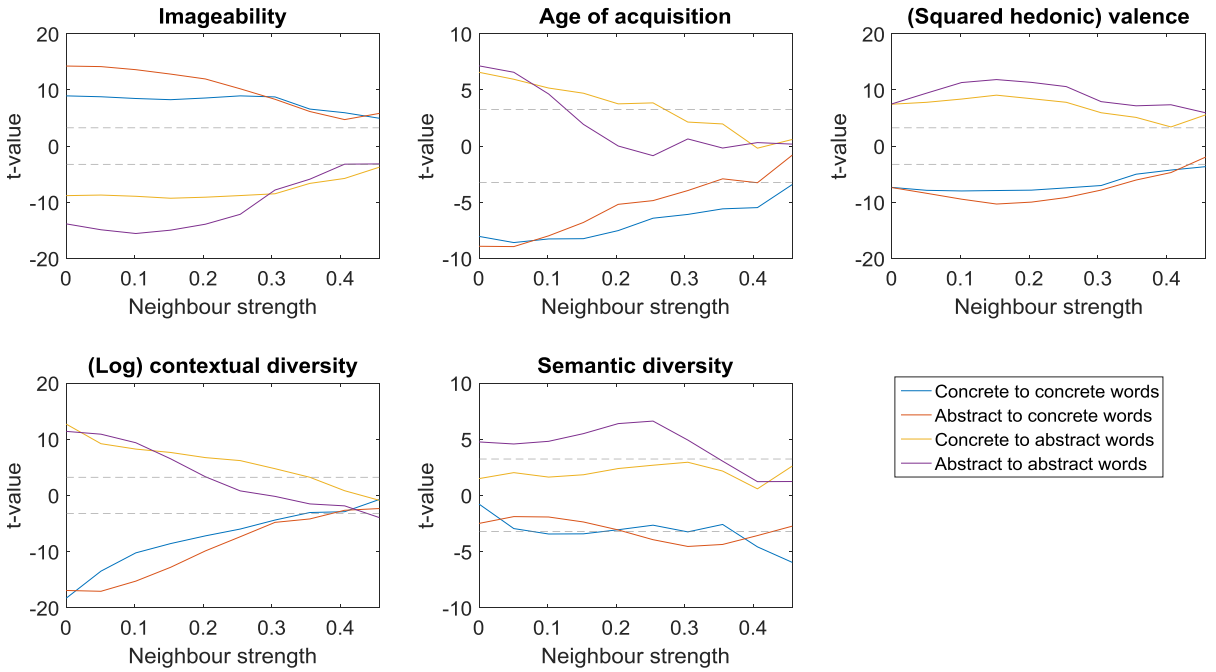


Figure 22: t-values for imageability, age of acquisition, squared hedonic valence, log contextual diversity, and semantic diversity, when entered simultaneously as linear predictors for the clustering coefficient, for both concrete and abstract words. Both the initial words and their semantic associates are divided into concrete and abstract words. The dashed, grey lines correspond to Bonferroni-corrected, one-tailed significance thresholds.

A cursory examination of the results indicates that the concreteness of the associates of a given word (rather than that of the word itself) is strongly related to the clustering coefficient corresponding to that word.

With respect to imageability, we find that concrete neighbourhoods are denser than abstract ones, which is not surprising, given that imageability and neighbourhood size are positively correlated.

For age of acquisition, it seems that early learned words favour associations to concrete neighbours, but not abstract ones, which again can be attributed to the particular structure of the early lexicon, where concrete words play a fundamental role in cognition.

Moving on to squared hedonic valence, once again, we discover that abstract neighbourhoods are more interconnected than concrete neighbourhoods. One potential reason might be the relatively low dimensionality of the affective space, when compared to the considerable complexity of the perceptual space. Reduced dimensionality implies increased network density (Karlgren, Holst, & Sahlgren, 2008), since the transitivity of semantic associations is strengthened.

In the case of contextual diversity, it seems that there is a non-trivial relation between the number of different linguistic contexts in which a word appears, and the density of that word's concrete and abstracts neighbourhoods. We speculate that one of the important factors behind this effect is related to differences in word learning strategies. More specifically, the semantics of concrete words can be reasonably constructed from perceptual and motor interactions with the physical environment, while the semantics of abstract words are very much dependent on language, rather than the actual physical objects, situations and events which language refers to. For concrete words, linguistic input is likely to be useful up to a certain point, after which it becomes a source of noise, given that it provides very little additional information. In contrast, for abstract words, it probably takes a considerable exposure to language in order to reach the same point. As a simple example, reading 100 news articles on football will not make someone a lot more knowledgeable than another person who reads only 10 articles, while the opposite is true when reading about democracy.

Finally, semantic diversity behaves very similarly to what we found in the previous analysis (see Figure 21). The only noteworthy characteristic, in our opinion, is the fact that the interconnectivity of the abstract-abstract neighbourhoods correlates

positively with semantic diversity, lending further support to the idea that the semantics of abstract words depends heavily on linguistic experience.

All in all, it seems that concrete words are indeed semantically richer than abstract ones, but this advantage is considerably more noticeable in the interconnectivity of the semantic neighbourhoods, rather than in the size of the respective neighbourhoods. Imageability seems to be the main factor that influences neighbourhood density, and the difference in clustering coefficients reflects the fact that the representation of concrete words is more strongly shaped by perceptual information, than in the case of abstract words. Another important factor is semantic diversity, which favours abstract words over concrete ones, as opposed to imageability. However, the effect of imageability is stronger than that of semantic diversity, which means that the interplay between these two factors is dominated by the former. Age of acquisition also plays a role (albeit a small one), which manifests itself almost exclusively for the concrete words belonging to the early lexicon. This role is mediated by the atypical structure of said lexicon, which consists overwhelmingly of concrete words. Valence is perhaps the most curious of the factors we tested, since it has no impact on neighbourhood size and interconnectivity, but instead dictates the strength of the semantic associations to valenced vs neutral words. Finally, contextual diversity appears to have a detrimental effect for both concrete and abstract words, since it indexes mostly “semantic noise”, when controlling for factors such as imageability, age of acquisition, valence and semantic diversity. Since the effect of all the aforementioned factors are very similar for concrete and abstract words, it seems that the contrast between the two classes of words is mostly one of quantity, rather than quality.

## 5.3. Conclusions

Our results, and that of other researchers (Recchia & Jones, 2012), seem to indicate that concrete and abstract words do not differ significantly, in terms of the number of neighbours for each word. This finding is very different from that of Hargreaves and Pexman (2014), which reveals a considerable numerical advantage

for abstract words (i.e., an average of 1,380 and 3,561 neigbours for concrete and abstract words, respectively). However, it is worth keeping in mind that the authors employ a very liberal definition of what constitutes a semantic neighbour, meaning that they lump together strong and weak neighbours, while the study by Recchia and Jones (2012) considers only close neighbours. Moreover, the model by Hargreaves and Pexman (2014) is trained over a text corpus which overestimates the amount of linguistic information to which a person is exposed in their lifetime. This translates into a psychologically unrealistic (i.e., too large) amount of semantic diversity, from which abstract words benefit more than concrete ones (see Figure 18, but also Figure 17).

Nevertheless, when richness is operationalised in the form of neighbourhood connectivity (i.e., clustering coefficient), concrete words clearly have the upper hand, most likely due to their increased imageability and lower age of acquisition. However, emotional valence and semantic diversity also play a role (see Figure 22), such that the neighbourhoods of highly valenced, abstract words are likely to be comparable to those of emotionally neutral, concrete words, in terms of interconnectivity. As discussed previously, our results seem to favour the idea that semantic richness is a rather heterogeneous, multifaceted concept, and that comparing concrete and abstract words along a single dimension of richness (e.g., neighbourhood size or interconnectivity) can yield incomplete, contradictory results.

Another conclusion that can be drawn from our study is that the close neighbourhood of a word is a relatively poor source of information. For instance, it offers only a partial picture of the difference between concrete and abstract words, suggesting that former are richer than the latter. However, in the case of (very) distant neighbours, this tendency disappears or is reversed (see Figure 17; also see Hargreaves & Pexman, 2014). Moreover, the difference in neighbourhood interconnectivity is greatly attenuated when using a broad, inclusive definition of semantic neighbours (see Figure 20). All in all, our findings seem to indicate that it might be profitable to look beyond neighbourhood size when analysing semantic richness, and to consider neighbourhoods at various semantic distances, rather than focusing only on very close neighbours.

# 6. Simulating semantic impairments in Developmental Language Disorder[22]

## 6.1. Introduction

Computational models of semantics, predominantly artificial neural networks (i.e., connectionist models), have been used in order to account for a variety of disorders with a semantic component, such as deep dyslexia (Hinton & Shallice, 1991), Alzheimer's disease (Devlin, Gonnerman, Andersen, & Seidenberg, 1998), semantic dementia (Rogers et al., 2004), schizophrenia (Cohen & Servan-Schreiber, 1992), and other disorders (Farah & McClelland, 1991).

Another prominent candidate for modelling is Developmental Language Disorder (DLD; previously known as Specific Language Impairment, or SLI; Leonard, 2014). DLD is an impairment in language production and comprehension, which cannot be attributed to hearing deficits (e.g., otitis media with effusion), anomalies of the oral-motor system, low nonverbal intelligence, poor interaction with people and objects (e.g., as is the case with autism spectrum disorder), or neurological damage. Some of the most noticeable and well-studied symptoms of DLD refer to the use of grammar, and consist of frequent omission of function words (e.g., "Why you need key for?"), and grammatical inflections (e.g., "Mimi help me blow out candles."), inappropriate use of past-tense (e.g., "Drawed picture.") and pronoun forms (e.g., "Him lost it."), as well as difficulties in comprehending and repeating syntactically complex sentences (e.g. "The dogs that are running are at the beach.")[23].

A better understanding of the mechanisms behind DLD, such as that obtained via computational modelling, would have important clinical implications, given the prevalence of the disorder (i.e., it affects more than 7% of the population; Tomblin et al., 1997), its persistence into adult age (Law, Rush, Schoon, & Parsons, 2009;

[22] Adapted from (Ponari, Norbury, Rotaru, Lenci, & Vigliocco, 2018).
[23] The examples are taken from Chapter 1 of (Leonard, 2014), with the exception of the last example, which is taken from (Leonard, Deevy, Fey, & Bredin-Oja, 2013).

Mawhood, Howlin, & Rutter, 2000), as well as its detrimental effects on social integration and emotion regulation (St. Clair, Pickles, Durkin, & Conti-Ramsden, 2011). However, although the grammatical aspects of DLD have been the subject of computational modelling studies (e.g., Hoeffner & McClelland, 1993; Joanisse & Seidenberg, 2003), the semantic characteristics have received little attention, most likely due to the fact that they are subtler than the phonological and morpho-syntactic symptoms. It is known that children with DLD have a vocabulary that is reduced in both breadth (i.e., how many words are known) and depth (i.e., how well words are known), as compared to typically developing children (e.g., McGregor, Oleson, Bahnsen, & Duff, 2013). This disadvantage can be seen in a variety of tests for vocabulary breadth (e.g., word definition, lexical decision, picture naming) and depth (e.g., word definition, selecting synonyms).

In this chapter, we employ distributional models, trained on psychologically realistic corpora, in terms of size, in order to examine some of the factors that might contribute to the semantic deficits associated with DLD. We achieve this goal by mapping model parameters to psychological factors (e.g., the size of the sliding window can be put in correspondence with working memory capacity), and then investigating the effects of changing the values of the parameters, away from their optimal values. We then compare our findings with experimental results from a number of meta-analyses and other studies.

We also test whether the magnitude of our simulated impairments depends on word concreteness. According to the Dual Coding theory of semantics (e.g., Paivio, 1971, 1986), the meaning of abstract words is learned predominantly from linguistic information, whereas that of concrete words is derived from both linguistic and perceptual experience. Since DLD is associated with linguistic, but not perceptual deficits, it seems likely that children with DLD should have more problems processing abstract words, as opposed to concrete ones. However, Ponari, Norbury, Rotaru, Lenci, and Vigliocco (2018) found that children with DLD do not show a greater impairment for abstract than for concrete words, when tested using verbal definition and lexical decision. A potential explanation might be that the learning of abstract words is also strongly supported also by non-linguistic factors, such as emotion (Ponari, Norbury, & Vigliocco, 2018).

The rest of the chapter is structured as follows. We begin by summarizing a number of findings with respect to differences in task performance between children with DLD and typically-developing children, in terms of semantic cognition. We then present several meta-analyses and other relevant studies, concerning potential sources of impairment for DLD, as well as their effects on task performance, in both semantic and non-sematic tasks. Next, we simulate semantic impairments in DLD by "damaging" two distributional models (i.e., Skip-gram and CBOW), and investigate the resulting "impaired" model by using Representational Similarity Analysis and linear mixed-effects models. Finally, we describe the results of a pilot experiment, ran in order to test the behavioural validity of certain results from the modelling study, and derive a number of conclusions from our computational and experimental work.

## 6.2. Semantic impairments in DLD

Relatively few studies have examined the qualitative and quantitative nature of the associative structure of semantic memory, in the case of children suffering from DLD. An early study (Kail & Leonard, 1986) looked at similarity ratings produced by language-impaired children, using names of animals and occupations. The study found that the performance for both children with DLD and age control children can be largely explained using the same two semantic dimensions per category (i.e., size and "predativity", for animals; production of goods vs services and "excitingness", for occupations), and that there are no significant difference between the groups in terms of the weights associated with each dimension (i.e., the relative importance of each dimension).

More recent studies (e.g., Brooks, Maouene, Sailor, & Seiger-Gardner, 2017; McGregor et al., 2012; Sheng & McGregor, 2010) employed continued and discrete versions of the free association task. The studies found that, in contrast to children in the age and/or linguistic ability control groups, children in the DLD group produced fewer correct answers (i.e., associations semantically related to the cue), and more wrong answers (i.e., no responses, repetitions of the cue, inflections of the cue, and associations unrelated to the cue). Also, the first study revealed that, in comparison to

the age control group, children in the DLD group provided less diverse associates, and the semantic networks derived from those cue-associate pairs exhibited less separation between semantic clusters.

Finally, a number of studies used the semantic fluency task (e.g., with categories such as animals, body parts, foods, occupations, furniture, etc.). One study (Kail & Leonard, 1986) found almost no differences between the DLD group and the age control group, in terms of temporal structure of retrieval, prototypically of responses, cluster size and number of clusters. Two other studies (Henry, Messer, & Nash, 2015; Weckerly, Wulfeck, & Reilly, 2001) revealed that children in the DLD group produced fewer correct responses than typically-developing children in the language-matched group. Cluster sizes were equal between the groups, but children in the DLD group generated fewer clusters. In addition, the first study showed that vocabulary size and working memory capacity correlated with percentage of correct responses, while inhibition correlated with percentage of errors.

## 6.3. Potential sources for the semantic impairments

A number of potential explanations for the causes behind DLD have been put forward, some of which aim to offer a global explanation for the deficits associated with DLD, while others focus on accounting for more specific areas of pronounced impairment. In the current study, we examine three potential causes for certain difficulties encountered by children with DLD, related to their use of semantic processes and representations, namely working memory, statistical learning, and attention.

Working memory (see Baddeley, 2003, for a review) refers to the ability to briefly store and subsequently manipulate limited amounts of relevant information, in order to support the cognitive tasks being performed at a given moment. The storage aspect of working memory is typically subdivided into three distinct components, namely the phonological loop and the visuospatial sketchpad, which hold verbal and visual/spatial representations, respectively, and the episodic buffer, which integrates long-term knowledge with information from the other two components. Several studies

found that working memory capacity appears to be reduced in children with DLD (for a recent review, see Henry & Botting, 2017). Verbal short-term memory has received perhaps the most attention, and has been investigated in a variety of tasks, employing either nonwords (e.g., nonword repetition) or words (e.g., listening span). A meta-analysis by Graf Estes, Evans, and Else-Quest (2007), covering 23 studies using the nonword repetition task, revealed that children diagnosed with DLD performed considerably poorer than typically developing children (i.e., the mean effect size was 1.27). With respect to visuospatial short-term memory, a meta-analysis by Vugs, Cuperus, Hendriks, and Verhoeven (2013), involving 18 studies, indicated a moderate deficit for children with DLD (i.e., a mean effect size of 0.49), but the pattern of results across studies was significantly less consistent than in the case of verbal short-term memory. Although several studies confirm the association between DLD and working memory capacity, there is mixed evidence that individual differences in working memory are predictive of language abilities. Of the studies that examined this issue, some found a significant correlation (e.g., for the verbal domain, see Ellis Weismer & Thordardottir, 2002; Leonard et al., 2007; for the visuospatial domain, see Kleemans, Segers, & Verhoeven, 2011), while others did not (e.g., for the verbal domain, see Briscoe, Bishop, & Norbury, 2001; Ellis Weismer, Evans, & Hesketh, 1999; for the visuospatial domain, see Lum, Conti-Ramsden, Page, & Ullman, 2012).

Statistical learning (see Romberg & Saffran, 2010, for a review) is a form of implicit learning, which involves discovering and extracting statistical regularities that characterize a wide range of sensory inputs, such as letters, phonemes, words, shapes, and faces, to name just a few. The patterns acquired through statistical learning can vary along multiple dimensions, such as complexity (e.g., frequency counts vs conditional probabilities), sensory modality (e.g., visual vs auditory), and domain (e.g., spatial vs temporal). A potential connection between DLD and deficits in statistical learning is made explicit by the procedural deficit hypothesis (Ullman, 2004; Ullman & Pierpont, 2005). According to this hypothesis, linguistic processes and representations are supported by two distinct types of memory, namely declarative and procedural memory, where the former stores semantic information about words, while the latter encodes the rules that make up the mental grammar. Children suffering from DLD are assumed to have impairments in procedural learning, which affect procedural memory, but not declarative memory. This assumption is in line with the

finding that children with DLD experience significant difficulties in mastering the morphological and syntactic aspects of language, which depend crucially on the learning of rules, whereas their knowledge of semantics is relatively comparable to that of typically developing children. Moreover, in a recent meta-analysis of 14 published studies, Obeid, Brooks, Powers, Gillespie-Lynch, and Lum (2016) found that language-impaired children perform more poorly in statistical learning tasks than typically developing children (i.e., the effect size was 0.46), and that this disadvantage is visible in a number of tasks (i.e., serial reaction time, speech stream segmentation, artificial grammar learning, and probabilistic classification learning), spanning the visual, auditory, and motor modalities. Like in the case of working memory, statistical learning and language abilities show mixed patterns of correlation, at the individual level. For instance, one study using the speech stream segmentation task (Evans, Saffran, & Robe-Torres, 2009) found a significant correlation between vocabulary size and statistical learning performance, in both language-impaired and typically developing children, while another study (Haebig, Saffran, & Ellis Weismer, 2017), following a very similar methodology, did not reveal any such relationship. Contrasting evidence is also provided by studies employing the serial reaction time task (e.g., Gabriel, Maillart, Guillaume, Stefaniak, & Meulemans, 2011; Gabriel et al., 2013). Furthermore, not all aspects of grammatical skill seem to depend on statistical learning to the same degree: in a series of studies on typically developing children, Kidd and collaborators showed that statistical learning abilities are associated with levels of performance in syntactic tasks (Kidd, 2012; Kidd & Arciuli, 2016), but not in morphological tasks (Kidd & Kirjavainen, 2011; Lum & Kidd, 2012).

Attention (see Nobre, K., Nobre, A. C, & Kastner, 2014, for a review) can be defined as the selective, enhanced processing of certain elements of exogenous and/or endogenous information. The majority of studies focusing of potential impairments of attention in DLD describe attention in terms of three or more components (Mirsky, Anthony, Duncan, Ahearn, & Kellam, 1991), including (1) focus (i.e., the selection of information), (2) sustain (i.e., the maintenance of the focus over time), shift (i.e., the redirection of focus towards other information). While there is mixed evidence with respect to the status of attentional focus (e.g., Stevens, Fanning, Coch, Sanders, & Neville, 2008; Stevens, Sanders, and Neville; 2006) and attentional shift (Lum, Conti-Ramsden, & Lindell, 2007; Schul, Stiles, Wulfeck, & Townsend,

2004), it appears that sustained attention is indeed suboptimal in children with DLD. A meta-analysis by Ebert and Kohnert (2011), covering 17 studies, found that language-impaired children are less able to maintain attention during continuous performance tasks than typically developing children (i.e., a mean effect size of 0.69), and that the difference between the two groups is more visible in tasks that employ auditory (verbal or nonverbal) stimuli, as opposed to visual stimuli. Moreover, performance in sustained attention task was been shown to partially account for individual differences in language abilities (e.g., Finneran, Francis, & Leonard, 2009; Montgomery, 2008; Montgomery, Evans, & Gillam, 2009; Spaulding, Plante, & Vance, 2008).

A number of studies also looked at whether certain subclasses of stimuli (e.g., infrequent vs frequent words) are processed differently in DLD, but with contradictory results: one study (Beckage et al., 2011) revealed that children with abnormally small vocabularies (many of which are later diagnosed with DLD) have semantic networks that are considerably less interconnected than those of typically developing children, compatible with the idea that language-impaired children allocate more attentional resources to “oddball”, unusual words (which are likely to be low in frequency), whereas another study (Jimenez & Hills, 2017), following a very similar methodology, reached the opposite conclusion.

## 6.4. Simulating impairments using computational models

As a complement to experimental studies, we believe that a better insight into DLD and its causes can also be obtained by using distributional models. These models are especially interesting because they provide an implicit mechanism for learning a rich vocabulary beyond children’s direct experiences with objects, actions, emotions, mental states, and so on. Inspired by connectionist studies of semantic deficits (see Chapter 4), we begin by assigning a psychological interpretation for the various parameters of two state-of-the-art distributional models, namely Skip-gram and CBOW. Then, by moving parameters away from their optimal values, we create virtual lesions within the models, and assess their impact on performance, for both concrete and abstract words. This allows us to determine the importance of each parameter,

and, implicitly, that of psychological factors associated with the parameters, which have been described in the clinical literature.

Particularly relevant for our purposes are "predict" models, such as the CBOW model (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), which use neural networks to derive representations that best predict the linguistic contexts (i.e., co-occurring words) in which words appear. CBOW was developed within computational linguistics, with a focus on improving performance in natural language processing tasks, rather than providing a psychologically informed model of semantic cognition. However, recent research argues in favour of the model's cognitive plausibility, by showing that the learning mechanism in CBOW is related to that of the classical Rescorla-Wagner model of associative learning (Mandera et al., 2017), which has been validated in a large number of animal and human studies. In addition, CBOW's learning objective is consistent with Anderson's rational analysis of memory framework (Hollis, 2017), an influential normative account of human memory. CBOW also shows consistently better fit to adult behavioural data than competitor models in a variety of semantic tasks (see Chapter 2). Therefore, we employ CBOW for the analyses presented in this subchapter.

We derive our semantic representations by training the model on a subset of the combined TASA (Zeno et al., 1995) and CBBC (Van Heuven et al., 2014) corpora. The TASA corpus is a collection of short texts extracted from textbooks, works of literature, and popular works of fiction and nonfiction, designed to offer a "quantitative summary of the printed vocabulary encountered by students in American schools". The CBBC corpus consists of subtitles for a variety of TV shows aired on the Children's BBC, which aims "to provide a wide range of high quality, distinctive content for 6-12 year olds, including drama, entertainment, comedy, animation, news and factual"[24]. In order to make our computational experiments more comparable with the behavioural experiments described in the first part of the current study, we filter the two corpora and keep only texts that can be relatively easily understood by children aged 11, based on the Degree of Reading Power (i.e., an index of text difficulty), for the TASA corpus, and the subjective judgement of the primary supervisor and her son, for the CBBC

[24] See http://www.bbc.co.uk/bbctrust/our_work/services/television/service_licences/cbbc.html for more details.

corpus. This results in a combined corpus of slightly over 12 million words. As a means of increasing the quality of the representations, we first pre-process the corpus by converting all the words to lowercase, eliminating punctuation marks and removing words whose absolute frequencies are less than five. We then construct 300-dimensional vector representations for the words in our combined corpus. We consider only words for which concreteness ratings are available (Brysbaert et al., 2014).

Our approach to simulating the effects of damage to the semantic system involves manipulating three model parameters that affect the quality of the linguistic representations, namely (1) the size of the sliding window over which word co-occurrences are considered, (2) the learning rate for the algorithm used in training the model, and (3) the degree of subsampling for the occurrence of frequent words. Each of these factors can be given a clear psychological interpretation: the size of the sliding window is linked to the capacity of verbal working memory; the learning rate is related to the efficiency and precision of the statistical learning mechanisms involved in the acquisition of word meanings from exposure to language; finally, the subsampling of frequent words, which we refer to as "novelty bias", is associated with the amount of attentional resources allocated to processing of words which are encountered relatively often. In order to damage our linguistic model, we follow a 3 (window size) x 5 (learning rate) x 5 (degree of subsampling) factorial design, and create 75 versions of the initial model. The "healthy" version of the model uses parameter values that have been shown to provide the best fit to behavioural data and/or are recommended by the authors of the models. The values for the window size are 5, 3 and 1, such that 5 corresponds to the "healthy" case. The values for the learning rate and the novelty bias are set to 180, 140, 100, 60 and 20 percent of the value for the "healthy" case (i.e., 0.025 for the learning rate, and 0.001 for the degree of subsampling).

Given the sparsity of data easily amenable to distributional modelling, we estimate the performance of each of the "damaged" models, relative to that of the "healthy" one, by means of Representational Similarity Analysis (RSA; Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008), a technique typically employed in cognitive neuroscience. RSA allows researchers to compute the similarity between patterns of brain activation (e.g., fMRI, MEG, EEG, single-cell or electrode-array recordings), and corresponding representations generated by computational models, primarily in order to assess the cognitive plausibility of the models, as well as how the

models relate to each other. Importantly, this technique can also be used to compare representations across different regions in the same brain, the same region in different participants or even different species, different types of brain-imaging data, and different computational models, among other applications. RSA avoids the problem of directly testing the correspondence between representations, but instead focuses on evaluating the similarity of the relations that hold between representations (i.e., it tries to compare representational geometries).

As an example of how RSA works, let us assume that that we have a set of stimuli $S$, with $n$ elements (i.e., $S(i)$, for $1 \leq i \leq n$), and two sets of representations for those stimuli, namely $R_1$ (i.e., $R_1(i)$, for $1 \leq i \leq n$) and $R_2$ (i.e., $R_2(i)$, for $1 \leq i \leq n$). We are interested in testing whether $R_1$ and $R_2$ encode the same information about S. The traditional method would be to train a model $M$[25], in order to predict the representations in $R_2$, based on the representations in $R_1$. Given a properly chosen similarity measure (e.g., correlation), namely *sim,* we would then compute and aggregate the similarities $sim(M(R_1(i)),R_2(i))$, for $1 \leq i \leq n$, typically by averaging or summing them. If the aggregated similarity is high, then $R_1$ and $R_2$ capture roughly the same information about $S$. Otherwise, $R_1$ and $R_2$ reflect different aspects of $S$.

Unlike this traditional approach, in RSA we would select two distance measure, namely $dist_1$ and $dist_2$, and then build two distance matrices, namely $D_1$ and $D_2$, such that $D_1(i,j) = dist_1(R_1(i),R_1(j))$ and $D_2(i,j) = dist_2(R_2(i),R_2(j))$, for $1 \leq i \leq n$, $1 \leq j \leq n$. Finally, we would compute the Spearman correlation between the matching values in $D_1$ and $D_2$, and interpret the resulting correlation value in the same way as we did for the aggregated similarity value described previously. Thus, the main difference between the traditional approach and the RSA approach is that, in the former, we are evaluating the match between the individual representations in $R_1$ and $R_2$, whereas in the latter, we are evaluating the match between the relationships of the individual representations $R_1$ and $R_2$ with the rest of the representations in $R_1$ and $R_2$, respectively. Put differently, in the traditional approach it is the individual representations that matter most, whereas in the RSA approach it is the relationships

[25] For instance, when mapping distributional representations to fMRI activation patterns, *M* is usually a set of multiple regression models, such that the activation for each voxel is predicted as a (different) weighted sum of the values in the corresponding vector representations generated by the distributional model (e.g., Mitchell et al., 2008).

between the individual representations that are critical. Some of the advantages of RSA are that it is very simple to understand and apply, it does not require to select and train a model *M*, linking $R_1$ and $R_2$, and it uses a richer set of information (i.e., the pairwise relationships between the representations in $R_1$ and $R_2$, repectively).

In our simulations, we employ the following procedure. For each version *Ver* of the CBOW model, we compute two matrices (one for concrete and one for abstract words) of word dissimilarities, such that the dissimilarity of any two vectors is computed as one minus the vector cosine:

$$dist(W_1,W_2) = 1 - cos(vec_{Ver}(W_1), vec_{Ver}(W_2))$$

where $vec_{Ver}(W_1)$ and $vec_{Ver}(W_2)$ correspond to the vectors associated with the words $W_1$ and $W_2$, respectively. These two matrices, namely $D_{CONC}$ (*Ver*) and $D_{ABS}$ (*Ver*), contain the dissimilarities for all the concrete-concrete and abstract-abstract word pairs, where words were classified as being either concrete or abstract, based on a median split on the ratings. In order to assess the variability of the Spearman correlations between the dissimilarity matrices, as explained previously, we use the method suggested in (Kriegeskorte et al., 2008) and recalculate the correlation for 100 bootstrap resamplings of the words in each dissimilarity matrix. The results of analysis are shown in Figure 23.

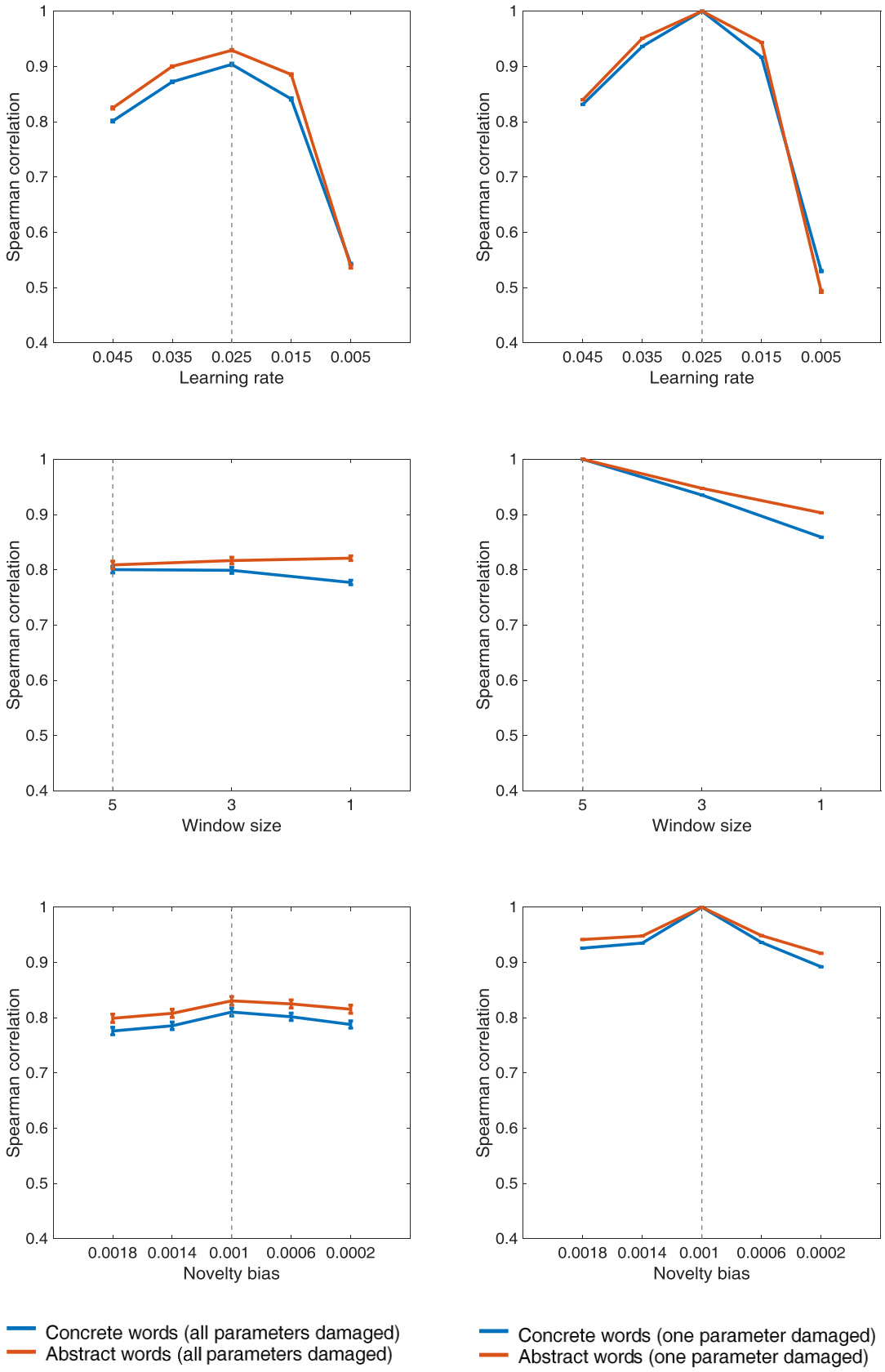


Figure 23. Results of the representational similarity analysis for concrete and abstract words. Error bars indicate 95% confidence intervals. The dashed line corresponds to the reference (“healthy”) value for each parameter. Left: correlations for the set of simulations involving all the valid combinations of values for the three parameters. Right: correlations for the set of simulations where only one parameter was

allowed to vary, while the other two were set to their reference values.

Based on visual inspection, there appears to be no advantage for concrete words, as compared to abstract ones. In fact, the graphs seem to reveal a small effect in the opposite direction. Also, it seems that learning rate has by far the largest impact on model performance, followed by window size and novelty bias. In order to formally test the effects of damaging the "healthy" model, we employed a linear mixed-effects model. We first considered a set of very concrete words (i.e., the top 1/3 most concrete words in our model, namely 5,111 words), and a set of very abstract words (i.e., the top 1/3 most abstract words in our model, namely 5,108 words). From each of these two sets we selected the top 10,000 word pairs, in terms of cosine similarity between the words in the pairs. For each word pair *WP* and each version *V* of the CBOW model, our dependent variable consisted of the difference *cosDiff*(*WP*, *V*) = *cos*(*WP*)|*RM* - *cos*(*WP*)|*V*, where *RM* is our reference model, corresponding to typically developing children. This difference represents an estimation of the amount of damage inflicted upon the strength of the semantic association between the words in *WP*, when comparing the "healthy" model (i.e., *cos*(*WP*)|*RM*), to one of its "damaged" versions (i.e., *cos*(*WP*)|*V*). The fixed effects consisted of window size (i.e., *winSize*(*V*)), learning rate (i.e., *learnRate*(*V*)), novelty bias (i.e., *novelBias*(*V*)), concreteness class (i.e., *concClass*(*WP*), equal to 1 for pairs of concrete words, and to 0 for pairs of abstract words), and the interaction between concreteness class and the other three fixed effects. The random effect, limited only to intercepts, consisted of word pair (i.e, *WP*). The formula for our main linear mixed-effects model is the following:

*cosDiff* ~ *winSize* + *learnRate* + *novelBias* + *concClass* + *winSize* * *concClass* + *learnRate* * *concClass* + *novelBias* * *concClass* + (1|*WP*)

In order to quantify the magnitude of the effects, we compare the previously described model to four simpler models, which do not include (1) concreteness class and its interactions with the other predictors, (2) window size and its interaction with concreteness, (3) learning rate and its interaction with concreteness, and (4) novelty bias and its interaction with concreteness. These models are the following:

1. *cosDiff ~ winSize + learnRate + novelBias* + (1|*WP*)
2. *cosDiff ~ learnRate + novelBias + concClass + learnRate * concClass + novelBias * concClass* + (1|*WP*)
3. *cosDiff ~ winSize + novelBias + concClass + winSize * concClass + novelBias * concClass* + (1|*WP*)
4. *cosDiff ~ winSize + learnRate + concClass + winSize * concClass + learnRate * concClass* + (1|*WP*)

The intercepts varied significantly across words pairs, *SD* = 0.025 (95% CI = [0.0246, 0.0254]). There was a significant main effect of window size, *F* = 79,557.7, *p* < .0001, learning rate, *F* = 496,770.3, *p* < .0001, novelty bias, *F* = 39,765.7, *p* < .0001, and concreteness class, *F* = 7,832.5, *p* < .0001. Furthermore, there were significant interactions between concreteness class and window size, *F* = 4,425.4, *p* < .0001, learning rate, *F* = 1,057.3, *p* < .0001, and novelty bias, *F* = 536.9, *p* < .0001. These results are very much in agreement with those revealed by the representational similarity analysis. *F* values are very large, owing to the considerable number of items entered in the model (i.e., 2 * 10,000 * 3 * 5 * 5 = 1,500,000 word pairs), it is difficult to quantify the size of the reported effects, which is why we consider it more informative to examine the predictive power of the "full" model and of its "simpler" versions. By looking at the “full” model ($R^2$ = .7589), and the models where we remove concreteness class and its interactions ($R^2$ = .7536), window size ($R^2$ = .7216), learning rate ($R^2$ = .1336), and novelty bias ($R^2$ = .6991), respectively, we find that the results are comparable with those suggested by the representational similarity analysis, especially with regards to the very limited role played by concreteness class.

**Additional analysis:**

In order to test the generality of the results obtained from the analysis presented above, as well as to further examine the factors that can influence the learning process, we run an additional analysis, with a number of important differences. Firstly, given that RSA provides a very indirect measure of model performance, we decided to use similarity ratings for the analyses described in this subchapter. Given that currently there is very little data available on semantic task behavior for children with DLD, as described in the introduction to this chapter, we employ ratings collected from

adults, as found in the SimLex-999 and SimVerb-3500 datasets. In contrast to other popular norms, such as MEN and SL, the two datasets we chose cover a broader range of word concreteness, and model performance for distributional models is far below ceiling, allowing for a better evaluation of the models.

Secondly, we replace the CBOW model with the Skip-gram model, since we wish to determine whether our findings are strongly dependent on the use of CBOW as our model of choice.

Thirdly, we employ the BNC corpus, instead of parts of the TASA and CBBC corpora. This change is motivated by findings from a study by Brysbaert, Stevens, Mandera, and Keuleers (2016), which presents various estimates for the average number of words encountered every year by a typical person. The estimates depend on the source of the linguistic input, namely social interactions (11.7 million words per year), watching TV programmes (27.3 million words per year), and reading (105 million words per year). This means that by the age of 11, a typical child encounters at least 11 * 11.17 = 128.7 million words, whereas the filtered TASA and CBBC corpora contain only slightly over 12 million words. In contrast, the written part of the BNC corpus consists of approximately 87 million words.

Finally, we investigate the effects of additional model parameters related to the statistical learning mechanism of our linguistic model, by including the number of negative samples and the amount of noise in the input to the output layer, as new parameters. Furthermore, we also include the probability of predicting a wrong word during learning: for instance, in a sentence like “the dog barks”, the “healthy” Skip-gram model would always attempt to correctly predict the word “barks” upon processing the word “dog”, whereas a “damaged” Skip-gram model would occasionally wrongly predict a different word (e.g, “jumps”), instead of “barks”. From a psychological point of view, the first factor is a measure of inhibitory, attentional control, whereas the second parameter corresponds to the efficiency and precision of the statistical learning mechanism for word learning. Finally, the third parameter can be associated with the overall level of focused attention, where poor attention can result in the retrieval of an incorrect word from semantic memory.

Given the large number of model parameters, it would be impractical for us to follow a factorial design in damaging the linguistic model, like in our previous analyses. Instead, we start from a “healthy” model (i.e., with values of 5 for window size, 0.025

for learning rate, 0.001 for novelty bias, 5 for number of negative samples, 0 for noise added to the input of the output layer, and 0 for probability of wrong prediction), and then modify only one parameter at a time, while keeping all the other parameters at the reference value.

The values for the window size are 5, 3 and 1, such that 5 corresponds to the "healthy" case. The values for the learning rate, novelty bias, and number of neagitve samples are set to 20, 60, 100, 140 and 180 percent of the value for the "healthy" case. For both the amount of noise introduced in the output layer and the probability of making a wrong prediction, the reference value is 0, which means that it makes no sense to try various proportions of that value. Instead, we first estimated that a reasonable amount of noise for the output layer might be the variance of the inputs received by the output layer during a typical run of the "healthy" model, trained on the BNC. Then, as values for the noise, we decided to employ 0, 50, 100, 150 and 200 percent of the resulting estimate. With respect to the probability of a wrong prediction, when processing each word, we decided to use values of 0, 20, 40, 60, and 80 percent[26]. The results are shown in Figure 24, Figure 25, Figure 26, Figure 27, Figure 28, Figure 29, and Table 14.

We found no significant differences in performance for any two factors, or between model performance with respect to concrete and abstract words, for any of the factors. However, given the very small sample size, the absence of significant effects is perhaps not surprising. Based on visual inspection, instead, it does seem that window size stands out from the other factors, since it appears that it more strongly affects abstract words, in comparison to concrete words. In order to test whether the size of the context in which words are encountered indeed has an effect on task performance, we decided to run a pilot study. We used a continued free association task, where we asked participant to record the first three words that come to mind after reading a word stimulus, presented in a linguistic context. Crucially, we manipulated the length of the context, which can be seen as the equivalent to window size, as well

[26] We corrupted the output layer by adding an independent and identically distributed error term to each of its inputs, sampled from a normal distribution N(0, $v * sc$), where $v$ is the variance estimated from the "healthy model", as described in the text, and $sc$ is a scaling factor (i.e., 0, 0.5, 1, 1.5, or 2). We impaired the prediction mechanism by taking each training word pair ($W_1$, $W_2$), where the model learns to predict $W_2$ after encountering $W_1$, and replacing $W_2$ with a different, incorrect word $W_3$, with a probability of $pr$ (i.e., 0, 0.2, 0.4, 0.6, or 0.8).

as the concreteness of the stimuli. The experiment and its results are decribed in the next subchapter.

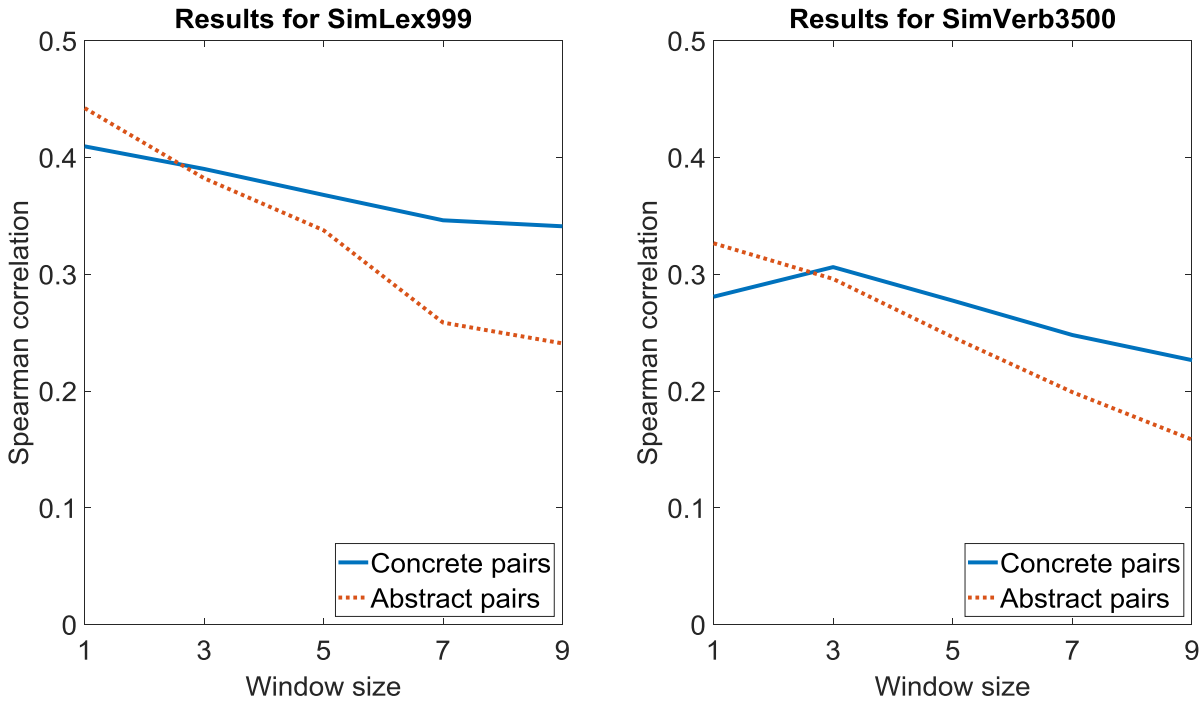


Figure 24. Model performance in predicting similarity/relatedness ratings from the SimLex-999 and SimVerb-3500 datasets, as a function of window size.

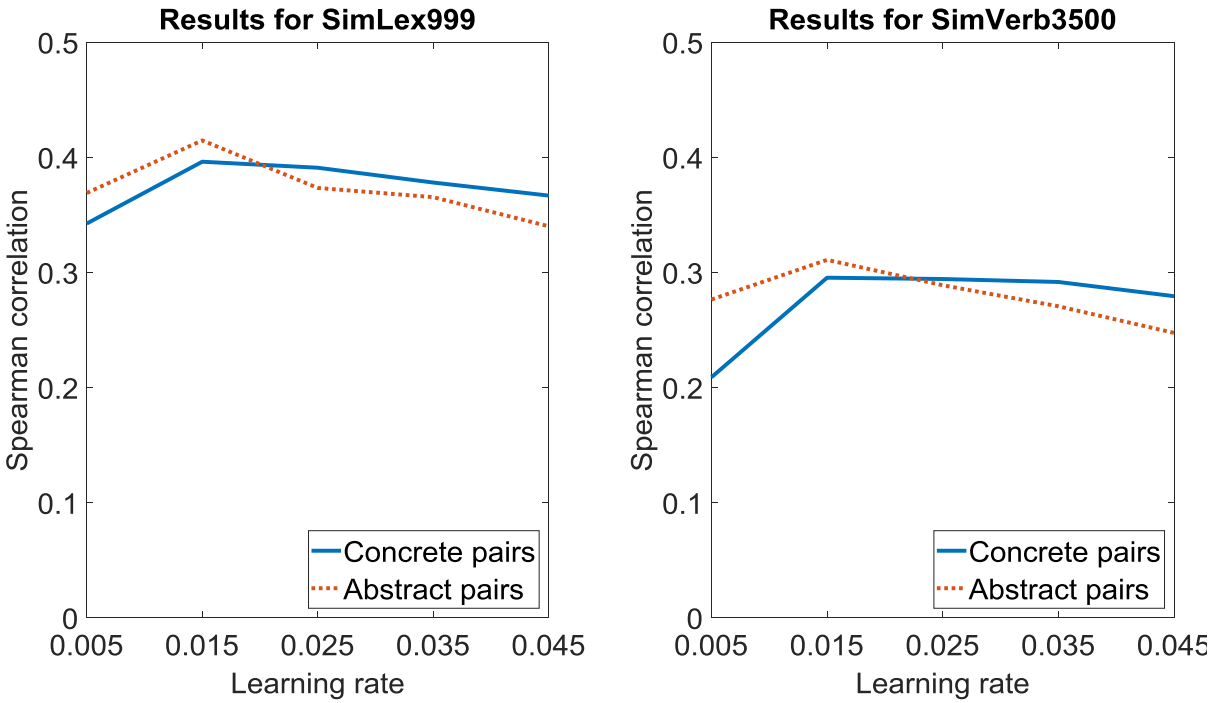


Figure 25. Model performance in predicting similarity/relatedness ratings from the SimLex-999 and SimVerb-3500 datasets, as a function of learning rate.

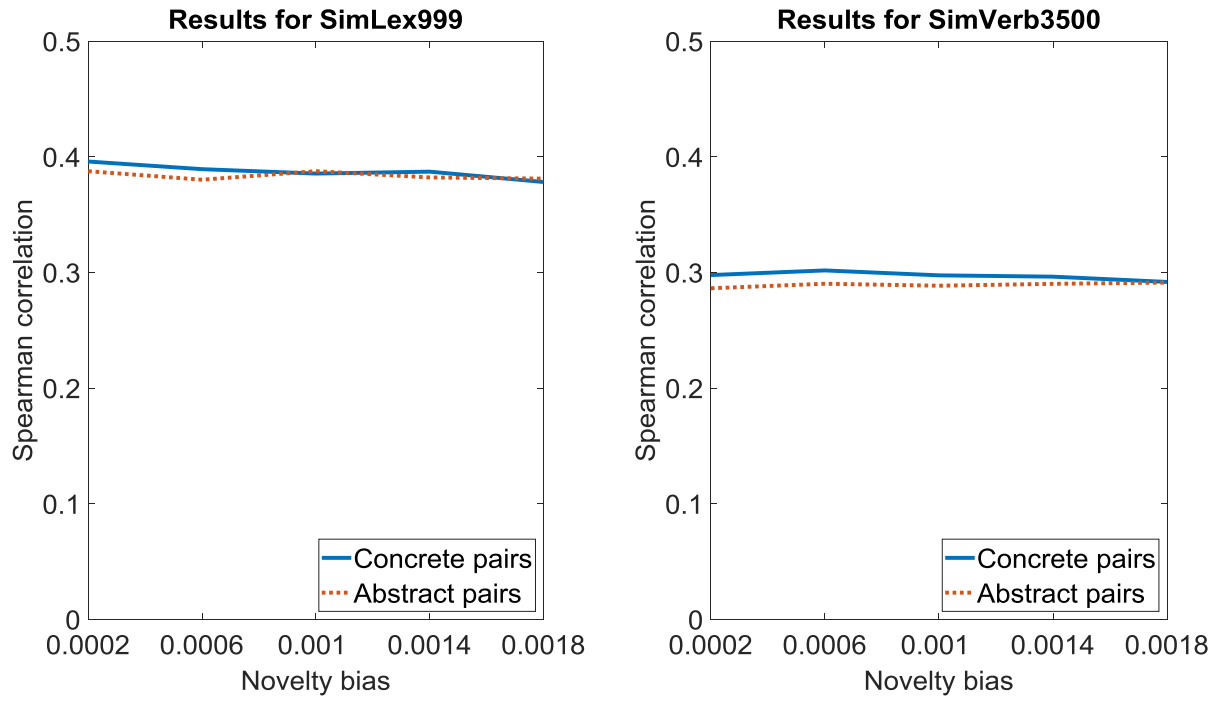


Figure 26. Model performance in predicting similarity/relatedness ratings from the SimLex-999 and SimVerb-3500 datasets, as a function of novelty bias.

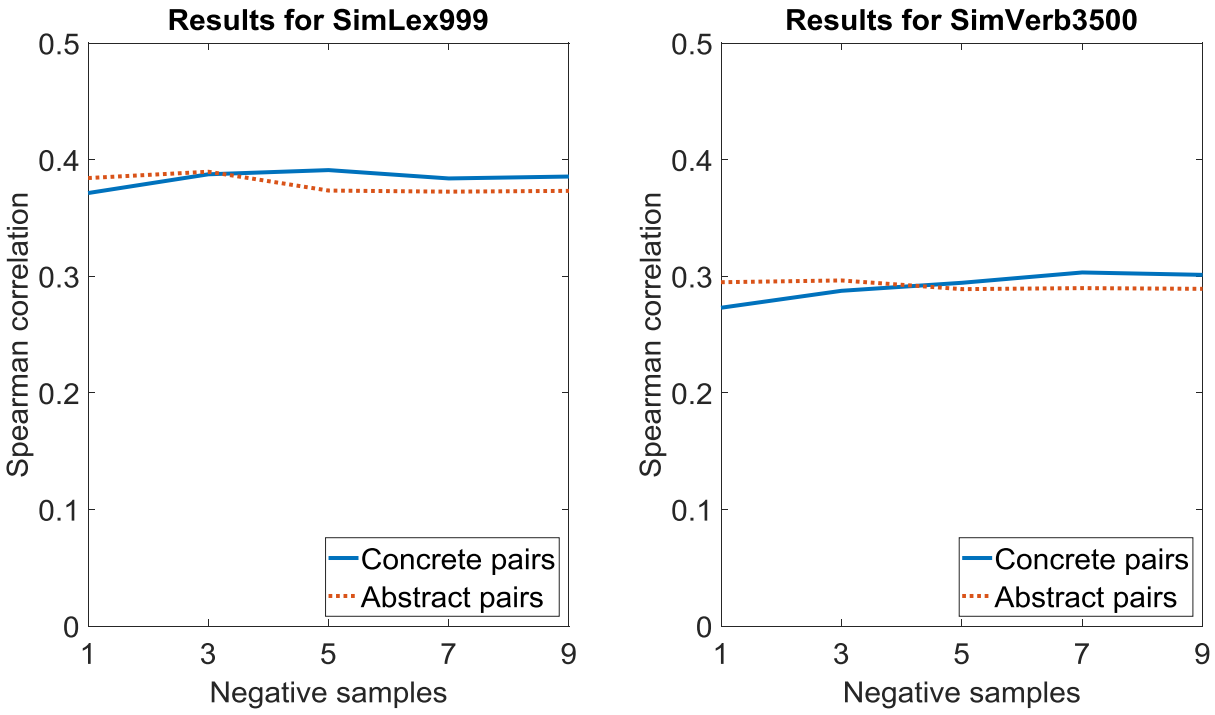


Figure 27. Model performance in predicting similarity/relatedness ratings from the SimLex-999 and SimVerb-3500 datasets, as a function of the number of negative samples.

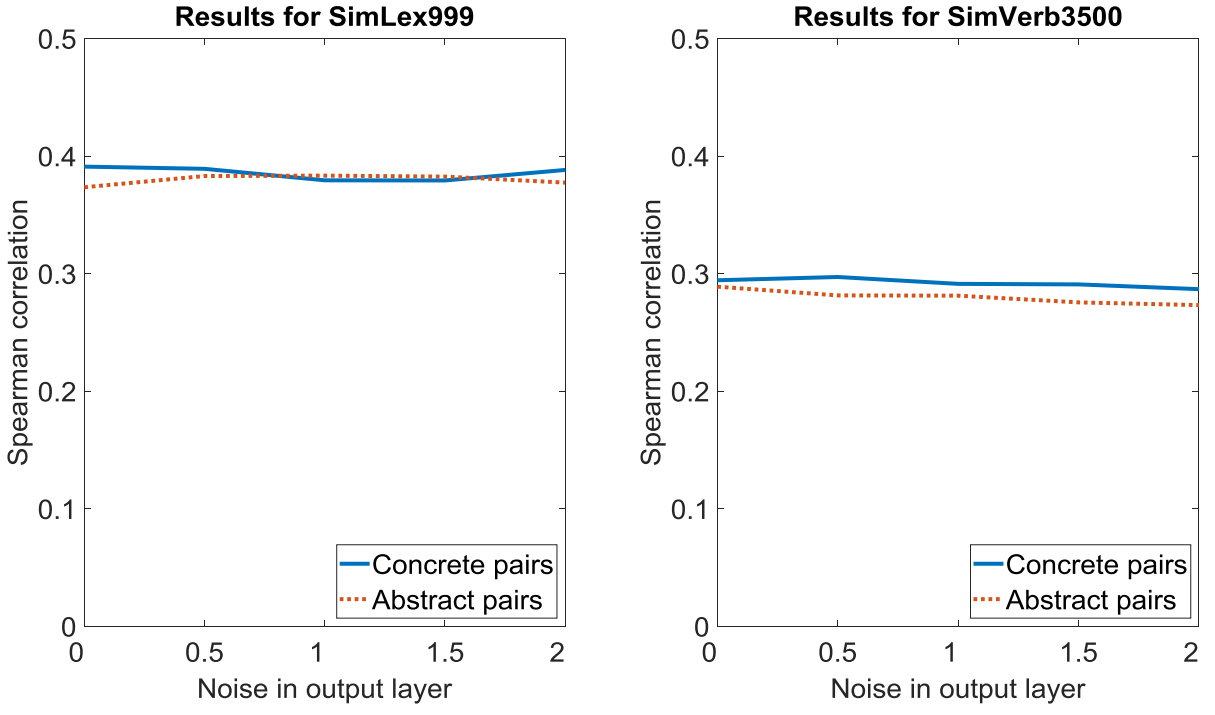


Figure 28. Model performance in predicting similarity/relatedness ratings from the SimLex-999 and SimVerb-3500 datasets, as a function of the amount of noise in the output layer.

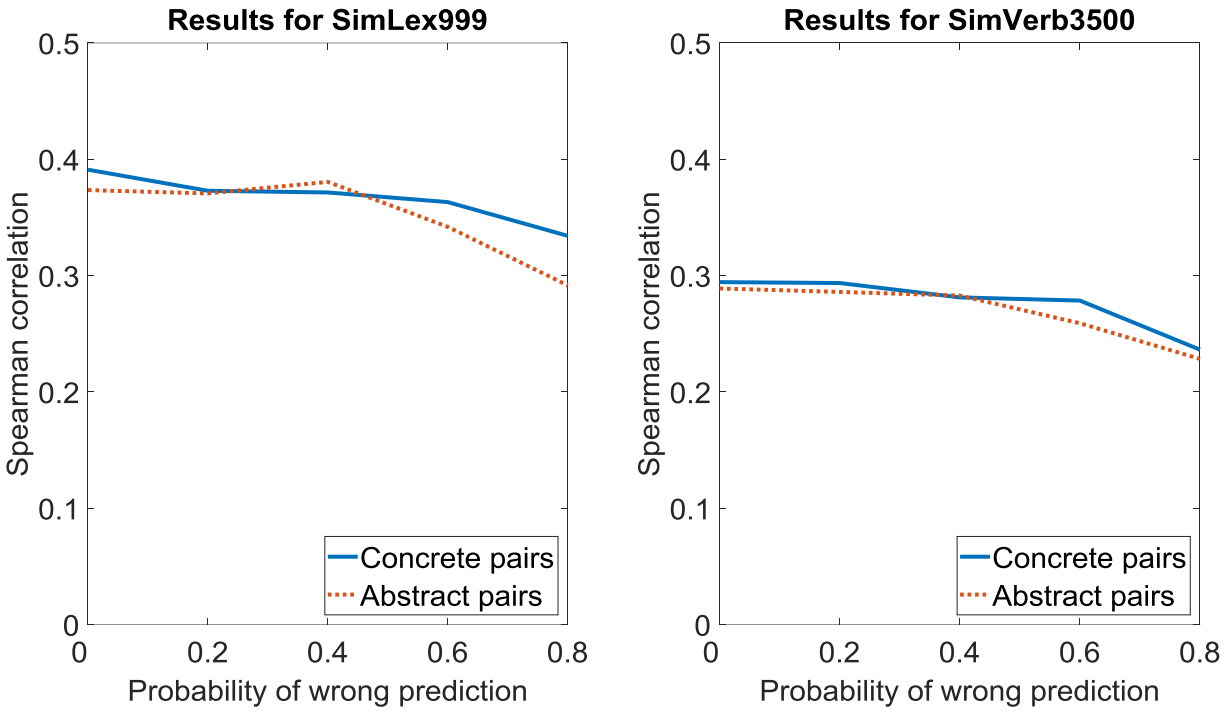


Figure 29. Model performance in predicting similarity/relatedness ratings from the SimLex-999 and SimVerb-3500 datasets, as a function of the probability of making a wrong prediction.

Table 14. Means and standard deviations for the Spearman correlations between subjective similarity ratings (i.e., from the SimLex-999 and the SimVerb-3500 datasets) and vector cosines, for the "healthy" and "damaged" models. Larger correlations indicate better model performance. The results for concrete and abstract words are presented separately.

| Concreteness (statistic) | Window size | Learning rate | Novelty bias | Negative samples | Noise level | P(wrong prediction) |
|---|---|---|---|---|---|---|
| | | | Similarity / relatedness rating (SimLex-999 dataset) | | | |
| Concrete (M) | .371 | .375 | .387 | .384 | .385 | .367 |
| Abstract (M) | .332 | .373 | .384 | .379 | .380 | .352 |
| Concrete (SD) | .029 | .021 | .006 | .007 | .006 | .021 |
| Abstract (SD) | .084 | .027 | .004 | .008 | .006 | .037 |
| | | | Similarity / relatedness rating (SimVerb-3500 dataset) | | | |
| Concrete (M) | .268 | .274 | .297 | .292 | .292 | .277 |
| Abstract (M) | .245 | .279 | .289 | .292 | .280 | .269 |
| Concrete (SD) | .031 | .037 | .004 | .012 | .004 | .024 |
| Abstract (SD) | .069 | .023 | .002 | .004 | .004 | .026 |

## 6.5. Pilot study

**Participants:**

Twenty adults (8 female, 10 male, 2 non-binary; mean age = 27.65), all native English speakers, were recruited via the crowdsourcing platform Prolific Academic. The participants were briefly informed about the purpose of the experiment, that they had the option of withdrawing from the study at any time, that their data would be kept confidential, and that they would be paid £6 for their participation. After agreeing to participate in the study, the participants were asked to provide demographic data concerning their gender, age, knowledge of foreign languages, potential hearing or vision impairment, and main occupation.

**Materials:**

We used 80 target words, of which 40 were abstract and 40 were concrete, taken from the study by Vigliocco and collaborators (2014). The two sets of words were matched for a large number of factors, including imageability, context availability, familiarity, age of acquisition, mode of acquisition, (log) frequency, number of letters, as well as a variety of phonological and orthographic measures. The list of all the words can be found in the Appendix C.

For each target word, we first extracted all the contexts in which that word appeared, within the BNC, and then removed the contexts in which the target word was a proper name. The contexts consisted of either 1, 5, 7 or 9 words, both preceding and following the target word. Out of all the contexts, we kept only 10 contexts per word and window size.

**Procedure:**

We used a continued free association task, where we asked participants to read a word in context and produce the first three associates that came to mind. Each

participant was shown all the target words, each within its own context. For 32 abstract words and 32 concrete words, half of the words were presented in a 1-word context, while the other half were presented in a 7-word context. For the remaining 8 abstract words and 8 concrete words, half of the words were presented in a 5-word context, while the other half were presented in a 9-word context, and the participants were also required to provide an answer to a simple comprehension, in order to ensure that they were paying attention to the task.

Each target word was displayed in context and written in boldface (e.g., “said Bobby playfully. ‘Don’t even **joke** to me about it. A’”). Three text boxes and a “Next” button were provided underneath the text, for entering the responses and proceeding to the next trial, respectively. If a participant left one or more of the text boxes unfilled, upon pressing “Next” they were informed (via a dialogue box) that they needed to provide all three associates, in order to continue. In addition, in 20 percent of the trials, after filling in the associates, the participants were also asked to answer a comprehension question related to the text that they had just read (e.g., “Which person is mentioned in the fragment?”), in order to proceed. Similar to a regular trial, when pressing the “Next” button, a participant was not allowed to advance to the next trial, unless they provided an answer to the comprehension question.

**Results:**

In order to evaluate the effects of manipulating the window size, as well as to test for any differences between abstract and concrete words, we first need to define a measure of task performance. Our measure of choice is the overlap between the responses produced by the participants, and the responses recorded in the SWoW continued association norms (De Deyne et al., 2019). In comparison to other popular norms, such as USF (Nelson et al., 2004) and EAT (Kiss et al., 1973), the SWoW dataset has the advantage of covering a larger number of cues (i.e., 12,000 words, vs 5,000 words, for USF, and 8,400 words, for EAT), and of being based on a continued association task (i.e., three associations generated per cue), rather than a discrete association task (i.e., one association generated per cue), thus providing a rich collection of responses. For each cue, we measure overlap in both an absolute and

relative manner, as the number and percentage of associations produced by the participants that can also be found in the SWoW norms.

For abstract words, the 1-word context condition and the 7-word context condition produced an average of 20.66 (66.75%) and 18.78 (65.59%) associations shared with the SWoW norms, respectively. No significant differences were found between the two conditions, for both the count measure ($t(62) = 1.45$, $p = .15$), and the percentage measure ($t(62) = 0.34$, $p = .73$).

For concrete words, the 1-word context condition and the 7-word context condition generated an average of 19.31 (63.99%) and 19.81 (66.21%) associations shared with the SWoW norms, respectively. The differences between the two conditions were not significant, for both the count measure ($t(62) = -0.33$, $p = .74$), and the percentage measure ($t(62) = -0.88$, $p = .38$).

When comparing abstract and concrete words, for the 1-word context condition there were no significant differences between the two classes of words, in terms of both the count measure ($t(62) = 0.88$, $p = .38$), and the percentage measure ($t(62) = 0.98$, $p = .33$). Similarly, the differences for the 7-word context condition were not significant, in terms of both the count measure ($t(62) = -0.80$, $p = .42$), and the percentage measure ($t(62) = -0.20$, $p = .85$).

To sum up, with respect to the (contextualized) continued free association task, we found no significant effect of context size and word concreteness, and of the interaction between the two factors. These results might be explained by having low statistical power, caused by the relatively small sample size (i.e., 20 participants). Another possibility is that, since the cue was more salient than the context, being written in boldface, the participants might have initially read only the cue, responded to it, and only then read the context.

## 6.6. Conclusions

We employed two state-of-the art distributional models (i.e., CBOW and Skip-gram), in order to simulate potential causes of semantic impairments in DLD. In order to do this, we first put forward a correspondence between important model parameters

(e.g., window size and learning rate) and psychological factors (i.e., the capacity of verbal working memory and the efficiency of the statistical learning mechanisms involved in word learning, respectively). Then, we assessed the impact of each parameter, by "damaging" a "healthy" model (i.e., by using suboptimal parameter values). This also allowed us to test whether the simulated impairments had different effects on the learning of concrete vs abstract words.

In our first set of analyses, based on the CBOW model, trained over the combined TASA and CBBC corpora, we found that reducing window size and changing the subsampling of frequent words do not affect the semantic representations as much as decreasing learning rate. The reason for this might be the fact that changes in the first two parameters have a strong impact only on specific categories of words, namely words that do not immediately precede or follow a given word, in the case of window size, and relatively frequent words, in the case of subsampling of frequent words. In contrast, changes in learning rate affect the processing of all the words encountered by the model. In addition, it is not the case that lesions have larger impact for abstract than concrete words, thus, demonstrating that despite differences in the associative structure for abstract and concrete words (e.g., Hoffman et al., 2013), parameter changes equally affect the two types of words. Thus damaging our artificial learners in their ability to extract information from the linguistic input impairs the abstract and the concrete domain alike.

In our second set of analyses, based on the Skip-gram model, trained over the BNC corpus, we found no statistical significant differences between the effects of any two factors, or between the effects for concrete vs abstract words. However, based on visual inspection, it appears that window size has the largest effect, and that the effect is greater for abstract words, than for concrete words. In order to test the validity of this observation, we ran a pilot study, based on a continued free association task. The participants were shown word stimuli in short linguistic contexts, and they were asked to record the first three words that came to mind upon reading the stimuli. Crucially, we manipulated the length of the contexts (i.e., the equivalent of window size), as well as the concreteness class of the stimuli. The results indicated no significant effect of context length, word concretness, or the interaction between the two factors.

In conclusion, we have illustrated a possible method of virtually lesioning distributional models, as a means of simulating semantic impairments associated with

DLD. The results of our analyses suggest that the induced lesions have comparable effects for concrete vs abstract words, in line with previous behavioural findings (Ponari, Norbury, Rotaru, et al., 2018). However, the effects of each model hyperparameter on model performance appear to depend on one or more factors, such as the model architecture, the size and nature of the training corpus, and the particular method(s) employed in evaluating model performance. For instance, decreasing the learning rate might have a strong detrimental effect on performance when the training corpus is small, given the relative scarcity of contextual information available for each word, whereas it might make little difference when the training corpus is large, since the corpus would contain a large amount of redundant information, compensating for the small learning rate. Further investigation is needed both to understand better the actual cognitive plausibility of distributional models like CBOW and Skip-gram, and to explore alternative modes of lesioning a model and their effect on semantic learning.

# 7. Final remarks

Distributional models of semantics, which learn semantic representations from word co-occurrence patterns in large text corpora, have become central tools in psycholinguistic and computational linguistic studies on semantics. For instance, a quick search on Google Scholar reveals that the two papers introducing the Skip-gram and CBOW models (i.e., Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) have been cited in at least 38,000 studies, in 7 years. In light of their ubiquity, the research presented in our dissertation aims to provide a better understanding of distributional models.

We begin by reviewing a wealth of empirical data regarding the psychological plausibility of distributional models, which demonstrates that such models can serve as valid instruments in the study of semantic cognition, even though they are more strongly tied to computational linguistics and artificial intelligence, than to psycholinguistics. Then, we show that the cognitive plausibility of linguistic and/or linguistic-visual models can be increased by including emotional representations, derived from emojis and accompanying texts, as predicted by embodied theories of semantics. Next, we explore the potential of joining the connectionist and distributional approaches to modelling semantics, and find that adding a spreading activation mechanism to distributional models can significantly increase their performance in accounting for behavioural data. We also investigate potential differences between concrete and abstract words, based on the structure of their model-based local neighbourhoods, and on the correlations between psychologically importantant factors (i.e., imageability, age of acquisition, hedonic valence, contextual diversity, and semantic diversity) and neighbourhood structure. Finally, we illustrate how distributional models can be employed in simulating semantic impairments associated with Developmental Language Disorder, by “lesioning” a number of “healthy” models, based on a psychological interpretation of certain model hyperparameters. The findings in each chapter are given a more detailed presentation in the rest of this chapter.

In Chapter 2, we provide a survey of studies that have directly or indirectly tested the psychological plausibility of distributional models. We begin by going over

some the reasons why distributional representations might be preffered over namely feature-based representations (McRae et al.; Vinson & Vigliocco, 2008), and free association-based representations (De Deyne et al., 2019; Nelson et al., 2004). The advantages of distributional representations include the fact that they can be automatically learned for a very large number of words (i.e., in the order of millions, when the models are trained over huge text corpora), they are derived from ordinary linguistic experience (rather than non-naturalistic tasks), and they allow researcher to easily test hypotheses regarding the influence of learning materials and processes, on the resulting semantic representations.

However, these advantages rely on the often-implicit assumption that distributional models have a high degree of cognitive plausibility. In order to present evidence in favour of this claim, we review a large number of relevant computational studies. The studies can be grouped into three main categories, depending on whether they investigate (1) model performance in fitting behavioural data from semantic tasks, (2) the structural properties of model-based semantic networks, as compared to those of free association-based networks, or (3) the ability to predict fMRI activation patterns, by using distributional representations. In selecting the studies, we looked only at those focusing on primarily semantic tasks, involving large and diverse behavioural datasets. Moreover, we selected studies that cover a wide range of distributional models, both unimodal (i.e., linguistic, visual), and multimodal models (i.e., linguistic-visual), whenever possible. Furthermore, in the interests of diversity, we typically included only one or two studies, per research team.

In the subchapter on modelling semantic task performance, we look at the similarity/relatedness rating, free association, and semantic categorization tasks. For the similarity/relatedness rating task, we find that model architecture plays an important role, such that, with respect to linguistic models, “predict” models outperform “count” models (e.g., Baroni et al., 2014; Pereira et al., 2016), whereas with respect to visual models, convolutional neural networks have a better performance than bag-of-visual-words models (Kiela & Bottou, 2014; Lazaridou et al., 2015; Silberer et al., 2017). Moreover, in the case of the visual models, with respect to model goodness of fit, the source of the images appears not be essential (Kiela et al., 2016), but the mechanism through which the linguistic and visual representations are combined does matter (i.e., models that blend the two modalities into a common, multimodal space

fare better than models with simply concatenate the two modelities; Bruni et al., 2014; Bruni, Uijlings, et al., 2012; Silberer & Lapata, 2014), and concreteness is an important factor (i.e., multimodal integration is typically beneficial for concrete words, but not abstract ones; Bruni, Boleda, et al., 2012; Bruni et al., 2014; Kiela et al., 2014). The particular choice of stimuli is highly significant, such that model performance is considerably better for similarity datasets, as opposed to relatedness datasets (Hill, Reichart, & Korhonen, 2015), and noun ratings are easier to predict than verb ratings (Hill et al., 2015). Also, despite its popularity as a "gold standard" in evaluating distributional models, the similarity/relatedness rating task has a number of shortcomings (Batcharov et al., 2016; Schnabel et al., 2015), such as low inter-rater agreement, and the fact that model performance in fitting ratings does not correlate strongly with performance in other tasks.

For the free association task, we find qualitative differences in performance similar to those for the similarity/relatedness task, with respect to the superiority of "predict" models over "count" models (e.g., Cattle & Ma, 2017; Thawani et al., 2019), and the influence of word concreteness (e.g., Hill & Korhonen, 2014). However, as opposed to the case for to the previously described task, studies that model free association probabilities employ a wide variety of model-based similarity measures (Feng & Lapata, 2010; Pereira et al., 2016) and behavioural datasets, making it difficult to compare studies. In addition, by not including including semantic processes that are likely to be involved in generating word associations, such as competition-based memory retrieval (Nelson et al., 1998; Raaijmakers & Shiffrin, 1981), we believe that nearly all the studies underestimate the predictive power of distributional models.

For the semantic categorization task, we find results that are quite different from those corresponding to the previous two tasks. More specifically, distributional models produce very high levels of performance (Baroni et al., 2014; Bruni et al., 2011, 2014; Riordan & Jones, 2011). Moreover, even though the vast majority of stimuli consist of concrete words, model goodness of fit does not increase significantly when adding visual representations (Bruni et al., 2011, 2014; Silberer & Lapata, 2014). Somewhat surprisingly, we did not come across any study comparing "predict" and "count" models, but within the "count" class of models it appears that "word-as-context" models slightly outperform "document-as-context" models (Riordan & Jones, 2011).

In the subchapter on the structure of model-based and free-association based semantic networks, we find that networks derived from distributional models have a “small-word”, “scale-free” structure, similar to that of networks obtained from free association norms, even though the match for the LSA-based networks seems to be significantly poorer than that for other models (Griffiths et al., 2007; Utsumi, 2015). Interestingly, it appears that hybrid models, obtained by combing two distributional models, are better than single models at matching the free association data, suggesting that the generation of word associations might depend on multiple types of linguistic information (e.g., associative and contextual; Gruenenfelder et al., 2016). Unfortunately, however, it is not straightforward to evaluate the generality of the results, given the large methodological differences between the studies, with respect to whether the networks are directed or undirected, the method used for defining the local neighbourhoods for each word (e.g., by applying the *k*-nn method, the *cs*-method, or the ε-method; Utsumi, 2015), and the particular structural network properties being investigated (e.g., the degree to which local neighhood sizes follow a truncated vs regular power-law distribution), among other factors.

In the subchapter on predicting fMRI data based on distributional models, we find that models successfully accounted for neural activation patterns, regardless of the type of method employed in testing the models. In whole brain analyses, perhaps unsurprisingly, multimodal models outperform unimodal ones (Anderson et al., 2013, 2017; Bulat et al., 2017), and the relative performance of the three classes of models (i.e., linguistic, visual, and linguistic-visual) is largely consistent across participants (Anderson et al., 2017; Bulat et al., 2017). However, when looking at specific brain regions, the modality associated with each region, as informed by the neuroscientific literature, coincided with the modality of the best performing model (Anderson et al., 2015). Unfortunately, given that almost all the studies use the same dataset (Mitchell et al., 2008), it is not obvious whether these findings generalize to different tasks and/or datasets.

In conclusion, a large number of studies, employing both behavioural and fMRI data, provide quantitative evidence in favour of the claim that distributional models are psychologically plausible, at least to a reasonable degree. Nevertheless, model performance can differ markedly as a function of task (e.g., free association probabilities are considerably harder to predict than semantic categories and

relatedness ratings), and model class (e.g., “predict” models are usually better than “count” models). Multimodal models, which employ both linguistic and visual representations, outperform purely linguistic models with respect to abstract words, but not concrete ones. Furthermore, model-based and free association-based semantic networks are structurally similar (i.e., they have “small-world” and “scale-free” properties). Additionally, distributional models appear to be good at capturing the type of semantic information reflected by fMRI activation patterns, although additional studies are required in order to assess the generality of more specific findings.

In Chapter 3, we investigate whether it is possible to improve the performance of linguistic and linguistic-visual models of semantics, by adding affective information. Our approach is motivated by two observations. Firstly, as discussed in Chapter 2, models which combine visual and linguistic representations typically outperform purely linguistic models in a variety of tasks, such as free association, similarity/relatedness rating, and semantic categorization. This finding lends support to embodied theories of semantics (e.g., Glenberg et al., 2008), according to which sensory-motor information, acquired through our interaction with the physical word, plays a significant role is the representation and processing of words. Moving from linguistic to multimodal, linguistic-visual models also offers a solution to the symbol grounding problem (e.g., Harnad, 1990), which occurs when word meaning is derived exclusively from linguistic sources, without linking it to real-world referents.

Secondly, almost all studies involving multimodal models of semantics have focused only on vision, as a source of extralinguistic data. However, it has been known for a relatively long time that emotion is another important factor in human cognition (e.g., Dolan, 2002), as shown by studies on attention (e.g., Öhman et al., 2001), perception (e.g., Gasper & Clore, 2002), and memory (e.g., Eich et al., 1994), among other topics. With respect to word processing, more recent studies have found that, as opposed to neutral words, emotionally valenced words are learnt earlier (Ponari, Norbury, & Vigliocco, 2018) and processed faster (Kousta et al., 2011).

Given that emotion appears to have a substantial contribution to word meaning, we test the hypothesis that combining affective representations with linguistic and linguistic-visual representations should increase the performance of the linguistic and linguistic-visual models, in accounting for subjective similarity/relatedness ratings, from four datasets (i.e, SimLex-999, SimVerb-3500, MEN, and SL). We begin by

selecting the linguistic, visual, and emotional models. The linguistic model we choose is GloVe, which has state-of-the-art performance in a number of tasks, such as analogy completion, similarity/relatedness rating, and named entity recognition. In order to choose the other model, we compare some of the most popular and successful visual models (i.e., SIFT, K&B, AlexNet, GoogLeNet, and VGG-19) and emotional models (i.e., CNN, GRU, LSTM, and DeepMoji), on the combination of the four rating datasets. We find that the best visual model is GoogLeNet, and the best emotional model is DeepMoji. Regarding the emotional model, rather than using valence, arousal, and dominance ratings, like in other studies (De Deyne et al., 2018), we prefer to employ the DeepMoji model, which has advantages of being a distributional model (i.e., it learns to predicts emojis from co-occurring text), having a large vocabulary (i.e., 50,000 words), and generating high-dimensional representations (i.e., vectors with 256 dimensions). We also provide quantitative evidence in favour of the assumption that DeepMoji representations actually capture emotional information, by using PCA and showing that the first 10 principal components correlate significantly with subjective ratings of emotion (i.e., valence, arousal, and dominance).

In our first set of analyses, we compare linguistic-visual and linguistic-emotional models, to the purely linguistic model. For the multimodal models, obtained by concatenating representations two representations, we vary the weights assigned to each extralinguistic component, since it is not clear beforehand how strong the contribution of the visual/emotional components should be. We find that adding a visual component significantly improves performance only for the SL dataset ($p < .001$), for weights between 0.6 and 1.2, and significantly decreases performance for the MEN dataset ($p < .001$), for weights between 1.6 and 2. These results seem to be at odds with previous studies showing that linguistic-visual models always perform better than purely linguistic ones. The discrepancy is likely due to the fact that other studies either do not use multiple weights (e.g., Kiela et al., 2014), or report only the results for the optimal choice of weights (e.g., Bruni et al., 2014), thus providing a very incomplete set of results. In addition, the linguistic corpus we use is considerably larger than those employed in other studies (e.g., Kiela & Bottou, 2014), which means that the linguistic model already has a very good performance, leaving little room left for improvement. We also find that adding affective information significantly improves

performance only for the SimVerb-3500 dataset ($p < .00125$), for weights ranging from 1.2 to 1.6, while it significantly decreases performance for the MEN dataset ($p < .001$), for weights between 1.4 and 2, and for the SL dataset ($p < .001$), for weights between 0.6 and 2. These results might be explained by the fact that the SimVerb-3500 dataset consists only of verbs, which rarely appear in other datasets. Verbs are usually more abstract than nouns, which makes them more likely to rely on affective information (Kousta et al., 2011).

In our second set of analyses, we extend the previous analyses by comparing trimodal, linguistic-visual-emotional models, to bimodal, linguistic-visual and linguistic-emotional model, as well as the unimodal, linguistic model. Once again, we vary the weights assigned to each extralinguistic component, independently. By looking at the best and worst trimodal models, for each set of ratings, we find that the addition of visual information is generally beneficial for datasets consisting mostly of concrete words (e.g., SL), but detrimental for datasets consisting mostly of abstract words (e.g., SimVerb-3500), while the opposite pattern of results becomes evident when adding affective information.

In our final set of analyses, we test whether indeed the effect (i.e., beneficial vs detrimental) of including visual and affective information depends on word concreteness. To do this, we combine the SimLex-999 and SimVerb-3500 datasets (i.e., the two datasets spanning the broadest range of concreteness), and select only the top 25% most abstract pairs and the top 25% most concrete pairs. For simplicity, we also assign equal weights to the linguistic, visual, and emotional components. The results reveal that the visual and linguistic models perform best on the most concrete pairs, the emotional models perform best on the most abstract pairs, and the performance of the linguistic-emotional and linguistic-visual-emotional models is not affected by concreteness.

In conclusion, we find that whether the addition of non-linguistic increases or decreases model performance, or instead has no effect, is determined by the weights attributed to the different types of information, which may have practical value for future modelling. Also, it appears that this impact depends on whether the dataset includes predominantly concrete or abstract words, such that bringing in visual information is particularly beneficial for more concrete concepts, whereas bringing in emotional information is particularly beneficial for more abstract concepts.

In Chapter 4, we test the potential benefits of starting with structural, distributional models of semantics, and then adding to them a spreading activation mechanism (Anderson, 1983; Collins & Loftus, 1975), as a means of capturing certain aspects of automatic semantic processing. We start from the observation that, within psycholinguistics and computational linguistics, there are two main classes of approaches to modelling semantic cognition. The first tradition is that of connectionist modelling (e.g., McClelland et al., 2010; Zorzi, Testolin, & Stoianov, 2013), where the emphasis is on exploring various types of semantic processing, without paying much attention to the cognitive plausibility of the underlying semantic representations (i.e., often, the representations consist of handcrafted features, and cover only a very limited set of words). In contrast, the second tradition is that of distributional modelling (e.g., Turney & Pantel, 2010), is almost entirely concerned with deriving rich semantic representations from large corpora, without including specific process that make use of the representations.

Given that these two traditions rarely interact, we explore whether it is possible to better account for a wide range of behavioural data, by creating "dynamical" models where semantic activation spreads within semantic networks, derived from "structural", distributional models of semantics. Our model allows us to examine the role of indirect, mediated relations between words, which are known to be reliable predictors of performance in tasks such as lexical decision, semantic similarity rating, and extralist cued recall (De Deyne et al., 2013; Steyvers et al., 2005). By considering the spreading of activation within the network, we can also allow test the effect of weak semantic relations (i.e., between relatively dissimilar words; Chen & Mirman, 2012; Mirman & Magnuson, 2008), which are considerably more numerous than strong semantic relations, and, therefore, might play a significant role in semantic processing.

Our structural models are based on the LSA, CBOW and GloVe models, trained over the BNC corpus. In order to create a structural model, we select the vector representations provided by one of the distributional models (e.g., LSA), and compute a matrix of cosine similarities between each pair of words. After we set to zero all the negative values, under the assumption that they carry little or no information, we obtain a structural model (i.e., SM). Next, in order to construct a dynamic model, we reweight the rows of SM, such that they sum to 1. This transformation is based on the following simplifying principles:

- the amount of activation passed between two words is proportional to the level of similarity between the words, such that words send more activation to other, closely related words
- a fixed part of any word's activation is retained within that particular word, without being propagated to other words
- the total amount of activation within the semantic network remains constant

After applying the transformation, the resulting matrix is the dynamic model (i.e., DM), and its entries denote the percentages of activation distributed from one word, to another. We assume that the spreading of activation begins with all the activation being concentrated within a single, initial word (i.e., the word that is the focus of attention), after which, based on the values in DM, part of this activation is passed on to the neighbours of the initial word, then to the neighbours of the neighbours, and so on, in a series of discrete steps. Technically, our dynamic model corresponds to a discrete-time Markov chain.

The datasets on which we evaluate the performance of the structural and distributional models consist of reaction times and accuracies for both lexical and semantics decision, concreteness and imageability ratings, as well as similarity/relatedness ratings, from four datasets (i.e, SimLex-999, SimVerb-3500, MEN, and SL). In the case of non-relational tasks (i.e., lexical or semantic decision, imageability or concreteness rating), our predictors consist of the number of neighbours of a given word, based on cosine similarities, for the structural models, and on activation values, for the dynamic models. For both types of models, we split the neighbourhoods based on their distance (i.e., going from very close neighbours to very distant ones). In the case of relational tasks (i.e., similarity/relatedness rating), our predictors consist of the cosine similarity between the words in a pair, for the structural models, and the activation value of either of the two words in a pair, taking the other word as the initial source of activation, for the dynamic models. When employing the dynamic models, we compute a separate set of predictors for each of the five time steps within the spreading activation process. In order to more precisely estimate the contributions of the structural and dynamic models, we also employ a task-specific baseline, which includes a number of relevant psychological factors (e.g.,

age of acquisition, valence, familiarity, frequency, semantic diversity, number of orthographic and phonological neighbours, etc.), for each task.

For the lexical and semantic decision data, the addition of the structural predictors almost always increases performance. However, there are certain task-specific differences, with respect to the dynamical predictors: for lexical decision, most of the predictors have a significant effect for accuracy, whereas only a few contribute significantly for response time; in contrast, for semantic decision, the pattern is reversed. For imageability and concreteness ratings, almost all the structural and dynamical predictors are significant predictors of task performance. For the similarity/relatedness ratings, the contributions of all the structural predictors are statistically significant, whereas those of the dynamical predictors depend on the dataset being employed: for the norms, all the predictors are significant; for the SL and MEN norms, the majority of the predictors are significant; for SimVerb-3500 and SimLex-999 norms and, relatively few predictors are significant.

In addition, of the three types of dynamic models, the ones based on the CBOW and GloVe models perform better than the ones based on the LSA model, in nearly all the tasks. This finding provides new evidence to the claim that “predict” models usually outperform “count” models (see the discussion in Chapter 2).

In conclusion, our dynamic models can reliably predict behavioural data from a variety of tasks, which require different levels of semantic information. The tasks include offline (untimed) tasks, such as imageability, concreteness, and similarity/relatedness rating, as well as online (timed) tasks, such as lexical and semantic decision. Also, importantly, most of the predictors from the dynamic models remain significant even after controlling for the contribution of the structural models and of the baseline models, consisting of several lexical, semantic, orthographic, and phonological variables. These results suggest that it is possible to improve the predictive power of distributional models, by adding processing assumptions, such as the automatic spreading of activation.

In Chapter 5, we examine potential differences between concrete and abstract words, by using semantic networks obtained from distributional models. Our approach is motivated by the lack of a comprehensive study on the structure of semantic neighbourhoods of concrete vs abstract words, especially since certain prominent theories of semantics make divergent predictions: for instance, according to the Dual

Coding theory (e.g., Paivio, 1971, 1986), concrete words should have more neighbors that abstract words, by virtue of having a richer semantic content; in contrast, theories that emphasize contextual diversity (e.g., Hoffman et al., 2013; Jones et al., 2012) predict that concrete words should have fewer neighbors that abstract words, given that abstract words appear in a wider variety of contexts.

We begin by taking a broader perspective, and noting that, in the psychological literature, the concept of "semantic richness" (i.e., the number of semantic neighbours, for a given word), has been operationalized by using mainly three types of representations, namely featural (i.e., derived from feature generation norms), associative (i.e., derived from free association norms), and distributional (i.e., derived from distributional models). We then review a large number of behavioural studies that employ such representations, focusing on reported differences in richness between concrete and abstract words, as well as on whether such differences have an effect on task performance, in tasks such as lexical decision, naming, semantic decision and progressive demasking. The findings paint a complex and sometimes surprising picture: the direction of the differences depends on the type of semantic representation (e.g., concrete words are richer than abstract ones, based on featural representations, whereas the opposite result holds for associative and distributional representations); semantic richness can be a significant predictor of task performance for one class of words (e.g., concrete), but not the other (i.e., abstract); semantic richness effects can be detected in certain tasks, but not others, and sometimes, even for a given task, they are not consistently found across multiple studies.

In order to reach a better understanding of richness differences between concrete and abstract words, as well as their potential causes, we first compare the neighbourhood structure for the two classes, where the neighbourhoods are obtained from a state-of-the-art, distributional model (i.e., CBOW), trained over a representative corpus of British English (i.e., BNC). Our measures of interest are neighbourhood size (i.e., the number of neighbours), and clustering coefficient (i.e., the fraction of neighbours that are also neighbours of one another). The results indicate that concrete words have slightly larger neighbourhoods than abstract words, and their neighbourhoods are considerably more interconnected.

We then test the degree to which neighbourhood size and clustering coefficient are sensitive to a number of important factors, namely imageability, age of acquisition,

squared hedonic valence, log contextual diversity, and semantic diversity. The main results are presented below, and they apply to both concrete and abstract words.

Words high in imageability have larger and denser neighbourhoods than less imageable words. This finding might be explained by the fact that perceptual grounding strengthens semantic associations, and that physical similarity/relatedness/co-occurrence, which contribute to semantic associations, are transitive relations, up to a certain degree.

Words acquired early have larger and denser neighbourhoods than words learned later in life. This tendency might be accounted for by using the process of differentiation (e.g., Steyvers & Tenenbaum, 2005), according to which the meanings of new words are derived as variations on the meanings of words already known, and the process of preferential attachment (e.g., Barabási & Albert, 1999), according to which words with many neighbours are more likely to be selected as targets from differentiation. Since early acquired words participate in more instances of differentiation, they are likely to have a richer neighbourhood structure than words acquired later in life.

Words with high hedonic valence are very similar to words with low hedonic valence, in terms of neighbourhood structure. This finding is particularly difficult to explain, given that emotion has been shown to play a significant role in the representation of abstract words (e.g., Kousta et al., 2009, 2011). In contrast, it is the neighbourhoods of concrete words, rather than abstract words, which seem to be more affected by hedonic valence, although the effects only barely pass the threshold for statistical significance.

Words with high contextual diversity have fewer near and distant neighbours, as well as more very distant neighbours, than words with low contextual diversity. In addition, contextual diversity has a strong negative correlation with clustering coefficient, regardless of distance. These results are likely to arise from the fact that, since we control for meaningful variation in semantic context (by using semantic diversity), words with high contextual diversity are likely to co-occur with a large number of other words, purely by chance, and thus become associated with them.

Words with high semantic diversity have more near and distant neighbours, as well as fewer very distant neighbours, than words with low semantic diversity. Also, semantic diversity has a weak negative correlation with clustering coefficient, at all

distances. Since semantic diversity is an index of polysemy (e.g., Hoffman et al., 2013), this pattern of findings might be explained by the fact that having more senses increases the number of neighbors for a given word, while also decreasing their strength, since each neighbour has fewer co-occurrences with that word. For near and distant neighbours, the first effect seems to dominate, whereas for very distant neighbours, the second effect appears to be stronger.

Several conclusions can be drawn from the results presented in this chapter. Firstly, concrete words are semantically richer than abstract words, in terms of both neighbourhood size and interconnectivity. However, the difference between the two classes of words is very pronounced for clustering coefficient, but barely detectable for number of neighbours. Secondly, the structure of the neighbourhoods (especially for near neighbours, which are the only ones considered in most studies) is influenced mainly by the interplay of imageability/age of acquisition, and contextual/semantic diversity. Given that concrete words have a sizeable advantage over abstract words, with respect to clustering coefficients, this suggests that the contribution of imageability/age of acquisition outweighs that of contextual/semantic diversity.

In Chapter 6, we employ distributional models in order to examine various factors that could play a role in the poor linguistic performance of children suffering from DLD. The most prominent symptoms of DLD are related to the use of grammar (e.g., Leonard, 2014), such as the frequent omission of function words and grammatical inflections, the inappropriate use of past-tense and pronoun forms, and difficulties in comprehending and repeating syntactically complex sentences. Consequently, impairments related to semantics haven't been explored at length, so far. In general, such impairments are reflected in the children's vocabulary (e.g., McGregor et al., 2013), in terms of how many words are known (i.e., vocabulary breadth), and how well they are known (i.e., vocabulary depth). Various studies have shown that the differences between children with DLD and typically developing children can be quite subtle. In a similarity rating task (Kail & Leonard, 1986), the performance of the two groups of children was well accounted for by using the same semantic dimensions, and the same relative contributions of the dimensions. In free association tasks (e.g., Brooks et al., 2017), children with DLD generated fewer correct answers, and more incorrect answers, than typically developing children. In semantic fluency tasks (e.g., Henry et al., 2015; Weckerly et al., 2001), children with DLD

produced fewer clusters than typically developing children, although clusters size were equal across the two groups.

Several potential causes for the behavioural effects associated with DLD have been suggested. One such cause might be an impairment of working memory (e.g., Baddeley, 2003). Children with DLD typically have a reduced working memory capacity (Henry & Botting, 2017), which might explain their relatively poor performance in linguistic tasks (e.g., Graf Estes et al., 2007), and in visual tasks (e.g., Vugs et al., 2013). However, individual differences in working memory performance correlate significantly with task performance in certain studies (e.g., Ellis Weismer & Thordardottir, 2002; Kleemans et al., 2011), but not in others (e.g., Briscoe et al., 2001; Lum et al., 2012). Another cause might be an impairment in statistical learning (e.g., Romberg & Saffran, 2010). Children with DLD have a lower level of performance than their typically developing peers (Obeid et al., 2016), in a variety of tasks relying on statistical learning, across the visual, auditory, and motor modalities. Similarly to the case of working memory, individual differences in statistical learning performance are significantly correlated with vocabulary size in some studies (e.g., Evans et al., 2009), but not in others (e.g., Haebig et al., 2017). Yet another cause might be an impairment in attention (e.g, Nobre et al., 2014). Children with DLD have relatively poor performance in tasks that require sustained attention (e.g., Ebert & Kohnert, 2011). Furthermore, individual differences in task performance are significant predictors of linguistic abilities (e.g., Finneran et al., 2009; Montgomery et al., 2009).

In order to test the plausibility of the aforementioned, potential causes, we use a modelling approach, based on the CBOW and Skip-gram models. Here, we rely on the observation that certain parameters of these models can be assigned a psychological interpretation. More specifically, the size of the sliding window reflects the capacity of working memory, the learning rate reflects the efficiency and accuracy of the (statistical) word learning mechanisms, and the subsampling of frequent words reflects the amount of attentional resources allocated to frequent vs infrequent words.

In our first set of analyses, we employ the CBOW model, trained on a corpus of child-oriented language. In order to evaluate the effects of each factor, we start from a “healthy” model, where the values of the parameters are the ones recommended in the literature. Then, we create numerous “damaged” models, by factorially manipulating the values for window size, learning rate, and novelty bias (i.e., the

subsampling of frequent words). Next, we employ both representational similarity analysis (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008), and linear-mixed effects models, as a means of comparing the “healthy” and “damaged” models, as well as estimating the amount of damage per factor. The results show that the learning rate has by far the largest effect on performance, followed by window size and novelty bias.

We also conduct a second set of analysis, in which we replace the CBOW model with the Skip-gram model (i.e., in order to evaluate the generality of the initial results), and switch from the child-oriented corpus to an adult-oriented corpus (i.e., in order to have a more realistic corpus, in terms of size). More importantly, we include additional model parameters, namely the number of negative samples (i.e., linked to the degree of inhibitory, attentional control), the amount of noise in the neural network (i.e., linked to the efficiency and precision of the statistical learning mechanism), and the probability of predicting a wrong word during word learning (i.e., linked to the overall level of focused attention). The results reveal a large effect of window size, and little of no effect of the other parameters. Also, the effect is stronger for abstract words, than for concrete words. As an extension of this set of analyses, we also run a free association experiment, to see whether the interaction between window size and word concreteness is also supported by behavioural data. In the experiment, the participants have to generate three associates for a set of concrete and abstract words, where the cues were presented in either a short linguistic context, or a long linguistic context. The amount of damage is quantified as the overlap between the associations produced by the participants, and the associations recorded in the SWoW free association norms, such that low overlap translates into large damage. We find no effect of context length or concreteness class, and no significant interaction between the two factors.

In conclusion, the modelling experiments suggest that some of most important factors that might be responsible for the effects of DLD are the capacity of the working memory (i.e., window size), and the robustness of the word learning mechanism (i.e., learning rate). Which of the two factors prevails is likely to depend on the amount of linguistic input (i.e., corpus size), such that the efficiency of the statistical learning process becomes less consequential as the amount of linguistic experience increases, most likely due to increased redundancy of the linguistic input. Moreover, the two computational experiments, the behavioural experiment, and another experimental

study (Ponari , Norbury,  Rotaru, et al., 2018), seem to indicate that there are no major differences between children with DLD and typically developing children, in the processing of concrete vs abstract words.

# References


Aakerlund, & L., Hemmingsen, R. (1998). Neural networks as models of psychopathology. *Biological Psychiatry, 43,* 471-482.

Abdul-Mageed, M., & Ungar, L. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (*Vol. 1: Long* Papers, pp. 718-728). Stroudsburg, PA: Association for Computational Linguistics.

Abnar, S., Ahmed, R., Mijnheer, M., & Zuidema, W. (2018). Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In A. Sayeed, C. Jacobs, T. Linzen, & M. van Schijndel (Eds.), *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 57-66). Stroudsburg, PA: Association for Computational Linguistics.

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814-823.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and Wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19-27). Stroudsburg, PA: Association for Computational Linguistics.

Albert, R., Jeong, H., & Barabási, A. L. (2000). Error and attack tolerance of complex networks. *Nature*, *406*(6794), 378-382.

Almuhareb, A., & Poesio, M. (2005). Concept learning and categorization from the Web. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 103-108). Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, A. J., Bruni, E., Bordignon, U., Poesio, M., & Baroni, M. (2013). Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1960-1970). Stroudsburg, PA: Association for Computational Linguistics.

Anderson, A. J., Bruni, E., Lopopolo, A., Poesio, M., & Baroni, M. (2015). Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, *120*, 309-322.

Anderson, A. J., Kiela, D., Clark, S., & Poesio, M. (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, *5*, 17-30.

Anderson, A. J., Zinszer, B. D., & Raizada, R. D. (2016). Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, *128*, 44-53.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261-295.

Andrews, M., & Vigliocco, G. (2010). The hidden Markov topic model: A probabilistic model of semantic representation. *Topics in Cognitive Science*, *2*(1), 101-113.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463-498.

Auguste, J., Rey, A., & Favre, B. (2017). Evaluation of word embeddings against cognitive processes: Primed reaction times in lexical decision and naming tasks. In S. Bowman, Y. Goldberg, F. Hill, A. Lazaridou, O. Levy, R. Reichart, & A. Søgaard (Eds.), *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP* (pp. 21-26). Stroudsburg, PA: Association for Computational Linguistics.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 2200-2204). European Language Resources Association.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189-208.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 3rd International Conference on Learning Representations* (pp. 1-15). Stroudsburg, PA: Association for Computational Linguistics.

Bakarov, A. (2018). A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Balota, D. A. (1990). The role of meaning in word recognition. In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 9-32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445-459.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Baronchelli, A., Ferrer i Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *17*(7), 348-360.

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, *10*(1), 3-13.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673-721.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Vol. 1: Long Papers*, pp. 238-247). Stroudsburg, PA: Association for Computational Linguistics.

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, *34*(2), 222-254.

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 129–163). Cambridge, United Kingdom: Cambridge University Press.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and Embodiment: Debates on meaning and cognition* (pp. 245-283). Oxford, United Kingdom: Oxford University Press.

Batchkarov, M., Kober, T., Reffin, J., Weeds, J., & Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 7-12). Stroudsburg, PA: Association for Computational Linguistics.

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLOS One*, *6*(5), e19348.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, *2*(1), 1-127.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527-536.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767-2796.

Blaney, P. H. (1986). Affect and memory: A review. *Psychological Bulletin*, *99*(2), 229-246.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Bookheimer, S. Y., Strojwas, M. H., Cohen, M. S., Saunders, A. M., Pericak-Vance, M. A., Mazziotta, J. C., & Small, G. W. (2000). Patterns of brain activation in people at risk for Alzheimer's disease. *New England Journal of Medicine*, *343*(7), 450-456.

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

Briscoe, J., Bishop, D. V., & Norbury, C. F. (2001). Phonological processing, language, and literacy: A comparison of children with mild-to-moderate sensorineural hearing loss and those with specific language impairment. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(3), 329-340.

Brooks, P. J., Maouene, J., Sailor, K., & Seiger-Gardner, L. (2017). Modeling the semantic networks of school-age children with specific language impairment and their typical peers. In M. LaMendola & J. Scott (Eds.), *Proceedings of the 41st Annual Boston University Conference on Language Development* (pp. 114-127). Somerville, MA: Cascadilla Press.

Bruce, V., Green, P. R., & Georgeson, M. A. (2003). *Visual perception: Physiology, psychology, & ecology*. Hove, United Kingdom: Psychology Press.

Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in Technicolor. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (*Vol. 1: Long Papers*, pp. 136-145). Stroudsburg, PA: Association for Computational Linguistics.

Bruni, E., Tran, G. B., & Baroni, M. (2011). Distributional semantics from text and images. In S. Pado & Y. Peirsman (Eds.), *Proceedings of the GEMS 2011*

*Workshop on Geometrical Models of Natural Language Semantics* (pp. 22-32). Stroudsburg, PA: Association for Computational Linguistics.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*(1), 1-47.

Bruni, E., Uijlings, J., Baroni, M., & Sebe, N. (2012). Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In N. Babaguchi, K. Aizawa, & J. Smith (Eds.), *Proceedings of the 20th ACM International Conference on Multimedia* (pp. 1219-1228). New York, NY: Association for Computing Machinery.

Bruza, P., Kitto, K., Nelson, D., & McEvoy, C. (2009). Is there something quantum-like about the human mental lexicon? *Journal of Mathematical Psychology*, *53*(5), 362-377.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, 1116.

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.

Bulat, L., Clark, S., & Shutova, E. (2017). Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1081-1091). Stroudsburg, PA: Association for Computational Linguistics.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510-526.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, *44*(3), 890-907.

Bullinaria, J. A., & Levy, J. P. (2013). Limiting factors for mapping corpus-based semantic representations to brain activity. *PLOS One*, *8*(3), e57191.

Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J., & Gabrieli, J. D. (2002). Immature frontal lobe contributions to cognitive control in children: Evidence from fMRI. *Neuron*, *33*(2), 301-311.

Cambria, E., Olsher, D., & Rajagopal, D. (2014). SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In C. E. Brodley & P. Stone (Eds.), *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 1515–1521). Palo Alto, CA: AAAI Press.

Cappallo, S., Svetlichnaya, S., Garrigues, P., Mensink, T., & Snoek, C. G. (2019). New modality: Emoji challenges in prediction, anticipation, and retrieval. *IEEE Transactions on Multimedia*, *21*(2), 402-415.

Carlson, T. A., Simmons, R. A., Kriegeskorte, N., & Slevc, L. R. (2014). The emergence of semantic meaning in the ventral temporal pathway. *Journal of Cognitive Neuroscience*, *26*(1), 120-131.

Cattle, A., & Ma, X. (2017). Predicting word association strengths. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1283-1288). Stroudsburg, PA: Association for Computational Linguistics.

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, *2*(10), 913-919.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417-430.

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In D. Wu, M. Carpuat, X. Carreras, & E. M. Vecchi (Eds.), *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103-111). Stroudsburg, PA: Association for Computational Linguistics.

Choudhury, M., & Mukherjee, A. (2009). The structure and dynamics of linguistic networks. In A. Deutsch, N. Ganguly, & A. Mukherjee (Eds.), *Dynamics on and of complex networks* (pp. 145-166). New York, NY: Birkhäuser Boston.

Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). NUS-WIDE: A real-world web image database from National University of Singapore. In S. Marchand-Maillet & I. Kompatsiaris (Eds.), *Proceedings of the ACM International Conference on Image and Video Retrieval* (pp. 1–9). New York, NY: Association for Computing Machinery.

Clark, S. (2015). Vector space models of lexical meaning. In S. Lappin & C. Fox (Eds.), *Handbook of contemporary semantics* (2nd ed., pp. 493-522). Malden, MA: John Wiley & Sons.

Cohen, J. D., Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review, 99*(1)*,* 45-77.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407-428.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In W. W. Cohen, A. K. McCallum, & S. T. Roweis (Eds.), *Proceedings of the 25th Annual International Conference on Machine Learning* (pp. 160-167). New York, NY: Association for Computing Machinery.

Çöltekin, Ç., & Rama, T. (2018). Tübingen-Oslo at SemEval-2018 Task 2: SVMs perform better than RNNs in emoji prediction. In M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, & M. Carpuat (Eds.), *Proceedings of the 12th International Workshop on Semantic Evaluation* (pp. 34-38). Stroudsburg, PA: Association for Computational Linguistics.

Coltheart, M., Davelaar, E., Jonasson, J.T. and Besner, D. (1977). Access to the internal lexicon. In S. Dornick (ed.), *Attention and performance (Vol. 6*, pp. 535-556). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coman, A. C., Nechaev, Y., & Zara, G. (2018). Predicting emoji exploiting multimodal data: FBK participation in ITAmoji Task. In T. Casselli, N. Novielli, V. Patti, & P.

Rosso (Eds.), *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.
Coman, A. C., Zara, G., Nechaev, Y., Barlacchi, G., & Moschitti, A. (2018). Exploiting deep neural networks for tweet-based emoji prediction. In P. Basile, V. Basile, D. Croce, F. Dell'Orletta, & M. Guerini (Eds.), *Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence co-located with 17th International Conference of the Italian Association for Artificial*. Aachen, Germany: CEUR.
Cong, J., & Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews*, *11*(4), 598-618.
Cramer, P. (1968). *Word association*. New York, NY; London, United Kingdom: Academic Press.
Danguecan, A. N., & Buchanan, L. (2016). Semantic neighborhood effects for abstract vs concrete words. *Frontiers in Psychology*, *7*(1034), 1-15.
De Deyne, S., & Storms, G. (2008a). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*(1), 213-231.
De Deyne, S., & Storms, G. (2008b). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*(1), 198-205.
De Deyne, S., Kenett, Y. N., Anaki, D., Faust, M., & Navarro, D. J. (2016). Large-scale network representations of semantics in the mental lexicon. In M. N. Jones (ed.), *Big Data in Cognitive Science* (pp. 174-202).
De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*(2), 480-498.
De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2018). Visual and affective grounding in language and mind. https://doi.org/10.31234/osf.io/q97f8
De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2012). Strong structure in weak semantic similarity: A graph based account. In N. Miyake (Ed.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1464-1469). Red Hook, NY: Curran Associates.
De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987-1006.
De Deyne, S., Verheyen, S., & Storms, G. (2016). Structure and organization of the mental lexicon: A network approach derived from syntactic dependency relations and word associations. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard, & B. Job (Eds.), *Towards a theoretical framework for analyzing complex linguistic networks* (pp. 47-79). New York, NY; Dordrecht, Netherlands; London, United Kingdom: Springer-Verlag Berlin Heidelberg.
Deese, J. (1966). *The structure of associations in language and thought*. Baltimore, MD: Johns Hopkins University Press.
Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283-321.
Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255). Red Hook, NY: Curran Associates.
Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience*, *33*(48), 18906-18916.

Devlin, J., Gonnerman, L., Andersen, E., & Seidenberg, M. S. (1997). Category specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, *10*, 77–94.

Dhillon, P., Rodu, J., Foster, D., & Ungar, L. (2012). Two step CCA: A new spectral method for estimating vector models of words. In J. C Langford & J. Pineau (Eds.), *Proceedings of the 29th International Conference on Machine Learning* (pp. 67-74). Madison, WI: Omnipress.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, *298*(5596), 1191-1194.

Duñabeitia, J. A., Avilés, A., & Carreiras, M. (2008). NoA's ark: Influence of the number of associates in visual word recognition. *Psychonomic Bulletin & Review*, *15*(6), 1072-1077.

Ebert, K. D., & Kohnert, K. (2011). Sustained attention in children with primary language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1372-1384.

Eich, E., Macaulay, D., & Ryan, L. (1994). Mood dependent memory for events of the personal past. *Journal of Experimental Psychology: General*, *123*(2), 201-215.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3-4), 169-200.

Ellis Weismer, S., & Thordardottir, E. (2002). Cognition and language. In P. Accardo, A. Capute, & B. Rogers (Eds.), *Disorders of Language Development* (pp. 21-37). Timonium, MD: York Press.

Ellis Weismer, S., Evans, J., & Hesketh, L. J. (1999). An examination of verbal working memory capacity in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *42*(5), 1249-1260.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Plunkett, K., & Parisi, D. (1998). *Rethinking innateness: A connectionist perspective on development* Cambridge, MA: MIT Press.

Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in Cognitive Sciences*, *12*(10), 388-396.

Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the. Mathematical Institute of the Hungarian Academy of Sciences*, *5*(1), 17-60.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635-653.

Ettinger, A., & Linzen, T. (2016). Evaluating vector space models using human semantic priming results. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 72-77). Stroudsburg, PA: Association for Computational Linguistics.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *52*(2), 321-335.

Fairhall, S. L., & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience*, *33*(25), 10552-10558.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality- specificity and emergent category-specificity. *Journal of Experimental Psychology: General, 120*(4)*,* 339-357.

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 30-35). Stroudsburg, PA: Association for Computational Linguistics.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1616–1626). Stroudsburg, PA: Association for Computational Linguistics.

Fellbaum, C. (1998). A semantic network of English: The mother of all WordNets. In P. Vossen (Ed.), *EuroWordNet: A multilingual database with lexical semantic networks* (pp. 137-148). Dordrecht, Netherlands: Kluwer Academic Publishers.

Feng, Y., & Lapata, M. (2010). Visual information in semantic representation. In R. Kaplan, J. Burstein, M. Harper, & G. Penn (Eds.), *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 91-99). Stroudsburg, PA: Association for Computational Linguistics.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short form versions of the MacArthur communicative development inventories. *Applied Psycholinguistics*, *21*, 95-115.

Ferrer i Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*(1482), 2261-2265.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116-131.

Finneran, D. A., Francis, A. L., & Leonard, L. B. (2009). Sustained attention in children with specific language impairment (SLI). *Journal of Speech, Language, and Hearing Research*, *52*(4), 915-929.

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin*, *117*(1), 39-66.

Fountain, T., & Lapata, M. (2010). Meaning representation in natural language categorization. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1916-1921). Red Hook, NY: Curran Associates.

Gabriel, A., Maillart, C., Guillaume, M., Stefaniak, N., & Meulemans, T. (2011). Exploration of serial structure procedural learning in children with language impairment. *Journal of the International Neuropsychological Society*, *17*(2), 336-343.

Gabriel, A., Maillart, C., Stefaniak, N., Lejeune, C., Desmottes, L., & Meulemans, T. (2013). Procedural learning in specific language impairment: Effects of sequence complexity. *Journal of the International Neuropsychological Society*, *19*(3), 264-271.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. In A. F. Famili, J. N. Kok, J. M. Pena, A. Siebes, & A. Feelders (Eds.), *Proceeding of the 2005 International Symposium on Intelligent Data Analysis* (pp. 121-132). New York, NY; Dordrecht, Netherlands; London, United Kingdom: Springer-Verlag Berlin Heidelberg.

Gasper, K., & Clore, G. L. (2002). Attending to the big picture: Mood and global vs local processing of visual information. *Psychological Science*, *13*(1), 34-40.

Gee, N. R., Nelson, D. L., & Krawczyk, D. (1999). Is the concreteness effect a result of underlying network interconnectivity?. *Journal of Memory and Language*, *40*(4), 479-497.

Gentner, D. (2006). Why verbs are hard to learn. In K. A. Hirsh-Pasek & Roberta M. G. (Eds.), *Action meets word: How children learn verbs* (pp. 544-564). Oxford, United Kingdom: Oxford University Press.

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2173–2182). Red Hook, NY: Curran Associates.

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

Gladkova, A., & Drozd, A. (2016). Intrinsic evaluations of word embeddings: What can we do better?. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 36-42). Stroudsburg, PA: Association for Computational Linguistics.

Glenberg, A. M., Graesser, A. C., & de Vega, M. (Eds.). (2008). *Symbols and embodiment: debates on meaning and cognition*. Oxford, United Kingdom: Oxford University Press.

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *50*(1), 177-195.

Griffin, L. D., Wahab, M. H., & Newell, A. J. (2013). Distributional learning of appearance. *PLOS One*, *8*(2), e58074.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357-364.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science*, *18*(12), 1069-1076.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244.

Gruenenfelder, T. M. (1986). Relational similarity and context effects in category verification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(4), 587.

Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (2015). Graph-theoretic properties of networks based on word association norms: Implications for models of lexical semantic memory. *Cognitive Science*, *40*(6), 1460-1495.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006-1033.

Haebig, E., Saffran, J. R., & Ellis Weismer, S. (2017). Statistical word learning in children with autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry*, *58*(11), 1251-1263.

Hahn, U. E., & Ramscar, M. E. (2001). *Similarity and categorization*. Oxford, United Kingdom: Oxford University Press.

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: Controversies and consensus. *Trends in Cognitive Sciences*, *16*(9), 458-466.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*(12), 2639-2664.

Hargreaves, I. S., & Pexman, P. M. (2014). Get rich quick: The signal to respond

procedure reveals the time course of semantic richness effects during visual word recognition. *Cognition*, *131*(2), 216-242.
Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.
Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.
Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301-307.
Henry, L. A., & Botting, N. (2017). Working memory and developmental language impairments. *Child Language Teaching and Therapy*, *33*(1), 19-32.
Henry, L. A., Messer, D. J., & Nash, G. (2015). Executive functioning and verbal fluency in children with language difficulties. *Learning and Instruction*, *39*, 137-147.
Hill, F., & Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In A. Moschitti, B. Pang, & W. Daelemans. (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 255-265). Stroudsburg, PA: Association for Computational Linguistics.
Hill, F., Korhonen, A., & Bentz, C. (2014). A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*, *38*(1), 162-177.
Hill, F., Reichart, R., & Korhonen, A. (2014). Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, *2*, 285-296.
Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.
Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431-440.
Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science*, *7*(3), 513-534.
Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(6), 1331-1356.
Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*(1), 74-95.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.
Hoeffner, J. H., & McClelland, J. L. (1993). Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation. In E. V. Clark (Ed.), *Proceedings of the 25th Child Language Research Forum* (pp. 38-49). Stanford, CA: Center for the Study of Language and Information.
Hoffman, P., & Woollams, A. M. (2015). Opposing effects of semantic diversity in lexical and semantic relatedness decisions. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 385-402.
Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718-730.
Hollis, G. (2017). Estimating the average need of semantic knowledge from distributional semantic models. *Memory & Cognition*, *45*(8), 1350-1370.

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.

Horst, J. S., McMurray, B., & Samuelson, L. K. (2006). Online processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (p. 339-344). Mahwah, NJ: Lawrence Erlbaum Associates.

Houghton, G. (2005). *Connectionist models in cognitive psychology*. Hove, United Kingdom: Psychology Press.

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. (2011). Constructing semantic representations from a gradually changing representation of temporal context. *Topics in Cognitive Science*, *3*(1), 48-73.

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, *53*(2), 258-276.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (*Vol. 1: Long Papers*, pp. 873-882). Stroudsburg, PA: Association for Computational Linguistics.

Huettel, S. A., Song, A. W., & McCarthy, G. (2014). *Functional magnetic resonance imaging* (3rd ed.). Sunderland, MA: Sinauer Associated.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., Yap, M. J., Bengson, J. J., Niemeyer, D., & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099-1114.

Isen, A. M., Daubman, K. A., & Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology*, *52*(6), 1122-1131.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37-50.

Jimenez, E., & Hills, T. (2017). Network analysis of a large sample of typical and late talkers. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 2302-2307). Red Hook, NY: Curran Associates.

Joanisse, M. F., & Seidenberg, M. S. (2003). Phonology and syntax in specific language impairment: Evidence from a connectionist model. *Brain and Language*, *86*, 40-56.

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103-120.

Johns, B. T., Jamieson, R. K., & Jones, M. N. (2020). The continued importance of theory: Lessons from big data approaches to cognition. In S. E. Woo, R. Proctor, & L. Tay (Eds.), *Big data methods for psychological research: New horizons and challenges* (pp. 277-295). Washington, DC: APA.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review*, *26*(1), 103-126.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1-37.

Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2018). In defense of spatial models of semantic representation. *New Ideas in Psychology*, *50*, 54-60.
Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *66*(2), 115-124.
Jones, M. N., Willits, & J., Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of mathematical and computational psychology* (pp. 232-254). Oxford, United Kingdom: Oxford University Press.
Kacmajor, M., & Kelleher, J. D. (2019). Capturing and measuring thematic relatedness. *Language Resources and Evaluation*, 1-38.
Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition*, *30*(6), 823-840.
Kail, R., & Leonard, L. (1986). *Word-finding abilities in children with specific language impairment*. Rockville, MD: American Speech-Language-Hearing Association.
Kajić, I., & Eliasmith, C. (2018). *Evaluating the psychological plausibility of word2vec and GloVe distributional semantic models*. Ontario, Canada: University of Waterloo. (Report number: CTN-TR-20180824-012)
Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1476), 2109-2128.
Karlgren, J., Holst, A., & Sahlgren, M. (2008). Filaments of meaning in word space. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. White (Eds.), *Proceedings of 30th European Conference on Information Retrieval Research* (pp. 531-538). New York, NY; Dordrecht, Netherlands; London, United Kingdom: Springer-Verlag Berlin Heidelberg.
Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience, 8*(407), 1-16.
Kenett, Y. N., Wechsler-Kashi, D., Kenett, D. Y., Schwartz, R. G., Ben Jacob, E., & Faust, M. (2013). Semantic organization in children with cochlear implants: Computational analysis of verbal fluency. *Frontiers in Psychology*, *4*, 543.
Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287-304.
Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology*, *48*(1), 171-184.
Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, *87*(1), 184-193.
Kidd, E., & Kirjavainen, M. (2011). Investigating the contribution of procedural and declarative memory to the acquisition of past tense morphology: Evidence from Finnish. *Language and Cognitive Processes*, *26*(4-6), 794-829.
Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In A. Moschitti, B. Pang, & W. Daelemans. (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 36-45). Stroudsburg, PA: Association for Computational Linguistics.
Kiela, D., & Clark, S. (2014). A systematic study of semantic vector space model parameters. In A. Allauzen, R. Bernardi, E. Grefenstette, H. Larochelle, C. Manning, S. W. Yih (Eds.), *Proceedings of the 2nd Workshop on Continuous*

*Vector Space Models and their Compositionality* (pp. 21-30). Stroudsburg, PA: Association for Computational Linguistics.

Kiela, D., Hill, F., Korhonen, A., & Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Vol. 2: Short Papers*, pp. 835-841). Stroudsburg, PA: Association for Computational Linguistics.

Kiela, D., Verő, A. L., & Clark, S. C. (2016). Comparing data sources and architectures for deep visual representation learning in semantics. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 447-456). Stroudsburg, PA: Association for Computational Linguistics.

Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics* (p. 1367-1373). Stroudsburg, PA: Association for Computational Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, & W. Daelemans. (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1746-1751). Stroudsburg, PA: Association for Computational Linguistics.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, *42*(1), 21-40.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies* (pp. 153-165). Edinburgh, United Kingdom: University Press.

Kleemans, T., Segers, E., & Verhoeven, L. (2011). Precursors to numeracy in kindergartners with specific language impairment. *Research in Developmental Disabilities*, *32*(6), 2901-2908.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). Centrality indices. In U. Brandes & T. Erlebach (Eds.), *Network analysis: Methodological foundations* (pp. 16-61). New York, NY: Springer-Verlag Berlin Heidelberg.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14-34.

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.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401-412.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, *4*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 1097-1105). Neural Information Processing Systems Foundation.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990.

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), 211-240.

Lapesa, G., & Evert, S. (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, *2*, 531-546.

Lapesa, G., Evert, S., & Schulte Im Walde, S. (2014). Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In J. Bos, A. Frank, & R. Navigli (Eds.), *Proceedings of the Third Joint Conference on Lexical and Computational Semantics* (pp. 160-170). Association for Computational Linguistics and Dublin City University.

Law, J., Rush, R., Schoon, I., & Parsons, S. (2009). Modeling developmental language difficulties from school entry into adulthood: Literacy, mental health, and employment outcomes. *Journal of Speech, Language, and Hearing Research*, *52*, 1401-1416.

Lazaridou, A., Bruni, E., & Baroni, M. (2014). Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Vol. 1: Long Papers*, pp. 1403-1414). Stroudsburg, PA: Association for Computational Linguistics.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal Skip-gram model. In R. Mihalcea, J. Chai, & A. Sarkar (Eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 153-163). Stroudsburg, PA: Association for Computational Linguistics.

Lebret, R., & Collobert, R. (2014). Word embeddings through Hellinger PCA. In S. Wintner, S. Goldwater, & S. Riezler (Eds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 482–490). Stroudsburg, PA: Association for Computational Linguistics.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In M. Nagao & Y. Wilks (Eds.), *Proceedings of the 15th Conference on Computational Linguistics* (*Vol. 1*, pp. 622-628).

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151-171.

Lenci, A., Baroni, M., Cazzolli, G., & Marotta, G. (2013). BLIND: A set of semantic feature norms from the congenitally blind. *Behavior Research Methods*, *45*(4), 1218-1233.

Leonard , L. , Ellis Weismer , S. , Miller , C. , Francis , D. , Tomblin , J. B. , & Kail , R. (2007). Speed of processing, working memory, and language impairment in children. *Journal of Speech, Language, and Hearing Research*, *50*, 408-428.

Leonard, L. B. (2014). *Children with specific language impairment*. Cambridge, MA: MIT Press.

Leonard, L. B., Deevy, P., Fey, M. E., & Bredin-Oja, S. L. (2013). Sentence comprehension in specific language impairment: A task designed to distinguish between cognitive capacity and syntactic complexity. *Journal of Speech, Language, and Hearing Research*, *56*(2), 577-589.

Leshed, G., & Kaye, J. J. (2006). Understanding how bloggers feel: Recognizing affect in blog posts. In G. Olson & R. Jeffries (Eds.), *CHI'06 Extended Abstracts on*

*Human Factors in Computing Systems* (pp. 1019-1024). Stroudsburg, PA: Association for Computational Linguistics.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago, IL: University of Chicago Press.

Levy, O., & Goldberg, Y. (2014a). Dependency-based word embeddings. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Vol. 2: Short Papers*, pp. 302-308). Stroudsburg, PA: Association for Computational Linguistics.

Levy, O., & Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 2177-2185). Neural Information Processing Systems Foundation.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211-225.

Li, P., Hastie, T. J., & Church, K. W. (2006). Very sparse random projections. In T. Eliassi-Rad, L. H. Ungar, M. Craven, & D. Gunopulos (Eds.), *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 287-296). New York, NY: Association for Computing Machinery.

Li, P., Zhao, X., & Mac Whinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science, 31*(4), 581-612.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273-302.

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, *10*(3), 573-589.

Louwerse, M. M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 309-326). Oxford, United Kingdom: Oxford University Press.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91-110.

Lowe, W., & McDonald, S. (2000). The direct route: Mediated priming in semantic space. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 806-811). Mahwah, NJ: Lawrence Erlbaum Associates.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*(2), 81-95.

Lum, J. A., & Kidd, E. (2012). An examination of the associations among multiple memory systems, past tense, and vocabulary in typically developing 5-year-old children. *Journal of Speech, Language, and Hearing Research*, *55*(4), 989-1006.

Lum, J. A., Conti-Ramsden, G., & Lindell, A. K. (2007). The attentional blink reveals sluggish attentional shifting in adolescents with specific language impairment. *Brain and Cognition*, *63*(3), 287-295.

Lum, J. A., Conti-Ramsden, G., Page, D., & Ullman, M. T. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex*, *48*(9), 1138-1154.

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

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (*Vol. 1*, pp. 142-150). Stroudsburg, PA: Association for Computational Linguistics.

Maki, W. S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*(3), 598-603.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57-78.

Mawhood, L., Howlin, P., & Rutter, M. (2000). Autism and developmental receptive language disorder - A comparative follow-up in early adult life. I: Cognitive and language outcomes. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *41*, 547-559.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348-356.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology*, *11*(1), 1-37.

McDonald, S. (2000). *Environmental determinants of lexical processing effort*. Unpublished PhD Thesis, University of Edinburgh.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295-322.

McGregor, K. K., Berns, A. J., Owen, A. J., Michels, S. A., Duff, D., Bahnsen, A. J., & Lloyd, M. (2012). Associations between syntax and the lexicon among children with or without ASD and language impairment. *Journal of Autism and Developmental Disorders*, *42*(1), 35-47.

McGregor, K. K., Oleson, J., Bahnsen, A., & Duff, D. (2013). Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. *International Journal of Language & Communication Disorders*, *48*(3), 307-319.

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.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99-130.

Mehler, A. (2008). Large text networks as an object of corpus linguistic studies. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 328-382). Berlin, Germany: De Gruyter Mouton.

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000). Characterizing the hemodynamic response: Effects of presentation rate,

sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, *11*(6), 735-759.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In J. Bengio & Y. LeCun (Eds.), *Proceedings of Workshop at the International Conference on Learning Representations* (pp. 1–12).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 3111-3119). Neural Information Processing Systems Foundation.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In L. Vanderwende, H. Daumé III, & K. Kirchhoff (Eds.), *Proceedings of the 2013 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746-751). Stroudsburg, PA: Association for Computational Linguistics.

Miller, E. K., Freedman, D. J., & Wallis, J. D. (2002). The prefrontal cortex: Categories, concepts and cognition. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1424), 1123-1136.

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 65-79.

Mirman, D., Landrigan, J. F., & Britt, A. E. (2017). Taxonomic and thematic semantic systems. *Psychological Bulletin*, *143*(5), 499-520.

Mirsky, A., Anthony, B., Duncan, C., Ahearn, M., & Kellam, S. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review*, *2*, 109-145.

Mishne, G. (2005). Experiments with mood classification in blog posts. In S. Argamon, J. Karlgren, & J. Shanahan (Eds.), *Proceedings of the ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access* (pp. 321-327). New York, NY: Association for Computing Machinery.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191-1195.

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 2265-2273). Neural Information Processing Systems Foundation.

Moffat, M., Siakaluk, P. D., Sidhu, D. M., & Pexman, P. M. (2015). Situated conceptualization and semantic processing: Effects of emotional experience and context availability in semantic categorization and naming tasks. *Psychonomic Bulletin & Review*, *22*(2), 408-419.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*Vol. 1: Long Papers*, pp. 174-184). Stroudsburg, PA: Association for Computational Linguistics.

Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301-326.

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods*, *45*(2), 440-461.

Montgomery, J. W. (2008). Role of auditory attention in the real-time processing of simple grammar by children with specific language impairment: A preliminary investigation. *International Journal of Language & Communication Disorders*, *43*(5), 499-527.

Montgomery, J. W., Evans, J. L., & Gillam, R. B. (2009). Relation of auditory attention and complex sentence comprehension in children with specific language impairment: A preliminary study. *Applied Psycholinguistics*, *30*(1), 123-151.

Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cognitive Science*, *37*(1), 125-145.

Murphy, B., Talukdar, P., & Mitchell, T. (2012). Selecting corpus-semantic models for neurolinguistic decoding. In E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, & D. Yuret (Eds.), *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (*Vol. 1*, pp. 114-123). Stroudsburg, PA: Association for Computational Linguistics.

Nelson, D. I., & Goodmon, L. B. (2002). Experiencing a word can prime its accessibility and its associative connections to related words. *Memory & Cognition*, *30*(3), 380-398.

Nelson, D. L., Dyrdal, G. M., & Goodmon, L. B. (2005). What is preexisting strength? Predicting free association probabilities, similarity ratings, and cued recall probabilities. *Psychonomic Bulletin & Review*, *12*(4), 711-719.

Nelson, D. L., Kitto, K., Galea, D., McEvoy, C. L., & Bruza, P. D. (2013). How activation, entanglement, and searching a semantic network contribute to event memory. *Memory & Cognition*, *41*(6), 797-819.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review*, *105*(2), 299-324.

Newcombe, P. I., Campbell, C., Siakaluk, P. D., & Pexman, P. M. (2012). Effects of emotional and sensorimotor knowledge in semantic processing of concrete and abstract nouns. *Frontiers in Human Neuroscience*, *6, 1245*.

Nobre, K., Nobre, A. C., & Kastner, S. (Eds.). (2014). *The Oxford handbook of attention*. Oxford, United Kingdom: Oxford University Press.

Obeid, R., Brooks, P. J., Powers, K. L., Gillespie-Lynch, K., & Lum, J. A. (2016). Statistical learning in specific language impairment and autism spectrum disorder: A meta-analysis. *Frontiers in Psychology*, *7*, *1245*.

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*(3), 466-478.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1717-1724). Red Hook, NY: Curran Associates.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, *97*(3), 315-331.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space

models. *Computational Linguistics*, *33*(2), 161-199.
Paivio, A. (1971). *Imagery and verbal processes*. New York, NY: Holt, Rinehart, and Winston.
Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, England: Oxford University Press.
Palermo, D. S., & Jenkins, J. J. (1964). *Word association norms: Grade school through college.* Minneapolis, MN: University of Minnesota Press
Pecher, D., Boot, I., & Van Dantzig, S. (2011). Abstract concepts: Sensory-motor grounding, metaphors, and beyond. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 54, pp. 217-248). Cambridge, MA: Academic Press.
Peelen, M. V., & Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, *93*(1), 603-608.
Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532-1543). Stroudsburg, PA: Association for Computational Linguistics.
Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3-4), 175-190.
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., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: Concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, *49*(2), 407-417.
Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience* (*Vol. 1*, pp. 3-31). New York, NY: Academic Press
Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science*, *21*(2), e12549.
Ponari, M., Norbury, C. F., Rotaru, A. S., Lenci, A., & Vigliocco, G. (2018). Learning abstract words and concepts: Insights from developmental language disorder. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170140.
Poria, S., Gelbukh, A., Cambria, E., Hussain, A., & Huang, G. B. (2014). EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, *69*, 108-123.
Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93-134.
Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*(3), 195-237.
Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, *29*(9), 2352-2449.
Recchia, G., & Jones, M. N. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, *6, 315*.
Reddy, L., Tsuchiya, N., & Serre, T. (2010). Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, *50*(2), 818-825.
Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In R. Witte, H. Cunningham, J. Patrick, E. Beisswanger, E. Buyko, U. Hahn, K. Verspoor, & A. R. Coden (Eds.), *Proceedings of the LREC 2010*

*Workshop on New Challenges for NLP Frameworks* (pp. 46-50). Paris, France: European Language Resources Association.

Reisinger, J., & Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In R. Kaplan, J. Burstein, M. Harper, & G. Penn (Eds.), *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 109-117). Stroudsburg, PA: Association for Computational Linguistics.

Riener, C. R., Stefanucci, J. K., Proffitt, D. R., & Clore, G. (2011). An effect of mood on the perception of geographical slant. *Cognition and Emotion*, *25*(1), 174-182.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303-345.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*(1-3), 107-142.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205-235.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, *8*, 627-633.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906-914.

Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Rossini, P. M., Caltagirone, C., Castriota-Scanderbeg, A., Cicinelli, P., Del Gratta, C., Demartin, M., Pizzella, V., Traversa, R., & Romani, G. L. (1998). Hand motor cortical area reorganization in stroke: A study with fMRI, MEG and TCS maps. *NeuroReport*, *9*(9), 2141-2146.

Rotaru, A. S., & Vigliocco, G. (2020). Constructing semantic models from words, images, and emojis. *Cognitive Science*, *44*(4), e12830.

Rotaru, A. S., Vigliocco, G. (2019). Modelling semantics by integrating linguistic, visual and affective information. In In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the Forty-First Annual Conference of the Cognitive Science Society* (pp. 2681-2687). Red Hook, NY: Curran Associates.

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2016). From words to behaviour via semantic networks. In A. Papafragou, D. J. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the Thirty-Eighth Annual Conference of the Cognitive Science Society* (pp. 2207-2212). Red Hook, NY: Curran Associates.

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the structure and dynamics of semantic processing. *Cognitive Science*, *42*(8), 2890-2917.

Rubin, D. C., & Talarico, J. M. (2009). A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory*, *17*(8), 802-808.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(3), 273-294.

Sack, A. T. (2009). Parietal cortex and spatial cognition. *Behavioural Brain Research*, *202*(2), 153-161.

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52-61.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics, 20*(1), 33-53.

Salle, A., Idiart, M., & Villavicencio, A. (2016). Enhancing the LexVec distributed word representation model using positional contexts and external memory. *arXiv preprint arXiv:1606.01283*.

Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flow from one brain to another during gestural communication. *Proceedings of the National Academy of Sciences*, *107*(20), 9388-9393.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In L. Màrquez, C. Callison-Burch, & J. Su (Ed.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 298-307). Stroudsburg, PA: Association for Computational Linguistics.

Schul, R., Stiles, J., Wulfeck, B., & Townsend, J. (2004). How 'generalized'is the 'slowed processing' in SLI? The case of visuospatial attentional orienting. *Neuropsychologia*, *42*(5), 661-671.

Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand?. In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 223-250). Hillsdale, NJ: Lawrence Erlbaum Associates.

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), 82-102.

Schwarz, N., & Clore, G. L. (1996). Feelings and phenomenal experiences. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 433-465). New York, NY: Guilford Press.

Schwarzkopf, D. S., Song, C., & Rees, G. (2011). The surface area of human V1 predicts the subjective experience of object size. *Nature Neuroscience*, *14*(1), 28-30.

Serfozo, R. (2009). *Basics of applied stochastic processes*. Berlin, Germany: Springer Science & Business Media.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*(2), 190-195.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*(2), 393-413.

Sheng, L., & McGregor, K. K. (2010). Lexical-semantic organization in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *53*(1), 146-159.

Siew, C. S., Wulff, D. U., Beckage, N. M., & Kenett, Y. N. (2019). Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2108423.

Sigman, M., & Cecchi, G. A. (2002). Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences*, *99*(3), 1742-1747.

Silberer, C., & Lapata, M. (2012). Grounded models of semantic representation. In J. Tsujii, J. Henderson, & M. Paşca (Eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and*

*Computational Natural Language Learning* (pp. 1423-1433). Stroudsburg, PA: Association for Computational Linguistics.

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 721-732). Stroudsburg, PA: Association for Computational Linguistics.

Silberer, C., Ferrari, V., & Lapata, M. (2013). Models of semantic representation with visual attributes. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*Vol. 1: Long Papers*, pp. 572-582). Stroudsburg, PA: Association for Computational Linguistics.

Silberer, C., Ferrari, V., & Lapata, M. (2017). Visually grounded meaning representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2284-2297.

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the International Conference on Learning Representations* (pp. 1-14).

Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision* (p. 1470-1477). Los Alamitos, CA: IEEE Computer Society.

Solé, R. V., Corominas-Murtra, B., Valverde, S., & Steels, L. (2010). Language networks: Their structure, function, and evolution. *Complexity*, *15*(6), 20-26.

Solomon, K. O., & Barsalou, L. W. (2004). Perceptual simulation in property verification. *Memory and Cognition*, *32*(2), 244-259.

Spaulding, T. J., Plante, E., & Vance, R. (2008). Sustained selective attention skills of preschool children with specific language impairment: Evidence for separate attentional capacities. *Journal of Speech, Language, and Hearing Research*, *51*(1), 16-34.

Speed, L J., Vinson, D. P., Vigliocco, G. (2015). Representing meaning. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (*Vol. 39*, pp. 190-211). Berlin, Germany: De Gruyter Mouton.

St. Clair, M., Pickles, A., Durkin, K., & Conti-Ramsden, G. (2011). A longitudinal study of behavioral, emotional and social difficulties in individuals with a history of specific language impairment. *Journal of Communication Disorders*, *44*, 186-199.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598-605.

Stevens, C., Fanning, J., Coch, D., Sanders, L., & Neville, H. (2008). Neural mechanisms of selective auditory attention are enhanced by computerized training: Electrophysiological evidence from language-impaired and typically developing children. *Brain Research*, *1205*, 55-69.

Stevens, C., Sanders, L., & Neville, H. (2006). Neurophysiological evidence for selective auditory attention deficits in children with specific language impairment. *Brain Research*, *1111*(1), 143-152.

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*(3), 234-243.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41-78.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer* (pp. 237-249). Washington, DC: American Psychological Association Press.

Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *Journal of Neuroscience*, *29*(5), 1565-1572.

Strapparava, C., & Valitutti, A. (2004). WordNet Affect: An affective extension of WordNet. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation* (*Vol. 4*, pp. 1083-1086). European Language Resources Association.

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review*, *18*(3), 605-611.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In H. Bischof, D. Forsyth, C. Schmid, & S. Sclaroff (Eds.), *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-9). Red Hook, NY: Curran Associates.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Vol. 1: Long Papers*, pp. 1556-1566). Stroudsburg, PA: Association for Computational Linguistics.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers*, pp. 1555-1565). Stroudsburg, PA: Association for Computational Linguistics.

Thawani, A., Srivastava, B., & Singh, A. (2019). SWOW-8500: Word association task for intrinsic evaluation of word embeddings. In A. Rogers, A. Drozd, A. Rumshisky, & Y. Goldberg (Eds.), *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP* (pp. 43-51). Stroudsburg, PA: Association for Computational Linguistics.

Thomas, M. S., & Karmiloff-Smith, A. (2003). Modeling language acquisition in atypical phenotypes. *Psychological Review, 110*(4), 647-682.

Thomas, M. S., & McClelland, J. L. (2008). Connectionist models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 23-58). Cambridge, United Kingdom: Cambridge University Press.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1245-1260.

Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., & Dyer, C. (2015). Evaluation of word vector representations by subspace alignment. In L. Màrquez, C. Callison-Burch, & J. Su (Ed.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2049-2054). Stroudsburg, PA: Association for Computational Linguistics.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141-188.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352.
Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1), 231-270.
Ullman, M. T., & Pierpont, E. I. (2005). Specific language impairment is not specific to language: The procedural deficit hypothesis. *Cortex*, *41*(3), 399-433.
Utsumi, A. (2015). A complex network approach to distributional semantic models. *PLOS One*, *10*(8), e0136277.
Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, *44*(6), e12844.
Van Den Heuvel, M. P., & Pol, H. E. H. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, *20*(8), 519-534.
Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.
Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289-335.
Van Rensbergen, B., Storms, G., & De Deyne, S. (2015). Examining assortativity in the mental lexicon: Evidence from word associations. *Psychonomic Bulletin & Review*, *22*(6), 1717-1724.
Vigliocco, G., Kousta, S. T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. *Cerebral Cortex*, *24*(7), 1767-1777.
Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(2), 219-247.
Vilnis, L., & McCallum, A. (2014). Word representations via Gaussian embedding. In In J. Bengio & Y. LeCun (Eds.), *Proceedings of the 3rd International Conference on Learning Representations* (pp. 1-12).
Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183-190.
Vinson, D. P., Ponari, M., & Vigliocco, G. (2014). How does emotional content affect lexical processing?. *Cognition & Emotion*, *28*(4), 737-746.
Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In E. Dykstra-Erickson & M. Tscheligi (Eds.) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 319-326). New York, NY: Association for Computing Machinery.
Vugs, B., Cuperus, J., Hendriks, M., & Verhoeven, L. (2013). Visuospatial working memory in specific language impairment: A meta-analysis. *Research in Developmental Disabilities*, *34*(9), 2586-2597.
Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2016). Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (*Vol. 1: Long Papers*, pp. 1671-1682).
Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191-1207.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440-442.

Weckerly, J., Wulfeck, B., & Reilly, J. (2001). Verbal fluency deficits in children with specific language impairment: Slow rapid naming or slow to name?. *Child Neuropsychology*, *7*(3), 142-152.

Wiemer-Hastings, K., & Xu, X. (2005). Content differences for abstract and concrete concepts. *Cognitive Science*, *29*(5), 719-736.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In R. J. Mooney (Ed.), *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 347-354). Stroudsburg, PA: Association for Computational Linguistics.

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.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979.

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

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

Zorzi, M., Testolin, A., & Stoianov, I. P. (2013). Modeling language and cognition with deep unsupervised learning: A tutorial overview. *Frontiers in Psychology*, *4, 515*.

# Appendix A

## LSA

The LSA model (Landauer & Dumais, 1997) was created as a computational solution to the "poverty of the stimulus" problem. When applied to semantics, this problem refers to the fact that humans have a much better knowledge of word meanings, than what can be derived from direct linguistic experience. As a concrete example, the authors note that a typical American seventh grader learns the meaning of 10-15 new words each day, while vocabulary tests suggest that he/she should acquire at most 3 new words each day, based on the amount of new text to which he/she is exposed. In order to account for this phenomenon, the authors put forward the hypothesis that word learning is largely based on inductive reasoning, which is supported by a process of dimensionality reduction over a semantic space. This process is the core of the LSA model, and its role is to compute latent, indirect, and higher-order associations between words.

The distributional information used by the LSA model, in order to create vector representations, is stored in a term-document matrix $\boldsymbol{A}$. The rows of $\boldsymbol{A}$ correspond to the words in the model's vocabulary, while the columns of $\boldsymbol{A}$ index the documents that make up the training corpus, such that each entry $\boldsymbol{A}_{\mathbf{ij}}$ counts how many times word *i* appears in document *j*. The element of $\boldsymbol{A}$ are then weighted using two functions, one local and one global. The local weighting function is a logarithmic transformation of the local occurrence frequency, meant to ensure that terms which are repeated very often within a document do not have too large an effect on the semantic representations. The global weighting function depends on the frequency of a given word within the entire corpus, as well as on how the occurrences of that word are distributed across all the documents in the corpus, and is meant to measure how informative the occurrence of that word is, in the context of a particular document. By combining and applying the two functions, the matrix $\boldsymbol{A}'$ is computed, the elements of which are defined as follows:

$$\boldsymbol{A'_{ij} = \log_2(A_{ij} + 1)\left(1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}\right)}$$

where $\boldsymbol{p_{ij} = A_{ij} / \sum_k A_{ik}}$, and *n* denotes the total number of documents.

The matrix $\boldsymbol{A'}$ is then decomposed using singular value decomposition:

$$\boldsymbol{\mathbf{A'}_{m \times n} = U_{m \times z} S_{z \times z} (V_{n \times z})^T}$$

where *z* = min(*m*, *n*), the diagonal values of $\boldsymbol{S}$ are the singular values of $\boldsymbol{A'}$, and the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are the left-singular vector and right-singular vectors of $\boldsymbol{A'}$, respectively. For the purpose of dimensionality reduction, only the first *d* columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are retained, resulting in the matrices $\boldsymbol{U'}$ and $\boldsymbol{V'}$. The rows of $\boldsymbol{U'}$ correspond to word representations, while the rows of $\boldsymbol{V'}$correspond to vector representations of the documents. The final word representations are obtained as the rows of the matrix $\boldsymbol{U' \cdot S'}$, where $\boldsymbol{S'}$ consists of the first d rows and columns of $\boldsymbol{S}$.

## HAL and related models

The HAL model (Hyperspace Analogue to Language; Lund & Burgess, 1996) was proposed as a solution to the problem of building psychologically plausible semantic spaces. Its authors note that constructing such spaces traditionally involves selecting a number of cognitively meaningful semantic dimensions, and then asking participants to rate a set of word along each chosen dimension. This approach has (at least) two problems. Firstly, the researcher must choose a set of dimensions that provide a sufficiently detailed picture of semantics, and hope that participants can reliably assign a position for each word, within the resulting space. Secondly, collecting such behavioural norms is expensive, in terms of both time and money, which means that the researcher must settle for a relatively small number of words and/or dimensions. Both these problems can be solved by using a distributional model, since running such a model requires minimal intervention from the researcher, the dimensions of meaning are guaranteed to be meaningful and are selected automatically by the model, and the vocabulary of the model can easily include tens to hundreds of thousands words.

The representations generated by the HAL model are stored in a matrix $\boldsymbol{X}$, such that each entry $\boldsymbol{X}_{\mathbf{ij}}$ counts the co-occurrences of words *i* and *j*, where *j* appears before *i*. Moreover, each co-occurrence is weighted inversely proportional to the number of words that separate *i* and *j*. The use of this linear ramp weighting scheme is motivated by the assumption that words which are strongly semantically related are produced in closer proximity than words word which share only a weak semantic connection. The model computes co-occurrences both before and after each target word, which means that the actual word vectors are obtained by concatenating the corresponding rows and columns of $\boldsymbol{X}$. Finally, in order to reduce the level of noise in the word representations, the model retains only a few hundred of the most variant vector dimensions, and the vectors are normalised to constant length.

One important problem for “context word” models, such as HAL, is related to the fact that the distribution of word frequencies in language is highly skewed. More specifically, the most frequent words by far are function words, such as “the”, “a” and “I”, which carry very little semantic information. As a result, distributional models will

indicate a high similarity between semantically unrelated words, such as “book” and “plank”, simply because they frequently co-occur with the same function words. This means that the models do not distinguish between syntax-based co-occurrences (“the food”), and semantics-based co-occurrences (“spicy food”). In order to solve this problem, almost all “context word” models employ a lexical association function, as a means of reducing the effect of uninformative co-occurrences. In the rest of our discussion, we will look at various models derived from the HAL model, and describe their lexical association function.

HiDEx (High Dimensional Explorer; Shaoul & Westbury, 2006) was created starting from the observation that, in HAL, the structure of a word’s semantic neighbourhood is highly correlated with the frequency of that word. Since different corpora have different frequency distributions, especially for words of medium and low frequency, this means that semantic neighbourhoods depend strongly on the choice of corpus, which is an undesirable feature. As a solution for this shortcoming, the authors first apply the linear ramp weighting scheme from the HAL model, resulting in the matrix $\boldsymbol{X}$, and then concatenate the corresponding rows and columns of $\boldsymbol{X}$, in order to account for co-occurrences both before and after each target word. Next, they normalize the elements in each word vector, by dividing each term $\boldsymbol{X}_{\mathbf{ij}}$ by the frequency of word *j*. The result of the normalisation is that the structure of the semantic neighbourhoods becomes only very weakly correlated with word frequency.

COALS (Correlated Occurrence Analogue to Lexical Semantics; Rohde et al., 2005) is another attempt to reduce the effects that very frequently occurring function words have on semantic representations. Similar to the HAL model, a linear ramp weighting scheme is used, but the distinction between words occurring before a target word, and words occurring after it, is no longer retained. The lexical association function employed by the authors is based on Pearson correlation, and is defined as follows:

$$\boldsymbol{X}'_{\mathbf{ij}} = \frac{\boldsymbol{S}\boldsymbol{X}_{\mathbf{ij}} - \sum_{\mathbf{b}} \boldsymbol{X}_{\mathbf{ib}} \cdot \sum_{\mathbf{a}} \boldsymbol{X}_{\mathbf{aj}}}{\sqrt{\left(\sum_{\mathbf{b}} \boldsymbol{X}_{\mathbf{ib}} \cdot (\boldsymbol{S} - \sum_{\mathbf{b}} \boldsymbol{X}_{\mathbf{ib}}) \cdot \sum_{\mathbf{a}} \boldsymbol{X}_{\mathbf{aj}} \cdot \left(\boldsymbol{S} - \sum_{\mathbf{a}} \boldsymbol{X}_{\mathbf{aj}}\right)\right)}}$$

where $\boldsymbol{S} = \sum_{\mathbf{a},\mathbf{b}} \boldsymbol{X}_{\mathbf{ab}}$. If $\boldsymbol{X}'_{\mathbf{ij}}$ is positive, then it is replaced with $\sqrt{\boldsymbol{X}'_{\mathbf{ij}}}$, in order to increase the importance of small values, relative to large ones; otherwise, $\boldsymbol{X}'_{\mathbf{ij}}$ is replaced with 0, since negative values carry little semantic information. The authors find that the

COALS model is considerably better than the HAL model, in accounting for word-pair similarity ratings and performance in multiple choice vocabulary tests, which suggests a diminished negative effect of high frequency, syntax-based co-occurrences.

## Topic

The Topic model (Griffiths, Steyvers, & Tenenbaum, 2007) provides a probabilistic solution to three important and pervasive problems in semantic processing, namely those of (1) prediction (i.e., predicting the next word in a sequence, in order to facilitate its retrieval), (2) disambiguation (i.e., identifying the sense of a word, that is implied by the context of that word), and (3) gist extraction (i.e., determining the gist of a set of words).

In order to tackle these problems, they are first given a formal definition, as follows. Let $w = (w_1, w_2, \ldots, w_n)$ be a sequence of $n$ words. It is assumed that this sequence is generated by a latent semantic structure $l = (g, z)$, where $g$ corresponds to the gist of $w$, and $z = (z_1, z_2, \ldots, z_n)$ are the senses attributed to the words in $w$. The senses correspond to topics, where a topic is a probability distribution defined over words, and $g$ is a probability distribution defined over topics. In this context, prediction refers to predicting $w_{n+1}$ from $w$ (i.e., computing $\boldsymbol{P}(\boldsymbol{w}_{\mathbf{n+1}} |\boldsymbol{w})$) while disambiguation and gist extraction require inferring $z$ and $g$ from $w$ (i.e., computing $\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{w})$ and $\boldsymbol{P}(\mathbf{g}|\boldsymbol{w})$, respectively. All three probabilities can be calculated based on the joint distribution over words and latent semantic structures, $\boldsymbol{P}(\boldsymbol{w}, \boldsymbol{l}) = \boldsymbol{P}(\boldsymbol{w}|\boldsymbol{l})\boldsymbol{P}(\boldsymbol{l})$.

The Topic model computes this joint distribution by following a generative modelling approach, which involves learning structured probability distributions. Generative models describe how a particular set of data is obtained by following a causal chain consisting of probabilistic steps. Through statistical inference, the steps of the model can be reversed, in order to derive the most likely structure $l$ that generates the data w. Using Bayes's rule, $\boldsymbol{P}(\boldsymbol{l}|\boldsymbol{w})$ can be expressed as:

$$\boldsymbol{P}(\boldsymbol{l}|\boldsymbol{w}) = \frac{\boldsymbol{P}(\boldsymbol{w}|\boldsymbol{l})\boldsymbol{P}(\boldsymbol{l})}{\boldsymbol{P}(\mathbf{w})}, \mathbf{where}\ \boldsymbol{P}(\boldsymbol{w}) = \sum_{\boldsymbol{l}} \boldsymbol{P}(\mathbf{w}|\boldsymbol{l})\boldsymbol{P}(\boldsymbol{l})$$

Starting from this formula, the probabilities for prediction, disambiguation and gist extraction can be derived as follows:

$$P(w_{n+1}|w) = \sum_{l} P(w_{n+1}|l,w)P(l|w)$$

$$P(z|w) = \sum_{g} P(l|w)$$

$$P(g|w) = \sum_{z} P(l|w)$$

The model outputs the observed data *w*, based on the latent structure *l*, by using a sequence of probabilistic steps. More specifically, Topic assumes that the gist *g* of a sequence of words (or document) w is a probability distribution over *T* topics (i.e., word senses), where each topic is a probability distribution over words. Each word from the sequence *w* is generated by first sampling a gist *g*, then sampling a topic $z_i$ from the distribution provided by *g*, and finally by sampling the word $w_i$ from the distribution defined by $z_i$. This process is depicted in Figure 30, for a sequence consisting of four words.

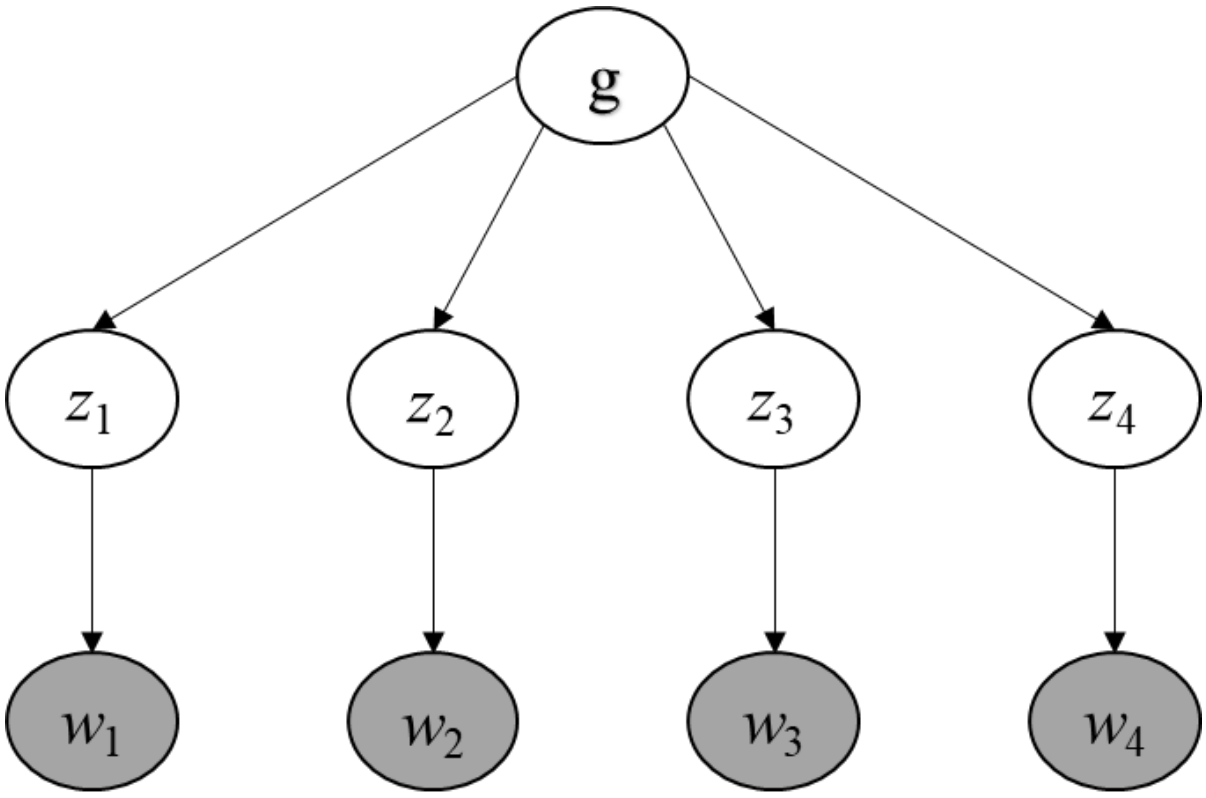


Figure 30. Toy example of the Topic model. A document is generated by choosing a distribution over topics that reflects the gist of the document, *g*, choosing a topic $z_i$ for each potential word from a distribution determined by *g*, and then choosing the actual word $w_i$ from a distribution determined by $z_i$. Adapted from Griffiths, Steyvers, & Tenenbaum (2007).

Given the gist *g* of a document, the probability of the *i*th word can be computed as follows:

$$\boldsymbol{P}(\boldsymbol{w}_\mathbf{i}|\boldsymbol{g}) = \sum_{\boldsymbol{z}_\mathbf{i}=\mathbf{1}}^{\boldsymbol{T}} \boldsymbol{P}(\boldsymbol{w}_\mathbf{i}|\boldsymbol{z}_\mathbf{i})\boldsymbol{P}(\boldsymbol{z}_\mathbf{i}|\boldsymbol{g})$$

The probability $\boldsymbol{P}(\boldsymbol{w}|\boldsymbol{z})$ measures how relevant word *w* is to the topic *z*, while the probability $\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{g})$ indicates the prevalence of topic *z* within the document. For instance, in a topic about computers, relevant words such as “data”, “e-mail” and “website” would have high probabilities $\boldsymbol{P}(\boldsymbol{w}|\boldsymbol{z})$, whereas irrelevant words such as “cat”, “dog” and “pet” would be assigned low probabilities. If computers are one of the main topics for a given document, then that topic would receive a high probability $\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{g})$. Otherwise, the value of $\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{g})$ would be low.

In order to generate the observed data (i.e., the documents), the Topic model relies on latent Dirichlet allocation (Blei, Ng, & Jordan, 2003). Given a set of documents, the gist of each document, *g*, is sampled from a multinomial distribution over the *T* topics, with parameters $\boldsymbol{\theta}^{(\mathbf{d})}$, such that for a word in document *d*, $\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{g}) = \boldsymbol{\theta}_\mathbf{z}^{(\mathbf{d})}$. The *z*th topic is sampled from a multinomial distribution over the *W* words in the vocabulary, with parameters $\boldsymbol{\phi}^{(\mathbf{z})}$, such that $\boldsymbol{P}(\boldsymbol{w}|\boldsymbol{z}) = \boldsymbol{\phi}_\mathbf{w}^{(\mathbf{z})}$. Finally, $\boldsymbol{\theta}^{(\mathbf{d})}$ is a symmetric $\mathbf{Dirichlet}(\boldsymbol{\alpha})$ prior over all the documents, while $\boldsymbol{\phi}^{(\mathbf{z})}$ is a symmetric $\mathbf{Dirichlet}(\boldsymbol{\beta})$ prior over all the topics. The complete generative model can be described as follows:

$$\boldsymbol{w}_\mathbf{i}|\boldsymbol{z}_\mathbf{i}, \boldsymbol{\phi}^{(\mathbf{z_i})} \sim \mathbf{Discrete}\left(\boldsymbol{\phi}^{(\mathbf{z_i})}\right)$$

$$\boldsymbol{\phi}^{(\mathbf{z})} \sim \mathbf{Dirichlet}(\boldsymbol{\beta})$$

$$\boldsymbol{z}_\mathbf{i}|\boldsymbol{\theta}^{(\mathbf{d_i})} \sim \mathbf{Discrete}(\boldsymbol{\theta}^{(\mathbf{d_i})})$$

$$\boldsymbol{\theta}^{(\mathbf{d})} \sim \mathbf{Dirichlet}(\boldsymbol{\alpha})$$

Assuming a set of topic parameters $\boldsymbol{\phi}$, the probabilities computed for prediction, disambiguation and gist extraction can be expressed in the following manner:

$$\boldsymbol{P}(\boldsymbol{w}_{\mathbf{n+1}}|\boldsymbol{w};\boldsymbol{\phi}) = \sum_{\mathbf{z},\mathbf{z}_{\mathbf{n+1}}} \boldsymbol{P}(\boldsymbol{w}_{\mathbf{n+1}}|\boldsymbol{z}_{\mathbf{n+1}};\boldsymbol{\phi})\boldsymbol{P}(\boldsymbol{z}_{\mathbf{n+1}}|\boldsymbol{z})\,\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{w};\boldsymbol{\phi})$$

$$\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{w};\boldsymbol{\phi}) = \frac{\boldsymbol{P}(\boldsymbol{w},\boldsymbol{z}|\boldsymbol{\phi})}{\sum_{\mathbf{z}} \boldsymbol{P}(\boldsymbol{w},\boldsymbol{z}|\boldsymbol{\phi})}$$

$$\boldsymbol{P}(\boldsymbol{g}|\boldsymbol{w};\boldsymbol{\phi}) = \sum_{\mathbf{z}} \boldsymbol{P}(\boldsymbol{g}|\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z}|\boldsymbol{w};\boldsymbol{\phi})$$

## CBOW and Skip-gram

The CBOW (Continuous Bag-Of-Words) and Skip-gram models (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) are shallow neural networks, consisting of an input layer, a hidden layer, and an output layer. The objective of the CBOW model is to predict the occurrence of a given word, based on that word's context (i.e., the words that precede and follow it). Conversely, the objective of the Skip-gram model is to predict the context of a given word, based on its occurrence. More exactly, assuming a sequence of words $w_1$, $w_2$, …, and a sliding window of size $k$, the learning objective of the CBOW model, with respect to the word $w_t$, is to maximize the probability:

$$\boldsymbol{P(w_t | w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k})}$$

The learning objective of the Skip-gram model is to maximize the probability:

$$\boldsymbol{P(w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k} | w_t)}$$

The structure of the CBOW model is shown in Figure 31 (left).

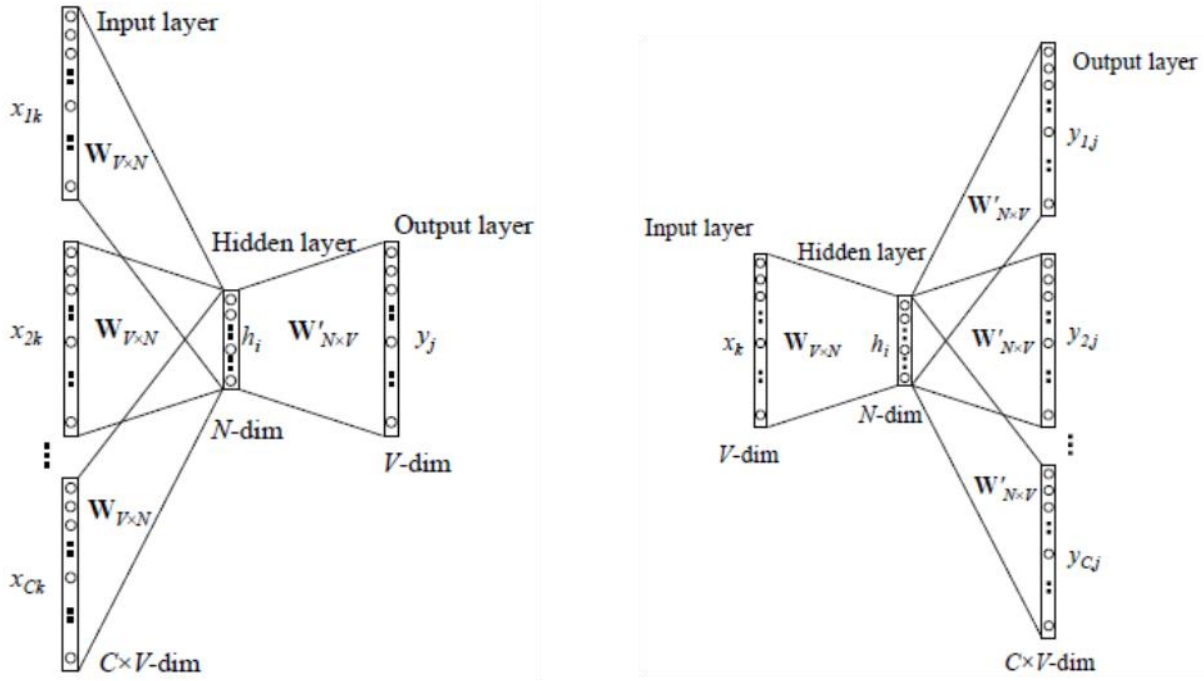


Figure 31. The CBOW model (left) and the Skip-gram model (right). The models have a vocabulary of size *V* and a context of size *C*. The weights from the input layer to the hidden layer are stored in the matrix *W*, while the weights from the hidden layer to the output layer are stored in the matrix *W'*. Adapted from Rong (2014).

The input layer is made up of $C$ vectors, each of size $V$. Each context word $cw_i$ is represented as a one-hot $V \times 1$ vector $x_i$, which has a value of 1 at the $i^{th}$ position, and a value of 0 at all the other positions. The weights between the input layer and the hidden layer are stored in the $V \times N$ matrix W. For every context word $cw_i$, the $i^{th}$ row of W contains the $N$-dimensional vector representation of that word, denoted by $vec_i$.

The hidden layer has a linear activation function (i.e., $g(x) = x$), and its activation $h$ can be computed as follows:

$$\boldsymbol{h} = \frac{\mathbf{1}}{\boldsymbol{C}} \boldsymbol{W^T} \left(\boldsymbol{x_{i_1}} + \boldsymbol{x_{i_2}} + \cdots + \boldsymbol{x_{i_C}}\right)$$
$$= \frac{\mathbf{1}}{\boldsymbol{C}} \left(\boldsymbol{vec_{i_1}} + \boldsymbol{vec_{i_2}} + \cdots + \boldsymbol{vec_{i_C}}\right)$$

In other words, $h$ is the average of the word vectors corresponding to the $C$ input (i.e., context) words, namely $\boldsymbol{cw_{i_1}}, \boldsymbol{cw_{i_2}}, \ldots, \boldsymbol{cw_{i_C}}$. The weights between the hidden layer and the output layer are stored in the $N \times V$ matrix $W'$.

The output layer has a softmax activation function, which is defined as follows, for any vector $z = [z_1, z_2, \ldots, z_F]$:

$$\mathbf{softmax(z)} = \left[\frac{\mathbf{exp(z_1)}}{\sum_j \mathbf{exp(z_j)}}, \frac{\mathbf{exp(z_2)}}{\sum_j \mathbf{exp(z_j)}}, \ldots, \frac{\mathbf{exp(z_F)}}{\sum_j \mathbf{exp(z_j)}}\right]$$

The activation y of the output layer can be computed as follows:

$$\boldsymbol{y} = \mathbf{softmax}(\boldsymbol{W'^T h})$$

The activation in the $i^{th}$ unit (i.e., position) in $y$ represents the estimated probability of occurrence for the $i^{th}$ word in the model's vocabulary, given a context made up of the words $\boldsymbol{cw_{i_1}}, \boldsymbol{cw_{i_2}}, \ldots, \boldsymbol{cw_{i_C}}$.

The structure of the Skip-gram model is shown in Figure 31 (right). The matrices $W$ and $W'$ have the same meaning as in the CBOW model, namely that of representing the weights from the input layer to the hidden layer, and from the hidden layer to the output layer, respectively. Also, the activation functions are the same as for the CBOW model, namely linear, for the hidden layer, and softmax, for the output layer.

The input to the network is a one-hot $V \times 1$ vector $x_i$, which has a value of 1 at the $i^{th}$ position, and a value of 0 at all the other positions.

The activation h of the hidden layer can be computed as follows:

$$\boldsymbol{h} = \boldsymbol{W}^{\boldsymbol{T}}\, \boldsymbol{x}_{\mathbf{i}} = \boldsymbol{vec}_{\mathbf{i}}$$

where $vec_i$ is the $i^{th}$ row of $W$, which contains the $N$-dimensional vector representation of the input word $w_i$. Finally, the output layer consists of $C$ vectors, $y_1$ through $y_C$, each of size $V$. The values in each vector $y_i$ can be computed as follows:

$$\boldsymbol{y}_{\mathbf{i}} = \mathbf{softmax}(\mathbf{W}'^{\boldsymbol{T}}\mathbf{h})$$

The activation in the $j^{th}$ unit (i.e., position) in $y_i$ corresponds to the estimated probability of that the $i^{th}$ context word, namely $cw_i$, is the $j^{th}$ word in the model's vocabulary.

The empirical success and computational efficiency of CBOW and Skip-gram depend crucially on fine-tuning a number of model hyperparameters (Levy, Goldberg, & Dagan, 2015). The first two hyperparameters are window size and learning rate. Window size refers to the maximum possible distance between the co-occurrence of any two words in a given text, such as $w_1$ and $word_2$, in order for the pair ($w_1$, $w_2$) to be counted as a co-occurrence by the models and to be processed during training (i.e., by teaching the model to predict the occurrence of $w_2$, given that of $w_1$, and vice-versa). Learning rate quantifies the amount of information extracted from each training example: small rates mean that the model needs to encounter many instances of the pair ($w_1$, $w_2$) in order for it to learn that $w_1$ and $w_2$ are related, whereas large rates have the opposite effect[27].

Other important parameters are the number of negative samples and the frequency-based subsampling threshold. Negative sampling refers to how the models employ "negative" information. For instance, when encountering a word pair such as ("bee", "hive"), besides strengthening the association between "bee" and "hive" (i.e., by incorporating "positive" information), the models also weaken the association between "bee"/"hive" and other words which do not co-occur with them, such as "guerilla" or "axiom" (i.e., by incorporating "negative" information). Therefore, the number of negative samples is a measure of the amount of "negative" information generated and employed when training the model. Frequency-based subsampling refers to manner in which the models process frequent words during training. For

[27] The latter outcome might appear desirable, but it creates a very strong recency effect, since encountering new information effectively "deletes" the old information, rather than integrating the two types of information.

infrequent words (e.g., “hierophany”), each context in which that word occurs provides valuable information. In contrast, for frequent words (e.g., “kitchen”), after encountering a relatively large number of contexts, the subsequent contexts add very little new information (for instance, after reading a thousand books on the topic of “football”, you are unlikely to learn anything novel and of significance by reading yet another volume). As a result, each time a given word (e.g., $w_i$) is encountered in context, the two models discard the resulting training information with a probability given by the following formula:

$$\boldsymbol{P_{subsamp}(w_i) = \min\left(1 - \sqrt{\frac{thr}{f(w_i)}}, 1\right)}$$

where $\boldsymbol{f(w_i)}$ is the relative frequency of $w_i$, computed over the training corpus, and *thr* is the subsampling threshold, such that frequent words (i.e., with relative frequency greater than *thr*) are strongly subsampled. High values for *thr* mean that the vast majority of words are not subsampled (i.e., the model is very sensitive to frequency, with a preference towards learning about frequent words), whereas low values mean that most of the words are subsampled (i.e., the model is relatively insensitive to frequency, incorporating comparable amounts of information about both frequent and infrequent words).

## GloVe

The GloVe model (Global Vectors; Pennington et al., 2014) was proposed as a solution to certain (potential) shortcomings of two classes of popular distributional models, namely global matrix factorization models, such as LSA, and local context window models, such as CBOW and Skip-gram. According to the authors, the first class of models perform poorly on word analogy tasks, denoting a sub-optimal vector space structure, while the second class of models fail to exploit global co-occurrence information.

The information used by the GloVe model is stored in the global co-occurrence matrix $\boldsymbol{X}$, where the values $\boldsymbol{X_{ij}}$ indicate the number of times that word $w_j$ occurs in the context of word $w_i$. Let $\boldsymbol{X_i} = \sum_k \boldsymbol{X_{ik}}$ denote the number of times any word appears in the context of $w_i$. Also, let $\boldsymbol{P_{ij}} = \boldsymbol{P}(\boldsymbol{j}|\boldsymbol{i}) = \boldsymbol{X_{ij}} / \boldsymbol{X_i}$ be the probability that $w_j$ appears in the context of $w_i$.

One of the design principles of the GloVe model is that it should be sensitive to probability ratios, since they contain useful semantic information. For instance, let $w_i$ = “ice” and $w_j$ = “steam”. In order to determine how the two words relate, we can look at their co-occurrence probabilities with other (probe) words, denoted by $w_k$. For words $w_k$ related to “ice” but not “steam” (e.g., $w_k$ = “solid”), the ratio $\boldsymbol{P_{ik}} / \boldsymbol{P_{jk}}$ should be large, while for words $w_k$ related to steam but not ice (e.g., $w_k$ = “gas”), the same ratio should be small. Instead, for words $w_k$ that are either related or unrelated to both “ice” and “steam” (e.g., $w_k$ = “water” and $w_k$ = “fashion”, respectively), the ratio should be close to 1. Therefore, these ratios are useful for discriminating between $w_i$ and $w_j$.

Another guiding principle is that vector offsets correspond to directions of meaning. This was demonstrated by Mikolov and collaborators (Mikolov, Chen, et al., 2013; Mikolov, Yih, & Zweig, 2013), who found that simple word vector arithmetic, such as vector subtraction, can capture meaningful relations between words. This has been shown for semantic relations (e.g., *vec*(“king”) - *vec*(“queen”) ≈ *vec*(“man”) - *vec*(“woman”)), as well as for syntactic ones (e.g., *vec*(“biggest”) - *vec*(“big”) ≈

*vec*(“smallest”) - *vec*(“small”))[28]. Additionally, a later study (Vylomova et al., 2016) found that certain lexical relations can be predicted with nearly perfect accuracy, based on vector differences.

Based on these two principles, the authors define the GloVe model starting from the following equation:

$$\boldsymbol{F}\left(\boldsymbol{v}_{\mathbf{i}}, \boldsymbol{v}_{\mathbf{j}}, \boldsymbol{v}'_{\mathbf{k}}\right) = \frac{\boldsymbol{P}_{\mathbf{ik}}}{\boldsymbol{P}_{\mathbf{jk}}}$$

where $\boldsymbol{v}_{\mathbf{i}}, \boldsymbol{v}_{\mathbf{j}} \in \mathbb{R}^{\mathbf{d}}$ are *d*-dimensional vector representations of $w_i$ and $w_j$, while $\boldsymbol{v}'_{\mathbf{k}} \in \mathbb{R}^{\mathbf{d}}$ is a *d*-dimensional context representation of (probe) word $w_k$. Given a function $\boldsymbol{F}$, computing the vectors of interest $\boldsymbol{v}_{\mathbf{i}}$, $\boldsymbol{v}_{\mathbf{j}}$, and $\boldsymbol{v}'_{\mathbf{k}}$ simply amounts to solving the previous equation. The function $\boldsymbol{F}$, which defines the model, is obtained by establishing a number of desirable properties that such a function should have.

The function should be sensitive to vector offsets, and relate them to probability ratios, which means that $\boldsymbol{F}$ should only depend on the difference between the vectors $\boldsymbol{v}_{\mathbf{i}}$ and $\boldsymbol{v}_{\mathbf{j}}$. This leads to the following new equation:

$$\boldsymbol{F}\left(\boldsymbol{v}_{\mathbf{i}} - \boldsymbol{v}_{\mathbf{j}}, \boldsymbol{v}'_{\mathbf{k}}\right) = \frac{\boldsymbol{P}_{\mathbf{ik}}}{\boldsymbol{P}_{\mathbf{jk}}}$$

Furthermore, the function *F* should depend on a linear relation between $\boldsymbol{v}_{\mathbf{i}} - \boldsymbol{v}_{\mathbf{j}}$ and $\boldsymbol{v}'_{\mathbf{k}}$, resulting in the following refinement:

$$\boldsymbol{F}\left((\boldsymbol{v}_{\mathbf{i}} - \boldsymbol{v}_{\mathbf{j}})^{\boldsymbol{T}} \boldsymbol{v}'_{\mathbf{k}}\right) = \frac{\boldsymbol{P}_{\mathbf{ik}}}{\boldsymbol{P}_{\mathbf{jk}}}$$

where $(\boldsymbol{v}_{\mathbf{i}} - \boldsymbol{v}_{\mathbf{j}})^{\boldsymbol{T}} \boldsymbol{v}'_{\mathbf{k}}$ computes the dot product between $\boldsymbol{v}_{\mathbf{i}} - \boldsymbol{v}_{\mathbf{j}}$ and $\boldsymbol{v}'_{\mathbf{k}}$. Also, within the word co-occurrence matrix $\boldsymbol{X}$, there is no formal distinction between a word and a context word, which means that the two should have interchangeable roles. As a result, it should be possible to exchange $\boldsymbol{v}$ for $\boldsymbol{v}'$, as well as $\boldsymbol{X}$ for $\boldsymbol{X}^{\boldsymbol{T}}$. In order to enforce this, the authors first impose that $\boldsymbol{F}$ must be a group homomorphism between $(\mathbb{R}, +)$ and $(\mathbb{R}_{>\mathbf{0}}, \cdot)$, (i.e., $\boldsymbol{F}(\boldsymbol{a} + \boldsymbol{b}) = \boldsymbol{F}(\boldsymbol{a}) \cdot \boldsymbol{F}(\boldsymbol{b})$, for any real numbers *a* and *b*). This restriction leads to the following equation:

[28] The vector representation of a word *w* in a given distributional model is denoted as *vec*(*w*).

$$F\left((v_i - v_j)^T v'_k\right) = \frac{F(v_i{}^T v'_k)}{F\left(v_j{}^T v'_k\right)} = \frac{P_{ik}}{P_{jk}}$$

One solution to the previous equation is to have:

$$F(v_i{}^T v'_k) = \exp(v_i{}^T v'_k) = P_{ik} = \frac{X_{ik}}{X_i}$$

By taking the logarithm of the previous expression, $v_i{}^T v'_k$ can be computed as:

$$v_i{}^T v'_k = \log(P_{ik}) = \log\left(\frac{X_{ik}}{X_i}\right) = \log(X_{ik}) - \log(X_i)$$

This last equation still does not have the desired exchange symmetries, namely $v \leftrightarrow v'$ and $X \leftrightarrow X^T$, due to the term $\log(X_i)$. However, since this term does not depend on *k*, it can be incorporated into a bias $b_i$ for $v_i$. By also adding a bias $b'_k$ for $v'_k$, the symmetries now hold, leading to the new expression:

$$v_i{}^T v'_k + b_i + b'_k = \log(X_{ik})$$

Learning the representation of the current model can be achieved by minimizing the following cost function:

$$J = \sum_{i,j=1}^{|V|} \left(v_i{}^T v'_j + b_i + b'_j - \log(X_{ij})\right)^2$$

where *V* denotes the model's vocabulary. A shortcoming of this cost function is that it assigns equal weight to all co-occurrences, including those that are very rare, or even not present at all. The problem can be remedied by introducing a weighting function $f(X_{ij})$, such that the cost function becomes the following:

$$J = \sum_{i,j=1}^{|V|} f(X_{ij})\left(v_i{}^T v'_j + b_i + b'_j - \log(X_{ij})\right)^2$$

In order for it produce good results, the weighting function needs to satisfy certain conditions. Firstly, since the term $\log(X_{ij})$ is ill-defined when $X_{ij} = 0$ , it should be case that $f(0) = 0$ , and that $f(0)\log(0) = 0$ , by convention. Secondly, $f$ should be an increasing function, such that infrequent co-occurrences are associated with small weights, given that they carry little semantic information (i.e., they are likely to be noisy). Thirdly, frequent co-occurrences should not be overweighed, not to

overestimate their informativeness. Of all the possible functions that fit this description, the authors of GloVe opt for the following, based on good empirical results:

$$\boldsymbol{f(x) = \begin{cases} (x/x_{\max})^{\alpha} & \textbf{if } x < x_{\max} \\ 1 & \textbf{otherwise} \end{cases}}$$

where $\boldsymbol{x_{\max}}$ and $\boldsymbol{\alpha}$ are model parameters. The model is now complete.

One last important aspect related to the model is its time complexity. Since the cost function effectively ignores the terms where $\boldsymbol{X_{ij} = 0}$, the complexity will depend only on the number of non-zero elements in $\boldsymbol{X}$. For a vocabulary of size $|\boldsymbol{V}|$, this means that the worst-case complexity is $\boldsymbol{O}(|\boldsymbol{V}|^{\boldsymbol{2}})$. Given that $|\boldsymbol{V}|$ is usually in the order of hundreds of thousands, this would make the GloVe model more difficult to train than local context window models, such as CBOW and Skip-gram, which scale with the corpus size, $|\boldsymbol{C}|$. However, the authors show that, by finding a tighter bound on the number of non-zero elements of $\mathbf{X}$, a worst-case complexity of $\boldsymbol{O}(|\boldsymbol{C}|^{\boldsymbol{0.8}})$ is obtained, which is much better than both $\boldsymbol{O}(|\boldsymbol{V}|^{\boldsymbol{2}})$ and $\boldsymbol{O}(|\boldsymbol{C}|)$, for typical values of $|\boldsymbol{V}|$ and $|\boldsymbol{C}|$.

## Convolutional Neural Networks (CNNs) and other visual models

A wide variety of visual models, trained over image datasets, have been developed in the last 20 years. One of the earliest, as well as most popular approaches to creating visual representations (e.g., Bruni et al., 2011; Feng & Lapata, 2010; Kiela et al., 2014) is the bag-of-visual-words (Sivic & Zisserman, 2003) approach, inspired by work on linguistic, distributional models of semantics. The process of extracting visual representations involves the following steps. Firstly, highly distinctive regions within an image are detected automatically, and a low-level feature vector is associated with each region, containing information regarding elements such as edges, textures, or colors. These vectors, both within and across images, are then assigned to a number of clusters (e.g., using K-means clustering), which constitute "visual words". The final representations for each image is a count vector, such that the value at each position counts how many feature vectors, derived from that image, belong to a given cluster. Within the bag-of-visual-words approach, some of the most widely used feature vectors are obtained from SIFT descriptors (Lowe, 2004), which have the advantage of being invariant to translations, rotations and rescalings, as well as to changes in perspective and illumination.

More recent approaches to image representation rely on deep convolutional networks (e.g., AlexNet; Krizhevsky et al., 2012; GoogLeNet; Szegedy et al., 2015; VGG-19; Simonyan & Zisserman, 2014), typically trained on supervised object recognition tasks. Convolutional networks are inspired by the neuroscience of visual perception. More specifically, within the visual system, neurons are organised in a hierarchy of layers. Each neuron responds only to stimuli from a limited region of the visual field (i.e., the neuron's receptive field). Neurons from the lower levels are sensitive to relatively simple, local visual elements (e.g., oriented edges, patches of colour), while neurons higher up in the hierarchy encode more complex, global information (e.g., the outline of an object). Convolutional networks mimic this hierarchical process, by using convolutional layers. As opposed to the fully connected layers traditionally used in classical (i.e., non-convolutional) networks, where each unit of a given layer receives input from all the elements in the previous layer, convolutional

layers have units that receive input only from limited, spatially contiguous regions of the previous layer (i.e., the receptive fields of the units).

Convolutional Neural Networks (CNNs) typically consist of three types of layers, namely convolutional layers, pooling layers, and fully connected layers. Unlike in the case of classical (i.e., non-convolutional) neural networks, where each unit of a given layer typically receives input from all the elements in the previous layer, convolutional and pooling layers have units that receive input only from limited, spatially contiguous regions of the previous layer (i.e., the receptive fields of the units).

Convolutional layers have a three-dimensional structure, consisting of one or more feature maps, stacked across the depth of the layer. Each feature map is created using a filter (or kernel), represented as a three-dimensional array of weights. To calculate the activations in a given feature map, the filter associated with that feature map is slid over the previous layer, and the dot product is computed between the weights in the filter and the values in each volume covered by the filter. Let us assume that the $n^{th}$ layer has width $w_N$, height $h_n$, and depth $d_n$. Also, let the filter $\boldsymbol{F}$ in layer $n$+1 have width $w_F$, height $h_F$, and depth $d_N$. In the simplest case, the feature map $\boldsymbol{FM}$ generated by the filter $\boldsymbol{F}$ in layer $n$+1, has width $w_{FM}$ and height $h_{FM}$, and its activations can be expressed as follows:

$$\boldsymbol{FM}_{\mathbf{ij}} = \boldsymbol{b}_{\mathbf{F}} + \sum_{\mathbf{i_F=1}}^{\mathbf{h_F}} \sum_{\mathbf{j_F=1}}^{\mathbf{w_F}} \sum_{\mathbf{k_F=1}}^{\mathbf{d_n}} \mathbf{F}_{\mathbf{i_F j_F k_F}} \cdot \mathbf{A}^{[\mathbf{n}]}_{\mathbf{i_F+i-1, j_F+j-1, k_F}}$$

where $\boldsymbol{A}^{[\mathbf{n}]}$ denotes the activations from the $n^{th}$ layer, and $b_F$ is a bias associated with the filter $\boldsymbol{F}$. This expression corresponds to horizontal and vertical strides of 1, meaning that at each horizontal step, the filter is slid one position rightward, while at each vertical step, the filter is slid one position downward. In the general case, the horizontal and vertical strides can differ from 1, and the activation array $\boldsymbol{A}^{[\mathbf{n}]}$ can be padded in all directions with zeros, in order to obtain a desired size for the feature map $\boldsymbol{FM}$. Also, as mentioned previously, the activations $\boldsymbol{A}^{[\mathbf{n+1}]}$ of layer $n$+1 are obtained by stacking a given number of feature maps, depthwise. Convolutional layers implement a form of weight sharing (i.e., the weights corresponding to a filter do not change when the filter is slid over the previous layer), under the assumption that if a feature detector is useful at a given location, then the same detector should be useful at any other location. Moreover, weight sharing drastically reduce the number of free parameters for

convolutional layers, which greatly improves the computational efficiency of CNNs, making it possible to efficiently train convolutional networks with tens or even hundreds of layers.

Pooling layers are similar to convolutional layers, in that they have a three-dimensional, composed of a number of feature maps. Moreover, the feature maps are obtained by sliding a pooling window over the feature maps in the previous layer. However, there are two significant differences between pooling and computational layers. Firstly, unlike filters, which are three-dimensional, pooling operators are two-dimensional, meaning that each feature map in a pooling layer is obtained from a single, corresponding feature map in the previous layer. Secondly and most importantly, pooling operators have no weights associated with them, and they do not use the dot product in order to aggregate the activations falling within the pooling window. Instead, such operators usually compute the average or maximum of the input values. The role of pooling layers is to subsample the activations of the previous layers, in order to make the reduce the layer sizes and number of parameters for a given model, making it less computationally demanding and less prone to overfitting.

Finally, fully connected layers correspond to the typical layers of traditional neural networks. They are one-dimensional and their activations are computed as a linear combination of the activations in their previous layers.

Three of the most well-known and widely used CNNs are AlexNet (Krizhevsky et al., 2012), VGG-19 (Simonyan & Zisserman, 2014), and GoogLeNet (Szegedy et al., 2015).

## AlexNet

AlexNet (Krizhevsky et al., 2012) won first prize at ILSVRC 2012 (i.e., ImageNet Large Scale Visual Recognition Challenge), a machine learning competition which involved recognizing objects in more than one million images, covering 1000 categories. The CNN obtained an error rate of 15%, while the next best model produced a considerably higher error rate of 26%. AlexNet was one of the first models to show that deep CNNs can easily outperform non-convolutional models, and that they can be trained in a reasonable amount of time, through the use of GPUs. As a result, deep CNNs quickly became the most popular type of models used in computer vision. Some of the distinguishing characteristics of AlexNet are its use of Rectified Linear Units (ReLUs) and local response normalization. The authors show that employing the ReLU activation function results in a dramatic improvement in the number of steps needed to train their neural network, as compared to other popular activation functions, such as the logistic sigmoid or the hyperbolic tangent. Also, they find that the performance of their model can be slightly increased by performing local response normalization, a technique inspired by the lateral inhibition associated with biological neurons. This generates a competition for activation in neurons within a layer, which have the same receptive field, but belong to different feature maps. The architecture of AlexNet is summarised in Table 15.

Table 15. Main hyperparameters for the AlexNet model.

| Layer | Type | Feature maps | Size | Kernel size | Stride |
|---|---|---|---|---|---|
| 1 | Input | 3 (RGB) | 224 × 224 | – | – |
| 2 | Convolution | 96 | 55 × 55 | 11 × 11 | 4 |
| 3 | Maximum pooling | 96 | 27 × 27 | 3 × 3 | 2 |
| 4 | Convolution | 256 | 27 × 27 | 5 × 5 | 1 |
| 5 | Maximum pooling | 256 | 13 × 13 | 3 × 3 | 2 |
| 6 | Convolution | 384 | 13 × 13 | 3 × 3 | 1 |
| 7 | Convolution | 384 | 13 × 13 | 3 × 3 | 1 |
| 8 | Convolution | 256 | 13 × 13 | 3 × 3 | 1 |
| 9 | Fully connected | - | 4,096 | - | - |
| 10 | Fully connected | - | 4,096 | - | - |
| 11 | Fully connected | - | 1,000 | - | - |

## GoogLeNet

GoogLeNet (Szegedy et al., 2015) won first place at ILSVRC 2014, with an error rate of 7%. The main distinguishing characteristics of GoogLeNet are the use of inception modules, as well as that of 1×1 convolutions. Inception modules are a means of increasing the flexibility of the CNN. They combine convolutional layers with filters of different sizes (i.e., 1 × 1, 3 × 3, 5 × 5) and pooling layers, in parallel, thus capturing features at different scales. Inception modules allow information to flow through the network via a large number of partially overlapping routes, meaning that the network can learn to select the most successful path for predicting the output. Within each module, the use of 1×1 convolutions serves to reduce the dimensionality of the convolutional layers, in a non-linear manner. The structure of an inception module is shown in Figure 32.

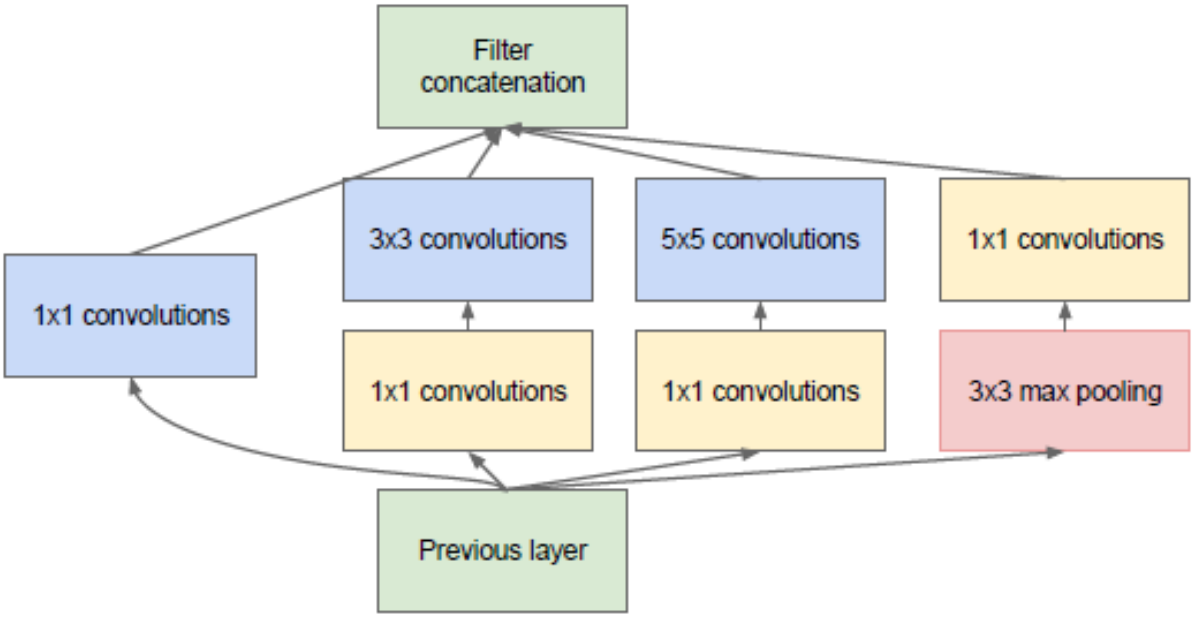


Figure 32. Structure of an inception module. Adapted from Szegedy et al. (2015).

The architecture of GoogLeNet is summarised in Table 16.

Table 16. Main hyperparameters for the GoogLeNet model.

| Layer | Type | Feature maps | Size | Kernel size | Stride |
|---|---|---|---|---|---|
| 1 | Input | 3 (RGB) | 224 × 224 | - | - |
| 2 | Convolution | 64 | 112 × 112 | 7 × 7 | 2 |
| 3 | Maximum pooling | 64 | 56 × 56 | 3 × 3 | 2 |
| 4 | Convolution | 192 | 56 × 56 | 3 × 3 | 1 |
| 5 | Maximum pooling | 192 | 28 × 28 | 3 × 3 | 2 |
| 6 | Inception (3a) | 256 | 28 × 28 | - | - |
| 7 | Inception (3b) | 480 | 28 × 28 | - | - |
| 8 | Maximum pooling | 480 | 14 × 14 | 3 × 3 | 2 |
| 9 | Inception (4a) | 512 | 14 × 14 | - | - |
| 10 | Inception (4b) | 512 | 14 × 14 | - | - |
| 11 | Inception (4c) | 512 | 14 × 14 | - | - |
| 12 | Inception (4d) | 528 | 14 × 14 | - | - |
| 13 | Inception (4e) | 832 | 14 × 14 | - | - |
| 14 | Maximum pooling | 832 | 7 × 7 | 3 × 3 | 2 |
| 15 | Inception (5a) | 832 | 7 × 7 | - | - |
| 16 | Inception (5b) | 1,024 | 7 × 7 | - | - |
| 17 | Average pooling | 1,024 | 1 × 1 | 7 × 7 | 1 |
| 18 | Fully connected | - | 1,000 | - | - |

## VGG-19

VGG-19 (Simonyan & Zisserman, 2015) won second place at ILSVRC 2014, with an error rate of 7%. The most notable characteristic of VGG-19 is its use of stacked 3 × 3 convolutions, which are the smallest that can represent the important notions of left/right, up/down, and center. Stacking them provides two main benefits. Firstly, since the receptive fields of neurons increase with each layer, a stack of small convolutions can efficiently mimic the behaviour of a single, larger convolution, by using considerably fewer parameters. Secondly, a non-linear activation function is applied to each convolutional layer in the stack, therefore increasing the complexity of the patterns that the stack can detect. The architecture of VGG-19 is summarised in Table 17.

Table 17. Main hyperparameters for the VGG-19 model.

| Layer | Type | Feature maps | Size | Kernel size | Stride |
|---|---|---|---|---|---|
| 1 | Input | 3 (RGB) | 224 × 224 | - | - |
| 2 | Convolution | 64 | 224 × 224 | 3 × 3 | 1 |
| 3 | Concolution | 64 | 224 × 224 | 3 × 3 | 1 |
| 4 | Maximum pooling | 64 | 112 × 112 | 2 × 2 | 2 |
| 5 | Convolution | 128 | 112 × 112 | 3 × 3 | 1 |
| 6 | Convolution | 128 | 112 × 112 | 3 × 3 | 1 |
| 7 | Maximum pooling | 128 | 56 × 56 | 2 × 2 | 2 |
| 8 | Convolution | 256 | 56 × 56 | 3 × 3 | 1 |
| 9 | Convolution | 256 | 56 × 56 | 3 × 3 | 1 |
| 10 | Convolution | 256 | 56 × 56 | 3 × 3 | 1 |
| 11 | Convolution | 256 | 56 × 56 | 3 × 3 | 1 |
| 12 | Maximum pooling | 256 | 28 × 28 | 2 × 2 | 2 |
| 13 | Convolution | 512 | 28 × 28 | 3 × 3 | 1 |
| 14 | Convolution | 512 | 28 × 28 | 3 × 3 | 1 |
| 15 | Convolution | 512 | 28 × 28 | 3 × 3 | 1 |
| 16 | Convolution | 512 | 28 × 28 | 3 × 3 | 1 |
| 17 | Maximum pooling | 512 | 14 × 14 | 2 × 2 | 2 |
| 18 | Convolution | 512 | 14 × 14 | 3 × 3 | 1 |
| 19 | Convolution | 512 | 14 × 14 | 3 × 3 | 1 |
| 20 | Convolution | 512 | 14 × 14 | 3 × 3 | 1 |
| 21 | Convolution | 512 | 14 × 14 | 3 × 3 | 1 |
| 22 | Maximum pooling | 512 | 7 × 7 | 2 × 2 | 2 |
| 23 | Fully connected | - | 4,096 | - | - |
| 24 | Fully connected | - | 4,096 | - | - |
| 25 | Fully connected | - | 1,000 | - | - |

## Recurrent Neural Networks (RNNs)

Recurrent neural networks (RNNs) have been developed primarily as a means of modeling the semantics of sentences and phrases. In this regard, they are very different from bag-of-words models (e.g., LSA, HAL, Skip-gram, CBOW, GloVe, etc.), which focus only on the representation of individual words and are not very sensitive to syntax. RNNs processes text in a sequential manner, and have an internal representation (via hidden layers and memory cells) of the words encountered before the current word (and, for certain architectures, after the current word). In addition, some of the most successful types of models can accurately manipulate the flow of information through the model, by employing various types of gates, which regulate the information stored in the memory cells, and the contribution of that information to the activation of the hidden layers.

Two of the most well-known and widely used RNNs are Long Short-Term Memory networks (LSTM; Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit networks (Cho, van Merriënboer, Bahdanau, & Bengio, 2014). More recenct models can also focus especially on certain important words, while discarding largely irrelevant ones, by using an attention mechanism (Bahdanau, Cho, & Bengio, 2014).

## Long Short-Term Memory (LSTM)

LSTMs distinguish themselves from classical RNNs by using cells, which implement a form of memory, as well as three types of gates, namely forget, input, and output gates.

Let $x_t$, $h_t$, and $c_t$ denote the activations of the input, hidden, and cell layers, respectively, at time step $t$. The activation of the forget gate can be computed as:

$$\boldsymbol{f_t} = \boldsymbol{\sigma}(\boldsymbol{W_f x_t} + \boldsymbol{U_f h_{t-1}} + \boldsymbol{b_f})$$

where $\boldsymbol{W_f}$ and $\boldsymbol{U_f}$ are weight matrices, and $b_f$ is a bias vector. Similarly, the activation of the input gate can be calculated as follows:

$$\boldsymbol{i_t} = \boldsymbol{\sigma}(\boldsymbol{W_i x_t} + \boldsymbol{U_i h_{t-1}} + \boldsymbol{b_i})$$

where $\boldsymbol{W_i}$ and $\boldsymbol{U_i}$ are weight matrices, and $b_i$ is a bias vector. Next, the activation of the output gate can be expressed as:

$$\boldsymbol{o_t} = \boldsymbol{\sigma}(\boldsymbol{W_o x_t} + \boldsymbol{U_o h_{t-1}} + \boldsymbol{b_o})$$

where $\boldsymbol{W_o}$ and $\boldsymbol{U_o}$ are weight matrices, and $b_o$ is a bias vector. The activation of the cell layer can be computed as follows:

$$\boldsymbol{c_t} = \boldsymbol{f_t} \odot \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \odot \mathbf{tanh}(\boldsymbol{W_c x_t} + \boldsymbol{U_c h_{t-1}} + \boldsymbol{b_c})$$

where the operator $\odot$ denotes the Hadamard product (i.e., component-wise multiplication), $\boldsymbol{W_c}$ and $\boldsymbol{U_c}$ are weight matrices, and $b_c$ is a bias vector. Finally, the activation of the hidden layer can be computed as $\boldsymbol{h_t} = \boldsymbol{o_t} \odot \mathbf{tanh}(\boldsymbol{c_t})$.

The forget gate specifies what information can be removed from the cell layer, the input gate determines what new information can be added to the cell layer, while the output gate controls what information from the cell layer contributes to the activation of the hidden layer.

## Gated Recurrent Units (GRU)

Gated Recurrent Units (GRUs; Cho et al., 2014) are a variation on LSTMs, created with the aim of being more computationally efficient. To achieve this goal, the GRU model uses only two types of gates, namely update and reset gates.

Let $x_t$ and $h_t$ denote the activations of the input and hidden layers, respectively, at time step *t*. The activation of the update gate can be computed as follows:

$$\boldsymbol{z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)}$$

where $\boldsymbol{W_z}$ and $\boldsymbol{U_z}$ are weight matrices, and $b_z$ is a bias vector. Similarly, the activation of the reset gate can be computed as follows:

$$\boldsymbol{r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)}$$

where $\boldsymbol{W_r}$ and $\boldsymbol{U_r}$ are weight matrices, and $b_r$ is a bias vector. Finally, the activation of the hidden layer can be expressed as:

$$\boldsymbol{h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h'_t}$$

where $\boldsymbol{h'_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)}$, $\boldsymbol{W_h}$ and $\boldsymbol{U_h}$ are weight matrices, and $b_h$ is a bias vector.

The reset gate determines which information from the current hidden state is no longer useful for the future, and can be reset using only the current input. Therefore, the reset gate can be employed in learning short-term dependencies. In contrast, the update gate controls which information should be carried over from the previous hidden state, to the current hidden state. Thus, the update gate allows the model to learn long-term dependencies.

## RNNs with attention

Traditionally, RNNs, such as GRUs and LSTMs, compute the representation of a sequence $x = (x_0, x_1, \ldots, x_T)$ in a step-by-step, sequential manner, by first generating the intermediate representations $h_0, h_1, \ldots, h_{T-1}$, and then using the final representation $h_T$ in order to predict the output value $y$. This process is shown in Figure 33.

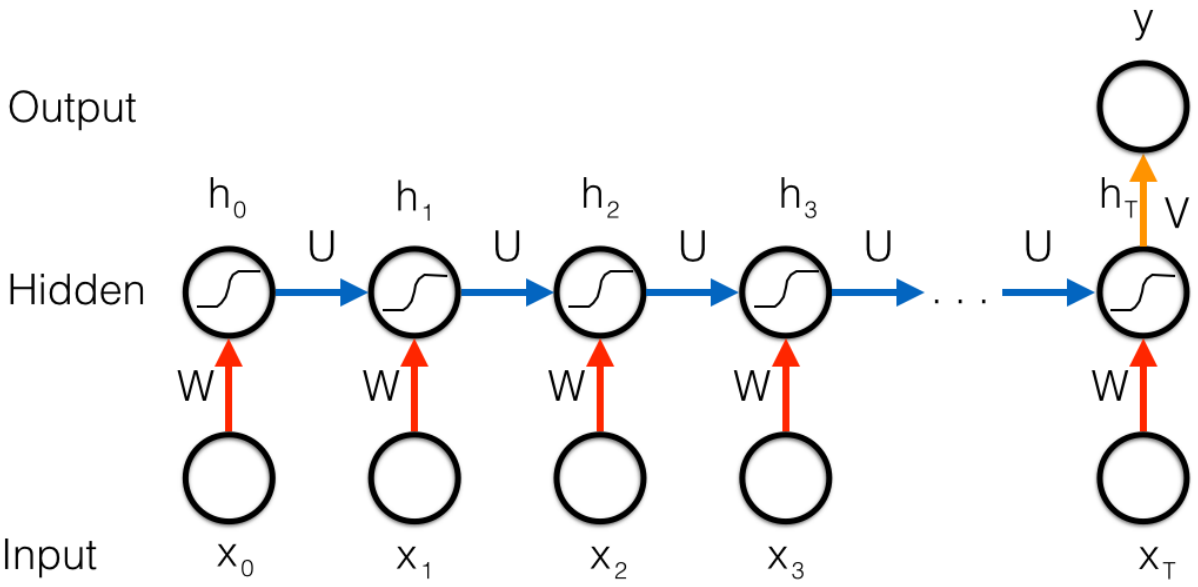


Figure 33. Flow of information through a typical RNN.

One notable shortcoming of this approach is that, for relatively long input sequences, the information carried by the first elements ($x_0$, $x_1$, $x_2$, etc.) is likely to be poorly captured by the final representation $h_T$. A straightforward solution to this problem is to employ an attention mechanism (Bahdanau et al., 2014), such that the output value $y$ is predicted based on not only the final representation $h_T$, but also all the intermediate representations $h_0, h_1, \ldots, h_{T-1}$. Also, the contributions are weighted, with the largest weights being assigned to the representations that are best at predicting $y$.

More precisely, the output value $y$ is predicted from the attention-based representation $h_{att} = a_0h_0 + a_1h_1 + \ldots + a_Th_T$, where $a_0, a_1, \ldots a_T$ are strictly positive

weights, which sum to 1. The weights can be computed using an attention layer, parameterized by the weight vector *W* and the bias value *b*, as follows:

$$\boldsymbol{a_i} = \frac{\mathbf{exp}(\boldsymbol{e_i})}{\sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{exp}(\boldsymbol{e_t})}, \mathbf{where}\ \boldsymbol{e_i} = \mathbf{tanh}(\boldsymbol{W h_i} + \boldsymbol{b})$$

# Appendix B

In order to see whether our structural and dynamic models might have a different impact for concrete and abstract words, we performed a median split on concreteness for the 2,328 words from the lexical decision task, and tested our models separately on the subset of concrete words, and that of abstract ones. The results for the lexical decision task are shown in Figure 34, Table 18, and Table 19. For log response time, the addition of the structural model improved the fit of the regression model for the concrete words (CBOW, GloVe, and LSA), as well as for the abstract words (CBOW). In the case of the dynamic model, concrete words only benefited from the inclusion of the second step (LSA). In contrast, for accuracy, the fit was significantly increased by the addition of the structural model for concrete words (CBOW, GloVe), and for abstract words (GloVe). For the dynamic model, the fit was improved only by the inclusion of step two (GloVe), for abstract words only.

The results seem to suggest that the structural models are more predictive of task performance for concrete words than for abstract words. One potential explanation for this finding is that understanding abstract words requires deeper, more elaborate processing than for concrete words.

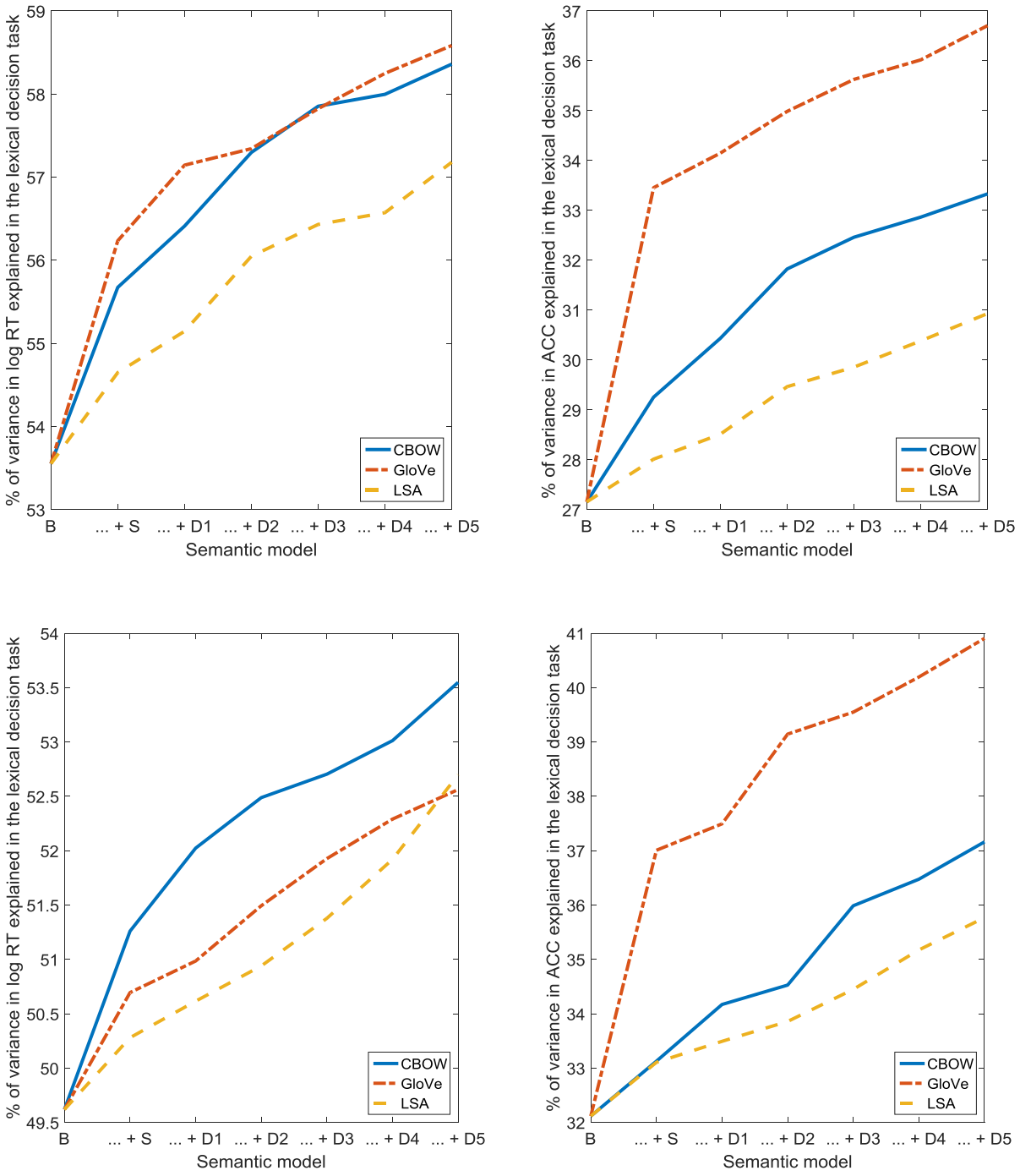


Figure 34. Percentage of variance in log response time (RT) and accuracy (ACC) in the lexical decision task, accounted for by the baseline model (B), the combination of the baseline model and the structural model (… + S), and the combination of the baseline model, the structural model, and consecutive steps of the dynamic model (… + D1 through … + D5). Top: results for concrete words only. Bottom: results for abstract words only.

Table 18. Results of model comparisons for predicting log response time and accuracy in the lexical decision task, for concrete words. B = baseline model; S = structural model; $D_{1...k}$ = first *k* individual steps of the dynamic model.

| | | Enhanced vs simple model | | | | | |
|---|---|---|---|---|---|---|---|
| Model | Statistic | B vs B + S | B + S vs B + S + $D_1$ | B + S + $D_1$ vs B + S + $D_{1...2}$ | B + S + $D_{1...2}$ vs B + S + $D_{1...3}$ | B + S + $D_{1...3}$ vs B + S + $D_{1...4}$ | B + S + $D_{1...4}$ vs B + S + $D_{1...5}$ |
| | | Log response time (lexical decision) - concrete words | | | | | |
| CBOW | F | 5.47 | 1.91 | 2.60 | 1.43 | 0.37 | 0.73 |
| | (p) | (< .0001) | (.04) | (.004) | (.16) | (.96) | (.70) |
| | df | 10, 1142 | 10, 1132 | 10, 1122 | 10, 1112 | 10, 1102 | 10, 1092 |
| GloVe | F | 7.00 | 2.71 | 0.65 | 1.30 | 1.35 | 0.88 |
| | (p) | (< .0001) | (.003) | (.77) | (.23) | (.20) | (.55) |
| | df | 10, 1142 | 10, 1132 | 10, 1122 | 10, 1112 | 10, 1102 | 10, 1092 |
| LSA | F | 2.76 | 1.09 | 2.31 | 0.96 | 0.36 | 1.53 |
| | (p) | (.002) | (.37) | (.001) | (.48) | (.96) | (.12) |
| | df | 10, 1142 | 10, 1132 | 10, 1122 | 10, 1112 | 10, 1102 | 10, 1092 |
| | | Accuracy (lexical decision) - concrete words | | | | | |
| CBOW | F | 3.39 | 1.92 | 2.30 | 1.10 | 0.66 | 0.54 |
| | (p) | (.0002) | (.04) | (.01) | (.36) | (.76) | (.86) |
| | df | 10, 1142 | 10, 1132 | 10, 1122 | 10, 1112 | 10, 1102 | 10, 1092 |
| GloVe | F | 10.81 | 1.25 | 1.47 | 0.78 | 0.63 | 0.88 |
| | (p) | (< .0001) | (.26) | (.15) | (.64) | (.79) | (.55) |
| | df | 10, 1142 | 10, 1132 | 10, 1122 | 10, 1112 | 10, 1102 | 10, 1092 |
| LSA | F | 1.36 | 0.81 | 1.46 | 0.61 | 0.78 | 0.97 |
| | (p) | (.19) | (.62) | (.15) | (.81) | (.65) | (.47) |
| | df | 10, 1142 | 10, 1132 | 10, 1122 | 10, 1112 | 10, 1102 | 10, 1092 |

Table 19. Results of model comparisons for predicting log response time and accuracy in the lexical decision task, for abstract words. B = baseline model; S = structural model; $D_{1...k}$ = first *k* individual steps of the dynamic model.

| | | Enhanced vs simple model | | | | | |
|---|---|---|---|---|---|---|---|
| Model | Statistic | B vs B + S | B + S vs B + S + $D_1$ | B + S + $D_1$ vs B + S + $D_{1...2}$ | B + S + $D_{1...2}$ vs B + S + $D_{1...3}$ | B + S + $D_{1...3}$ vs B + S + $D_{1...4}$ | B + S + $D_{1...4}$ vs B + S + $D_{1...5}$ |
| | | Log response time (lexical decision) - abstract words | | | | | |
| CBOW | F | 3.85 | 1.87 | 1.10 | 0.60 | 0.81 | 1.20 |
| | (p) | (< .0001) | (.05) | (.36) | (.81) | (.62) | (.28) |
| | df | 10, 1144 | 10, 1134 | 10, 1124 | 10, 1114 | 10, 1104 | 10, 1094 |
| GloVe | F | 2.50 | 0.75 | 1.18 | 1.03 | 0.96 | 0.55 |
| | (p) | (.006) | (.68) | (.30) | (.41) | (.48) | (.85) |
| | df | 10, 1144 | 10, 1134 | 10, 1124 | 10, 1114 | 10, 1104 | 10, 1094 |
| LSA | F | 1.52 | 0.77 | 1.73 | 0.88 | 1.23 | 1.76 |
| | (p) | (.13) | (.66) | (.07) | (.55) | (.27) | (.06) |
| | df | 10, 1144 | 10, 1134 | 10, 1124 | 10, 1114 | 10, 1104 | 10, 1094 |
| | | Accuracy (lexical decision) - abstract words | | | | | |
| CBOW | F | 1.73 | 1.77 | 0.60 | 2.52 | 1.21 | 1.26 |
| | (p) | (.07) | (.06) | (.81) | (.005) | (.28) | (.25) |
| | df | 10, 1144 | 10, 1134 | 10, 1124 | 10, 1114 | 10, 1104 | 10, 1094 |
| GloVe | F | 8.88 | 0.91 | 3.43 | 0.96 | 1.09 | 1.34 |
| | (p) | (< .0001) | (.52) | (.0002) | (.48) | (.37) | (.20) |
| | df | 10, 1144 | 10, 1134 | 10, 1124 | 10, 1114 | 10, 1104 | 10, 1094 |
| LSA | F | 1.69 | 0.61 | 0.65 | 0.86 | 1.21 | 1.16 |
| | (p) | (.08) | (.80) | (.77) | (.57) | (.28) | (.31) |
| | df | 10, 1144 | 10, 1134 | 10, 1124 | 10, 1114 | 10, 1104 | 10, 1094 |

# Appendix C

The word stimuli employed in the pilot study from Chapter 6 are listed in Table 20.

Table 20. Word stimuli for the pilot study described in Chapter 6. The words are grouped based on whether the stimulus presentation was followed by a comprehension question.

| Concrete words | | | Abstract words | | |
|---|---|---|---|---|---|
| Regular trials | | Comprehension trials | Regular trials | | Comprehension trials |
| author | garment | staff | adultery | happiness | slumber |
| belt | insect | starch | angel | horror | sneer |
| book | liquor | thong | bargain | joke | space |
| bureau | lobby | tribe | beauty | joy | strength |
| cable | machinery | troop | burden | love | thrill |
| carriage | manure | university | calmness | luxury | triumph |
| channel | medicine | weapon | concert | magic | welcome |
| chlorine | mountain | widow | crime | minute | woe |
| column | ounce | | danger | number | |
| corridor | package | | dream | panic | |
| cue | plate | | expanse | pleasure | |
| dent | pocket | | fashion | protest | |
| disease | product | | flutter | quest | |
| drape | rod | | frenzy | reflection | |
| estate | rye | | fun | romance | |
| freight | sound | | fury | seduction | |