THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2015
http://dx.doi.org/10.1080/17470218.2014.994098

Routledge
Taylor & Francis Group

The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts

Simon De Deyne, Steven Verheyen, and Gert Storms
Department of Psychology, University of Leuven, Belgium

One of the most significant recent advances in the study of semantic processing is the advent of models based on text and other corpora. In this study, we address what impact both the quantitative and qualitative properties of corpora have on mental representations derived from them. More precisely, we evaluate models with different linguistic and mental constraints on their ability to predict semantic relatedness between items from a vast range of domains and categories. We find that a model based on syntactic dependency relations captures significantly less of the variability for all kinds of words, regardless of the semantic relation between them or their abstractness. The largest difference was found for concrete nouns, which are commonly used to assess semantic processing. For both models we find that limited amounts of data suffice in order to obtain reliable predictions. Together, these findings suggest new constraints for the construction of mental models from corpora, both in terms of the corpus size and in terms of the linguistic properties that contribute to mental representations.

Keywords: Semantic memory; Text corpora; Syntactic dependency; Word associations; Similarity.

The key idea behind lexico-semantic models is that the meaning of any word can be inferred from the context in which it is used. If the semantic models are text-corpus-based, other words in the sentence or document in which the word appears are usually thought of as the context (so-called bag-of-words models). The impressive scale and comprehensive scope of corpus-based semantic models have proven instrumental in studying the acquisition and structure of the lexicon of children (Denhière & Lemaire, 2004; Monaghan, Chater, & Christiansen, 2005), healthy adults (M. N. Jones & Mewhort, 2007), psychiatric patients (Elvevåg, Foltz, Rosenstein, & DeLisi, 2010) and patients with various semantic lesions (Vinson, Vigliocco, Cappa, & Siri, 2003) using a large variety of semantic tasks such as priming, picture-word interference, and classification.

In this study we investigate two alternatives to the traditional bag-of-words models: a text-based syntactic dependency model and an association-based model. We regard both the syntactical annotated text-corpus model and the word-association model as semantic networks of the mental lexicon (De Deyne, Navarro, & Storms, 2013). We compare their ability to capture different types of semantic relations (thematic vs categorical) for different types of concepts (concrete vs abstract)

Correspondence should be addressed to Simon De Deyne, Faculty of Psychology and Educational Sciences, University of Leuven, B-3000 Belgium. E-mail: simon.dedeyne@ppw.kuleuven.be

Our gratitude goes to Kris Heylen, Dirk Speelman, and Dirk Geeraerts for making the LeNC corpus available, to Yves Peirsman for collaboration during the early stages of this project, and to Marc Brysbaert for his suggestions on lexicon size. This work was supported by Research Grant G.0436.13 from the Research Foundation - Flanders (FWO) to the first author and by the interdisciplinary research project IDO/07/002 awarded to Dirk Speelman, Dirk Geeraerts, and Gert Storms. Steven Verheyen is a postdoctoral fellow at the Research Foundation - Flanders.

at various levels of abstraction (domain vs basic level). In keeping with the theme of this special issue, we also investigate how the models' performance depends on corpus size. The models' relative performance is interpreted in light of the different linguistic and mental constraints of the representations they deliver. We first describe which types of concepts and relations between concepts will be studied, before going into the details of the models.

Types of words and the semantic relations between them

Despite the large number of studies that have focused on different methods for deriving semantic information from corpora it remains unclear whether representations of lexico-semantic models derived from written or spoken language are general enough to capture the meaning of all kinds of words (Vigliocco & Vinson, 2007), including concrete (e.g., EAGLE) and abstract (e.g., IDEA) ones. Our limited knowledge about differences between types of words and types of semantic relationships has important repercussions, as it might explain inconsistent findings obtained in a variety of experimental tasks that access semantics (Hutchison, Balota, Cortese, & Watson, 2008).

In contrast to concrete words, which can be understood in isolation, abstract words are believed to be relational in nature: it is their relationship with other words that determines abstract words' meaning (Goldstone, 1996; Verheyen, Stukken, De Deyne, Dry, & Storms, 2011; Vigliocco, Vinson, Lewis, & Garrett, 2004; Wiemer-Hastings & Xu, 2005). If the meaning of abstract words is predominantly derived from the context in which they are used, we expect lexico-semantic models to do better for these words than for concrete entities. The ability to capture the meaning of concrete words will also depend on the degree to which their perceptual properties are adequately encoded through language (Barsalou & Wiemer-Hastings, 2005; Wiemer-Hastings & Xu, 2005). The role of pragmatics in particular gets in the way of inferring simple perceptual properties such as the fact that bananas are yellow, which is considered common knowledge and thus would occur less frequently than expected in the everyday language that many semantic models rely on. However, abstract words present a different challenge to lexico-semantic models as they are more complex (Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005) and later acquired (M. N. Jones & Mewhort, 2007; Morrison, Chappell, & Ellis, 1997) than concrete words.

In a similar vein one might question whether the representations of lexico-semantic models are detailed enough to adequately differentiate between two closely related words such as GOOSE and DUCK, which both belong to the same taxonomic category of birds and are only distinct by a small degree. Distinguishing both birds requires detailed attributional information of a mainly perceptual nature which might not become fully encoded through language (Rosch, Mervis, Grey, Johnson, & Boyes-Braem, 1976; Tversky & Hemenway, 1984). Despite their comprehensive vocabulary, the models might only capture differences at a more general level, such as the domain of animals, where distinctions are more pronounced. In other words, rather than distinguishing structure at the detailed basic category level, lexico-semantic models might only allow a more coarse structure that distinguishes entities *between* rather than *within* categories.

Categorical relations are not the only type of semantic relation of significance in the lexicon (Schank & Abelson, 1977). While categorical relations are likely to be the main determinant for natural kinds of categories like birds or mammals, such a categorical structure is not as clearly defined for many artefact (Ceulemans & Storms, 2010; Goldstone, 1996; Verheyen, De Deyne, Dry, & Storms, 2011) and abstract categories (e.g., Crutch & Warrington, 2005; Hampton, 1981; Verheyen, Stukken, et al., 2011). Instead, an increasing number of studies suggest that thematic relations (e.g., DOCTOR–HOSPITAL) are just as important as categorical relations (Gentner & Kurtz, 2005; Lin & Murphy, 2001; Wisniewski & Bassok, 1999) and should therefore be included as well. This implies that any lexico-semantic model must also be able to account for a possible thematic structure in the mental lexicon.

We will use judgements of semantic relatedness to evaluate the two models' semantic representations of the various concepts and relations between them. The main reason for choosing this is because of the way semantic relatedness maps directly onto the notion of similarity, which provides the foundation for explaining many semantic phenomena, including categorisation, induction, and word retrieval from the mental lexicon (Goldstone & Son, 2005) and can be derived from both models with a minimum of assumptions.

MODELS

Syntactic dependency model

The first lexico-semantic model that will be considered is derived from written and spoken text corpora. This model uses co-occurrences derived from a stream of words similar to *Hyperspace Analogue to Language* (HAL; Lund & Burgess, 1996) and *Latent Semantic Analysis* (LSA; Landauer & Dumais, 1997). These early models apply the bag-of-words assumption: co-occurrences are simply defined for adjacent words without regard for word order or syntax. Clearly, this represents a minimal assumption on how mental models can be built from language. Other information from part-of-speech and syntactic relations also contributes to our understanding of sentences and their constituents. Understanding something about the part-of-speech of a word allows us to infer whether this word refers to an action (verb), an entity (noun) or properties of these entities (adjective). Similarly, syntactic relations also inform us about the meaning of words, for instance when the relation between a subject and an object allows us to infer agency.

More recent models have addressed this limitation to some extent. This is the case in the *Bound Encoding of the Aggregate Language Environment* model (BEAGLE; M. N. Jones & Mewhort, 2007), which accounts for word order, and further enhancements of the topic model, which include notions of syntax (Griffiths, Steyvers, Blei, & Tenenbaum, 2004). In the syntactic dependency model we propose, the context in which a word occurs is determined by a set of predefined syntactic relations or dependencies such as the modification of a noun by an adjective (e.g., an *interesting idea*) or the dependency between a subject and an object (the *bear* eats *honey*). In addition, since it represents each sentence as a hierarchical tree, it also accounts for the nested structure of sentences which occur quite frequently. In other words, it also takes into account relations of words which might be separated by dependent clauses (e.g., the *bear* sitting in the tree is eating *honey*). Given previous work in computational linguistics we expect that including this information will provide better results in predicting relatedness than a simple bag-of-words approach where word order and syntax are ignored (Heylen, Peirsman, & Geeraerts, 2008; Lapata, McDonald, & Keller, 1999; Peirsman, Heylen, & Speelman, 2007, August).

In terms of the different types of concepts we consider, we expect the syntactic dependency model to result in more accurate representations for abstract concepts compared to concrete ones, as the role of syntax might be crucial for bootstrapping our understanding of these words (Gleitman et al., 2005). Furthermore, several researchers have pointed out that document-based models that consider co-occurrence in larger text units than sentences tend to emphasise thematic relatedness such as CIGARETTE–SMOKER, while word-based models tend to emphasise synonymous relatedness such as CIGARETTE–CIGAR (Hutchison, 2003; M. Jones & Love, 2007). This would suggest that the current dependency model should adequately capture categorical relations as it extracts meaning from sentences rather than documents. The predictions for thematic relations are less clear, as the addition of syntax and word order might be beneficial, as illustrated in the *honey* and *bear* example presented above.

Association-based model

The second model is based on word associations. It is chosen because it relies on a unique novel data set that includes a comprehensive vocabulary of more than 12,000 words, covering most of the human lexicon. Associations to these cues were obtained through a continued procedure in which three associations per cue were collected from each participant (De Deyne et al., 2013). The combination of an extensive number of cues and a large and heterogeneous number of responses for each of the cues allows for accurate representations of meaning as these factors reduce the sparsity in the response distributions (De Deyne et al., 2013).

There are various reasons to assume that a word-association model is likely to encode mental representations differently compared to text-based representations. For the most part this is because word associations are not merely propositional but tap directly into the semantic information of the mental lexicon (McRae, Khalkhali, & Hare, 2011; Mollin, 2009). They are considered to be free from pragmatics or the intent to communicate some organised discourse, and thought to be simply the expression of thought. Moreover, these associations do not necessarily reflect a linguistic system, but might reflect imagery, knowledge, beliefs, attitudes, and affect (De Deyne & Storms, 2008; Simmons, Hamann, Harenski, Hu, & Barsalou, 2008; Szalay & Deese, 1978; Van Rensbergen, De Deyne, & Storms, 2014). If these claims are correct, we would expect the association-based model to perform better for all types of concepts and semantic relations. One obvious limitation of the word-association model is that it is much sparser than a text-based model as it currently only includes 300 different responses at most. Whether this is enough to capture differences in categories with basic-level items or is more appropriate for capturing the larger differences within a domain is an open question. This question will be explicitly addressed by manipulating the size of the corpus.

The role of corpus size

Most psychological studies do not consider whether the amount of information encoded through the corpus is adequate for the studied behaviour. This is quite surprising, as text corpus size tends to vary widely, from 5 million words in the carefully compiled Touchstone Applied Science Associates Inc. (TASA) corpus that is used in LSA, to over 100 million words in the British National Corpus (BNC; Aston & Burnard, 1997), to 1 trillion words in the Google *n*-gram corpus (Michel et al., 2011). When a corpus is too small or too large, it could affect the representations that are derived from it and cause certain phenomena to remain undetected. An illustrative case is the study by M. N. Jones, Kintsch, and Mewhort (2006), where HAL failed to account for mediated priming (when prime and target are only indirectly related through a mediator: LION → TIGER → STRIPES) with only the first 1000 dimensions derived from the relatively small TASA corpus, while denser models like BEAGLE and LSA did account for the mediated priming effects. It is also supported by evidence suggesting that for word-based models in particular, larger corpora systematically perform better on tasks such as semantic similarity rating (e.g., Recchia & Jones, 2009). Unfortunately, most of the comparisons that involved corpus size have used simple word-based models instead of models that also encode syntax (Bullinaria & Levy, 2007; Recchia & Jones, 2009).

An important goal of our study is to evaluate the role of sample size in both the text- and association-based models. One possibility is that many of the existing accounts simply underestimate how related certain concepts are, because their sample size is too small. We already mentioned this possibility for the association-based model, but it also applies to the text-based model. Especially for concrete entities such as animals or plants, the amount of information in text corpora tends to be smaller than for other types of concepts. Using the extensive recent SUBTLEX-NL word frequency norms from Keuleers, Brysbaert, and New (2010) for example illustrates this, as lemmas for concrete words such as *ladybug* and *broccoli* occur less than once per million words. From these observations, it is not entirely clear if such information is sufficiently represented in text corpora and whether or not corpus size is a factor. Once again, including different types of words allows us to investigate whether concrete, abstract, domain or thematic pairs require distinct amounts of knowledge.

NETWORKS

In this section and the following ones, we present each model as a network or graph indicated with the symbol G following our previous work (De Deyne et al., 2013). In this context the difference between a space-based model and a network model is mostly notational. However, addressing the text-based syntax model as a network highlights the close connection between rows (words) and columns (also words) connected through a weighted syntactic relation. Furthermore, it directly shows the analogy with the word-association model, where a network interpretation is more common, and puts them on equal grounds.

Syntax dependency network

We compiled a new Dutch corpus that is of adequate size for conducting psycholinguistic research. The size is adequate as it approximates the exposure to language for an average adult (see the General Discussion section below). It consists of three language resources spanning different registers. A first source uses text derived from Dutch articles in newspapers and magazines, which consists of a combination of the Twente Nieuws Corpus of Dutch (Ordelman, 2002) and the Leuven Newspaper corpus (Heylen et al., 2008). A second source consists of more informal language retrieved from Internet web pages. This corpus consists of 1000 documents for each of 8568 search terms retrieved using the Google and Yahoo Search API collected between 2005 and 2007 and the Dutch Wikipedia retrieved in 2008. Additional details can be found in De Deyne et al. (2008). A final source consists of spoken text, which includes Dutch movie subtitles and the Corpus of Spoken Dutch (Oostdijk, 2000). The total number of tokens in these sources after removing stop words and proper names is 79 million. The majority of these tokens were obtained from newspaper material: about 62% of them were taken from newspapers and magazines from Belgium, 12% from newspapers from the Netherlands, 25% from less formal online text, and 1% from spoken materials.

Each sentence in the corpus was parsed using the Alpino dependency parser for Dutch (Bouma, vanNoord, & Malouf, 2000). Similar to Pereira, Tishby, and Lee (1993) and Padó and Lapata (2007), two words were connected by a small number of predefined dependency paths (see Table 1). To reduce sparsity, part-of-speech tagged lemma forms provided by Alpino were used instead of word forms. In other words, plurals and inflections were all reduced to a more basic form. Next, all lemmas were counted and only adjectives, adverbs, nouns, and verbs occurring at least 60 times were retained. Since the dependency paths are undirected, each directed path resulted in two co-occurrence counts, from word a to b and vice versa.

The resulting corpus vocabulary consisted of 157 million co-occurrence tokens derived from undirected dependency paths and 103,842 different lemma types; 82.7% were nouns, 12.6% adjectives, 4.5% verbs, and 0.2% adverbs. A separate dependency matrix was constructed for each dependency pattern p, for a total of eight $N \times N$ dependency matrices G_p, where each cell corresponds to the frequency count of pattern p consisting of lemma a and lemma b. Each dependency matrix can be interpreted as a weighted directed graph, where two words are connected by a weight corresponding to the frequency of their dependency relationship, providing a straightforward interpretation and shared lexicon for nouns, adjectives, verbs, and so on.

The resulting number of types (i.e., the number of unique combinations between lemma a and lemma b given pattern p) and tokens (the count or frequency of occurrence of each type in the corpora) for each matrix are shown in Table 1. The total number of tokens varied considerably, from 57.9 million for ObjHd to 2.7 million for HdPredc. As a result of the large number of lemmas, the density of the matrices G_p was very low, with only 0.08% of the cells in the matrix for the most frequent dependency (G_{ObjHd}) different from zero.

Table 1. *Overview of the syntactic relations p used to construct dependency paths with examples in English for the target* COFFEE. *For each relation type, the number of observed dependency pattern types* F_{typ} *and tokens* F_{tok} *(*$\times 10^6$*) are listed in the fourth and fifth columns*

Relation p	*Path*	*Example*	F*typ*	F*tok*
ObjHd	$N\xrightarrow{\text{object of head}}\underline{\text{V}}$	We need some more *coffee.*	8.6	57.9
HdMod	$N\xleftarrow{\text{modification}}\underline{\text{A}}$	This is, excuse me, damn good *coffee.*	6.0	43.6
HdModObj	$N\xrightarrow{\text{modification}}\text{NP}\xrightarrow{\text{object of}}\underline{\text{N}}$	Lucy takes a loud *sip* of coffee.	7.0	22.9
SuObj	$N\xrightarrow{\text{subject of object}}\underline{\text{N}}$	*Coffee* contains lots of caffeine.	4.0	10.7
SuHd	$N\xrightarrow{\text{subject of head}}\underline{\text{V}}$	This *coffee* tastes delicious!	2.5	9.0
Cnj	$N\xleftrightarrow{\text{conjuction}}N$	Norma arrives with Cooper's *pie* and *coffee.*	2.2	7.3
SuPredc	$N\xrightarrow{\text{subject of predicative pharse}}\underline{\text{N}}$	*Coffee* is a drink.	1.2	3.3
HdPredc	$V\xrightarrow{\text{predicative completement}}\underline{\text{A}}$	This coffee *tastes* delicious!	0.8	2.7

Table 1 gives an example of each of the eight paths. With the exception of the HdModObj pattern, which is an indirect path with length 2 through a modifier, all paths have a length of 1. For each pattern a reverse path was created by transposing the path-dependent graph. For example, for pattern HdMod, the weight of a path for the adjective GOOD and the noun COFFEE is derived from the transposed dependency matrix G'_{HdMod}. An example of the obtained dependencies based on the sum of the original and transposed paths described in Table 1 for the word COFFEE is shown in Table 2. As can be seen from this table, the most frequent relations uncovered by the syntactic dependencies are interpretable as corresponding to distinctions in terms of function, attributes, and related entities.

Word-association network

Following the ideas from Deese (1965) and Steyvers, Shiffrin, and Nelson (2004), we use word associations as a distributional model of meaning. We are able to do so because unlike extant word-association norms, which tend to be small and sparse, we collected three associations per cue from each participant instead of one (De Deyne et al., 2013) which leads to more reliable distributions. A total of 71,380 native Dutch speakers provided associations. The cues were initially selected from a small set of 338 mostly concrete nouns (see Ruts et al., 2004). This set was gradually expanded using a snowball procedure where the most frequent responses were added at different points of time during the course of the project. Each participant generated three different responses to a cue word. Each cue was presented to 100 different participants, thus resulting in 100 primary, 100 secondary, and 100 tertiary responses. For a set of 12,581 cues, a total of 3.77 million responses were collected this way.[1]

The network is constructed from a weighted adjacency matrix where both the rows and the columns correspond to different cue words and the entries represent the association frequencies observed between a cue and a response. In other words, only responses that were also presented as cues are encoded in the network. Restricting the network to words that were present both as a cue and as a response reduced the number of nodes from 12,581 to 12,418.

Two networks were derived: G_{asso1}, a network based on the primary responses and comparable to other single-response datasets (e.g., Nelson, McEvoy, & Schreiber, 2004) and G_{asso123}, a network including the secondary and tertiary responses as well. The network G_{asso1} had a density of 0.22%. Including the secondary and tertiary responses increased the density considerably, to 0.64% for G_{asso123}. This confirms that the continued procedure draws on a more heterogeneous response set through the inclusion of weaker links that might go undetected in single-response procedures (see De Deyne et al., 2013, for further

[1] The study is ongoing at http://www.smallworldofwords.com/nl and currently contains data for over 16,000 cue words.

Table 2. *Dutch examples and English translations for the five most frequent syntax dependencies derived for* COFFEE

HdPredc	*HdMod*	*HdModObj*	*Cnj*
klaar (ready)	gratis (free)	thuisploeg (hometeam)	thee (tea)
koud (cold)	sterk (strong)	hand (hand)	taart (cake)
gratis (free)	vers (fresh)	versnapering (snack)	gebak (cake)
op (finished)	eerlijk (fair)	man (man)	water (water)
heerlijk (delicious)	zwart (black)	suiker (sugar)	pannenkoek (crepe)
SuObj	*SuPredc*	*SuHd*	*ObjHd*
bezoeker (visitor)	drank (drink)	drinken (to drink)	drinken (to drink)
mens (human)	thee (tea)	serveren (to serve)	zetten (to make)
man (man)	water (water)	schenken (to pour)	gaan (to go)
team (team)	product (product)	zetten (to make)	schenken (to pour)
iemand (someone)	leven (life)	maken (to make)	krijgen (to get)

discussion). While the density increases through continued responses, it should be noted that the average number of links per cue is still quite low. For the network based on a single response G_{asso1} this was 27.1, while for the network including all responses G_{asso123} it was 79.8.

RELATEDNESS STUDIES

The following studies were conducted to evaluate the effect of concept and semantic-relation type on the ability of lexico-semantic models to predict relatedness. The studies are organised into three main distinctions. First, concrete entities are compared with abstract entities. In this first series of studies, similarity judgements for all pairwise combinations in concrete categories (e.g., mammals, clothing) and abstract basic-level categories (e.g., emotions, sciences) were collected. Because these comparisons are performed between basic level items, they require an evaluation of nuanced and detailed attributes (for instance, when comparing HAMSTER and MOUSE) which might require access to perceptual or other non-linguistic represented information. Since abstract concepts rely primarily on relational information derived from language rather than perceptual properties, predictions for these concepts should be relatively more accurate than for concrete ones in the syntax dependency model.

To investigate the possibility that lexico-semantic models are more sensitive to domain-level differences than to differences among basic-level category exemplars, a second series of studies was included where items from various animal or artefact categories were paired, leading to pairs such as BUTTERFLY and EAGLE or ACCORDION and FRIDGE.

To assess the extent to which lexico-semantic models can predict thematic rather than categorical relationships, a third series of studies was undertaken with items that were thematically related such as UMBRELLA and RAIN or stumble and PAIN.

Materials

There were 13 concrete sets, comprising exemplars of 6 Artefact categories (clothing, kitchen utensils, musical instruments, tools, vehicles, and weapons), 5 Animal categories (birds, fish, insects, mammals, and reptiles), and 2 Food categories (fruit and vegetables). The list of items is available in De Deyne et al. (2008).

There were 7 abstract sets, comprising exemplars of the categories art forms, crimes, diseases, emotions, media, sciences, and virtues. The list of items is available in Verheyen, Stukken, et al. (2011).

The domain sets consisted of exemplars from all 6 concrete Artefact sets or all 5 concrete Animal sets. Since it is not feasible to present all the pairwise combinations of the combined set of Artefact or Animal items, we selected 5 items from each of the Artefact and Animal sets. Both items that were central to the set (e.g., SWALLOW is a typical bird and thus a central member) and items that were not (e.g., BAT is an atypical member of the mammals set, and is closely related to birds) were included. The resulting domain sets consisted of 6 × 5 Artefact items and 5 × 5 Animal items, respectively. To increase the generalisability of the results, two replications of the above procedure were performed, resulting in an A and B set. See Appendix B for a list of the items.

The thematic set consisted of pairs from two different studies. The first study was a replication of the study by Miller and Charles (1991), a widely used benchmark test in computational linguistics. The second study was similar to that of Miller and Charles and consisted of 100 pairs of thematically related words such as RABBIT and CARROT, including 20 words that were weakly related to cover the entire range. The set will henceforth be referred to as the Thematic (mixed) set and translations for the 100 pairs are available in Appendix C.

Procedure

All participants were affiliated to the University of Leuven, either as students or as staff. They were paid the equivalent of $10/h (concrete, domain), received course credit (abstract, thematic), or volunteered (abstract). The participants were requested to perform a pairwise rating task for one or more sets of items. They were asked to rate the similarity (concrete, abstract, domain) or relatedness (thematic) of each item pair on a scale ranging from 1 (no similarity) to 20 (maximum similarity) The item pairs within a set, the items within a pair, and—where applicable—the sets were presented in random order.

For each set, Appendix A shows the number of item pairs, raters, and reliability. The obtained average ratings were all very reliable, with Spearman Brown split-half correlations ranging from .85 to .99. The averages were based on the judgements from participants that correlated at least .45 with the total average. Note that for the text model a total of 12 words (4 concrete, 6 abstract, and 2 thematic) were missing from the respective data sets. Since we will only use pairs with words that were present in both data sets, these items were removed. In all sets, the words were nouns, except for the Thematic (mixed) set. In this set a total of 69 pairs consisted of nouns only, whereas the remaining 31 pairs comprised at least one verb or adjective. For comparability, only the data for the nouns were analysed.

RESULTS

Deriving relatedness from the networks

Before we assess how well relatedness derived from the syntax dependency network and the word-association network can account for the human relatedness judgements, we briefly describe how relatedness indices are derived from both networks. Both the syntax dependency network and the word-association network represent weighted graphs, with the weights reflecting the co-occurrence frequency of two words (either as a function of their syntactic relation or as a response to an association cue). The weight of each edge is generally chosen to be a function of this frequency that either transforms the distribution (e.g., through a logarithmic transformation) or reflects how specific the information encoded in the edge is (based on heuristic, information theoretic, or statistical criteria). Here we applied the *positive point-wise mutual information* (PMI) weighting as proposed by Church, Gale, Hanks, and Hindle (1991) because of its systematic good performance in the context of word co-occurrence models (Bullinaria & Levy, 2007).[2]

[2] In past studies we have applied *t*-score weighting, as this consistently improved the estimates of the word-association measures over a range of tasks. In this study, PMI was nevertheless chosen to increase the comparability of the text-model where PMI is applied as standard.

A commonly used measure of similarity is the cosine measure (e.g., Landauer & Dumais, 1997; Lund & Burgess, 1996; Padó & Lapata, 2007; Steyvers et al., 2004). While it is often applied in spatial models such as LSA, it also has a straightforward network interpretation. In a network or graph, it functions as a distributional overlap measure that captures the extent to which two nodes in the network share the same immediate neighbours. Two nodes that share no neighbours have a similarity of 0, and nodes that are linked to the exact same set of neighbours have a similarity of 1. For each item pair within a set, the cosine similarity between the items was calculated, based on the syntax dependency network (G_{syn}) and on the networks derived from the first association response (G_{asso1}) and all three responses (G_{asso123}).

Predicting human relatedness

Judged relatedness was standardised to calculate the correlations over different concrete, abstract, domain, and thematic sets. The correlations between empirical relatedness and relatedness derived from G_{asso1}, G_{asso123}, and G_{syn} are shown in Table 3. *N* indicates the number of item pairs across which the correlations are taken. The procedure from Zou (2007), which estimates the confidence intervals for the difference between two dependent correlations, was used to compare the differences in correlations between different types of models, and these results are presented in Table 4.[3]

The results of this analysis in Table 4 suggest that in all sets G_{asso123} outperforms the syntax dependency model G_{syn} and the model based on a single associate G_{asso1}. In line with our hypotheses, domain judgements yielded higher correlations than basic-level judgements for abstract and concrete basic-level items. Both the syntax dependency model and the word-association model represented thematic judgements better than concrete and abstract ones. The strongest differentiating data set was the concrete one, which was on par with abstract concepts in the association models, but considerably harder to model than any other data set in the syntax dependency model. The similar results for abstract and concrete concepts using word associations support the idea that perceptual properties might be adequately encoded in the association model, but not in the text-based syntax dependency model. The better performance for abstract words in the dependency model confirmed the hypothesis that in contrast to concrete words, the language environment provides the primary source to derive meaning from. However, the absolute strength of the correlation was still lower compared to the association model and this suggests that other types of knowledge are potentially extracted at the sentence level than what is presently encoded through the syntactic dependencies.

One possible concern is that the performance of the association models depends on the direct associative strength between word-pairs. Especially in the thematic set, certain pairs are likely to be directly associated (e.g., CIGAR–SMOKING). Whereas the cosine measure only involves shared neighbours and thus does not take into account whether two words are directly associated, it might be the case that at least for some relations associative strength between two words suffices. To investigate this possibility associative strength for G_{asso123} was calculated as the average probability of generating a specific response *b* for a cue word *a* and a response *a* to cue *b*. For the thematic pairs associative strength and relatedness judgements were highly correlated, $r = .740$, $CI = [.632, .820]$, but this was still lower than the reported .823 and this difference $\Delta(r)$ was significant: $\Delta(r) = .144$, $CI = [.005,.174]$. Likewise, the differences for the other datasets favoured the relatedness measure: association strength Concrete, $r = .404$, $CI = [.379,.428]$, $\Delta(r) = .219, CI = [.192,.246]$; Abstract, $r = .461$, $CI = [.392,.524]$, $\Delta(r) = .156$, $CI = [.089,.225]$; and Domain, $r = .219$,

[3] The 95% confidence intervals used here encompass a significance test but also provide an estimate of the magnitude of the effect. If zero is included in the confidence interval, the result will not reach the 5% significance level.

Table 3. *Results of the correlation analyses for the four data sets (Concrete, Abstract, Domain, and Thematic) and the three network types. Confidence intervals at* $\alpha = .05$ *for cosine relatedness are indicated within square brackets*

	Concrete	*Abstract*	*Domain*	*Thematic*
N	4493	545	1470	94
G_{asso1}	.551 [.530, .571]	.515 [.451, .574]	.711 [.685, .735]	.650 [.515, .753]
G_{asso123}	.623 [.605, .640]	.617 [.562, .666]	.792 [.773, .811]	.823 [.744, .879]
G_{syn}	.366 [.340, .391]	.517 [.453, .576]	.679 [.651, .706]	.588 [.439, .707]

Table 4. *Comparison of the correlation strengths of the models. Values between brackets indicate the 95% confidence intervals for the difference between dependent correlations,* $\Delta(rG_A, rG_B) = r_{G_A} - r_{G_B}$. *Only the significant results excluding zero from this interval at* $\alpha = .05$ *are displayed.*

	Concrete		*Abstract*		*Domain*		*Thematic*	
$\Delta(G_{\text{asso1}}, G_{\text{asso123}})$	− .072	[−.087, −.058]	− .101	[−.147, −.058]	− .082	[−.100, −.065]	− .173	[−.285, −.085]
$\Delta(G_{\text{asso1}}, G_{\text{syn}})$	.185	[.156, .213]						
$\Delta(G_{\text{asso123}}, G_{\text{syn}})$	.257	[.230, .283]	.100	[.035, .166]	.114	[.086, .142]	.234	[.114, .376]

CI = [.170,.267], $\Delta(r)$ = .574, CI = [.526,.622]. Altogether, the results show that associative strength and the similarity scores do not necessarily measure the same thing, even for words that are thematically related. The dominant role of similarity confirms previous findings which show that participants find it very difficult to judge how strongly associated two words are without being influenced by their relatedness (De Deyne et al., 2013).

Manipulating the size of the networks

In this section we evaluate to what extent the above results are dependent upon the size of the networks. To this end, networks of different sizes are obtained through sampling, and their ability to account for human relatedness judgements is assessed.

For the syntax dependency model G_{syn}, 100 equal-sized weighted samples were drawn by sampling each of the eight dependency matrices G_p proportional to their raw syntax occurrence counts and then summing the results for all eight dependencies. To vary the size of the corpus, a total of 1 to 100 samples was summed, and the resulting data weighted using the PMI weighting function outlined previously. To obtain an estimate of the stability of the results derived from these samples, this procedure was repeated 10 times with permuted sample orders. For G_{asso123} a similar procedure was used. Instead of using a weighted sample from the complete data, the samples were determined based on participants' responses. Since a total of 100 first, second, and third responses were generated by a total of 100 participants, 100 different samples were obtained by randomly assigning a particular response to one of the samples. Next, samples were summed in a cumulative fashion, converted to a unipartite network, and weighted using PMI-scores. This procedure was repeated 10 times, each time using a permuted set of samples.

The results of manipulating corpus size for the syntax dependency model are shown in the left-hand panel of Figure 1. The slope of the curve is initially steep, which indicates a fast improvement, but flattens off when more than 30% of the data is added to G_{syn}. The results of manipulating the number of responses for G_{asso123} are shown in the right-hand panel of Figure 1. In contrast to G_{syn}, the results show continuous improvement, although here as well the slope flattens when data from more than 40 to 50 participants are included. In both parts of Figure 1, the variability decreases as

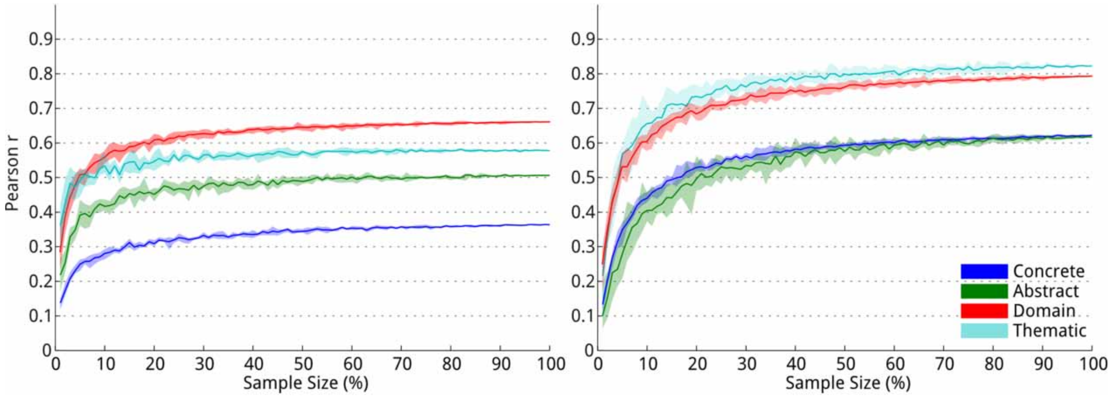

Figure 1. *Correlation between judged and derived relatedness as a function of data sample size for* G_{syn} *(left) and* G_{asso123} *(right). Shaded areas approximate 95% of the correlations based on 10 repeated draws from 100 samples.*

a function of sample size which further indicates that the correlations become reliable with just a portion of the data. The correlations including all samples of course correspond to those in Table 3. The results in Figure 1 also indicate that the differences between types of concept and semantic relationships that were described in the previous section are largely unaffected by sample size.

While the difference between the correlations of a corpus that is just a fraction of the full size is rather small, the number of observations influences whether or not they are different. For instance, in the case of the concrete set with more than 4000 pairs the results continue to improve G_{asso} using up to 97% of the data ($r = .619$ compared to the full dataset with $r = .623$, $\Delta(r) = .004$, CI = [.001,.007], while for the thematic set with 94 pairs, the improvement stopped after 59% explained will not be higher with just a small sample of the set. For larger studies, this difference might be statistically reliable, but a difference of less than 1% of extra variance is likely to have limited influence on the interpretation of the phenomena studied in our field.

Finally, note that the rate of increased prediction in Figure 1 is slightly different depending on the data set, especially in the case of concrete and abstract entities (see the right-hand panel of Figure 1). The lower rate for abstract compared to concrete entities could be interpreted as support for the hypothesis that the acquisition of abstract words lags behind that of concrete ones, due to the former's dependence on considerable exposure to language (M. N. Jones & Mewhort, 2007). However, as indicated by the shaded area for the word-association network on the right-hand panel of Figure 1, strong conclusions in this respect are likely to be preliminary.

GENERAL DISCUSSION

In this study we tested two key aspects of the use of large-scale lexico-semantic models based on language. First, we tested the ability of these models to account for the structure in the mental lexicon accessed through human relatedness judgements. The most striking result was that this ability varies widely, depending on what kind of relatedness is measured. A syntax dependency model only yields weak to moderate correlations for category with basic-level items. This is especially the case for concrete categories such as birds and tools, which by definition involve information that is mostly attributional and of a sensory nature. As expected, the situation is better for thematic relations and for abstract categories such as sciences and virtues, where the distributional semantics in the linguistic environment are the primary source to derive meaning from. These findings are in line with previous results by Vigliocco et al. (2004), who found better results for event words than for abstract words in LSA. The best results are found when judgements at a domain level are considered. This shows that the distinction between entities such as birds and mammals or vehicles and tools is well represented in text-based syntactic dependency models. The inclusion of an alternative semantic model based on word associations allows us to put these findings into a larger context. Relatedness derived from a continued association task systematically improves predictions for all kinds of concepts and relations considerably. In contrast to the text-based syntactic dependency model, the difference between concrete and abstract concepts was less pronounced.

The second aspect we tested concerned the quantitative aspects of language exposure. The findings indicate that only a small subset of information available in language contributes to word meaning (as measured through relatedness). More precisely, regardless of the type of concepts and comparisons, a corpus that was less than half the size of our complete text corpus performed equally well, indicating that ever larger corpora do not always offer much improvement. Moreover, word-association data, representing only a fraction of the number of tokens present in the text corpus, showed that even smaller sample sizes based on around 40 to 50 persons generating three words captures relatedness for all kinds of concepts and semantic relations between them.

These findings have a number of theoretical and methodological implications and provide a different perspective on classic issues in our field. These include (i) the role of language exposure and the mechanisms of acquisition, and (ii) the contributions of non-linguistic information to word meaning for a variety of concepts. In the following sections, we will elaborate on both points.

Natural language input during the lifespan versus quantitative and qualitative corpus properties

One of the challenges in acquiring a new language is that the input is very sparse. The sparsity at the input side is a manifestation of the poverty of the stimulus argument, according to which the knowledge acquired from language far outstrips the information that is available in the (linguistic) environment (see Laurence & Margolis, 2001, for a discussion of Chomsky's classic argument). Sparsity is also a potential problem for most lexico-semantic models. There are at least three strategies that can be combined to tackle this problem. A first one would be to consider additional information through huge corpora, as is the case with the trillion-word Google *n*-grams project (Michel et al., 2011). We have explored this possibility here and found that more data does not necessarily result in better representations. This finding contrasts with previous results that used word-based models but did not include syntax, where the performance on the Test of English as a Foreign Language (TOEFL) test increased from around 50 to 85% when corpus size increased from 1 million to 100 million words (Bullinaria & Levy, 2007). As the number of syntactic paths in a sentence is smaller than the number of co-occurrences that can be derived even for small windows (e.g., one or two neighbouring words), this might indicate that selecting the right kind of (syntactic) information surpasses some limitations that can be attributed to sparsity.

Next, the signal-to-noise ratio could be improved by selecting materials that closely align with the type of knowledge humans are exposed to. An example of such an approach is the TASA corpus, which is based on hand-picked reading text at different grade levels. Previous research indeed suggests that well-balanced corpora such as the TASA or BNC corpus perform better than corpora based on newsgroup text (Bullinaria & Levy, 2007). In this paper we did not address the issue of corpus quality explicitly, but following the above reasoning expect good performance given the composition of the Dutch corpus which relies mostly on magazines and newspapers.

Finally, one could try to infer new meaning from the linguistic environment through various unsupervised techniques. It is this last strategy that has received the most attention and has provided the basis of numerous strong claims about fundamental properties of lexico-semantic models. The vast number of unsupervised techniques include singular value decomposition (Landauer & Dumais, 1997), random projections (Sahlgren, 2005), holographic projections (M. N. Jones & Mewhort, 2007), non-negative dimension reduction (Hoyer, 2004), self-organising maps (Vinson et al., 2003), probabilistic inference over topics (Griffiths, Steyvers, & Tenenbaum, 2007), and random graph walks (De Deyne, Navarro, Perfors, & Storms, 2012; Hughes & Ramage, 2007). Each of these techniques reduces the sparsity, either at a representational level (through singular value decomposition for instance) or online through random graph walks. They often improve the results, especially for small-to-moderately sized corpora, but not necessarily for large corpora (see Bullinaria & Levy, 2007; Louwerse & Connell, 2011; Recchia & Jones, 2009). Especially in word-based models, dimensionality reduction does not lead to substantial improvements (Bullinaria & Levy, 2007). Despite the contentious nature of dimensionality reduction, it is likely that inferring additional structure from language input combining both supervised and unsupervised learning mechanisms is an important feat of humans. Further studies will need to show how principles like dimensionality reduction can be aligned with human constraints on language acquisition. For example, in the context of syntax dependency models, a logical step would involve studying what kind of relationships allow the inference of new information.

While the inference mechanisms may introduce unnecessary complexity, turning to ever-expansive corpora entails a different risk. The unlimited amount of linguistic data that is available nowadays could result in corpora that overestimate the redundancy encoded in the linguistic environment. For many words, the actual exposure through language can be quite small, yet large corpora do provide a stable representation through word co-occurrence or other measures that might surpass actual exposure (for example when using the most recent version of the English Wikipedia, which includes over 2.6 billion words). As such, it might be interesting to establish what the actual exposure to language is. This could be inferred by considering the vocabulary size of an average adult. Depending on how words are counted and what is understood as knowing a word, 40,000 is often quoted as the number of words known by the average American high school graduate (Aitchison, 2003). A recent large-scale study in Dutch showed that out of a sample of 52,847 words, the average percentage known was 71.6% or 38,000 words (Brysbaert, Keuleers, Mandera, & Stevens, 2014). The percentage differs depending on age (around 50% for 12-year-olds and 80% for 80-year-olds). To get an estimate of total linguistic exposure, researchers have recorded natural language samples in a systematic study among university students and found that about 16,000 words are spoken each day (Mehl, Vazire, Ramírez-Esparza, Slatcher, & Pennebaker, 2007). Extrapolating this number, one obtains around 88 million words spoken in 15 years. All in all, this suggests that the current corpus of 79 million content words is a more realistic starting point than a rather small corpus based on reading materials such as the TASA. In addition, real language exposure should also act as a constraint for future corpora which will be potentially much larger than present ones. However, there is an important caveat, as our results clearly show that providing a more realistic approximation of language input in terms of corpus size does not necessarily improve the quality of the mental representations that are derived from language.

Non-linguistic contributions to word meaning

Recent views on semantic cognition consider the meanings of words to be represented across a variety of modalities that differ in whether they are sensory or more symbolic language-based in nature (e.g., Barsalou, Santos, Simmons, & Wilson, 2008; Vigliocco et al., 2004) and some studies have actually tried to augment lexico-semantic models with perceptual information. In one study, text models were combined with a bag-of-visual features approach derived from a large set of images (Bruni, Uijlings, Baroni, & Sebe, 2012). Although the results in this study showed some contribution of visual features, the gain was rather limited. Another interesting possibility is the proposal by Andrews, Vigliocco, and Vinson (2005), which combines speaker-generated features with distributional information derived from text. This proposal distinguishes the acquisition of attributional information through concrete experience with objects and events in the world from information implicitly derived from exposure to language. One of the key issues raised by these studies is under which conditions certain types of information affect semantic processing. Answering this question is likely to be difficult since many studies show that grounded or perceptual information is redundantly encoded in text-based resources as well (Louwerse, Hu, Cai, Ventura, & Jeuniaux, 2005).

Our results provide additional clues about where language-derived representations are likely to contribute most. For instance, there is a clear difference between how abstract and concrete entities are represented in the model derived from text and the model derived from word associations. Even though the text-based results show that there are limitations specific to concrete entities, these limitations are not necessarily due to the fact that perceptual properties cannot be accurately encoded in a linguistic and symbolic system (Bruni et al., 2012). Instead, it likely reflects a limitation of spoken and written language resources, where efficient communication consists of finding common ground between speakers. This type of pragmatics explains why mentally central properties (e.g., the fact that bananas are yellow or apples are round) are very strong responses in word-association data but much less prominently expressed in text corpora. This is not unexpected if one assumes that word associations sample from both lexico-semantic representations and modality-specific representations. In fact, previous studies have shown that the continued response procedure used in our word-association task increasingly results in more attributional and thematic responses, whereas the first response tends to reflect lexico-semantic properties such as superordinate or contrast relations (De Deyne et al., 2013). Access to a lexico-semantic register combined with inspection of sensory properties (most pronounced in later responses) might be the main reason why the word-association approach is so successful in accounting for the relatedness of all kinds of concepts, whether they are concrete or abstract. Of course, the results from word associations were by no means perfect, especially with respect to the prediction of basic-level comparisons for both concrete and abstract words. A number of task-specific factors might contribute to this. First of all, in the human relatedness judgement, all stimuli belonged to a single category and since all pairwise combinations were shown, it is quite likely that the participants framed their judgements to focus on entity-features instead of also considering thematic relations. Since the prediction of thematic relations was much better, it is possible that in more natural settings, this type of information makes a larger contribution. Similarly, it might also be the case that participants were primed with the category-consistent sense for some of the stimuli that are homonyms like the Dutch words *raket* (refers to tennis or a rocket) and *bank* (refers to a financial institution or a piece of furniture).

Methodological implications and conclusion

Before closing, let us elaborate on at least one methodological implication of the current findings and a few other issues related to this. As noted earlier, in the word-association model, continued responses uncover many weak links that are absent in a single-response procedure. The lack of these responses might explain why traditional single-response data are not often used to measure distributional semantics. This is the case with the frequently-used Florida norms (Nelson et al., 2004), where on average 13 different responses are produced per cue, compared to an average of 78 in our association norms. Additional evidence supports the idea that these weak links are important in other tasks as well. A case in point is a study by Maki (2007) who used a judgement task of associative strength and tried to explain why participants overestimate the associative strength of word pairs that never co-occurred in single-response tasks. Moreover, a recent study of our own showed that even for randomly chosen triads, which by definition exhibit very weak indirect relations, combining a network account with a random walk-based spreading activation mechanism could reliably recover the preferences among the triad choices (De Deyne et al., 2012). Interestingly, while this remote and weak structure in the lexicon seems shared among speakers of a language, text-based models only capture a fraction of what is covered by the association model. Combining these studies with the current results suggests that semantic representations for both close and distal relations are more accurately captured by association data than representative text corpora of language input. This is the case for abstract concepts, which presumably are mostly acquired through language exposure, and particularly for concrete words, where the sparse linguistic input does not allow for an accurate encoding of sensory-based attributes.

A type of data that is of particular interest in the study of the mental lexicon is priming data, as it provides an online measurement of how mental representations are accessed over time. For this area of research, lexico-semantic models have been instrumental in solving long-lasting debates. One of these debates concerns whether an associative or semantic relation is necessary between prime and target to exhibit a processing advantage (Hutchison, 2003). Another one concerns the status of mediated priming (e.g., Ratcliff & McKoon, 1994). In many cases it is nearly impossible to draw a firm distinction between direct and mediated priming, just like it is difficult to draw a distinction between pure associative and semantic priming, especially as both the quantity and quality of linguistic data and mental representations abstracted from it become more realistic. So far, results show that LSA fails to predict priming at an item level, while associative strength measures succeed (Hutchison et al., 2008). It would be interesting to see whether these conclusions still hold with improved measures, provided by a syntax dependency or word-association model of the kind we have introduced here. This will undoubtedly be a subject for future investigations.

One often-heard criticism is that association strength is an empty variable as it does not tell how these strengths themselves are acquired through language exposure (see, for instance, Hutchison et al., 2008). Because of this, one could argue that association strength should only be treated as a dependent variable. Although this criticism has appeal, this kind of reasoning in the domain of psycholinguistics (but perhaps not philosophy) requires explanations that for the time being are not necessarily more meaningful. For example, if associations are a special kind of propositions restrained from pragmatics, it is apparent that text corpora cannot fill this explanatory role either, as we have no understanding of how the text is generated. Clearly, what is needed is a model that captures the physical environment, the structure of which is reflected in a non-arbitrary way through language, combined with a deep understanding of physiology and evolutionary dynamics that allow us to backtrack to the origins of language. Beyond any doubt this is a worthwhile and encompassing project which has already taken root, for instance through evolutionary linguists where studies focus on how artificial agents co-evolve with language (see Steels & Hild, 2012). However, once the right conditions have been installed to have embodied artificial agents learn this language, it might be less straightforward to understand how and what representations are mentally represented *in silico* than initially imagined.

While associative strength is likely to continue a double career as both independent and dependent variable, it should be noted that the strength of the semantic network metaphor does not lie at the physical or perhaps algorithmic level, but as convincingly advocated by Deese (1965) its explanatory power resides at a computational level, where we learn new things about word meaning by simultaneously looking at the macro-, meso- and micro-level structure of the network rather than focusing on the strength of a single pair of words (Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013).

To conclude, in studying word meaning what seems to be missing from text-based models are exactly those non-linguistic mental properties which for now remain primarily accessible through the word-association procedure. The choice and size of the corpus (text or word associations) will affect the predictive power of our semantic models and the conclusions we can draw from a variety of studies on semantic processing, including priming. For many studies involving single-word semantics, a word-association corpus will be more appropriate. However, the current text-based models can also be improved in many ways and other phenomena that go beyond a single word, such as how humans extract the gist of a story and other aspects of discourse, might be more suited for text-based approaches in general. Regardless of the specific questions, the strength of both approaches lies in how extensive data can generate original hypotheses and open up many new areas to systematic inquiry. From the myriad of alternative approaches out there, a priori not all are equally suited to be applied as a model for word meaning or a measure of relatedness instrumental in memory and language studies. In this domain, the success of the models derived from them will increasingly reflect the degree to which they capture the mental properties of language.

Original manuscript received 11 February 2014
Accepted revision received 28 October 2014

REFERENCES

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon*. Chichester: Wiley-Blackwell.

Andrews, M., Vigliocco, G., & Vinson, D. P. (2005). The role of attributional and distributional information in representing meaning. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Meeting of the Cognitive Science Society* (pp. 127–132). Mahwah, NJ: Lawrence Erlbaum.

Aston, G., & Burnard, L. (1997). *The BNC handbook exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

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

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols, embodiment, and meaning* (pp. 245–283). Oxford: Oxford University Press.

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

Bouma, G., van Noord, G., & Malouf, R. (2000). Alpino: Wide coverage computational analysis of Dutch. In W. Daelemans, K. Sima'an, J. Veenstra, & J. Zavrel (Eds.), *Computational Linguistics in the Netherlands 2000* (pp. 45–59). Amsterdam: Rodopi.

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. R. Smith, S. Satoh, T. Plagemann, X.-S. Hua, & R. Yan (Eds.), *Proceedings of the 20 th ACM Multimedia Conference, MM '12, Nara, Japan, October 29–November 2, 2012* (pp. 1219–1228). New York, NY: ACM Press.

Brysbaert, M., Keuleers, E., Mandera, P., & Stevens, M. (2014). *Woordenkennis van Nederlanders en Vlamingen anno 2013: Resultaten van het Groot Nationaal Onderzoek Taal* (Tech Rep.). Ghent: University of Ghent.

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

Ceulemans, E., & Storms, G. (2010). Detecting intra- and inter-categorical structure in semantic concepts using HICLAS. *Acta Psychologica*, *133*, 296–304.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon* (pp. 115–164). Hillsdale, NJ: Lawrence Erlbaum.

Crutch, S., & Warrington, E. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, *128*, 615–627.

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2012). Strong structure in weak semantic similarity: A graph based account. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Meeting of the Cognitive Science Society, Sapporo, Japan, August 1–4, 2012* (pp. 1464–1469). Austin, TX: Cognitive Science Society.

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*, 480–498.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*, 213–231.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*, 1030–1048.

Deese, J. (1965). *The structure of associations in language and thought*. Baltimore, MD: Johns Hopkins University Press.

Denhiere, G., & Lemaire, B. (2004). A computational model of children's semantic memory. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Meeting of the Cognitive Science Society, August 4–7, 2004, Chicago, Illinois, USA* (pp. 297–302). Mahwah, NJ: Lawrence Erlbaum.

Elvevåg, B., Foltz, P. W., Rosenstein, M., & DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics*, *23*, 270–284.

Gentner, D., & Kurtz, K. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 151–175). Washington, DC: American Psychology Association.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*, 23–64.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*, 608–628.

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. Holyoak & R. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 13–36). Cambridge: Cambridge University Press.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating topics and syntax. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17 (NIPS 2004)* (pp. 537–544). Cambridge, MA: MIT Press.

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

Hampton, J. A. (1981). An investigation of the nature of abstract concepts. *Memory & Cognition*, *9*, 149–156.

Heylen, K., Peirsman, Y., & Geeraerts, D. (2008). Automatic synonymy extraction. In S. Verberne, H. van Halteren, & P.-A. Coppen (Eds.), *A comparison of syntactic context models: LOT computational linguistics in the Netherlands 2007* (pp. 101–116). Utrecht: Netherlands National Graduate School of Linguistics.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, *5*, 1457–1469.

Hughes, T., & Ramage, D. (2007). Lexical semantic relatedness with random graph walks. In J. Eisner (Ed.), *EMNLP-CoNLL: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28–30, 2007, Prague, Czech Republic* (pp. 581–589). Stroudsburg, PA: Association for Computational Linguistics.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap?. *Psychonomic Bulletin & Review*, *10*, 785–813.

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*, 1036–1066.

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology*, *55*, 196–231.

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

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

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*, 643–650.

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*, 211–240.

Lapata, M., McDonald, S., & Keller, F. (1999). Determinants of adjective-noun plausibility. In H. S. Thompson & A. Lascarides (Eds.), *EACL '99: Proceedings of the ninth conference on European chapter* *of the Association for Computational Linguistics* (pp. 30–36). Stroudsburg, PA: Association for Computational Linguistics.

Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *The British Journal for the Philosophy of Science*, *52*, 217–276.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, *130*, 3–28.

Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, *35*, 381–398.

Louwerse, M. M., Hu, X., Cai, Z., Ventura, M., & Jeuniaux, P. (2005). The embodiment of amodal symbolic knowledge representations. In I. Russell & Z. Markov (Eds.), *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference* (pp. 542–547). Menlo Park, CA: AAAI Press.

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

Maki, W. S. (2007). Judgments of associative memory. *Cognitive Psychology*, *54*, 319–353.

McRae, K., Khalkhali, S., & Hare, M. (2011). Semantic and associative relations: Examining a tenuous dichotomy. In V Reyna, S. Chapman, M. Dougherty, & J. Confrey (Eds.), *The adolescent brain: Learning, reasoning, and decision making* (pp. 39–66). Washington, DC: American Psychological Association.

Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men?. *Science*, *317*, 82.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*, 1–28.

Mollin, S. (2009). Combining corpus linguistics and psychological data on word co-occurrence: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory*, *5*, 175–200.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*, 143–182.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology: Section A*, *50*, 528–559.

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, and Computers*, *36*, 402–407.

Oostdijk, N. (2000). The Spoken Dutch Corpus: Overview and first evaluation. In S. Piperidis & G. Stainhaouer (Eds.), *Proceedings of Second International Conference on Language Resources and Evaluation* (Vol. 2, pp. 887–894). Paris: ELRA.

Ordelman, R. J. (2002). *Twente nieuws corpus (TWNC)* (Tech Rep.). Enschede: Parlevink Language Technology Group, University of Twente.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*, 161–199.

Peirsman, Y., Heylen, K., & Speelman, D. (2007). *Finding semantically related words in Dutch: Co-occurrences versus syntactic contexts.* Paper presented at the Sixth International and Interdisciplinary Conference on Modeling and Using Context, Roskilde University, Denmark.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In L. Schubert (Ed.), *ACL '93: Proceedings of the 31st annual meeting on Association for Computational Linguistics* (pp. 183–190). Stroudsburg, PA: Association for Computational Linguistics.

Ratcliff, R., & McKoon, G. (1994). Retrieving information from memory: Spreading-activation theories versus compound-cue theories. *Psychological Review*, *101*, 177–184.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*, 647–656.

Rosch, E., Mervis, C., Grey, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behaviour Research Methods, Instruments, & Computers*, *36*, 506–515.

Sahlgren, M. (2005). *The Word-Space Model: Using distributional analysis to represent syntagmatic and* *paradigmatic relations between words in high-dimensional vector spaces*. Unpublished doctoral dissertation, University of Stockholm, Sweden.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum.

Simmons, W., Hamann, S., Harenski, C., Hu, X., & Barsalou, L. (2008). fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology – Paris*, *102*, 106–119.

Steels, L., & Hild, M. (2012). *Language grounding in robots*. New York, NY: Springer.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 237–249). Washington, DC: American Psychological Association.

Szalay, L. B., & Deese, J. (1978). *Subjective meaning and culture: An assessment through word associations*. Hillsdale, NJ: Lawrence Erlbaum.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology*, *113*, 169–193.

Van Rensbergen, B., De Deyne, S., & Storms, G. (2014). *Examining assortivity in the mental lexicon: Evidence from word association data.*. (Manuscript submitted for publication).

Verheyen, S., De Deyne, S., Dry, M. J., & Storms, G. (2011). Uncovering contrast categories in categorization with a probabilistic threshold model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1515–1531.

Verheyen, S., Stukken, L., De Deyne, S., Dry, M. J., & Storms, G. (2011). The generalized polymorphous concept account of graded structure in abstract categories. *Memory & Cognition*, *39*, 1117–1132.

Vigliocco, G., & Vinson, D. P. (2007). Semantic representation. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 195–215). Oxford: Oxford University Press.

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*, 422–488.

Vinson, D. P., Vigliocco, G., Cappa, S., & Siri, S. (2003). The breakdown of semantic knowledge: Insights from a statistical model of meaning representation. *Brain and Language*, *86*, 347–365.

Wiemer-Hastings, K., & Xu, X. (2005). Content differences for abstract and concrete concepts. *Cognitive Science*, *29*, 719–736.

Wisniewski, E. J., & Bassok, M. (1999). What makes a man similar to a tie?. *Cognitive Psychology*, *39*, 208–238.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399–413.

APPENDIX A

OVERVIEW OF THE BASIC DESCRIPTIVES OF THE RELATEDNESS STUDIES

Study	*Set*	*Pairs*	*Raters*	*Reliability*
CONCRETE	Fruit	435	15	.91
	Vegetables	435	15	.88
	Birds	435	30	.91
	Insects	325	16	.89
	Fish	253	16	.87
	Mammals	435	17	.92
	Reptiles	231	22	.85
	Clothing	406	16	.92
	Kitchen Utensils	528	19	.90
	Musical Instruments	351	17	.92
	Tools	435	16	.86
	Vehicles	435	15	.96
	Weapons	190	22	.85
ABSTRACT	Art Forms	105	17	.95
	Crimes	105	17	.97
	Diseases	105	17	.95
	Emotions	105	17	.97
	Media	105	16	.94
	Sciences	105	18	.94
	Virtues	105	17	.94
DOMAIN	Animals A	300	12	.99
	Animals B	300	11	.99
	Artefacts A	435	18	.97
	Artefacts B	435	13	.97
THEMATIC	Miller Charles	30	18	.98
	Thematic (mixed)	100	33	.98

APPENDIX B

DOMAIN ITEMS

Artefacts A	*Artefacts B*	*Animals A*	*Animals B*
bass	accordion	boa	alligator
flute	drum	crocodile	cobra
harmonica	harpsichord	dinosaur	frog
piano	trumpet	iguana	lizard
tambourine	violin	salamander	tortoise
jeans	boots	gull	blackbird
pants	hat	ostrich	eagle
scarf	shirt	stork	parrot
sweater	skirt	swallow	peacock
swimsuit	suit	swan	rooster
axe	bow	bumblebee	butterfly
gun	dagger	caterpillar	cricket
spear	grenade	mosquito	dragonfly
stick	pistol	spider	grasshopper
sword	shield	wasp	moth
file	chisel	bat	cow
hammer	crowbar	monkey	dog
knife	grinding wheel	pig	donkey
nail	vacuum cleaner	rabbit	hedgehog
slicer	wheelbarrow	sheep	squirrel
apron	fridge	carp	eel
bottle	mixer	salmon	sardine
fork	scissors	shark	swordfish
grater	sieve	squid	trout
oven	stove	stingray	whale
balloon	moped		
bicycle	plane		
hovercraft	sled		
train	taxi		
tram	tractor		

APPENDIX C

THEMATIC PAIRS

angel–hat
author–theatre
avalanche–snow
bandit–fanfare
bed–mattress
bee keeper–honey
body part–leg
bones–fish
brewer–beer
burglary–abbey
bury–death
cake–pie
care–help
cigar–smoking
cradle–baby
cue–billiards
cultivate–grow
cynical–bitter
danger–profession
decadent–champagne
decisive–important
dromedary–desert
export–output
falcon–squirrel
field–dough
fire–flame
flowers–birthday
future–uncertain
gill–breathe
giraffe–neck
gland–swollen
gorilla–robber
gravel–red
handle–door
hot dog–food
ingenious–fantastic
injection–syringe
jar–grain
judge–points
juggle–conjure
key–treasure
kick-off–soccer
launch–rocket
lucid–clear
lump–sugar
mouth–river
oeuvre–work
pattern–regularity
percentage–discount
performance–reward
pinkie–finger
plain–sand
poodle–biscuit
prairie–wolf
prey–booty
prick–sting
principle–theorem
puff cake–pastry
quack–duck
quarter–test
queen–watch
raft–lion
rage–yell
rave–fever
recent–young
reed–grass
ring–call
romp–play
rotate–turn
rumour–gossip
servant–flour
shot–dark
smother–stench
snail–slow
song–fun
soon–swift
spontaneous–smile
stain–wash
stem–eel
step–stairs
stomach–intestines
strings–guitar
structure–dust
stub–cigarette
stubble–beard
stumble–pain
styrofoam–rubber
syndrome–disease
tame–circus
task–sin
thunder–lightning
tide–flood
tragedy–drama
twig–tree
umbrella–rain
volley-ball–net
voter–politics
wad–cable
wagon–train
weight–exercise