Running head: QUANTITATIVE AND QUALITATIVE CONTRIBUTIONS

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

University of Leuven, Department of Psychology, Tiensestraat 102, 3000 Leuven, Belgium

Simon De Deyne

Faculty of Psychology and Educational Sciences

University of Leuven

B-3000 Belgium

e-mail: simon.dedeyne@ppw.kuleuven.be

Abstract

One of the most significant recent advances in the study of semantic processing is the advent of models based on text and other corpora. The success of these lexico-semantic models often depends on a combination of large quantities of data combined with sophisticated ways (e.g., dimension reduction) to maximize knowledge encoded at the language input side. 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 word associations and a model based on syntactic dependency relationships perform very differently depending on the range of relatedness considered. In addition, we find that only certain types of linguistic input contribute to the prediction of semantic relatedness and that limited amounts of data suffice to obtain reliable predictions. This finding suggests 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, syntax, word associations, similarity.

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

Computational models of word meaning based on text-corpora, have played an important role in explaining how meaning is acquiring from the language environment. Their impressive scale and comprehensive scope have proven instrumental in uncovering a structure in the lexicon of both children (Denhière & Lemaire, 2004) and adults (Jones & Mewhort, 2007) using a large variety of experimental tasks (Vinson, Vigliocco, Cappa, & Siri, 2003). However, as to date it remains unclear whether representations in lexico-semantic models derived from written or spoken language are general enough to capture the meaning of all kinds of word representations, including those for concrete and abstract words (Vigliocco & Vinson, 2007). Moreover, we do not know if such models adequately differentiate between words belonging to the same basic level category such as birds or only capture differences at a more general level such as the domain of animals. In other words, even though the models are very comprehensive in terms of their vocabulary, we do not know how detailed the representations of these words are. This warrants an investigation of the ability of lexico-semantic models to capture the meaning of items from a vast range of domains and categories. Nor do we know whether the amount of information available through corpora is a good approximation of the amount of linguistic exposure of an individual. As the amount of information available for words is likely to have implications for the representation of these words, an investigation of the amount of data required for lexico-semantic models to capture meaning is warranted as well.

It is generally assumed that the structure in language is deeply embedded in the mental lexicon and thus that the lexico-semantic models that capture this structure are realistic models of how the mental lexicon is acquired. Presumably this assumption reflects the origins of many of these models outside of psychology, in domains such as information retrieval and computational linguistics, where assumptions regarding mental representations are often of secondary importance. Whether these assumptions are tenable remains to be tested empirically. This warrants a comparison

of models with different assumptions.

Outline

To address the above questions we will use two large-scale lexico-semantic models. The first model is derived from written and spoken text corpora. It is specifically selected because, unlike earlier models like Latent Semantic Analysis (LSA, Landauer & Dumais, 1997), it considers syntax and word order and has shown superior performance in previous studies (Heylen, Peirsman, & Geeraerts, 2008). The second model is based on word associations. It is chosen because it relies on a unique novel data set that includes a comprehensive set of cues, covering most of the human lexicon, while also addressing some of the shortcomings of previous approaches through a methodology that captures both strong and weak semantic links (De Deyne, Navarro, & Storms, 2012).

In the empirical part of the paper we evaluate these models by investigating to what degree they can predict semantic relatedness judgements. This task is chosen because the notion of relatedness is central to both models and crucial both at the level of acquisition of meaning and the manner in which it is represented. To capture relatedness for the majority of the words represented in the mental lexicon, we will include both abstract and concrete concepts and consider semantic relationships of a different scope as well. On the one hand this involves studying relations at the level of semantic domains, thus allowing to test the models at different resolutions. On the other hand, it includes thematic judgements, which reflects our consideration of the mounting evidence that thematic relations structure the lexicon (Lin & Murphy, 2001; Gentner & Kurtz, 2005; Wisniewski & Bassok, 1999). Contrasting these two models allows for further interpretation of their relative performance. While both types of models are related in the sense that they are derived from language output, they might highlight different types of information. Specific to the text-based model, we will show how adding additional assumptions about syntax provides a closer match to representations in the mental lexicon and offers insight in the role it plays for all kinds of words.

Following the theme of this special issue, we will then investigate the effect the amount of

information has on representations derived from corpora. More specifically, we will investigate the limits to what these models can represent as a function of the amount of information available.

Throughout the paper, our view of the mental lexicon is based on a semantic network. This choice reflects the parallels we wish to draw between early models of semantic memory (which often have their origin in cognitive science) and more recent lexico-semantic models (which tend to make less explicit claims about underlying mental properties). Moreover, while many lexico-semantic models are predominantly presented as spatial models, a network interpretation of these models closely aligns with existing psychological theories about the lexicon. Before proceeding, we will therefore briefly discuss the points of convergence between early network theories and more recent high-dimensional models, which allows us to grasp how the two large-scale lexico-semantic models introduced in this study are related to each other.

Adding linguistic and mental constraints to lexico-semantic models

Some of the shortcomings of early computational semantic models are addressed by text-based lexico-semantic models that make use of vast quantities of textual data derived from written or spoken resources that are often publicly available. One of the key advantages of these models is that they are comprehensive: They can be used to infer the meaning of any arbitrary word. In contrast to many of the early network approaches, the information derived from textual data does not suffer from a selection bias, but considers all possible linguistic utterances. The key idea behind these models is that the meaning of a word can be inferred from the context in which it is used. One of the main implications of learning from context is that two words are deemed more similar, if they co-occur in similar contexts.

Roughly, two types of models can be distinguished depending on the type of context. A first type uses a document-based representation. In these models, the meaning of a word is determined by the documents in which it occurs. Examples of such models are Latent Semantic Analysis (LSA, Landauer & Dumais, 1997) and topic models (Griffiths, Steyvers, & Tenenbaum, 2007). A second

type of models is word-based. These models typically use co-occurrences derived from a stream of words, where only words co-occurring in a window of 1 to 10 words around the target word are counted. Examples of a word-based approach are the Hypertext Analogue of Language (HAL) model (Lund & Burgess, 1996), the Word Association Space (WAS) model (Steyvers, Shiffrin, & Nelson, 2004) and more recently the Bound Encoding of the Aggregate Language Environment (BEAGLE) model (Jones & Mewhort, 2007).

Many text-based lexico-semantic models tend to be flexible and are able to explain a whole array of findings, ranging from low-level word processing to high level text comprehension (Landauer, 2007). Lexico-semantic models can be constrained in a number of ways to render them more plausible models of semantic memory. A first constraint is whether a document or word based representation is preferred. In this respect, it could be argued that word co-occurrence statistics derived from a short paragraph or sentence are more likely to be mentally encoded than word document statistics if humans are limited in tracking co-occurrence statistics over thousands of words uttered over a considerable time-span. Moreover, the document unit might not be the most natural context to express a single idea since most written text often introduces multiple independent subplots, which alternate even faster in spoken language. Finally, document-based models lead to a representation that has a split personality. In LSA, for instance, the rows of the input matrix correspond to words and are qualitatively different from the columns which represent abstracted context derived from documents or words co-occurring in a sentence. In network terms, such a representation corresponds to a bipartite graph. This can be avoided when the mental lexicon is implemented as a unipartite graph. In such a graph, each node represents a word and is connected through incoming and outgoing links, providing a straightforward interpretation and shared lexicon for nouns, adjectives, verbs, and so on.

Second, a major limitation of text-based lexico-semantic models is the application of the *bag-of-words* assumption. In models like LSA and HAL, word order or syntax are often omitted. Clearly, simple co-occurrence of words represents a minimal assumption on how mental models can

be build from language. More recent models have addressed this limitation to some extent. This is the case in BEAGLE, which accounts for word order (Jones & Mewhort, 2007), and further enhancements of the topic model, which include notions of syntax (Griffiths, Steyvers, Blei, & Tenenbaum, 2004). While syntax and word order may not have a primary role in deriving meaning (see Landauer, 2007), they clearly carry psychological meaning and might provide us with a unique opportunity to investigate what and how information becomes mentally represented. Previous work in computational linguistics indicates that including this information often provides better results than a simple bag-of-word approach (Lapata, McDonald, & Keller, 1999; Peirsman, Heylen, & Speelman, 2007).

Finally, most psychological studies do not directly address the properties of the text corpus. One of these properties, corpus size, tends to vary widely. It ranges from 5 million words in the carefully compiled Touchstone Applied Science Associates Inc. (TASA) corpus that is used in LSA over > 100 million words in the British National Corpus (Aston & Burnard, 1997) to the Google *n*-gram corpus based on 1 trillion words (Michel et al., 2011). When a corpus is too small or too large, this might affect the representations that are derived from it. This might explain why certain phenomena go undetected. One of them is mediated priming, in which a prime and target are only indirectly related through a mediator (e.g. LION → TIGER → STRIPES). Being able to predict such findings crucially depends on the type of context and corpus size. An illustrative case is the study by Jones, Kintsch, and Mewhort (2006), where HAL failed to account for mediated priming with only the first 1000 words derived from the relative small TASA corpus, while denser models like BEAGLE or LSA did account for the mediated priming effects. In the next section, we will describe two large-scale lexico-semantic models (based on syntactic dependency paths and word associations) that consider these constraints.

Syntax Dependency Network

Word meaning and part-of-speech characteristics are highly correlated in many languages. Generally speaking, verbs tend to refer to actions, nouns to entities, and adjectives to properties of these entities. Apart from the information about part-of-speech, other syntactic properties at the sentence level are likely to contribute to our understanding of the sentence and its constituents. Consider the following sentence:

“While watching the moon, a falling apple gave Isaac Newton the idea that both the apple and the moon are subject to the same gravity.”

There are many noteworthy facts about the linguistic properties of this sentence. We name just a few in order to illustrate how additional linguistic information can help us build more realistic mental models compared to a simple bag-of-words approach. First, this sentence exhibits a complex structure not uncommon to many written text sources. It has a nested structure, the meaning of which can be derived by rearranging various parts. Some parts like *while watching the moon* are dependent clauses, which indicates an idea that is subordinate to the main message. Second, many sentences like this one, contain some form of anaphora, where *he* refers to Isaac Newton. Moreover, *Isaac Newton* is a special word. It refers to a name, and although it consists of two words, these words belong together as a multi-word unit. This means that the words *give* and *idea*, which are syntactically related, do not necessarily need to be adjacent within an n -gram window. Below, we will present a procedure through which such rich linguistic structure can be derived from text.

Materials and Procedure

We compiled the first Dutch corpus that is of adequate size to conduct psycholinguistic research. It consists of three language resources spanning different registers. A first source consists of text derived from Dutch articles in newspapers and magazines. It consists of the Twente Nieuws Corpus of Dutch (Ordelman, 2002) combined with Flemish Dutch texts from the Leuven Newspaper corpus (LeNC, Heylen et al., 2008). A second source consists of more informal language retrieved

from internet web pages. This corpus consists of 1,000 documents for each of 8,568 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. It includes Dutch movie subtitles and the Corpus of Spoken Dutch (Oostdijk, 2000). The total number of tokens in these sources was 79 million. The majority of these tokens were obtained from newspaper material: About 62% of them were taken from Belgian Dutch newspapers and magazines, 12% from Dutch newspapers from the Netherlands. The remaining two sources consisted mainly of less formal online text (25%) and spoken materials (1%).

Each sentence in the corpus was parsed using the Alpino dependency parser for Dutch (Bouma, van Noord, & 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. 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.

Results

The resulting corpus vocabulary consisted of 157 million co-occurrence tokens derived from undirected dependency paths and 103,842 different lemmas; 82.7% were nouns, 12.6% adjectives, 4.5% verbs and 0.2% adverbs. A separate dependency matrix was constructed for each of eight dependency patterns. This resulted in eight $N \times N$ dependency matrices G_p , one for each pattern 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.

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 HdPrefc. 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 holds 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 had 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 C. 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

Association measures have been successful in the prediction of many semantic phenomena, such as the distance effects in free recall and cued recall (Steyvers et al., 2004) or semantic priming (Chumbley & Balota, 1984). We used the Leuven association norms collected between 2003 and 2006 (De Deyne, Navarro, & Storms, 2012). In contrast to other word association studies (Kiss, 1968; Nelson, McEvoy, & Schreiber, 2004), this study used a continued procedure in which three associations per cue were collected from each participant instead of one. This continued procedure has two advantages. First, weak(er) associations can be collected, which is especially important for cues with very dominant associations (e.g., BLOOD and RED). Second, the resulting representations are denser and therefore more suited for a distributional approach of meaning. In the following section we briefly describe the most important characteristics of this study and refer to De Deyne,

Navarro, and Storms (2012) for a full description.

Materials and Procedure

A total of 71,380 native Dutch speakers provided associations. Stimulus words or 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 way¹.

Results

The network is constructed from a weighted adjacency matrix where both the rows and 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.

The use of a continued procedure was primarily devised to measure weak links and increase the density of the graph. To measure the effectiveness of this procedure, graphs were derived from the data by selecting those responses that were in the set of 12,418 cues and counting the number of responses for these words. In effect this corresponds to a transformation from a bipartite graph representation to a unipartite representation. Two graphs were derived: G_{asso1} , a graph based on the primary responses and comparable to other single response datasets (cfr Nelson et al., 2004) and G_{asso123} , a graph including the secondary and tertiary responses as well. The graph 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, Navarro, & Storms, 2013, for further 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 Experiments

The following experiments evaluate the effect of domain and semantic relation-type on the ability of lexico-semantic models to predict relatedness. The experiments are organized along two main distinctions. First, concrete entities are compared with abstract entities. If the meaning of abstract words is predominantly derived from how these words are used in context, we expect both the syntactic and association models to do better for these words than for concrete entities. This is due to the fact that in contrast to concrete words, which can be understood in isolation, abstract words are relational in nature (Goldstone, 1996; Vigliocco, Vinson, Lewis, & Garrett, 2004; Verheyen, Stukken, De Deyne, Dry, & Storms, 2011; Wiemer-Hastings & Xu, 2005). In a first series of experiments, similarity judgements for all pairwise combinations in basic level concrete and abstract categories such as mammals, clothing, emotions and sciences are collected. Because these comparisons are performed at a basic level, they require an evaluation of nuanced and detailed properties (for instance, when comparing HAMSTER and MOUSE).

It might be the case that lexico-semantic models cover semantics at a wider range. In other words, rather than distinguishing structure at the detailed basic category level, they might allow a more coarse structure that distinguishes entities between categories. To investigate this possibility, a second experiment was included where items from various basic level animal or artefact categories were paired, leading to pairs such as BUTTERFLY and EAGLE or ACCORDION and FRIDGE. If lexico-semantic models are primarily sensitive to domain-level differences, we expect better predictions for these differences than those derived for basic-level categories.

Many psychological theories assume that semantic knowledge is organized according to

frames rather than hierarchical categories (Schank & Abelson, 2013). This implies that any lexico-semantic model must be able to account for a possible thematic structure in the mental lexicon. This hypothesis was evaluated in the remaining experiments.

A second goal of our study is to evaluate the role of sample size in both the text and association model. One possibility is that many of the existing accounts simply underestimate certain semantic relationships, because their sample size is too small. This is not an unlikely possibility for certain categories, like concrete entities at a basic level. Here again, including different types of words allows us to investigate whether the requirements of the amount of knowledge is distinct for concrete, abstract, domain or thematic pairs.

Finally, we have argued that syntactic dependencies offer a vast improvement over typical bag-of-word models and other extant frameworks. If this holds true, we expect to be able to capture relatedness better by explicitly taking into account these different relations. Perhaps more important than a pure quantitative evaluation, is the identification of the role that different types of syntactic dependency relations play. Do the different syntactic dependencies contribute equally to semantic relatedness? Here as well, the inclusion of four different types of concept pairs can learn us which dependency types are specifically important.

Materials and 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.

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 corresponds to that in Verheyen 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 generalizability 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 part of this set was the Thematic (mixed) experiment and similar to Miller and Charles 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. Translated pairs are available in Appendix C.

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 judgments 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) dataset. In this set a total of 69 pairs consisted of nouns only, while the remaining 31 pairs comprised at least one verb or adjective. To facilitate the interpretation of specific syntax effects, only the data for the nouns were analysed.

Results

Deriving relatedness from the networks

First we assess how well relatedness derived from the text-based and association-based network can account for the human relatedness judgments.

Both the text-based network and the association-based network represent weighted graphs, with the weights reflecting the co-occurrence frequency of two words (either as a function of their syntactic relation or 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 range (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

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 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 graph 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 similarity 1. For each item pair within a set, the cosine similarity between the items was calculated, both based on the text-based network (G_{syn}) and on the network based on the first (G_{asso1}) or all three associates (G_{asso123}).

Judged relatedness was standardized to calculate the correlations over different concrete, abstract, domain, or thematic sets. The correlations between empirical relatedness and relatedness derived from G_{asso1} , G_{asso123} , and G_{syn} are shown in Table 3.³ N indicates the number of item pairs across which the correlations are taken. Comparing the different types of models, the results in Table 3 suggest that in all sets G_{asso123} outperforms the text-based models and the model based on a single associate G_{asso1} . The differences between the correlations were tested using the procedure of (2, 2) for dependent correlations and are presented in Table Table 3.

In line with our hypotheses, domain judgements yielded higher correlations than judgements for abstract and concrete category items. Similarly, if we assume that thematic judgments are primarily defined as relational properties, we expect better results for these judgments than for concrete judgments. This was the case for the both the text and association model. They both represented thematic judgments 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 model.

One possible concern is that the performance of the association models depends on direct associates between word-pairs. Especially in the thematic set, certain pairs are likely to be directly associated. While 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. Associative strength 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 they were highly correlated, $r = .660$, $CI = [.,.]$, for G_{asso123} , but this was still lower than the reported .823. A similar situation holds for the other sets, concrete .321, abstract .389, and domain, .258. As can be expected, this pattern also holds for G_{asso1} , with thematic = .607, concrete .261, abstract .347 and domain .237. Altogether these results show that associative strength and the similarity scores don't necessarily measure the same thing, even for concepts that are thematically related. This is in line with previous reports that show that participants find it very

difficult to judge how strongly associated two words are (Judgments of associated strength).

Invariably, these judgments are better predicted by similarity judgments.

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 judgments is assessed.

For the syntax model G_{syn} , $n = 100$ 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 n 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 graph, 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 text-based model, shown in the left panel of Figure 1, show fast improvement that flattens off when more than 30% of the data is added to G_{syn} . For the Concrete set, sample sizes of 32% and larger were not significantly different from a sample using all data ($r_{32\%} - r_{100\%}$: $z = -1.64$, ns). The required sample size was even smaller for the abstract set ($r_{13\%} - r_{100\%}$: $z = -1.25$, ns) and domain set ($r_{33\%} - r_{100\%}$: $z = -1.47$, ns). For the thematic set, a two percent sample was just as good as a larger sample ($r_{2\%} - r_{100\%}$: $z = -1.37$, ns).

The results of manipulating corpus size for G_{asso123} are shown in the right 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 50 participants are included. This general pattern is confirmed by the minimal sample size required for a correlation to be indistinguishable from that based on the complete data. For the concrete set this was 57% ($r_{57\%} - r_{100\%}: z = -1.63, \text{ns}$), for the abstract set 36% ($r_{36\%} - r_{100\%}: z = -1.39, \text{ns}$), for the domain set 58% ($r_{58\%} - r_{100\%}: z = -1.42, \text{ns}$), and for the thematic set 19% ($r_{19\%} - r_{100\%}: z = -1.56, \text{ns}$).

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 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 (Jones & Mewhort, 2007). However, as indicated by the shaded area in Figure 1, strong conclusions in this respect are likely to be preliminary.

Weighted syntactic dependencies

Although the syntax model we used represents the state of the art in terms of text-based network models, one might object to the previous analyses that ignoring the dependency relationships underlying G_{syn} could underestimate the results. Instead of treating all dependency relations equal, we therefore propose a relatedness measure that allows for a different contribution of the dependency relations. This has the added benefit that it yields an indication of the sources of information the text-based model relies upon for the prediction of relatedness.

To allow a differential contribution depending on the type of dependency path, relatedness was derived from a weighted sum of the dependency similarity matrices $\mathbf{S}$ with pattern p as follows:

$$\mathbf{S}_w = \frac{1}{p} \sum_i^p \beta_i \mathbf{S}_i \quad (1)$$

with $\beta \in [0, 1]$, $\sum \beta = 1$, and $\mathbf{S}_i$ the similarity matrix for a specific dependency relation

calculated as before. The positive weights reflect the additional assumption that only positive evidence can contribute to mental representations based on syntactic dependency paths. This additional assumption is somewhat restrictive, but facilitates the interpretation of the results since two words cannot become more dissimilar by being observed in a syntactic dependency relation⁴.

The results in Table 3 indicate that distinguishing different types of dependency paths indeed improves the prediction of human ratings for all data sets. It represents the best results for G_{syn} across all four data sets, but the weighted dependence model is still outperformed by the association-based model.

To clarify the role of the different syntactic dependency paths, the estimated weights β_i are presented in Figure 2. This figure again shows that across different data sets the distribution across weights is far from uniform. Instead, the weights seem consistently distributed among a subset of dependency paths. Quite remarkably, a consistent contributor to the relatedness structure is due to a conjunctive pattern. This was particularly the case for the Abstract words. Moreover, attributional meaning, captured by the HdMod relation is important for all data sets except the Thematic data, where SuObj relations are most influential. This confirms our intuition about the relative role of attributes or frame information (e.g., actor – patient) and other information in this comparison.

As was shown in Table 1, the number of observations varied strongly depending on the type of dependency. To investigate to what degree the weights in Figure 2 reflect this number, we calculated for all the items in each dataset the number of observations for each dependency separately. The frequencies for each dependency were converted to sum to one and are shown in Figure 2. Comparing the weights and the frequency proportions shows only a modest agreement between both quantities, thus excluding the possibility that occurrence frequency is indicative of a syntactic dependency’s contribution to the relatedness prediction.

General Discussion

In this study we tested two key assumptions underlying 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 judgments. The most striking result was that this ability varies widely, depending on what kind of relatedness is measured. A syntax-based model only yields weak to moderate correlations at the basic level of a category. This is especially the case for concrete concepts such as birds or tools, which by definition involve information that is mostly attributional and of sensory nature. As expected, the situation is better for thematic relations and for abstract concepts such as SCIENCES or VIRTUES, where the distributional semantics in the linguistic environment is 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 judgments at a domain level are included. This shows that the distinction between entities such as birds and mammals or vehicles and tools is well represented through language. The inclusion of an alternative semantic model based on word associations allows us to put these findings in a larger context. First of all, relatedness derived from a continued association task systematically improves predictions for all kinds of concepts and relations considerably. In contrast to the text model, the difference between concrete and abstract concepts was less pronounced (see further) and better predictions of thematic judgments were obtained.

The second assumption we tested was that adding additional constraints can learn us something about the mental properties of word meaning. More precisely we focused on (a) the quantitative aspects of language exposure and (b) the role of syntax. In both cases the findings point towards the same conclusion: only a small subset of information available in language and syntax contributes to word meaning (as measured through relatedness). More precisely, regardless of the type of concepts and comparisons, a corpus that was less than one third the size of our complete text corpus performed equally well, indicating that ever larger corpora do not always offer much improvement.

In addition, while weighting different syntactic patterns showed some improvement of the relatedness predictions, only a small set of patterns were driving this result. Especially conjunctions played a big role, and indirectly this might cast a shadow on how general the claim is that meaning is encoded in language, as this type of pattern represents only a minority of the total syntactic paths. This indicates that there are limits to what can be learned from language and these limits are reached much faster than what we could expect given the amount of information available.

These findings have a number of theoretical and methodological implications, raise new questions, and provide a different perspective on classic issues in our field. These include (i) the contributions to word meaning of sources outside language and the differential representation of different types of concepts, and (ii) the role of language exposure and the mechanisms of acquisition. In the following sections, we will elaborate on these points.

Non-linguistic contributions to word meaning

A number of researchers who study embodied representations have criticized the lexico-semantic approach because the symbolic representations contained are ungrounded and fail to capture sensory or modal-specific information (Glenberg, 1997; Pecher & Zwaan, 2005). In recent years, this debate has become less polarized and most researchers would now agree that the truth lies somewhere in between. That is, it is considered likely that the meanings of words are represented across a variety of modalities that differ in whether they represent sensory representations or perhaps more symbolic language-based representations (e.g. Barsalou, Santos, Simmons, & Wilson, 2008; Vigliocco et al., 2004). In response, some studies have 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 models derived from text and those derived from word associations. Even though the text-based results clearly 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 a common ground between speakers. This type of pragmatics explain 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, while the first response tends to reflect lexico-semantic properties (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.

Dealing with sparsity of language input and an abundance of text materials

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 do not necessarily result in more human-like representations. Next, the signal-to-noise ratio could be improved by selecting materials that closely align 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. Finally, one could try to infer new meaning from the linguistic environment through various unsupervised techniques. It is this last strategy that has received 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 (Jones & Mewhort, 2007), non-negative dimension reduction (Hoyer, 2004), self organizing maps (Vinson et al., 2003), probabilistic inference over topics (Griffiths et al., 2007), and random graph walks (Hughes & Ramage, 2007). Each of these techniques reduces the sparsity, either at a representational level (through singular value decomposition for instance) or on-line (through random graph walks, e.g., De Deyne, Navarro, Perfors, & Storms, 2012). They often improve the results, especially for small to moderately sized corpora, but not necessarily for large corpora (see Recchia & Jones, 2009). Especially in syntax- and word-based models dimensionality reduction did not lead to substantial improvements (Bullinaria & Levy, 2007). Despite the contentious nature of dimensionality reduction, inferring additional structure from language input is likely 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.

While the inference mechanisms may introduce unnecessary complexity, turning to ever expansive corpora entails a different risk. The unlimited amounts of linguistic data that is available nowadays could result in corpora that overestimate the redundancy encoded in the linguistic environment as is potentially the case when a corpus of more than 2.6 billion words like the English Wikipedia is used. Instead, 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. 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, Mander, & Stevens, 2014). The percentage differs depending on age (around 50% for 12-year old persons and 80% for 80-year old persons). 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, Slater, & 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 157 million words is a more realistic starting point than a rather small corpus based on reading materials such as the TASA corpus and constraints future models that are potentially much larger. At the same time, our results clearly show that providing a more realistic approximation of language input does not necessarily improve the quality of the mental representations that can be derived from language.

Methodological implications and conclusion

Before closing, let us elaborate on at least one methodological implication of the current findings. 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. Additional evidence supports the idea that these weak links are important in other tasks as well. A case in mind is a study by Maki (2007) who used a judgment task of associative strength and tried to explain why participants over-estimate 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, Navarro, Perfors, & Storms, 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 an accurate encoding of sensory-based attributes.

At the same time, measures derived from text might be better suited to provide insight into phenomena that go beyond a single word such as how humans extract the gist of a story. Regardless of the specific questions, the availability of large data sets open up many new areas to systematic inquiry. From the myriad of alternatives out there, a priori not all are equally suited. The success of the models derived from them will increasingly reflect the degree to which they capture the mental properties of language. At least from a purely psychological point of view, the limits of our language

(derived from huge text corpora) are not the limits of our world (see Wittgenstein, 1921).

References

- Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon*. Wiley-Blackwell.
- Andrews, M., Vigliocco, G., & Vinson, D. P. (2005). The role of attributional and distributional information in representing meaning. In *Proceedings of the Twenty-Seventh Meeting of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Aston, G., & Burnard, L. (1997). *The BNC handbook exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
- Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Symbols, embodiment, and meaning. In M. De Vega, A. M. Glenberg, & A. C. Graesser (Eds.), (pp. 245–283). Oxford: Oxford University Press.
- Bouma, G., van Noord, G., & Malouf, R. (2000). Alpino: Wide coverage computational analysis of Dutch. In *Computational linguistics in the Netherlands CLIN*.
- Bruni, E., Uijlings, J., Baroni, M., & Sebe, N. (2012). Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM international conference on multimedia* (pp. 1219–1228).
- 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, Belgium: Dept of Psychology, 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.
- Chumbley, J. I., & Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Memory & Cognition*, 12, 590-606.
- Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. In U. Zernik (Ed.), (p. 115-164). Hillsdale, NJ: Lawrence Erlbaum Associates.
- De Deyne, S., Navarro, D., & Storms, G. (2012). Better explanations of lexical and semantic cognition using networks derived from continued rather than single word associations. *Behavior Research Methods*, Advance online publication. 10.3758/s13428-012-0260-7.
- De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2012). Strong structure in weak semantic similarity: A graph based account. In *Proceedings of the 34th annual conference of the cognitive science society*

- (pp. 1464–1469). Austin, TX: Cognitive Science Society.
- De Deyne, S., Navarro, D. J., & Storms, G. (2013). Associative strength and semantic activation in the mental lexicon: evidence from continued word associations. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (p. 2142-2147). Austin, TX: Cognitive Science Society.
- 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.
- Denhière, G., & Lemaire, B. (2004). A Computational Model of Children's Semantic Memory. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (p. 297-302). Mahwah, NJ: Erlbaum.
- Gentner, D., & Kurtz, K. (2005). Categorization inside and outside the lab. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), (p. 151-175). Washington, DC: American Psychology Association.
- Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, 20, 1-55.
- Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, 24, 608-628.
- Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating topics and syntax. In *Advances in neural information processing systems* (pp. 537–544).
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114, 211-244.
- Heylen, K., Peirsman, Y., & Geeraerts, D. (2008). Automatic synonymy extraction. *A Comparison of Syntactic Context Models. LOT Computational Linguistics in the Netherlands, 2007*, 101–116.
- 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 *EMNLP-CoNLL* (pp. 581–589).
- Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of memory and language*, 55(4), 534–552.
- Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite

- holographic lexicon. *Psychological Review*, 114(1), 1-37.
- Kiss, G. R. (1968). Words, associations, and networks. *Journal of Verbal Learning and Verbal Behavior*, 7, 707-713.
- Landauer, T. K. (2007). LSA as a theory of meaning. *Handbook of latent semantic analysis*, 3-35.
- 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 *Proceedings of the 9th conference of the European Chapter of the Association for Computational Linguistics* (pp. 30-36).
- Laurence, S., & Margolis, E. (2001). The Poverty of the Stimulus Argument. *The British Journal for the Philosophy of Science*, 52(2), 217-276. doi: 10.1093/bjps/52.2.217
- Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, 1, 3-28.
- 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 18th International Florida Artificial Intelligence Research Society* (p. 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(4), 319-353.
- Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317(5834), 82-82.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... others (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.
- 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 *Proceedings of second*

- international conference on language resources and evaluation (LREC)* (p. 887–894).
- Ordeman, R. J. (2002). *Twente nieuws corpus (TWN)* (Tech. Rep.). Parlevink Language Technology Group, University of Twente.
- Padó, S., & Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33, 161–199.
- Pecher, D., & Zwaan, R. A. (Eds.). (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. New York: Cambridge University Press.
- Peirsman, Y., Heylen, K., & Speelman, D. (2007). Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In *Proceedings of the CoSMO workshop*. Roskilde, Denmark.
- Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 31st annual meeting on association for computational linguistics* (pp. 183–190).
- 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.
- 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, and 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.
- Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum.
- Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Experimental Cognitive Psychology and its Applications. In A. Healy (Ed.), (chap. Word association Spaces for Predicting Semantic Similarity Effects in Episodic Memory.). Washington, DC: American Psychological Association.
- 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(6), 1117–1132.
- Vigliocco, G., & Vinson, D. P. (2007). The oxford handbook of psycholinguistics. In G. Gaskell (Ed.), (pp. 195–215). Oxford, UK: 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, Sep). 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.
- Wittgenstein, L. (1921). *Tractatus logico-philosophicus*. Routledge & Kegan Paul.

Appendix A

Overview of the basic descriptives of the relatedness experiments.

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

Author Note

My 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 post-doctoral fellow at the Research Foundation - Flanders. Comments may be sent to `simon.dedeyne@ppw.kuleuven.be`

Footnotes

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

²In 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 chosen nevertheless to increase comparability of the text-model where PMI is standard applied.

³For this table and the remainder of the text all correlations are $p < .05$ using a one-tailed t -test, unless stated otherwise.

⁴A regression-based model with less restrictions on β performed slightly better, but did not alter the pattern of results.

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 $F_{\text{tok}} (\times 10^6)$ are listed in the 4th and 5th column.

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

Table 2

Dutch examples and English translations for the 5 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	HdObj
bezoeker (visitor)	drank (drink)	drinken (to drink)	drinken (to drink)
mens (human)	thee (tea)	serveren (to serve)	zetten (to make)
man (man)	water (water)	schénken (to pour)	gaan (to go)
team (team)	product (product)	zetten (to make)	schénken (to pour)
iemand (someone)	leven (life)	maken (to make)	krijgen (to get)

Table 3

Results of the similarity analyses for the four data sets (Concrete, Abstract, Domain, and Thematic) and the four types of graphs. Confidence intervals for cosine relatedness (Cosine rel.) and weighted syntactic dependencies (w(Syntax)) are indicated by square brackets.

	Concrete	Abstract	Domain	Thematic
N	4562	545	1470	94
G_{asso1}	.549 [.528,.569]	.515 [.451,.574]	.711 [.685,.735]	.650 [.515,.753]
G_{asso123}	.622 [.604,.639]	.617 [.562,.666]	.792 [.773,.811]	.823 [.744,.879]
G_{syn}	.366 [.341,.391]	.517 [.453,.576]	.679 [.651,.706]	.588 [.439,.707]
G_{wsyn}	.378 [.353,.402]	.538 [.476,.595]	.690 [.662,.716]	.606 [.459,.720]

Table 4

Comparison of the correlation strenghts of the models. Values between brackets indicate the 95 % confidence intervals for the difference between dependent correlations. Only the significant results excluding zero from this interval at $\alpha = .05$ are displayed.

		Concrete		Abstract		Domain		Thematic	
G_{asso1}									
	G_{asso123}	-.072	[-.088, -.058]	-.101	[-.147, -.057]	-.100	[-.065, -.081]	-.081	[-.099, -.065]
	G_{syn}	.183	[.154, .212]			.032	[.002, .064]		
	G_{wsyn}	.170	[.142, .199]						
G_{asso123}									
	G_{syn}	.256	[.230, .282]	.100	[.035, .166]	.114	[.086, .142]	.234	[.114, .376]
	G_{wsyn}	.244	[.218, .270]	.078	[.013, .145]	.103	[.077, .130]	.217	[.102, .355]
G_{syn}									
	G_{wsyn}	-.012	[-.021, -.002]						

Figure Captions

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 indicate 95% of the correlations based on 10 repeated draws from 100 samples.

Figure 2. Estimated β weights for each of the eight syntax dependencies in the four data sets. A Stair plot shows the frequency contribution (proportional) for each dependency pattern.