Behav Res
DOI 10.3758/s13428-012-0260-7

Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations

Simon De Deyne · Daniel J. Navarro · Gert Storms

Abstract In this article, we describe the most extensive set of word associations collected to date. The database contains over 12,000 cue words for which more than 70,000 participants generated three responses in a multiple-response free association task. The goal of this study was (1) to create a semantic network that covers a large part of the human lexicon, (2) to investigate the implications of a multiple-response procedure by deriving a weighted directed network, and (3) to show how measures of centrality and relatedness derived from this network predict both lexical access in a lexical decision task and semantic relatedness in similarity judgment tasks. First, our results show that the multiple-response procedure results in a more heterogeneous set of responses, which lead to better predictions of lexical access and semantic relatedness than do single-response procedures. Second, the directed nature of the network leads to a decomposition of centrality that primarily depends on the number of incoming links or in-degree of each node, rather than its set size or number of outgoing links. Both studies indicate that adequate representation formats and sufficiently rich data derived from word associations represent a valuable type of information in both lexical and semantic processing.

S. De Deyne · G. Storms
Faculty of Psychology and Educational Sciences,
University of Leuven,
Tiensestraat 102,
3000 Leuven, Belgium

S. De Deyne · D. J. Navarro
School of Psychology, University of Adelaide,
5000 Adelaide, Australia

S. De Deyne (✉)
Laboratory of Experimental Psychology,
Tiensestraat 102 box 3711,
3000 Leuven, Belgium
e-mail: simon.dedeyne@psy.kuleuven.be

Published online: 10 October 2012

Keywords Word associations · Semantic network · Lexical decision · Semantic relatedness · Lexical centrality

Associative knowledge is a central component in many accounts of recall, recognition, and semantic representations in word processing. There are multiple ways to tap into this knowledge, but word associations are considered to be the most direct route for gaining insight into our semantic knowledge (Nelson, McEvoy, & Schreiber, 2004; Mollin, 2009) and human thought in general (Deese, 1965). The type of information produced by word associations is capable of expressing any kind of semantic relationship between words. Because of this flexibility, networks are considered the natural representation of word associations, where nodes correspond to lexicalized concepts and links indicate semantic or lexical relationships between two nodes. These networks correspond to an idealized localist representation of our mental lexical network. The properties derived from such a network have been instrumental in three different research traditions, which will be described below. These traditions have focused on (1) direct association strength, (2) second-order strength and distributional similarity, and (3) network topology and centrality measures.

The first tradition has used word associations to calculate a measure of associative strength and was inspired by a behaviorist view of language in terms of stimulus–response patterns. This notion of associative strength plays an important role in studies that have focused on inhibition and facilitation in list learning (e.g., Roediger & Neely, 1982), studies on episodic memory (e.g., Nelson et al., 2004), and studies that have tried to distinguish semantic and associative priming (for a recent overview, see Lucas, 2000). In these studies, the notion of an underlying network is of secondary importance, and the focus is on direct measures of strength that can be obtained relatively easily, since only the cues present in the task need to be considered.

A second tradition focuses on second-order strength of associations. Instead of looking at individual strengths between cues and responses, the interesting properties are found in the structure of the network itself (Deese, 1965). The simpler and small-scale version of this approach underlies mediated associations, which have been studied extensively in priming as well. More elaborate approaches are those where the complete association distributions are used instead of the first-order one. For example, the words *blood* and *accident* might not be directly associated, but the overlap between common associations, such as *red*, *wound*, *hurt*, and so forth, of these two words provides insight into the way both words are related. This idea was initially explored in the factor-analytic studies of nouns and adjectives by Deese, which predate many modern approaches to semantic memory that rely on distributional similarity, such as latent semantic analysis (Landauer & Dumais, 1997), word association spaces (Steyvers, Shiffrin, & Nelson, 2004), or association-based topic models (Andrews, Vinson, & Vigliocco 2008). Moreover, the added potential of this approach was indicated in the study by Steyvers and his colleagues, who found that a high-dimensional vector space model based on word associations gave a better account of cued recall than did any similar text-based model. The first two traditions were introduced some time ago, and the main results are summarized in Cramer (1968) and Deese.

The third and most recent development has used elaborated networks of associations, implemented as graphs, to learn about the development of language and the way words can be retrieved efficiently. Some of the initial ideas about the exploration of a semantic network through spreading activation go back to the seminal work of Collins and Quillian (1969). More recently, the global structure of large-scale networks has provided new evidence of possible mental processes and structures. In both English (Steyvers & Tenenbaum, 2005) and Dutch (De Deyne & Storms, 2008a), it has been shown that the mental lexicon represented as a word association network is characterized by a specific organization common to many networks that grow steadily over time, such as the World Wide Web. Similar to these naturally growing networks, word association networks do not have an arbitrary organization but show a small-world structure where, on average, two words are separated by less than four associated nodes. Large-scale modeling of growing networks has also resulted in a mechanistic account for numerous findings in word processing, including frequency and age-of-acquisition (AoA) effects (Steyvers & Tenenbaum, 2005). The importance of expanding this type of research is also acknowledged by Balota and Coane (2008), who argue that these procedures have taken a significant step toward capturing semantic memory within an empirically verified network (Balota & Coane, 2008, p. 516).

These three lines of research correspond to a shift from local interactions measured as direct strength, to interactions within a subnetwork (mediated and distributional measures) to global characteristics of the network (network topology and centrality) and coincide with recent computational advancements and theoretical developments such as the renewed interest in network theory (see Newman, 2010, for an introduction).

Naturally, there are limitations to the word association approach as well. For instance, according to Aitchison (2003), one of the most important limitations is the lack of weak links in networks derived from word associations. We agree that studying the structures from which a large part of our mental and verbal behavior originates requires a better approximation of the associative network. In what follows, we will argue that both the sample size and the coverage of the lexicon are important considerations in making inferences about the structure and other properties of such a network. Centrality in a network, for instance, will stabilize only if a large proportion of the important nodes in a network are present, and the reliability of distributional similarity will depend on whether the distribution of two words are sampled extensively enough to allow any overlap. This motivated us to invest considerable effort in building an extensive database for word associations.

Such a network is unique and valuable in several ways. First, few studies have attempted to compile a semantic network that covers a reasonably large part of the human lexicon. The largest word association databases are the Edinburgh Associative Thesaurus (EAT; Kiss, Armstrong, Milroy, & Piper, 1973) and the University of South Florida association data set (USF; Nelson et al., 2004). The EAT associations consist of responses made by 100 British English speakers to each of the 8,400 cues collected between 1968 and 1971. The USF associations are more recent and were collected among American English speakers from the late seventies onward. The USF associations are normed, meaning that each associate in this set was also presented as a cue. However, only 5,400 cues were presented. Moreover, in both data sets, a discrete free association task was used, meaning that only a single response per cue was generated by the participants. One of the consequences of the discrete procedure is that the response frequencies are reliable only for the very strong associates, while weaker responses are either unreliable or missing (Nelson, McEvoy, & Dennis, 2000). This lack of weak associations is seen as a general drawback of the word association procedure (Aitchison, 2003) and has been responsible for questioning of the results of previous findings in mediated priming (Chwilla, Kolk, & Mulder, 2000; Ratcliff & McKoon, 1994) and, presumably, affects still other semantic tasks. Furthermore, when word associations are represented in a network, the coverage of the lexicon in this network becomes important. Insufficient

coverage can be problematic for various network-based centrality and distance measures that depend on the directionality of the association between a cue and a response. Most often, this corresponds to a network where each node is presented as a cue and where the network indicates for each node the number of incoming and outgoing links. The number of incoming links for a node relies on the initial selection of words presented as cues. In most situations, it is not feasible to present all the association responses as cues. Instead, a small subset, such as the single most frequently generated response per cue, is presented. In this sense, the word association network is never complete and always slightly biased, since many responses were never incorporated as a cue and cannot contribute as an incoming or outgoing link. Ideally, all the collected associations are also presented as cues. While this is infeasible in practice, it is important to provide an estimate of the bias this introduces into the structure of the underlying network.

Outline

In this article, we present a new study that provides the most realistic approximation of the human lexicon to date by creating a lexico-semantic network that includes the responses of over 70,000 participants. In the first part, we will show how the current network addresses the problem of sample size by using a multiple response association procedure. Next, we will argue that the network extracted from these data covers a useful portion of the lexicon.

In the second part, we indicate how such a rich network leads to improved predictions in two key areas of cognition. The first test involves the evaluation of availability effects in word processing using network-derived centrality measures in the lexical decision task (LDT). One of the core advantages of a realistic size network is that we can derive distinct properties that can explain the centrality of words and see how these measures relate to lexical access or retrieval. A second key area is the study of semantic cognition, where the large-scale network allows us to infer the distributional similarity of various nodes in the network. As network sparsity has been problematic in semantic tasks such as mediated priming, we investigate these properties in a pure semantic context by predicting semantic relatedness and similarity judgments.

In both parts, we will particularly focus on the differential predictions derived from a discrete versus a multiple-response association task. We will conclude with a discussion on the use of the measures derived from the network and the extent to which this network covers the mental lexicon.

The Dutch Association Lexicon

In the following section, we will first describe the collection of data used to build the Dutch Association Lexicon. A small proportion of this lexicon of word associations, containing 1,424 words, is described in De Deyne and Storms (2008b). We will focus on the impact of the continued multiple response procedure, since this procedure sets our study apart from previous large-scale association procedures.

Method

Participants

The data were gathered in two stages. In the first stage, participants were primarily first-year students at the Belgium universities of Leuven and Ghent who volunteered or participated to fulfill the credit requirements of a course. In the second stage, the vast majority of participants were volunteers who participated online and forwarded the experiment URL[1] to other persons. The total group of participants consisted of 71,380 persons (47,459 females, 22,966 males, 955 unspecified). Age was specified by 70,786 participants and varied between 7 and 96 years ($M = 40$). In the course of the second phase, we also started recruiting participants in the Netherlands and began asking for additional information about the spoken Dutch variation (Dutch spoken in Flanders or the Netherlands). These variations of Dutch are closely related, in a similar way as British and American English. A total of 6,875 out of the 71,380 participants indicated that they were Dutch speakers from the Netherlands.

Materials

The collection of norms started in 2003. Although new data are still being collected at the moment of writing, this report includes only material up to November 2010. The initial set of cue words was taken from a set of semantic concepts reported in Ruts et al. (2004) and De Deyne and Storms (2008b). The Ruts et al. study includes words from 13 different semantic categories for artifacts, natural kinds such as various animal categories, and actions. This set of seeding words was further expanded in 2008 by De Deyne and Storms (2008b) using a snowball procedure in which the most frequent responses were gradually added to the list of cues. Simultaneously, this set of words was expanded with words that might be useful for a number of research lines. Examples of such sets are words commonly used in picture-naming norms (Severens, Van

[1] The Web site can be accessed at http://www.smallworldofwords.com/.

Lommel, Ratinckx, & Hartsuiker, 2005) and English–Dutch cognates (Brysbaert, personal communication, 2008). The final set includes 12,571 cues. This set of cues covered the important parts of speech and consisted primarily of nouns (64.5 %), followed by adjectives (16.9 %), verbs (15.7 %), adverbs (1.3 %), and prepositions (0.6 %).

Procedure

During the first stage of data collection, both pen-and-paper and online data collection procedures were used. Since De Deyne and Storms (2008b) have explained the pen-and-paper procedure and the subsequent collection of the data used a Web-based questionnaire, we will describe only the latter procedure. The task was a continued free word association task in which the participants provided three different associates to each cue presented. Each participant was asked to type the first three words that came to mind upon the presentation of a cue. They were instructed to avoid using full sentences as responses, abstain from using strictly personal associations, and consider only the cue word presented on top of the questionnaire form. If a cue was not known, they were asked to indicate this by pressing a button labeled "unknown." Next to these instructions, a questionnaire form was displayed containing a word and three blank spaces where the associations could be typed.

The use of a snowballing procedure implies that the cues were chosen from a list of words that varied depending on the time of the data collection. Every list that was generated for a particular participant was different. Furthermore, the order of presentation of the cues was randomized for each participant. The average list length contained 18 cues and varied between 7 and 30 cues. Generating the associations for a list of average size took less than 10 min.

Results

Preprocessing

To verify whether the responses for a particular participant were meaningful, a number of automated checks were performed. First, participants who gave more than 50 % "don't know the word" responses were discarded, resulting in the removal of 350 participants. Second, we checked whether the responses were actual Dutch words by comparing the responses with a wordlist obtained from the CELEX word frequency corpus (Baayen, Piepenbrock, & van Rijn, 1993). Only participants for whom more than 40 % of the responses were represented in the CELEX word list were retained. At the level of cue words, we aimed to have 100 participants provide at least one association per cue. For some words, the number of participants that provided associations was slightly larger. To aid interpretation and avoid the overrepresentation of these words, we kept only the 100 first participants who completed each cue. At this point, the data reflected the responses of 70,369 individuals. Finally, all the association responses were converted to lowercase words, and nonalphabetical characters were removed. Of the words that occurred more than once, 64 % were identified as a correctly spelled Dutch word, using the *OpenTaal* spelling dictionary (available from http://www.opentaal.org/bestanden).

Missing and unknown responses

A total of 3,771,300 responses were collected for a set of 12,571 cues. For each cue, 100 participants provided a primary, secondary, and tertiary association. If a word was not known, the primary, secondary, and tertiary association responses were coded as "x." Some participants provided only the first or the first two responses. These missing responses were considered to differ from the case where a word was unknown. A total of 15,885 secondary responses (accounting for 1.3 % of the data) and 55,000 tertiary (accounting for 4.4 % of the data) were considered missing. A total of 17,892 of the presented cue instances were unknown. The percentage unknown ranged between 0 % and 72 % (M = 1.42) per cue. These numbers also indicated that the vast majority of cues (97 %) were known by at least 90 % of the participants. In the remainder of the article, we will report statistics for the associations excluding missing and unknown responses.

Types and tokens

The number of types and tokens provide a description of the vocabulary size present in the association lexicon, and the rate at which new types are introduced as more responses are collected gives us an idea of how much information has already been collected for each word. Below, we report the statistics on the number of types and tokens tabulated for the entire data set and for each cue separately. Since the ratio of types and tokens can be interpreted as a measure of response heterogeneity, we propose a general measure that captures this information. For the entire database, the total number of response types was 201,356 for 3,646,739 valid response tokens. This result can be broken down into primary, secondary, and tertiary associations, resulting in 88,280 (1,239,208) types (tokens) for the primary response, 106,342 (1,223,323) for the secondary response, and 108,678 (1,184,208) for the tertiary response. At the level of cues, an average of 118 (SD = 23.6) different response types were generated for the cues. The number of response types varied considerably: from 33 (for the cue *article*) to 211 (for the cue *telepathy*). The entropy of the

distribution can express this heterogeneity in the distribution of responses:

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i),$$

where n is the size of the vocabulary or number of types, p_i is the probability for the ith type, and the sum is taken over all types with nonzero response probability. In other words, H increases as the responses become more heterogeneous. If all tokens are identical, H equals zero. If there is no overlap in the responses, all tokens are different, and the entropy is maximized. This upper bound depends on the number of responses: For 300 responses (i.e., 3 associations provided by 100 participants), the maximum entropy is $H = 8.22$. The change in entropy as a function of the number of collected tokens also indicates how many associations should be collected to have a stable representation. To investigate the relationship between increasing vocabulary size and H, we plotted the average entropy for each cue as function of the number of responses.

Figure 1a shows the entropy for the cues that were known to at least 95 % of the participants.[2] The figure shows the range in entropy as a gray surface and the average entropy as the curve plotted within this surface. The gray surface indicates that the rates at which new types are introduced as a function of the collected tokens varies widely between low- and high-entropy words. The entropy curves also provide a better intuition of how stable certain words are and are arguably more informative than measures of reliability, which reflect the end point of the heterogeneity in the responses (e.g., Nelson et al., 2000).

A second question pertains to the degree to which the second and third responses represent responses qualitatively similar to the first. If this is the case, the type–token ratio should be similar regardless of the serial position of the response. Figure 1b shows that this is not the case: The average entropy was higher for the secondary and tertiary response than for the primary response. These findings replicate and generalize the finding by Nelson et al. (2000), who performed a multiple-response procedure with two responses per cue for a small set of words and showed that the ranking of the response depended on the response position (i.e. whether it was a primary, secondary, or tertiary response).

A final case that needs further attention is that of idiosyncratic responses (types occurring only once), often referred to as *Hapax Legomena* types (a term from the Greek, "*said only once*"). For the total number of response types tabulated over the different cues, the majority (63 %) consist of Hapax Legomena or words occurring with a frequency of one. Despite the frequent occurrence of Hapax words, these words account for only 3.5 % of the response data. This is in line with Zipf's law, which states that the frequency of a type is inversely proportional to its rank in the frequency distribution. Although the most frequently used version of the USF data set excludes these idiosyncratic responses, Nelson and colleagues have indicated that most of these responses represent meaningful information (Nelson et al., 2004). Inspection of our data confirmed that this is the case for Dutch as well. For example, for the word *language* (Dutch: *taal*), the idiosyncratic responses included words like *body language*, *development*, *literature*, *text*, *travel*, *foreign*, *Swedish*, and so forth, indicating weak yet meaningful associative relationships. Furthermore, they reflect the long tail in the response generation distribution. On the basis of these observations, we decided to retain Hapax Legomena words as part of the data set.

The results of the data collection procedure can be summarized as follows. First, the continued version of the free association task results in an increasing number of different types, in comparison with a similar task where only one word association per cue is required. This indicates that the procedure provides a useful way of increasing the number of weak associations. Furthermore, inspection of the data showed that idiosyncratic responses encode meaningful relations with the cue. The usefulness of including a larger number of different types—in part, by using a continued version of the free association task and including idiosyncratic responses—needs to be verified empirically. This will be the topic of the last section of this article. However, first, we will derive a network model of the mental lexicon using the association norms presented earlier.

A semantic network of word associations

The richness of word associations is best captured by representing the associations as a weighted directed semantic network where the weights or strengths are determined by some function of association frequency and the direction is determined by the role of the node (either a cue or a response). When the nodes correspond to the cues, we obtain a unimodal network in which incoming and outgoing links are interpretable, given that the initial number of cues is adequately large. This unimodal network will be used as the substrate for representing our lexico-semantic knowledge. A number of additional steps are required to construct this unimodal cue by cue-directed network G. Because such a network only represents cue words, the number of different types is significantly reduced. Therefore it is important to consider the degree to which this network covers the human lexicon. The resulting network will be used as the foundation of the studies in the second part

[2] Both Fig. 1a and b are truncated because adding the full number of responses (300 and 100, respectively) would include missing and unknown responses, which would bias the entropy curves.

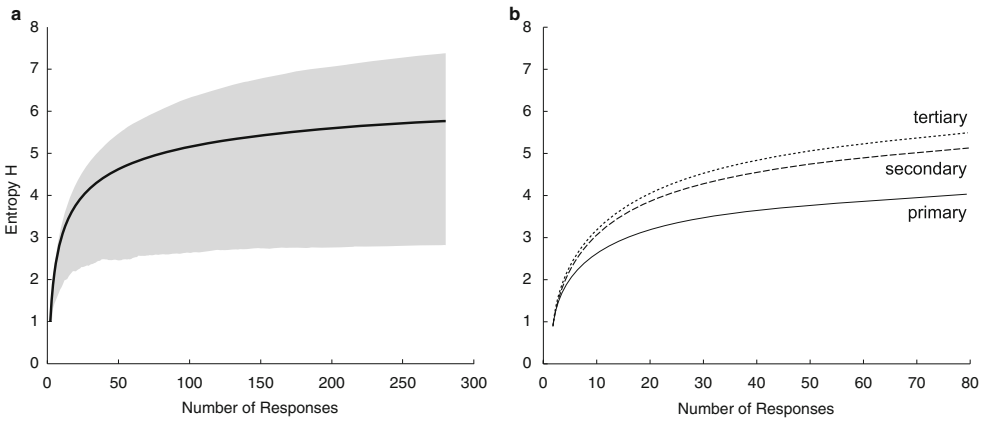

Fig. 1 Entropy as a function of increasing observed token counts over primary, secondary, and tertiary responses, with variability indicated by the gray area and average entropy by the black line (**a**) and average entropy separated for the three response positions (**b**)

of the article, but first we will describe the construction and description (including coverage and global statistics) of the network.

Method

Procedure

The network is constructed from a weighted adjacency matrix where both the rows and columns correspond to the different cues and contains 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. To allow the calculation of the various measures reported below, we removed all single loops in this network. For example, for the cue *hammer*, a loop would occur if the participants respond "*hammer*." Next, we obtained the largest weakly connected component. This component consisted of the maximal subset of nodes in a directed network that had at least one incoming or outgoing link. In addition, we also created a second set of structurally equivalent networks where the weights corresponded to estimated association strengths, using the procedure proposed by Nelson et al. (2000) and adapted by Maki (2008). Instead of treating the association frequencies as weights of the network, the frequencies are regarded as manifestations of strength, rather than strength itself. Nelson et al. argued that given a cue, a random sample of association strengths becomes activated, and the highest strength determines the ultimate response. The process captures the variability within and between participants and can be simulated using the gradient descent method proposed by Maki. This results in mean strength estimates of each cue–response pair based on the observed word association frequencies (see Maki, 2008, for technical details). To be able to compare the effect of the multiple-response procedure, we repeated the steps listed above to create networks based on the first (G_1), first and second (G_2), and all (G_3) responses.

Results and discussion

Network size

Restricting the network to words that were present both as a cue and as a response reduced the number of nodes from 12,571 to 12,428. The removed nodes included cue words like *dermatology* or *lentil*. These words were present in the set of cues as stimuli to be used in other experiments but turned out to be too infrequent to be produced as a response to any of the cues. All remaining words were present as part of the largest weakly connected component shared by all networks (G_1, G_2, and G_3). Since the network with weighted association strength differs only in terms of values of the edges and not in terms of the number of edges, the components for this network are identical to those of the original network.

Coverage

A potential problem with the conversion from two-mode association data to a directed weighted network is the fact that only the response tokens that are presented as a cue are retained. In other words, a large proportion of the data that were originally present in the two-mode data is not considered, since this procedure reduces the number of columns in the two-mode adjacency matrix from over 200,000 to 12,428. Although the Zipfian distribution of the response tokens implies that the majority of tokens should be retained for most words, it is possible that certain types of cues (i.e., those with low-frequent responses) are more affected than others. For those cues, the network-derived measures might be biased. To investigate this more systematically, we calculated for each cue the percentage of response tokens represented in each network. The results for each of the networks are shown in Fig. 2.

For each of the networks, the majority of the response tokens are represented in the network. The median number of response tokens, indicated as a bold line in Fig. 2,

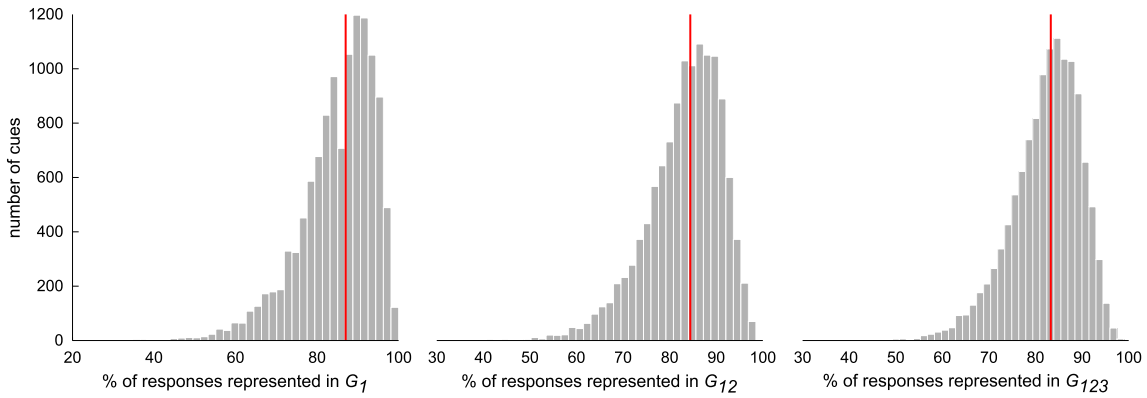

Fig. 2 Coverage of response types in the various directed networks based on primary, secondary and tertiary responses (clarification)

differed depending on the density of the network and corresponded to 87 for G_1, 84.4 for G_2, and 83.3 for G_3. The distribution of the percentage of responses shows that cues represented by less than 75 % of their response tokens is small (14.2 %, 15.3 %, and 15.9 % for the respective networks). The slight decrease of coverage for networks including secondary and tertiary responses can be explained by the additional heterogeneity present in these responses, which reduces the probability that they would occur as cues in the word association task. Moreover, the resulting distributions represent an underestimate of the number of responses that could be represented if all response word forms were converted to lemmas. For example, lemmatizing the various word forms for the word *dream* (e.g., *dreaming, dreams*) would significantly increase the number of responses encoded in the cue by cue adjacency matrix. Together, the loss of response types by considering a one-mode network representation indicates that even with only 12,428 cues, most of the association responses are still represented. This results in a unimodal network that approaches a stable sized lexico-semantic network as the majority of the associations themselves are present as cues.

Network properties

The multiple response procedure also affects the global topology of the network, such as its density, diameter, average path length, and clustering coefficient. To understand how the larger network compares with previous results and the contribution of multiple responses, we calculated these indices for the USF network containing 4,982 nodes (Nelson et al., 2004), the Leuven-2008 network with 1,424 nodes (De Deyne & Storms, 2008b), and the new networks that vary in terms of serial response position (G_1, G_2, and G_3). The density of the network is 1 if all nodes are connected to each other in the network. If multiple responses result in richer representations, we should see an increase in network density. The results in Table 1 show that this is indeed the case, with the density of G_3 about 3 times as dense as G_1.

In contrast to the much smaller Leuven-2008 network, the new networks are considerably less dense (0.64 % for G_3, in comparison with 2.40 % in the Leuven-2008 network). This might be due to the way the Leuven-2008 network was set up (starting from a small set of semantic classes). However, the new networks are denser than the USF network. Next, we considered average path length of the network. All possible paths were calculated using the Dijkstra algorithm (Dijkstra, 1959). Table 1 shows how adding more nodes to the network reduces the average path length even when the density is lower. This is indicated by the values for the Leuven-2008 network and G_3, where the former has an average path length of 3.27 and density of 2.4 % and the latter has an average path length of 3.06 and density of 0.64 %. Similarly, Table 1 shows how the diameter—the maximum path length connecting any two nodes in the network—decreases as a function of the number of nodes and density. Finally, we also calculated the clustering coefficient for these networks as an indicator of network organization. This measure corresponds to the probability that the neighboring nodes of a specific node in the network are connected themselves (cf. the Appendix). The pattern for this measure differs from all other measures because the probability that any two neighboring nodes are connected in a network drops much faster as new responses are added. This occurs both when comparing with the previous networks Leuven-2008 and USF and when comparing the single (G_1) against the multiple response networks (G_2 and G_3). Because the clustering coefficient depends on the density of the networks, we constructed three new networks based on a random permutation of the in- and out-going edges. The average clustering coefficients for the random networks were $C = .0189$ ($SD = 0.011$) based on G_1, $C = 0.032$ ($SD = 0.011$) for G_2, and $C = 0.0455$ ($SD = 0.014$) for G_3. Taking the ratio between the random and original networks (cf. Table 1) shows that the clustering in the network is 18, 10,

Table 1 Density, average path length (*L*), diameter (*D*), and clustering coefficient (*C*) network properties for the University of South Florida network (USF; Nelson, McEvoy, & Schreiber, 2004), Leuven network (Leuven-2008; De Deyne & Storms, 2008a, b), and G_1, G_2, and G_3

Network	n	Density (%)	L	SD(L)	D	C	SD(C)
USF	4,982	0.574	3.90	(2.14)	62	0.451	(0.374)
Leuven-2008	1,424	2.400	3.27	(1.77)	47	0.634	(0.314)
G_1	12,482	0.219	4.14	(1.22)	31	0.341	(0.263)
G_2	12,482	0.432	3.38	(0.87)	26	0.321	(0.174)
G_3	12,482	0.644	3.06	(0.69)	24	0.311	(0.147)

and 7 times larger than the corresponding random networks for G_1, G_2, and G_3, respectively. This indicates that adding secondary and tertiary responses reduces the tight clustering, as compared with networks based on a single response.

The global picture that emerges from this comparison is that larger networks necessarily result in sparser adjacency matrices with lower clustering coefficients, but the use of a multiple-response procedure reduces sparsity and the distance to reach any two nodes in the network.

Summarizing these results, we have found that collecting multiple responses in a continued word association task affects various properties of the semantic network derived from these word associations. A more extensive (and more varied in terms of cues) coverage of the mental lexicon results in a less strongly organized network in terms of the clustering coefficient, while the distance to any two particular nodes decreases only slightly (indicated by the average path length and diameter) when adding the secondary and tertiary responses to the network. The biggest difference is found for the density of these networks. This larger difference in density between G_1, G_2, and G_3 allow us to test and compare the influence of encoding secondary and tertiary responses, in comparison with the traditional single-response procedure most often used in previous word associations studies. In the next study, we will evaluate the effect of network density in word processing.

Study 1: Word processing

One of the ubiquitous findings in word processing is that some words are recognized and produced with less effort than are others. This is primarily explained in terms of word frequency. Other potentially influential measures include AoA (e.g., Turner, Valentine, & Ellis, 1998), familiarity (Balota & Chumbley, 1984; Gernsbacher, 1984), number of dictionary meanings (Azuma & Van Orden, 1997), imageability (e.g., Coltheart, Patterson, & Marshall, 1980), contextual diversity (e.g., Brysbaert & New, 2009), and arousal (e.g., Scott, Donnell, Leuthold, & Sereno, 2009). For a number of these variables, a structural hypothesis is provided in terms of the connectivity in a network, but only a few of these theoretical claims have been directly tested. A good example is the variety of different accounts for imageability effects. According to the contextual variety effect, highly imageable words have a processing advantage because they have a smaller set size (Galbraith & Underwood, 1973). A similar explanation based on context availability is given by Schwanenflugel and Shoben (1983) and by Schwanenflugel, Akin, and Luh (1992). A third theory, by Plaut and Shallice (1993) opposes this view and instead claims that concrete words have more semantic properties than do abstract concepts, thereby resulting in a processing advantage. This view is also supported by de Groot (1989) who claims that processing advantages for concrete concepts are due to larger associative sets compared to abstract words. Even for the effect of word frequency, which has been extensively documented and shown to affect nearly all tasks including word processing, there is no agreement on the precise mechanism (Balota, Yap, & Cortese, 2006). However, many of the interpretations assume that frequency is encoded as the weights among the connections between either units that correspond to words or abstract sublexical units (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996). A similar structural mechanism accounts for semantic effects of AoA. According to the semantic hypothesis, early-acquired words are positioned in a more central part of the semantic network, as compared with later acquired words (Steyvers & Tenenbaum, 2005). Network-derived measures might provide a direct means for testing many of these hypotheses. Although it seems unlikely that processing advantages can be entirely reduced to properties of the network interconnections, we want to take the network account seriously by investigating richer measures of centrality derived from the word association network. This offers a more direct approach to the problem and allows us to gain insight into word processing.

While there are numerous centrality measures that can be derived from networks, some are more important than others are, since they correspond to previously proposed psychological mechanisms. The following measures therefore do not represent an exhaustive set but are included on the basis of their correspondence to previously reported indices of word processing advantages. Importantly, many of these network measures address an important shortcoming of their

psychological proxies by providing a more principled way to distinguish different types of centrality by considering a more representative and directed network for the mental lexicon. A first variable that affects word processing is set size or number of features. Set size is known to affect numerous tasks, including performance in extra-list cued recall and recognition (e.g., Nelson, Cañas, & Bajo, 1987; Nelson, McEvoy, & Schreiber, 1990), while number-of-feature effects have been reported in numerous semantic judgment tasks (Pexman, Holyk, & Monfils, 2003). In the context of directed networks, set size corresponds to the out-degree of a node.

A second class of variables is the one that accounts for richness effects in word processing, such as semantic neighborhood size (Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008). This intuition is captured by the clustering coefficient measure, which can be calculated separately for each node in the network. This local clustering measure is related to in- and out-degree but is more sophisticated, since it also captures information about the connectivity of the neighboring nodes. Moreover, in contrast to the number-of-feature measures, the clustering coefficient is not as strongly related to degree measures, since words with relatively few neighbors can still have a high clustering coefficient.

The third type of measure takes centrality quite literally and provides a network proxy for centrality effects explained by variables such as AoA. Semantic explanations of the AoA effects posit that words are processed faster because they are acquired earlier and new information is added in an incremental way. As a result, early words provide the foundations of later acquired ones. *Betweenness* is an example of a measure that matches this definition quite well, since it captures how many times you can encounter a node by traversing paths in the network.

The last measure is inspired by the study by Griffiths, Steyvers, and Firl (2007) that used a letter phonological fluency task in which participants were asked to generate words starting with a certain letter during a short time span. They used a recursive centrality measure, in which centrality not only is influenced by the neighbors of a node, but also takes into account the centrality of these neighbors. This measure provided better estimates of word generation frequency than did other centrality measures, such as simple word frequency. The centrality measure they used is called PageRank and is better known for indexing the importance of a page of the World Wide Web by Google (Page, Brin, Motwani, & Winograd, 1998). The PageRank measure represents an example of a family of measures that include feedback information such as eigencentrality. These measures have applications beyond lexical retrieval and are used to describe the effect of feature correlations in concept representation as well (Sloman, Love, & Ahn, 1998).

One of the most widely used tasks to investigate visual word recognition is the LDT. While early theories assumed that the role of semantic information was minimal, a number of studies have shown that semantic characteristics affect lexical retrieval, even when this information is logically not necessary. The study by Chumbley and Balota (1984) is of particular interest since the semantic involvement in the LDT was determined by using a measure of associativeness. The study included two experiments. In an initial experiment, participants generated associates to a list of words, and reaction times (RTs) were recorded. The averaged association RTs for these cue words were then used in two follow-up experiments and were found to be the most important predictor of decision latencies in the LDT. One of the surprising findings was that the number of associates had an effect in both follow-up experiments, albeit of a magnitude smaller than the association latencies. More interesting, however, is the finding that the number of associations influenced association latencies as a suppressor. This means that including the number of associates in a regression, together with the association latencies, increases the contribution of the latter predictor. In the following section, we will show how network-based measures of meaningfulness or centrality not only provide an excellent account of the LDT, but also disentangles a puzzle in the results of word processing tasks that include number of associations such as the study by Chumbley and Balota.

Method

Materials

We used the large-scale lexical decision data set (Dutch Lexicon Project [DLP]) compiled by Keuleers, Diependaele, and Brysbaert (2010) to investigate (1) the predictive value of psycholinguistic variables derived from the word association database and (2) the effect of the multiple association procedure. The DLP database contains lexical decision times for 14,089 Dutch mono- and dissyllabic words and nonwords. Since some of the measures described in the next section can be more easily interpreted for nodes that have both in- and out-going links, we derived the largest strongly connected component for each of the networks. A strongly connected component is a subnetwork that consists of a maximal subset of nodes in a directed network that have an in- and out-degree of at least 1 (within the subset itself). The derivation of the strongly connected component network resulted in the deletion of 4,725 nodes for G_1, 2,237 for G_2, and 10 nodes for G_3. Four different measures of centrality were derived for each of these networks. A formal description of each of these measures is provided in the Appendix.

Results and discussion

The dependent measures consisted of the decision latencies for 5,918 words that were present both in the networks G_1, G_2, and G_3 and in the DLP database (Keuleers, Diependaele, & Brysbaert, 2010). Similar to Keuleers, Diependaele, and Brysbaert only the *z*-RT scores for items with accuracy larger than .66 were used. The network measures used in our analysis included weighted in-degree (k^{in}), out-degree (k^{out}), clustering coefficient (C), betweenness (b), and PageRank. To investigate the explanatory power of network-derived measures, as compared with traditional measures used as independent variables in the LDT, we added the word frequency and context diversity measures reported by Keuleers, Diependaele, and Brysbaert derived from Dutch subtitles (SUBTLEX-NL; Keuleers, Brysbaert, & New, 2010), since these measures captured most of the variance in their study.

Table 2 shows the correlations between the five centrality measures and the word frequency and context centrality measures. All measures are significantly correlated but do not represent a perfect correspondence. Although the correlations between the network measures are moderate to high, the correlations do show that these measures encode different information from the networks. Only k^{out} did not correlate strongly with most of the other centrality measures. Furthermore, the high correlation between PageRank and in-degree suggests that there may be problems differentiating between these two measures.

Next, we investigated how network centrality explains decision latencies in the LDT. A first question is whether including additional association response types originating from the later responses in the continued association task improves or dilutes the prediction of the RTs. To this end, we calculated the correlations between the RTs and the network measures derived from G_1, G_2, and G_3. The resulting coefficients are shown in the first four columns of Table 3.

Table 2 Correlations (ρ) between the network G_3 derived centrality measures (clustering coefficient [C], in-degree [k^{in}], out-degree [k^{out}], betweenness [b], and PageRank) and the SUBTLEX-NL measures for word frequency (WF) and contextual diversity (CD)

	N = 5,918	1	2	3	4	5	6	7
1	C		−.63	−.65	−.76	−.57	−.49	.51
2	k^{in}			.21	.70	.96	.60	.61
3	k^{out}				.68	.17	.21	.23
4	b					.67	.46	.47
5	PageRank						.57	.58
6	WF							.99
7	CD							

All correlations significant at $p < .001$ (two-tailed *t*-test)

Table 3 Correlations (ρ) between DLP reaction times (RTs) and centrality measures (clustering coefficient [C], in-degree [k^{in}], out-degree [k^{out}], betweenness [b], and PageRank derived for G_1, G_2, and G_3. The last two columns indicate partial correlations after removing effects of word frequency [WF] and context diversity [CD]

N = 5,918	G_1	G_2	G_3	G_3-WF	G_3-CD
C	.44	.47	.49	.25	.23
k^{in}	−.66	−.67	−.67	−.45	−.44
k^{out}	−.19	−.21	−.19	−.08	−.05
b	−.45	−.50	−.50	−.30	−.29
PageRank	−.63	−.64	−.65	−.45	−.44

All correlations significant at $p < .001$ (two-tailed *t*-test)

All measures were significantly correlated with the decision latencies. The best prediction was found for k^{in} and the related measure of PageRank. Out-degree, which corresponds to the set size of a node, did not correlate as strongly, as compared with the other variables. The degree-based measures k^{in}, k^{out}, betweenness, and PageRank show correlations that are quite similar regardless whether they are derived from G_1, G_2, or G_3. The added density in G_3 results in slightly higher correlations, as compared with G_1. This was most prominently the case for the clustering coefficient (C). Together, these findings suggest that adding additional links obtained through continued association does not hamper the prediction of decision latencies in the LDT but actually improves it.

A second issue pertains to the question of what network measures influence the LDT latencies. In many previous studies, centrality has been considered in terms of undirected networks, ignoring whether edges are in- or out-going. Instead, a nodes' centrality corresponds to the number of edges between nodes or the degree of that node. Directed networks do not conflate differential effects for in- and out-going edges but provide separate measures of centrality, using incoming and outgoing edges such as the in- or out-degree of a node. For example, De Deyne and Storms (2008b) have shown that centrality measures derived from undirected networks do not correspond as much with external centrality measures such as imageability and AoA, as compared with directed centrality measures. Comparing solutions for directed and undirected versions of the network allows us to constrain the meaning of accessibility or centrality processing advantages further by pinpointing these effects to the incoming edges of a lexico-semantic network. To investigate the effect of using degree instead of the directed measures of in- and out-degree, we calculated the degree of each node, ignoring whether the link was incoming or outgoing. The resulting correlation between degree and the DLP RTs, $\rho(5918) = -.64$, $p < .001$, was weaker than the value reported for in-degree, $\rho(5918) =$

−.67, $p < .001$ (see Table 3), [3] but in line with previous findings that degree measures perform suboptimal because they incorporate the out-degree of nodes (De Deyne & Storms, 2008b).

To investigate whether these network measures can account for any additional variance once frequency and context diversity are accounted for, we also calculated the partial correlations after removing these two variables separately. The results in columns 5 and 6 of Table 3 show that the centrality measures capture additional structure. The differences between the measures closely follow the previous pattern, with the best correspondence found for in-degree and PageRank and only weak effects of out-degree. In addition, we also performed a parametric analysis to investigate the relative contribution of the different predictors. To reduce nonlinearity, we took the square root of in-degree, betweenness, and clustering coefficient. Next, both the dependent variable (decision latencies) and the independent variables (except out-degree and clustering coefficient) were log (base 10) transformed. We performed a regression analysis on the decision latencies, with in-degree, out-degree, betweenness, and the SUBTLEX-NL measures of word frequency and context diversity included as predictors. The contribution to the multiple correlation coefficient was determined by using a relative importance estimate (Grömping, 2006).[4] These predictors accounted for 54 % of the variance in the decision latencies. The partitioning of the multiple correlation coefficient is shown in Fig. 3. These results confirm the basic pattern shown in Table 3. In-degree accounted for most of the variance, followed by the SUBTLEX measures and betweenness. Out-degree and the clustering coefficient explained hardly any additional variance.

Together, the results show that network measures derived from lexico-semantic networks give insight into which words in the lexicon are more important and will be processed more efficiently than other words. The present results indicate that considering only the primary association (G_1) provides a good approximation, as compared with using a more dense network consisting of both the secondary (G_2) and tertiary (G_3) associations as well. Furthermore, in terms of absolute strength, the correlation coefficients are found to

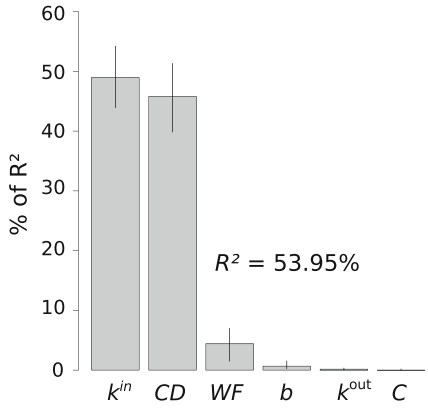

Fig. 3 Relative importance contribution and confidence intervals for the prediction of the LDT decision using the network measures in-degree (k^{in}), out-degree (k^{out}), betweenness (b) derived for G_3 and the SUBTLEX-NL derived measures of word frequency (WF) and context diversity (CD)

be similar to state-of-the-art frequency estimates (Keuleers et al., 2010a, b). Interestingly, the current measure of in-degree manages to explain LDT latencies that are not accounted for by either word frequency or contextual diversity. In contrast to Griffiths, Steyvers, and Firl (2007), our results failed to find convincing evidence for feedback effects as measured by PageRank, since this measure explained slightly less of the variance than that accounted for by in-degree. Still, richer effects of connectivity might play a role in the LDT. This idea is supported by the fact that more complex measures of centrality, such as betweenness and clustering coefficient, capture additional variance in the data beyond context diversity or word frequency.

At a theoretical level, contrasting directed and undirected networks shows us that naive interpretations of semantic effects such as interconnectedness based on counting the number of connections (i.e., the degree, if the representation is an undirected network) or considering only the out-degree (or forward strength) results in a degraded measure. This is caused by out-degree, since out-degree, or the heterogeneity of association responses, biases the number of incoming links as a measure of network centrality. Hence, it offers an explanation for the results of previous studies such as the one by Chumbley and Balota (1984), where analyses with out-degree suppress the effects of other measures of centrality.

Study 2: Semantic distance

The previous section described how the local interconnectedness of a word in the association network indicates how easily it is processed. All these interconnections themselves represent a meaningful structure in which two different lexicalized concepts can be semantically related. The structure captured by semantic proximtiy derived from the network is assumed to

[3] Here, the in-degree measure was defined as the count of in-coming edges, rather than the sum of the edge weights. This is a more appropriate choice in combination with the out-degree measure, which is also based on a count rather than a sum.

[4] The relative importance method does not suffer from the same problems as those associated with a traditional stepwise regression, where different orders of the predictors can result in different outcomes by taking into account the various permutations in which the predictors can appear in the equation. The *pmvd* procedure proposed by Feldman (2005) was used in determining the relative contribution of the predictors to the multiple correlation coefficient. See Grömping (2006) for more details about the beneficial performance of this measure.

be responsible for a host of findings in numerous tasks where semantic information is accessed, including semantic and associative priming (for recent overviews, see Hutchison, 2003; Lucas, 2000). In the following section, we opt for a more sensitive measure of semantic proximity in the form of direct similarity or relatedness judgments of concept pairs (e.g., Dry & Storms, 2009).

While we have shown that the multiple-response procedure results in more diverse associations, as compared with simply obtaining a larger number of single responses, it is not clear whether this increased diversity presents a more accurate approximation of the lexico-semantic information that might become activated in many semantic tasks. To investigate the effect of allowing multiple responses per cue, we derived distance measures for networks that include the primary (G_1), secondary (G_2), and tertiary (G_3) association responses. These network distances were compared with four sets of empirical similarity ratings. The first set aimed at replicating the relatedness judgment results of the classic Miller and Charles (1991) study, which replicated an earlier study by Rubenstein and Goodenough (1965). This study uses stimuli that are unconstrained in terms of semantic categories. It has been used as a standard benchmark in numerous previous studies (e.g., Budanitsky & Hirst, 2006; Durda & Buchanan, 2008; Jiang & Conrath, 1997; Resnik, 1999).

In the following studies, semantic similarity, rather than relatedness judgments, were used. In contrast to semantic relatedness, semantic similarity incorporates words that correspond to coordinate concepts (e.g., *apple* and *lemon*). In the second study, pairwise similarity judgments of concrete nouns for six artifact and animal categories (De Deyne et al., 2008; Ruts et al., 2004) were collected. For each category, the concepts belong to a well-defined category in which concept differences are fine-grained. In these categories, the participants have to decide on the similarity between highly similar entities such as types of birds or fish.

Unlike concrete categories, abstract categories are not hierarchically organized (Crutch & Warrington, 2005; Hampton, 1981). In addition, their representation will depend primarily on the way they are used in spoken or written discourse rather than on concrete, perceptual properties. Given that word associations incorporate both semantic and lexical co-occurrence information, the prediction of similarity of abstract concepts should be similar to those of other concepts. To verify this, a third data set was taken from Verheyen, Stukken, De Deyne, Dry, and Storms (2012), in which the similarity of abstract concepts was rated by participants.

A final data set was gathered to investigate semantic similarity in heterogeneous concepts that belonged to concepts of different categories in the artifact and the animal domains. For this purpose, an experiment was conducted to investigate the performance of the semantic network for a wider range of semantic concepts and to compare these results with the more detailed judgments made in a within-category context as presented in the other data sets.

A challenge for the word-association-based network measures is that, in contrast to many studies that use semantic features, the word associations are context nonspecific. This means that for a word like *chicken*, the associations will reflect the bird and the poultry sense of the animal, while this is not the case for feature ratings, which are often presented within a specific category context (e.g., *birds*). We therefore expect better performance on the first (Miller & Charles, 1991) and last data sets, which consist of an intermixed set of concepts belonging to different categories.

In all four data sets, we want to investigate to what extent word associations predict semantic relatedness and similarity, specifically focusing on the contribution of multiple-responses compared to data derived from discrete association responses.

Method

Participants

Thirty persons participated in the replication of the Miller and Charles (1991) study (indicated by MC), 97 persons participated in the concrete concepts study, 48 in the abstract concepts study, and 30 in the domain Study. All participants were students at the University of Leuven.

Materials and procedure

In the MC study, the stimuli consisted of 30 Dutch words that were close translations of the original English stimuli.[5] The 30 pairs were presented in a random order on a computer screen. The word position in the pairs was randomized. Participants were asked to enter a number between 1 (for *totally unrelated*) and 20 (for *totally related*). If one or two words of an exemplar pair were unknown, participants were asked to enter "−1."

The concrete categories consisted of members belonging to six different *artifact* categories (*clothing*, *kitchen utensils*, *musical instruments*, *tools*, *vehicles*, and *weapons*) and five *animal* categories (*birds*, *fish*, *insects*, *mammals*, and *reptiles*). The number of members per category varied from 20 (*reptiles*) to 33 (*kitchen utensils*). The seven abstract categories consisted of *virtues*, *art forms*, *media*, *diseases*, *sciences*, *crimes*, and *emotions*. Each category consisted of 15 members. For all 18 categories, the similarity between all possible pairs of exemplars was rated. The procedure was identical to the that in the MC study, except that the participants were instructed to rate similarity rather than relatedness. Full details are provided

[5] Since the word *woodland* does not have a translation in Dutch, two stimuli pairs containing this word were not used.

in De Deyne et al. (2008) and Verheyen, De Deyne, Linsen, and Storms (2012).

Finally, for the domain study, the stimuli consisted of members belonging to 6 different *artifact* categories and 5 *animal* categories from the concrete natural category study. Since it is not feasible to present all pairwise combinations of all exemplars of these categories, we selected five exemplars from each of the artifacts and animals categories that cover a wide range of typicality. This way, some members were central to the category representation (e.g., *sparrow* is a typical bird and, thus, a central member), while others were not (e.g., *bat* is an atypical member at the periphery of *mammals* and closely related to *birds*). To increase the generalizability of our results, two replications of the above procedure were performed, resulting in a set A and B, each consisting of 435 pairwise combinations using 30 artifact exemplars and 300 pairwise combinations derived from 25 animals. Each participant rated animal pairs or artifact pairs of either set A or B in multiple sessions, with the only restriction that no replication sets were allowed to be rated the same day. The procedure was identical to those in the previous studies, except that now a total of 435 for the artifacts pairs and 300 animal pairs was presented.

Next, we calculated semantic relatedness predictors by using the networks explained above. In contrast to the LDT study, the networks based on the weakly connected component were used. This guarantees that a maximal number of stimuli from the experiments are present, since not all of them were generated as a response. The edge weights represented the estimated association strength as outlined in the section on constructing the semantic networks. Similar to previous studies (e.g., De Deyne, Navarro, Perfors, & Storms, 2012), semantic relatedness was calculated using the cosine overlap between the pairs of words in each set. This measure was derived for the network consisting of the primary (G_1), primary and secondary (G_2), and all responses up to the tertiary response (G_3). If a network framework provides a useful way to investigate semantic relatedness, we expect that one of its distinguishing features—namely, the fact that the edges are directed—should be informative in semantic tasks as well. To investigate whether this was the case, we also derived an undirected version of the network with all responses (G'_3) by adding the network transpose to each individual directed network.

Results and discussion

For each data set, the similarity ratings were averaged over participants. Participants whose data correlated less than .45 with the average ratings were removed. This resulted in the removal of one participant in set B (*animals*) in the domain study. The reliability of the similarity ratings was calculated using the split half correlations with Spearman–Brown correction. The reliability was high in all experiments, $r_{splithalf}$ = .98 for the MC study, between .85 and .96 for the concrete categories, between .93 and .96 for the abstract categories, and between .94 and .97 for the domain set.

By calculating all pairwise cosine indices, a full similarity matrix S was obtained consisting of 12,428 × 12,428 similarity values. This matrix was used to extract similarity values for the pairs in each data set. To check whether the proposed solution leads to sensible results, the similarity values were sorted. As was expected, the most similar words were corresponding word forms, synonyms, or close-synonyms. Examples are the neighbors for the Dutch word for *cloud* (Dutch: *wolk*, *wolken*), *wound* (Dutch: *wond*, *wonde*, *wonden*), *to rest* (Dutch: *rust*, *uitrusten*), and *doughnut* (Dutch: *oliebol*, *smoutebollen*). This indicates that word associations correctly identify semantically similar entities at the very high end of the semantic relatedness scale. The agreement between the model-derived similarity and the empirical similarities for the different data sets is shown in Table 4.

Table 4 Correlations (ρ) between semantic relatedness (MC) or similarity and the model inferred semantic relatedness for directed networks G_1, G_2, and G_3 and undirected network G'_3

Set	Category	n	G_1	G_2	G_3	G'_3
MC		30	.88	.91	.91	.86
Abstract	Virtues	105	.70	.76	.77	.56
	Emotions	91	.52	.55	.57	.45
	Art forms	91	.55	.59	.65	.48
	Media	105	.55	.60	.66	.50
	Crimes	91	.49	.54	.59	.43
	Sciences	105	.36	.41	.47	.32
	Diseases	105	.24	.43	.61	.05†
	M		.49	.55	.62	.40
Animals	Birds	406	.39	.49	.49	.26
	Insects	300	.56	.62	.66	.45
	Fish	231	.67	.76	.75	.66
	Mammals	435	.44	.50	.54	.26
	M		.51	.59	.61	.41
Artifacts	Clothing	378	.63	.65	.65	.55
	Kitchen utensils	496	.41	.47	.51	.33
	Music instruments	325	.50	.51	.50	.21
	Tools	325	.48	.55	.59	.45
	Vehicles	435	.63	.68	.69	.57
	Weapons	171	.65	.65	.66	.65
	M		.55	.59	.60	.46
Animals	set A	253	.59	.67	.71	.58
	set B	253	.65	.70	.73	.56
Artifacts	set A	435	.61	.65	.63	.59
	set B	435	.44	.56	.61	.41

All correlations significant at $p < .05$, except for †, two-sided t

The highest correlations were found for domains with high variability in the set, such as the MC set (ρ between .86 and .91) and the domain sets with *animals* and *artifacts* (ρ between .41 and .73). The comparison between the networks G_1, G_2, and G_3 show systematic improvement by adding later responses, with sometimes substantially higher values for G_3 (e.g., *diseases*). In other words, the sparsity of the data affects semantic similarity, and adding responses that are more heterogeneous improves the result. However, sparsity is not the only factor. When the results for the directed network with all responses (G_3) and the undirected version (G_3'), which is considerably less sparse, are compared, the magnitude of the correlations with the similarity judgments are systematically lower for the undirected network (see the last column in Table 4). When the different data sets are compared, those with the most diverse pairs and largest range result in a better correspondence between the subjective judgment and the network-derived measures. This was indicated by the difference between the intracategory correlations for *animals* and *artifacts* and the experiment that compared members of these categories within a single task.

While we expected a relative better performance for abstract categories, as compared with categories that rely on a lot of sensorial information, such as *birds* or *mammals,* the results were very similar — on average, .62 for the abstract categories and .61 for *animals* with a network including all responses. Similarly, one could expect that fine-grained dimensions, especially for concepts that include numerous perceptual aspects, such as the *animal* categories, are not as well represented in the association network as those of *artifacts* that include situational aspects such as the agents, instruments, activities, and so forth. In this case, the results (an average of .61 for the *animal* categories and .66 for the *artifact* categories with G_3) are in line with our hypothesis.

Together, these results indicate that relatedness derived from directed associative networks gives a good account of judged relatedness and similarity, especially in networks with sufficient density. In our case, this was achieved by including asking participants multiple responses for each cue. Furthermore, the results suggest that these measures are applicable across a wide variety of concepts that are traditionally hard to capture using other measures, such as feature judgments (e.g., Dry & Storms, 2009). This is supported by (1) the results for the abstract concepts, which are at least as good as their concrete counterparts, and (2) the finding that similarity derived from the word association network also captures broad distinctions across categories in domains for artifacts and animals.

General discussion

One of the key criticisms on the use of word associations in word and semantic processing is the fact that the methodology does not allow the measurement of weak links (Aitchison, 2003). Furthermore, until now, all word association data sets have covered only a limited part of the human lexicon, and few attempts have been made to fully exploit using a large-scale network representation to approximate human lexico-semantic knowledge. Our results show that using a multiple-response procedure can solve some of the problems with weak links, since the resulting network encodes responses that are more heterogeneous and is denser than a similar network derived using a single-response free association procedure. Moreover, these heterogeneous responses encode useful information. First, in the case of word processing, results from a large-scale lexical decision experiment showed that measures of lexical centrality accounted for an equal amount of the variance in these tasks, as compared with state-of-the-art measures derived from spoken discourse (Keuleers, Diependaele, & Brysbaert, 2010) when later association responses were added. Moreover, there was no complete overlap between the predictors based on the word associations and spoken discourse. Second, in semantic tasks, the more heterogeneous networks that contain the later responses improve the prediction of relatedness judgments across a diverse set of concepts, including abstract and concrete concepts. An additional source of evidence for why a network with increased heterogeneity might be a more accurate description of the mental lexicon comes from the studies on judged association strength (Maki, 2008). In these studies, participants systematically "overestimate" the strength with which weakly or nonassociated pairs are related, in comparison with the associative strength observed in a single-response association procedure. By asking for multiple responses, weak associates that are qualitatively different from those given as a primary response (for instance, due to dominance effects) are made available.

Apart from incorporating more heterogeneous responses, extending the coverage of cues in word association studies allows us to construct a network that is arguably more representative of our mental lexicon. A network where the nodes are connected with each other by directed weighted link has numerous advantages. Focusing first on lexical processing, the distinction between incoming and outgoing links in the lexico-semantic network has proven to be very valuable in determining the processing advantage for words in the LDT. An accurate estimate of the number of incoming links depends on the number of cues, since only responses that are presented as a cue can be retained in a unimodal word association network. The number of different associates (i.e., the out-degree of a node) proved to be relatively unimportant, as compared with a node's in-degree.

While in-degree proved to be a good predictor of LDT decision latencies, other network measures sensitive to the

local density of the network indicate additional processing benefits. We expect measures such as betweenness or clustering coefficient to be more relevant in tasks that have been used to show a processing advantage based on semantic properties of the words used. For example, a number of researchers have argued that the number of features determines how easily a word is processed (Pexman et al., 2003; Pexman et al., 2008). Similarly, the density of a semantic neighborhood of a word also determines whether a word is recognized more quickly or not (Buchanan, Westbury, & Burgess, 2001; Mirman & Magnuson, 2006; Yates, Locker, & Simpson, 2003). If a degree or out-degree measure is used, these studies might fail to find the predicted centrality effects. For example, in the study of Mirman and Magnuson, semantic neighborhood measures defined using semantic features, word associations, and word co-occurrences from text were used to predict the decision latencies in a semantic categorization task. Only the semantic neighborhoods for word associations did not significantly predict the RTs from the categorization task. The measure used was the number of associates from the USF association corpus (Nelson, McKinney, Gee, & Janczura, 1998) and, thus, equaled the out-degree of a node. A similar measure ("number of different associates") was used as a predictor in the classic paper by Chumbley and Balota (1984). Even more recently, Locker, Simpson, and Yates (2003) used a lexical decision task in which semantic set size was determined by the number of associates of a word. In this study, words with a large semantic set size showed a processing advantage over words with a small set size. However, our results suggest that the effect found might have been underestimated in these studies, since the measures used were the out-degree of a word rather than its in-degree. While we also found a smaller but significant effect of out-degree, the interpretation of this effect will depend on the association procedure used. Most of the studies in English have relied on the USF database, in which a single response is asked for, resulting in an underestimate of set size or out-degree. In fact, ignoring in-degree as a measure of centrality is hardly surprising, since an accurate estimate of in-degree depends on the total number of the cues in the database and the USF database is the only recent word association collection that incorporated an acceptably large number of cues.

Together, this suggests that using a network-based approach allows the test of a number of new and explicit hypotheses in lexical and semantic aspects of word processing. Apart from centrality effects in lexical processing, using a directed network also helps to refine questions about how humans process the meaning of words. For instance, when directed and undirected networks are compared, our results show that the former result is a better prediction of semantic relatedness. It is important to note, however, that our similarity measure has been primarily chosen for ease of interpretation, and other measures might be better suited when network representations are used. Since similarity judgments can be asymmetric (Tversky, 1977) and the magnitude of semantic priming effects also indicate possible asymmetric strength relationships between primes and targets, it seems natural to use an asymmetric similarity measure as well. Furthermore, the similarity measure used here considers only direct overlap between the neighboring nodes of two nodes whose similarity we wish to quantify. Similar to spreading activation, it might be useful to consider indirect and mediating links in calculating this similarity. Both topics are currently under investigation in our lab, and preliminary results indicate that network-based similarity measures that incorporate these principles further improve results across a number of tasks (De Deyne et al., 2012).

Some thoughts on the size of the network

A network derived from word associations might provide a reliable approximation of the human lexico-semantic system if it represents a critical proportion of the words known by most humans. While the present study, consisting of more than 12,000 cues, represents the largest study of its kind, one could argue that even this number of cues does not suffice to approach the knowledge in adult speakers. Depending on the methods of counting (e.g., whether or not one distinguishes production and recognition), estimates of the English lexicon size in adults vary from 16,785 (D'Anna, Zechmeister, & Hall, 1991) to 58,000 (Nagy & Anderson, 1984) basic words. Nevertheless, there are numerous reasons why a number around 12,000 begins to approach the level of knowledge possessed by people.

First, the above estimated number of significant words in the lexicon of an individual might be exaggerated. After reviewing numerous studies, Hazenberg and Hulstijn (1996) concluded that when proper names, abbreviations, and compound words are not included, the vocabulary size of a Dutch university undergraduate is in the range of 14,000–17,000 words. Furthermore, the same researchers investigated the number of words that would have to be known in order to understand the content of first-year reading materials in Dutch. Their results showed that knowing 11,123 base words was sufficient to apprehend 95 % of word tokens in these materials. The coverage we obtained by reducing the number of types to the cues indicated that at least for the word associations, the coverage was around 80 %. Given the fact that we have not considered the lemmas for these responses or corrected any spelling mistakes, this number is an underestimate. In addition, some of the cues were added as part of other studies and varied in word frequency. This might also lower the coverage somewhat.

In sum, the proposed lexico-semantic network will not cover all possible words, but in terms of cumulative frequency, the coverage should allow the extraction of centrality measures of word processing that are less biased than ever before. We believe that this will be especially valuable for numerous studies in word recognition and semantic cognition. Our results show that a large portion of the variance in LDT can be explained by in-degree, a measure derived from a directed semantic network, which is distinct from other lexical availability measures such as word frequency. Given that these centrality measures affect nearly all word processing tasks, including word naming, we believe that the network-derived measures of availability might present a theoretical and practical alternative that deserves further investigation. Similarly, we expect that in semantic studies, such as priming, not only do word associations inform the forward association strength, but also, given a rich enough network, backward strength is informative. Moreover, as more mediated connections between two concepts are considered, it might turn out that previous distinctions between associative and semantic priming represent a continuum (cf., McRae, Khalkhali, & Hare, 2012). The networks that include secondary and tertiary responses should provide both better estimates of forward and backward strength effects in associative priming, while the distributional overlap measure of relatedness between two words has the potential to account for previously reported findings of semantic priming, especially when incorporating secondary and tertiary responses.

Author Note This work was supported by a research grant funded by 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. Daniel J Navarro was supported by ARC grant FT110100431. We want to thank Amy Perfors, Marc Brysbaert, and Douglas Nelson for their helpful suggestions. Comments may be sent to the author at simon.dedeyne@psy.kuleuven.be

Appendix Degree centrality

Each node has an in-degree k^{in} and an out-degree k^{out} corresponding to the number of incoming and outgoing arcs in a directed network. For the corresponding undirected network, nodes have a certain degree k, which is the number of edges of a node. The degree of the nodes is therefore also a measure of the importance of a certain node in the network. When only the adjacency structure is considered (i.e., the presence or absence of edges, regardless of their weight), k^{out} corresponds to the set size of the node. In our study, no weighted version of out-degree is calculated, since this measure reflects only the number of associations collected for each cue. This contrasts with our use of the in-degree measure, which is based on the weighted sum of incoming edges. Subsequent studies show that this measure is always more informative than the traditional in-degree measure derived from the adjacency matrix. For a network with N nodes, in-degree corresponds to:

$$k_i^{in} = \Sigma_{j \in N} w_{ij},$$

where w_{ij} indicates the connection weights between node i and j. Out-degree or set-size is calculated as the number of different outgoing links:

$$k_i^{out} = \Sigma_{j \in N} a_{ij},$$

where a_{ij} is equal to 1 if a link between i and j exists and 0 otherwise.

Clustering coefficients

A slightly more complex measure that is derived from the in- and out-degree of nodes is the clustering coefficient C (Watts & Strogatz, 1998). For a node i, this is the proportion of the edges with neighboring nodes divided by the maximal number of edges within this neighborhood. For an unweighted undirected network, C is derived as follows (cf. Watts & Strogatz, 1998):

$$C_i = \frac{1}{n} \Sigma_{i \in N} \frac{2t_i}{k_i(k_i - 1)},$$

where t_i corresponds to the number of triangles around node i calculated as follows:

$$t_i = \frac{1}{2} \Sigma_{j,h \in N} a_{ij} a_{ih} a_{jh}$$

For a weighted directed network, the formula can be extended by taking the weighted geometric mean of the directed triangles around i (see Fagiolo, 2007). Finally, a normalized coefficient C_i' is obtained by multiplication with the degree of i divided by the maximal possible number of triangles between a node i and all neighbors N. This is the measure used throughout the article.

Betweenness

Path centrality indicates how often a node is located on the shortest path between other nodes in the network. One measure of path centrality is betweenness (b). If a node with a high level of betweenness were to be deleted from a network, the network would fall apart into otherwise coherent clusters. Unlike degree, which is a count, betweenness is normalized by definition as the proportion of all shortest

paths that include the node under study. The betweenness centrality (b) for a node i is defined as (Freeman, 1978):

$$b_i = \frac{1}{(n-1)(n-2)} \sum_{\substack{h,j \in N, \\ h \neq j, h \neq i, j \neq i}} \frac{\rho_{hj}(i)}{\rho_{hj}},$$

where ρ_{hj} corresponds to the number of shortest paths from h to j passing through i and the denominator is the number of shortest paths from h to j for all possible pairs h,j.

Feedback centrality

The PageRank measure represents an example of a family of measures that include feedback information such as eigencentrality. For an unweighted network the measure is most commonly derived in matrix terms (see Langville & Meyer, 2006, for a full description):

$$PR = \left(I - \alpha A D^{-1}\right)^{-1} 1,$$

where A is the adjacency matrix and D is a diagonal matrix with the out-degree of each node as elements of the diagonal. This formula contains one free parameter, α, which was set at 0.85, according to convention (Griffiths et al., 2007; Langville & Meyer, 2006).

References

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

Andrews, M., Vinson, D., & Vigliocco, G. (2008). Inferring a probabilistic model of semantic memory from word association norms. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1941–1946). Austin: Cognitive Science Society.

Azuma, T., & Van Orden, G. C. (1997). Why safe is better than fast: The relatedness of a word's meaning affects lexical decision times. *Journal of Memory and Language, 36,* 484–504.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Balota, D. A., & Chumbley, J. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology. Human Perception and Performance, 10,* 340–357.

Balota, D. A., & Coane, J. H. (2008). Semantic memory. In J. H. Byrne, H. Eichenbaum, R. Menzel, H. L. Roediger III, & D. Sweatt (Eds.), *Handbook of learning and memory: A comprehensive reference* (pp. 512–531). Amsterdam: Elsevier.

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (A travel update). In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed.). Amsterdam: Academic Press.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990.

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

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics, 35,* 13–47.

Chumbley, J. I., & Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Memory & Cognition, 12,* 590–606.

Chwilla, D. J., Kolk, H. H. J., & Mulder, G. (2000). Mediated Priming in the Lexical Decision Task: Evidence from Event-Related Potentials and Reaction Time. *Journal of Memory and Language, 42*(3), 314–341.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 240–247.

Coltheart, M., Patterson, K., & Marshall, J. (1980). *Deep dyslexia*. London: Routledge and Kegan Paul.

Cramer, P. (1968). *Word Association*. New York: Academic Press.

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

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a Meaningful Definition of Vocabulary Size. *Journal of Literacy Research, 23,* 109–122.

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

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

De Deyne, S., & Storms, G. (2008b). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods, 40,* 198–205.

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.

de Groot, A. M. B. (1989). Representational Aspects of Word Imageability and Word Frequency as Assessed Through Word Association. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 824–845.

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

Dijkstra, E. W. (1959). A Note on Two Problems in Connection with Graphs. *Numerische Mathematik, 1,* 269–271.

Dry, M., & Storms, G. (2009). Similar but not the same: A comparison of the utility of directly rated and feature-based similarity measures for generating spatial models of conceptual data. *Behavior Research Methods, 41,* 889–900.

Durda, K., & Buchanan, L. (2008). Windsors: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods, 40,* 705–712.

Fagiolo, G. (2007). Clustering in complex directed networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 76,* 026107.

Feldman, B. (2005). Relative Importance and Value. Unpublished manuscript downloaded from http://www.prismanalytics.com/docs/RelativeImportance.pdf

Freeman, L. C. (1978). Segregation in social networks. sociological methods and research. *Sociological Methods and Research, 6,* 411–429.

Galbraith, R. C., & Underwood, B. J. (1973). Perceived frequency of concrete and abstract words. *Memory & Cognition, 1,* 56–60.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology. General, 113,* 256–281.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the Mind. *Psychological Science, 18,* 1069–1076.

Grömping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software, 17,* 1–27.

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

Hazenberg, S., & Hulstijn, J. (1996). Defining a Minimal Receptive Second-Language Vocabulary for Non-native University Students: An Empirical Investigation. *Applied Linguistics, 2,* 145–163.

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

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*. Taiwan: Taiwan University.

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

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, 1.

Kiss, G., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer Analysis. In A. Aitken, R. Bailey, & N. Hamilton-Smith (Eds.), *The Computer and Literacy Studies* (pp. 153–165). Edinburgh: University Press.

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.

Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton: Princeton University Press.

Locker, L., Simpson, G. B., & Yates, M. (2003). Semantic neighborhood effects on the recognition of ambiguous words. *Memory & Cognition, 31,* 505–515.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review, 7,* 618–630.

Maki, W. S. (2008). A database of associative strengths from the strength-sampling model: A theory-based supplement to the Nelson, McEvoy, and Schreiber word association norms. *Behavior Research Methods, 40,* 232–235.

McRae, K., Khalkhali, S., & Hare, M. (2012). Semantic and associative relations: Examining a tenuous dichotomy. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The Adolescent Brain: Learning, Reasoning, and Decision Making* (pp. 39–66). Washington, DC: APA.

Miller, G. A., & Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language & Cognitive Processes, 6,* 1–28.

Mirman, D., & Magnuson, J. (2006). The impact of semantic neighborhood density on semantic access. In *Proceedings of the 28th Annual Cognitive Science Society Meeting* (p. 1823–1828). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly, 19,* 304–330.

Nelson, D., Cañas, J., & Bajo, M. (1987). The effects of natural category size on memory for episodic encodings. *Memory & Cognition, 15,* 133–140.

Nelson, D., McEvoy, C., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition, 28,* 887–899.

Nelson, D., McEvoy, C., & Schreiber, T. (1990). Encoding context and retrieval conditions as determinants of the effects of natural category size. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 31–41.

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

Nelson, D., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review, 105,* 299–324.

Newman, M.E.J. (2010). *Networks: An introduction*. Oxford University Press.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank citation ranking: Bringing order to the web. (Tech. Rep.)*. Computer Science Department, Stanford University.

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

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

Plaut, D., McClelland, J., Seidenberg, M., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Plaut, D., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377–500.

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

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research, 11,* 95–130.

Roediger, H., & Neely, J. (1982). Retrieval blocks in episodic and semantic memory. *Canadian Journal of Psychology, 36,* 213–242.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8,* 627–633.

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

Schwanenflugel, P. J., Akin, C., & Luh, W. M. (1992). Context availability and the recall of abstract and concrete words. *Memory & Cognition, 20,* 96–104.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 82–102.

Scott, G., Donnell, P., Leuthold, H., & Sereno, S. (2009). Early emotion word processing: Evidence from event-related potentials. *Biological Psychology, 80,* 95–104.

Severens, E., Van Lommel, S., Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta Psychologica, 119,* 159–187.

Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22,* 189–228.

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

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science, 29,* 41–78.

Turner, J., Valentine, T., & Ellis, A. (1998). Contrasting effects of AoA and word frequency on auditory and visual lexical decision. *Memory and Cognition, 26,* 1282–1291.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Verheyen, S., De Deyne, S., Linsen, S., & Storms, G. (2012). Lexical and semantic norms for 1,000 Dutch adjectives. (Unpublished manuscript)

Verheyen, S., Stukken, L., De Deyne, S., Dry, M. J., & Storms, G. (2012). The generalized polymorphous concept account of graded structure in abstract categories. *Memory & Cognition, 39,* 1117–1132.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393,* 440–442.

Yates, M., Locker, L., & Simpson, G. B. (2003). Semantic and phonological influences on the processing of words and pseudohomophones. *Memory & Cognition, 31,* 856–866.