Using network clustering to uncover the taxonomic and thematic structure of the mental lexicon

Simon De Deyne
University of Adelaide
School of Psychology
5005 Adelaide, Australia
simon.dedeyne@adelaide.edu.au

Steven Verheyen
University of Leuven
Department of Psychology
Tiensestraat 102, 3000 Leuven, Belgium
steven.verheyen@ppw.kuleuven.be

While still influential, the view that concepts are organized as a hierarchical taxonomy as proposed by Rosch (1973) has been challenged on several occasions. For example, some studies have attributed a larger role to thematic relations (Gentner and Kurtz, 2005; Lin and Murphy, 2001), whereas others have stressed the role of affect in structuring word meaning (Niedenthal et al., 1999). A comprehensive account of how these different principles shape and structure meaning in the lexicon is missing, and most studies continue to be biased towards concrete noun categories that fit into hierarchical taxonomies (Medin and Rips, 2005). To capture mental or psychological properties that organize the lexicon for a wide range of concepts and semantic relations, we propose a large-scale semantic network derived from word associations as the basis to uncover what the structural principles are.

1 Network Clustering

Since this is one of the first times the mental lexicon is mapped in its entirety using an extremely extensive word association corpus, an exploratory approach is warranted. To achieve this, network clustering was used as a way to study how the mental lexicon can be structured at different scales and what type of semantic relations dominate its structure. At the basis lies a semantic network derived from a large scale word association corpus including over 12,000 cues and 3.77 million responses (De Deyne et al., 2013). For the purpose of this study, non-dominant word forms were removed (e.g., *apples* was removed if *apple* was also present) resulting in a network of 11,000 words. Next, the recent *Order Statistics Local Optimization Method* (OSLOM) was applied to identify statistically reliable clusters in a directed weighted word associations network (Lancichinetti et al., 2011). This method includes words in the final cluster solution on the basis of statistical criteria and allows for overlapping clusters. Similar to taxonomic theories of knowledge representation, words are grouped in progressively larger clusters, which allows us to evaluate structural properties of the lexicon at different scales. This hierarchical structure is also derived from the data by using a statistical criterion that involves a comparison with an appropriate null-model for the weighted directed graph.

Applying OSLOM to the semantic network resulted in a solution with five hierarchical levels. An overview of this solution is shown in Table 1. There was a large degree of variability in the number of clusters across the five hierarchical levels ranging between 2 and 506 clusters. On average, the p-value of the extracted clusters was low across all levels, indicating that the obtained clusters were unlikely to arise in a comparable random network[1]. There were few homeless nodes at any level, indicating that most words were reliably attributed to a specific cluster. There was also a considerable degree of overlap at all levels relative to the size of the clusters; clusters were more distinct at the more precise levels, where more clusters were obtained. For instance, at the lowest level 1,676 words appeared in multiple clusters, compared to 5,943 at the highest level.

Figure 1 illustrates the obtained clusters with the most prototypical examples of each cluster at various levels. At the most general level, Figure 1 shows two distinct clusters, with one of them containing highly central words with a negative connotation. In order to verify whether this interpretation is supported statistically, we used the valence judgments reported by Moors et al. (2012), which

[1]Default parameters were used in the OSLOM algorithm, except for the p cut-off value. Setting this value depends on the task as it affects the size of the clusters (Lancichinetti et al., 2011). In this application, the cutoff was set at 0.25, because the few clusters in the final solution with high p-values were easy to interpret. Other values of p did not alter the general pattern of results we report here.

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

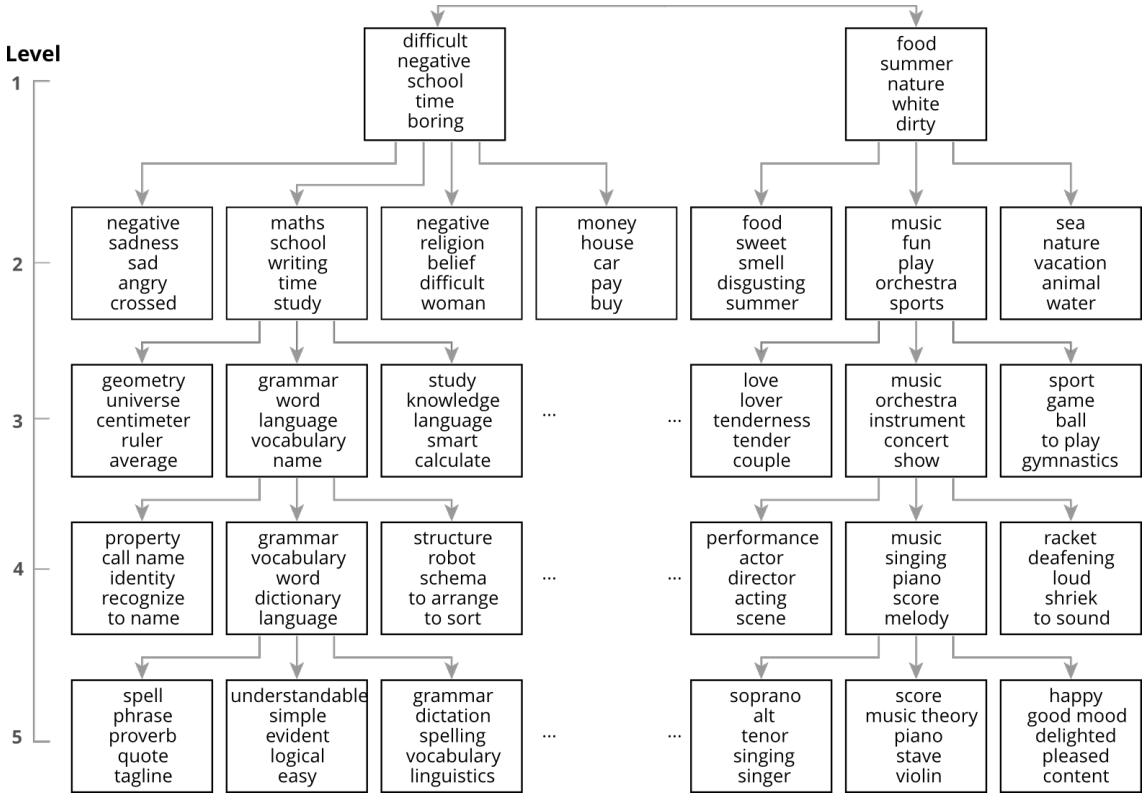

Figure 1: Hierarchical tree visualization of clusters in the lexicon with five most central members in terms of cluster in-strength.

Table 1: Overview of the hierarchical cluster structure showing five levels (Level 1 is broadest, Level 5 is most precise). The statistics include total number of clusters N, average cluster size $\langle N_c \rangle$ and its standard deviation, number of homeless nodes $N_{homeless}$, number of nodes member of multiple clusters $N_{overlapping}$, and the average p-value $\langle p \rangle$.

	1	2	3	4	5
N	2	7	37	161	506
$\langle N_c \rangle$	8588	3049	515	112	25
$sd(N_c)$	2112	973	364	66	12
$N_{homeless}$	18	18	39	86	380
$N_{overlapping}$	5943	6956	5263	4717	1676
$\langle p \rangle$	0	0.062	0.04	0.035	0.051

are applicable to 3,642 non-overlapping words in our clusters. The valence judgments differed significantly between our two clusters according to an independent t-test ($t(3640) = 7.367$, $CI = [0.190,0.327]$). This post-hoc test confirmed our interpretation of a valence difference between the clusters, which brings further support to studies that indicated valence is the most important dimension in semantic space (De Deyne et al., 2014; Samsonovic and Ascoli, 2010) and empirical findings highlighting affect-based category structure (Niedenthal et al., 1999).

At Levels 2 to 4, the meaning clusters become increasingly more concrete. For instance, Level 2 shows that the "negative" cluster in Level 1 includes clusters with abstract words or words related to human culture (*school*, *money*, *religion*, *time*,...) which are now differentiated from a purely negative cluster with central members like *negative*, *sad*, and *crossed*. The subdivisions of the "positive" cluster involve the central nodes *nature*, *music*, *sports*, and *food*, which might be interpreted as covering sensorial information and natural kinds.

At the lowest level, 506 clusters were identified, with an average size of 25 words. A total of 1,676 words occurred in multiple clusters; at least a part of them because of homonymy (e.g., *bank*) or polysemy (e.g., *language*, assigned to clusters about nationality, speech, language education, and communication). Most importantly, inspection of the content of all clusters exhibited a widespread thematic structure: the clusters were often composed of both nouns (*racket*), adjectives (*loud*), and verbs (*to sound*), which does not reflect a pure taxonomy of entitities, but also includes properties and actions.

2 Evaluating Taxonomic Structure

To test whether the clusters provide evidence for a hierarchical taxonomic view along the lines of Rosch and colleagues (Rosch, 1973) or support an alternative view based on thematic relations identified in the previous section, data from an exemplar generation task from Ruts et al. (2004) was used. In this task, 100 participants generated as many exemplars they could think of for six artifact categories (CLOTHING, KITCHEN UTENSILS, MUSICAL INSTRUMENTS, TOOLS, VEHICLES, and WEAPONS) and seven natural kinds categories (FRUIT, VEGETABLES, BIRDS, INSECTS, FISH, MAMMALS, and REPTILES). If the clusters in the word association network group together different types of birds, vehicles, fruits, and so on, this would indicate a taxonomic organization of semantic memory. For each category, we investigated the size of the best matching cluster and calculated precision and recall in terms of the F-measure for clustering performance.

A taxonomic-like organization would be evident in clusters with high precision and recall, resulting from many true positives and few false positives and false negatives. For instance, if the cluster corresponding to the category BIRDS contained *robin* (a true positive) and did not contain *spoon* (a true negative), that would increase the F-score. Conversely, if it contained *guitar* (a false positive) or did not contain *ostrich* (a false negative), that would decrease the F-score. This way, high F-scores should reflect categories that are not overly specific (many false negatives) or general (many false positives).

On average, the best matching clusters were found at Level 5. The results for each category are shown in Table 2. The average number of members in the exemplar generation task was on average 41 for the seven natural kinds categories, which is in the same range as the average best matching cluster size of 42. For artifacts the generated categories included on average 55 members, which was somewhat larger than the obtained average cluster size of 37.

The resulting F-values were on average 0.48 for the natural categories and 0.28 for the artifacts, indicating only limited support for the presence of taxonomic categories. The highest values were obtained for FISH (F = .57) and REPTILES (F = .65) where most items in the clusters were true category members.

Table 2: F-values and cluster sizes for items generated for 13 concrete noun categories. N_{human} is the category size based on the exemplar generation task; N_c is the size of the best-matching cluster; F captures precision and recall according to the human categories for the full network. F' is calculated from a network that excluded potential thematic information. F-values are fairly low, indicating lack of correspondence between the clusters and the taxonomic categories. Excluding thematic information results in F' values that do capture taxonomic information.

Category	N_{human}	N_c	F	F'
FRUIT	40	50	0.47	0.84
VEGETABLES	35	58	0.50	0.90
BIRDS	53	63	0.53	0.90
INSECTS	40	34	0.46	0.68
FISH	37	48	0.57	0.91
MAMMALS	61	21	0.20	0.76
REPTILES	21	22	0.65	0.51
Mean	41	42	0.48	0.79
CLOTHING	46	70	0.35	0.80
KITCHEN UT.	71	18	0.20	0.66
MUSIC INSTR.	46	24	0.37	0.89
TOOLS	73	56	0.25	0.76
VEHICLES	46	28	0.16	0.73
WEAPONS	46	25	0.37	0.88
Mean	55	37	0.28	0.79

Inspecting the false positives for each of the clusters in Table 3 confirms the validity of the approach as in the majority of the cases the superordinate label (e.g., *fruit*, *tools*, etc.) was the most central member of each cluster. The remaining intrusions were thematic in nature (e.g., FRUIT: pick, BIRDS: nest), thus confirming our earlier exploratory findings.

One potential response to the previous analyses relates to the nature of the data upon which they are based. Perhaps the word association task simply fails to capture taxonomic information, and if so, the results of these analyses are simply an artifact of the choice of task. Alternatively, perhaps the "failure" arises because the word association task is more general than the tasks typically used to study taxonomic categories.

There is some evidence that a different choice of task would produce different results. For instance, much of the work on taxonomic organization relies on tasks in which participants are asked to list features of entities (McRae et al., 2005; Ruts et al., 2004). One could argue that feature generation is

Table 3: Top 5 false positives ordered by cluster in-strength per category. Most of the false positives are thematic in nature. For instance, false positives for BIRDS include *beak*, *egg*, *nest*, and *whistle*.

Category	1	2	3	4	5
FRUIT	fruit	juicy	pit	pick	summer
VEGETABLES	vegetable	healthy	puree	sausage	hotchpotch
BIRDS	bird	beak	nest	whistle	egg
INSECTS	insect	vermin	beast	crawl	animal
FISH	fish	fishing	rod	slippery	water
MAMMALS	rodent	gnaw	tail	pen	marten
REPTILES	reptile	scales	animal	tail	amphibian
CLOTHING	clothing	fashion	blouse	collar	zipper
KITCHEN UT.	cooking	kitchen	stove	cooker hood	burning
MUSICAL INSTR.	wind instrument	to blow	fanfare	orchestra	harmony
TOOLS	tools	carpenter	carpentry	wood	drill
VEHICLES	speed	drive	vehicle	motor	circuit
WEAPONS	sharp	stab	blade	point	stake

a constrained version of the word association task, and the key difference is the number of thematic responses one gets in both procedures. Similarly, feature generation stimuli are usually restricted to concrete nouns, which places restrictions on what words *can* be grouped together. In other words, the tendency to find taxonomic categories may be a result of restricting the task.

To test this idea, we used the word association data to construct a network that included *only* those 588 words that belonged to one of the taxonomic categories. Moreover, in order to approximate the "shared features" measure that is more typical of feature generation tasks, we computed the cosine similarity between pairs of words. That is, words that have the same associates are deemed more similar, and this similarity was used to weight the edges in the restricted network.[2] We then applied the clustering procedure to this restricted network and repeated the analysis from the previous section. The F-statistics from this analysis are reported as the F'-values in Table 2. This time, the results of the clustering show a high degree of agreement with the taxonomic organization, with an average F-value of 0.79. The only exception was REPTILES, which upon inspection appears to reflect a failure to distinguish REPTILES from INSECTS.

The success of this analysis suggests two things. First, the word association task *does* encode taxonomic information, as evidenced by the fact that we are able to reconstruct taxonomic categories. However, the fact that the only way to do so is to mimic all the restrictive characteristics of a feature generation task (e.g., limited word set) is revealing. Taxonomic information is not the primary means by which the mental lexicon is organized: if it were, we should not have to resort to such drastic restrictions in order to uncover taxonomic categories.

In summary, even at the most detailed level of the hierarchy, only limited evidence for a taxonomic view along the lines of Rosch was found, even for typical taxonomic domains like animals. These results suggest that in much of the previous work the pervasive contribution of affective and thematic or relational knowledge structuring might be overlooked by a selection bias in terms of the concepts (nouns, mostly concrete) and semantic relations (predominantly taxonomic). This finding is in line with previous results indicating that network derived similarity estimates account better for human thematic relatedness judgments than for taxonomic relatedness judgments (De Deyne et al., in press). In priming studies, the dominance of thematic over taxonomic structure can also explain facilitation when thematic but not coordinate prime-target pairs are used (Hutchison, 2003). Finally, our findings converge with recent evidence that highlights the role of thematic representations even in domains such as animals (Gentner and Kurtz, 2006; Lin and Murphy, 2001; Wisniewski and Bassok, 1999) whereas previous reports that have stressed taxonomic organization might be more exceptional as they are heavily culturally defined (Lopez et al., 1997), a consequence of formal education (Sharp et al., 1979), or reflect different levels of expertise (Medin et al., 1997).

[2]Note that one could also derive such a similarity-based network for the complete lexicon, which would reflect the similarity between cues rather than their weighted associative strength. We did in fact do this. It produced similar results to the original analysis.

Acknowledgments

This research has been supported by an ARC grant DE140101749 awarded to SDD. SV is a postdoctoral fellow at the Research Foundation - Flanders. A longer version of this work was also submitted to the 37th Annual meeting of the Cognitive Science Society, Pasadena, 2015. We wish to express our gratitude to Dan Navarro and Amy Perfors, who contributed to the longer version of this work.

References

[De Deyne et al.2013] Simon De Deyne, Daniel J. Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single word associations. *Behavior Research Methods*, 45:480–498.

[De Deyne et al.2014] Simon De Deyne, Wouter Voorspoels, Steven Verheyen, Daniel J. Navarro, and Gert Storms. 2014. Accounting for graded structure in adjective categories with valence-based opposition relationships. *Language and Cognitive Processes*, 29(5):568–583.

[De Deyne et al.in press] Simon De Deyne, Steven Verheyen, and Gert Storms. in press. The role of corpus-size and syntax in deriving lexico-semantic representations for a wide range of concepts. *Quarterly Journal of Experimental Psychology*.

[Gentner and Kurtz2005] Dedre Gentner and Kenneth J. Kurtz. 2005. Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, and P. W. Wolff, editors, *Categorization inside and outside the lab.*, pages 151–175. American Psychology Association.

[Gentner and Kurtz2006] Dedre Gentner and Kenneth J. Kurtz. 2006. Relations, objects, and the composition of analogies. *Cognitive Science*, 30:609–642.

[Hutchison2003] Keith A. Hutchison. 2003. Is semantic priming due to association strength or feature overlap? *Psychonomic Bulletin and Review*, 10:785–813.

[Lancichinetti et al.2011] Andrea Lancichinetti, Filippo Radicchi, José J Ramasco, and Santo Fortunato. 2011. Finding statistically significant communities in networks. *PloS one*, 6(4):e18961.

[Lin and Murphy2001] Emilie L. Lin and Gregory L. Murphy. 2001. Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, 1:3–28.

[Lopez et al.1997] Alejandro Lopez, Scott Atran, John D Coley, Douglas L Medin, and Edward E Smith. 1997. The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive psychology*, 32(3):251–295.

[McRae et al.2005] Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547–559.

[Medin and Rips2005] Douglas L. Medin and Lance J. Rips. 2005. Concepts and categories: memory, meaning, and metaphysics. In K. Holyoak and R. Morrison, editors, *The Cambridge Handbook of Thinking and Reasoning*, pages 37–72. Cambridge University Press, Cambridge, UK.

[Medin et al.1997] Douglas L. Medin, Elizabeth B. Lynch, John D. Coley, and Scott Atran. 1997. Categorization and reasoning among tree experts: Do all roads lead to rome? *Cognitive psychology*, 32(1):49–96.

[Moors et al.2012] Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2012. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, pages 1–9.

[Niedenthal et al.1999] Paula M. Niedenthal, Jamin B. Halberstadt, and Åse H. Innes-Ker. 1999. Emotional response categorization. *Psychological Review*, 106(2):337.

[Rosch1973] Eleanor Rosch. 1973. Natural categories. *Cognitive Psychology*, 4:328–350.

[Ruts et al.2004] Wim Ruts, Simon De Deyne, Eef Ameel, Wolf Vanpaemel, Timothy Verbeemen, and Gert Storms. 2004. Dutch norm data for 13 semantic categories and 338 exemplars. *Behaviour Research Methods, Instruments, and Computers*, 36:506–515.

[Samsonovic and Ascoli2010] Alexei V. Samsonovic and Giorgio A Ascoli. 2010. Principal semantic components of language and the measurement of meaning. *PloS one*, 5(6):e10921.

[Sharp et al.1979] Donald Sharp, Michael Cole, Charles Lave, Herbert P Ginsburg, Ann L Brown, and Lucia A French. 1979. Education and cognitive development: The evidence from experimental research. *Monographs of the society for research in child development*, pages 1–112.

[Wisniewski and Bassok1999] Edward J. Wisniewski and M. Bassok. 1999. What makes a man similar to a tie? *Cognitive Psychology*, 39:208–238.