Extracting Clusters of Specialist Terms from Unstructured Text

Aaron Gerow
Computation Institute
University of Chicago
Chicago, IL, USA
gerow@uchicago.edu

Abstract

Automatically identifying related specialist terms is a difficult and important task required to understand the lexical structure of language. This paper develops a corpus-based method of extracting coherent clusters of satellite terminology — terms on the edge of the lexicon — using co-occurrence networks of unstructured text. Term clusters are identified by extracting communities in the co-occurrence graph, after which the largest is discarded and the remaining words are ranked by centrality within a community. The method is tractable on large corpora, requires no document structure and minimal normalization. The results suggest that the model is able to extract coherent groups of satellite terms in corpora with varying size, content and structure. The findings also confirm that language consists of a densely connected core (observed in dictionaries) and systematic, semantically coherent groups of terms at the edges of the lexicon.

1 Introduction

Natural language consists of a number of relational structures, many of which can be obscured by lexical idiosyncrasies, regional variation and domain-specific conventions. Despite this, patterns of word use exhibit loose semantic structure, namely that proximate words tend to be related. This distributional hypothesis has been operationalized in a variety of ways, providing insights and solutions into practical and theoretical questions about meaning, intention and the use of language. Distributional analyses rely primarily on observing natural language to build statistical representations of words, phrases and documents (Turney and Pantel, 2010). By studying dictionaries and thesauri, lexicographic and terminological research has proposed that a core lexicon is used to define the remaining portions of vocabulary (Itô and Mester, 1995; Massé et al., 2008). Though many words that comprise general language use reside in this core lexicon, even the most general language contains specialist or so-called “satellite” words. This paper introduces a method of extracting this peripheral structure, with co-occurrence networks of unstructured text.

The core-periphery structure has been observed in dictionaries where definitions tend to use a restricted vocabulary, repetitively employing a core set of words to define others (Sinclair, 1996; Picard et al., 2013). In the farther regions of the lexicon, it is more difficult to find systematic semantic definition with corpus-based techniques due to the overwhelming number of infrequent words. Unfortunately, the fringe of the lexicon can be more important than the core because this is where domain-specific terminology resides — features that may be more important than frequent.

Examining dictionaries, (Picard et al., 2013) propose that the lexicon consists of four main parts: a *core* set of ubiquitous words used to define other words, a *kernel* that makes up most of the lexicon, a *minimal grounding set* that includes most of the core and some of the kernel, leaving a set of *satellites* in the periphery. This topography, reproduced in Figure 1, has been found in the way dictionary entries use words to define one another. In networks of dictionary definitions, the core component tends to form a strongly connected component (SCC) leaving satellites in smaller SCCs with relatively weak links to the core. This paper explores whether these these satellites form systematic, cohesive groups and whether they are observable in natural language.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1426–1434,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

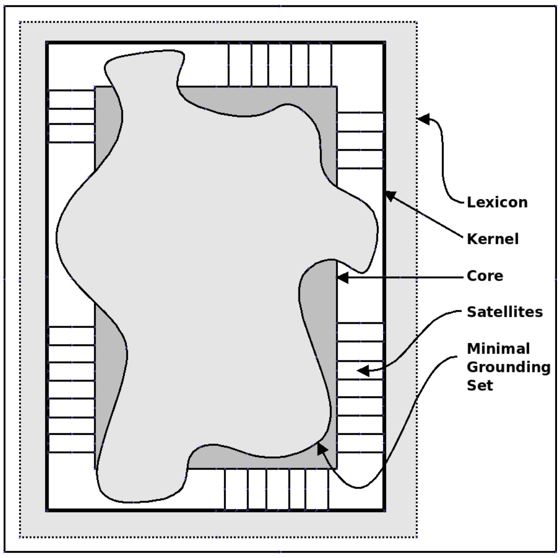

Figure 1: Dictionary studies have proposed that the lexicon consists of a strongly connected core, around which there is a kernel, an asymmetric grounding set and satellites. Adapted from (Picard et al., 2013).

Words with relatively specific definitions within subjects, referred to as *terms* in lexicographic research, are apparent in nearly all domains of discourse. Here, the goal is to explore structure among these peripheral terms without a dictionary. To do this, a method based on community detection in textual co-occurrence networks is developed. Such graph-based methods have become increasingly popular in a range of language-related tasks such as word clustering, document clustering, semantic memory, anaphora resolution and dependency parsing (see Mihalcea and Radev, 2011 for a review).

This paper seeks to address two important questions about the observed landscape of the lexicon in natural language: to investigate whether satellite clusters found in dictionaries can be observed in text, and more importantly, to explore whether statistical information in co-occurrence networks can elucidate this peripheral structure in the lexicon. We frame these questions as the task of extracting clusters of related terms. If satellites are systematically organized, then we can expect to find cohesive clusters in this region. Moreover, if the networked structure of dictionary entries supports the landscape in Figure 1, a similar structure may be present in co-occurrence patterns in natural text.

2 Method

Word clustering, as a means to explore underlying lexical structure, should accommodate fuzzy and potentially contradictory notions of similarity. For example, *red* and *green* are at once similar, being colors, but as colors, they are very different. Alternatively, the words *car*, *fast*, *wheel*, *export* and *motorists* share a thematic similarity in their relation to automobiles. One conception of word clustering is to construct a thesaurus of synonyms (Calvo et al., 2005), but clustering could allow other lexical semantic relationships. One such database, WordNet, defines specific semantic relationships and has been used to group words according to explicit measures of relatedness and similarity (Miller, 1995; Pedersen et al., 2004). Distributional, corpus-based techniques that define words as feature vectors (eg. word-document co-occurrences), can address many limitations of manually created lexicons (see Turney et al., 2010 for a review). Clustering nouns by argument structure can uncover naturally related objects (Hindle, 1990) and spectral methods can relate distinct classes of nouns with certain kinds of verbs to induce selectional preferences (Resnik, 1997; Sun and Korhonen, 2009; Wilks, 1975) and assist metaphor processing (Shutova et al., 2013).

A pervasive weakness of many existing approaches to word-clustering, is an underlying prioritization of frequent words. To help address this sparsity, many models collapse words into stems, preclude uncommon words, or underestimate the relevance of infrequent words (Dagan et al., 1999). Probabilistic topic models have emerged as a uniquely flexible kind of word-clustering used in content analysis (Steyvers and Griffiths, 2007), text classification (Wei and Croft, 2006) and provide an extensible framework to address other tasks (Blei, 2012). Because the structure of satellite terms is not likely to rely on specific (much less consistent) lexical semantic relationships, we adopt a measure of *semantic coherence*, commonly used to qualify the results of topic models, as an indirect measure of what people tend to view as a cohesive set of words. This measure, which is defined in the next section, is particularly attractive because it is corpus-based, does not assume any specific semantic relationship and correlates with expert evaluations (Mimno et al., 2011; Newman et al., 2010). Using semantic coherence provides a way of measuring the quality of word-associations without appeal to a dictionary or assuming rigid relationships among the clustered words.

The first step is to construct a co-occurrence graph from which communities are extracted. Then the centrality of each word is computed within a community to generate cluster-specific rankings. The goal is not to categorize words into classes, nor to provide partitions that separate associated words across a corpus. Instead, the method is designed to extract qualifiable sets of specialist terms found in arbitrary text. Crucially, the method is designed to require no document structure and minimal pre-processing: stop-words and non-words are not removed and no phrasal, sentence or document structure is required. Although stemming or lemmatization could provide more stream-lined interpretations, the minimal pre-processing allows the method to operate efficiently on large amounts of unstructured text of any language.

Co-occurrence networks have been used in a variety of NLP applications, the basic idea being to construct a graph where proximate words are connected. Typically, words are connected if they are observed in an n-word window. We set this window to a symmetric 7 words on either side of the target and did not use any weighting[1]. In the resulting network, edge frequencies are set to the number of times the given co-occurrence is observed. The resulting networks are typically quite dense and exhibit small-world structure where most word-pairs are only a few edges apart (Baronchelli et al., 2013; Ferror i Cancho and Solé, 2001). To explore the effect of this density, different minimum node- and edge-frequencies were tested (analogous to the word- and co-occurrence frequencies in text). It was found that not setting any thresholds provided the best results (see Figure 2), supporting our minimal pre-processing approach.

To extract clusters from the co-occurrence matrix, the Infomap community detection algorithm was used. Infomap is an information-theoretic method that optimizes a compression dictionary using it to describe flow through connected nodes (Rosvall and Bergstrom, 2008). By minimizing a description of this flow, the algorithm can also extract nested communities (Rosvall and Bergstrom, 2011). In our experiments, we used the co-occurrence frequencies as edge-weights and ran 50 trials for each run of the algorithm. Co-occurrence networks tended to form one monolithic community, corresponding to the lexicon's core SCC, surrounded by a number of smaller communities. The monolithic community is discarded out-right, as it represents the core of the lexicon where few specialist terms reside. As we will see, the community detection algorithm naturally identifies this SCC, distinguishing satellite clusters of terminology. Though we do not explore its effect, the sensitivity of Infomap can be tuned to vary the relative size of the core SCC compared to the satellites, effectively allowing less modular communities to be considered satellites.

Corpus	Docs	Tokens	Nodes	Edges
TASA	38,972	10.7M	58,357	1,319,534
NIPS	3,742	5.2M	28,936	1,612,659
enTenTen	92,327	72.2M	69,745	7,721,413

Table 1: Co-occurrence networks of each corpus.

To compare and interpret the resulting clusters, various measures of centrality were tested for ranking words within their communities. The goal of this ranking is to find words that typify or define their community without assuming its underlying semantics. The results in the next section show that a number of common centrality measures work comparably well for this task. The final output of the system is a set of communities, in which words are ranked by their centrality.

3 Results & Analysis

Three corpora were used for evaluation: the TASA, NIPS and enTenTen collections. TASA consists of paragraph-length excerpts from high-school level, American English texts (Landauer and Dumais, 1997). The NIPS collection contains 17 volumes of annual proceedings from the conference of the same name. The enTenTen corpus is a web-based collection of text-heavy, English web-sites. Table 1 summarizes the collections and their co-occurrence networks.

The extracted communities, which consist of word-centrality pairs, are similarly structured to the output of topic models. Because appeals to human judgement are expensive and can introduce issues of consistency (Chang et al., 2009; Hu et al., 2011), a corpus-based measure of semantic coherence has been proposed (Mimno et al., 2011). Co-

[1] 7 was found to be the optimal window-size in terms of coherence. These preliminary results are available at knowledgelab.org/docs/coherent_clusters-data.xls.

herence is used as a proxy for human judgments. A general form of semantic coherence can be defined as the mean pair-wise similarity over the top n words in a topic or cluster t

$$C(t) = \frac{1}{n} \sum_{\substack{(w_i,w_j)\in t \\ i<j}}^{n} S(w_i, w_j)$$

where S is a symmetric measure of similarity. Newman, et al. (2010) surveyed a number of similarity metrics and found that mean point-wise mutual information (PMI) correlated best to human judgements. PMI is a commonly used measure of how much more information co-occurring words convey together compared to their independent contributions (Church and Hanks, 1990; Bouma, 2009). Using PMI as S, we can define a version of coherence, known as *UCI Coherence*:

$$C_{UCI}(t) = \frac{1}{n} \sum_{\substack{(w_i,w_j)\in t \\ i<j}}^{n} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

where $p(w)$ is estimated as relative frequency in a corpus: $\frac{f(w)}{\sum_i f(w_i)}$. Using coherence to optimize topic models, Mimno et al. (2011) found that a simplified measure, termed *UMass Coherence*, is more strongly correlated to human judgments than C_{UCI}. For topic t, C_{UMass} is defined as follows:

$$C_{UMass}(t) = \frac{1}{n} \sum_{\substack{(w_i,w_j)\in t \\ i<j}}^{n} \log \frac{D(w_i, w_j) + 1}{D(w_j)}$$

where $D(w)$ is the number of documents containing w, and $D(w, w')$ is the number of documents containing both w and w'. Note that D relies crucially on document segmentation in the reference corpus, which is not encoded in the co-occurrence networks derived by the method described above. Thus, though the networks being analyzed and the coherence scores are both based on co-occurrence information, they are distinct from one another. Following convention, we compute coherence for the top 10 words in a given community. C_{UMass} was used as the measure of semantic coherence. and D was computed over the TASA corpus, which means the resulting scores are not directly comparable to (Mimno et al., 2011), though comparisons to other published results are provided below.

3.1 Ranking Functions & Frequency Thresholds

After communities are extracted from the co-occurrence graph, words are ranked by their centrality in a community. Six centrality measures were tested as ranking functions: degree centrality, closeness centrality, eigenvector centrality, Pagerank, hub-score and authority-score (Friedl et al., 2010). Degree centrality uses a node's degree as its centrality under the assumption that highly connected nodes are central. Closeness centrality measures the average distance between a node and all other nodes, promoting nodes that are "close" to the rest of the network. Eigenvector centrality favors well-connected nodes that are themselves connected to well-connected nodes. Pagerank is similar to eigenvector centrality, but also promotes nodes that mediate connections between strongly connected nodes. Hub and authority scores measure interconnectedness (hubs) and connectedness to interconnected nodes (authorities). Figure 2 shows the average coherence, across all communities extracted from the TASA corpus, for each centrality measure. The average coherence scores are highest using hub-score, though not significantly better than auth-score, eigenvector centrality or closeness centrality. In the results that follow, hub-scores were used to rank nodes within communities.

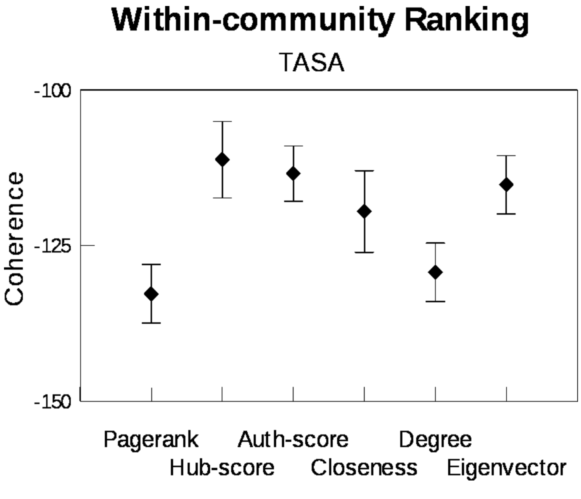

Figure 2: Mean coherence for six centrality measures. Error-bars are ±2 SE of the mean.

Imposing minimum node and edge frequencies in the co-occurrence graph was also tested. However, applying no thresholds provided the highest average coherence. Figure 3 shows the average coherence for eight threshold configurations. Though we used the TASA corpus for these tests, we have no reason to believe the results would differ significantly for the other corpora.

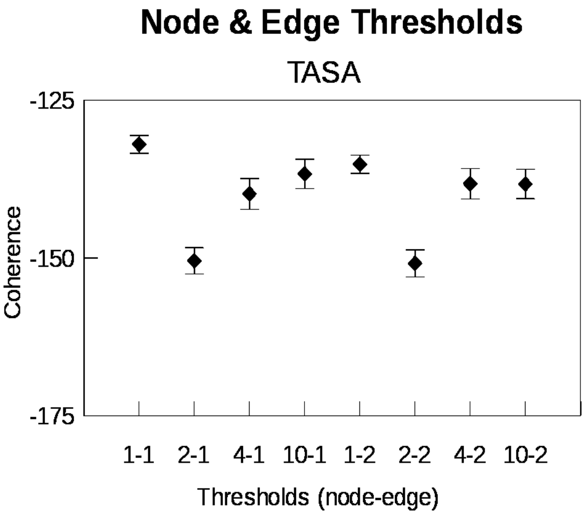

Figure 3: Mean coherence for different minimum node and edge frequencies, corresponding to thresholds for word and co-occurrence counts. Error-bars are ±2 SE of the mean.

3.2 Community Coherence

Table 2 shows three communities of specialist terms extracted from each text collection, with their normalized hub-scores. Normalizing the scores preserves their rank-ordering and provides an indication of relative centrality within the community itself. For example, compare the first and last words from the top TASA and NIPS clusters: the difference between *thou* and *craven* (TASA) is considerably more than *model* and *network* (NIPS). In general, higher ranked words appear to typify their communities, with words like *model*, *university* and *nuclear* in the NIPS examples. These clusters are typical of those produced by the method, though in some cases, the communities contain less than 10 terms and were not included in the coherence analysis. Note that these clusters are not systematic in any lexical semantic sense, though in almost every case there are discernible thematic relations (middle-English words, Latin America and seafood in TASA).

TASA		**NIPS**		**enTenTen**	
thou	1.00	model	1.00	cortex	1.00
shalt	0.72	learning	0.99	prefrontal	0.88
hast	0.49	data	0.96	anterior	0.41
thyself	0.26	neural	0.94	cingulate	0.33
dost	0.24	using	0.85	medulla	0.28
wilt	0.24	network	0.85	parietal	0.13
canst	0.12	training	0.73	insula	0.13
knowest	0.10	algorithm	0.66	cruciate	0.11
mayest	0.10	function	0.63	striatum	0.11
craven	0.01	networks	0.62	ventral	0.10
peru	1.00	university	1.00	pradesh	1.00
ecuador	0.84	science	0.85	andhra	0.67
bolivia	0.80	computer	0.83	madhya	0.56
argentina	0.67	department	0.74	uttar	0.50
paraguay	0.54	engineering	0.30	bihar	0.21
chile	0.52	report	0.30	rajasthan	0.19
venezuela	0.48	technical	0.29	maharashtra	0.16
uruguay	0.28	institute	0.26	haryana	0.12
lima	0.17	abstract	0.25	himachal	0.10
parana	0.11	california	0.23	arunachal	0.04
clams	1.00	nuclear	1.00	cilia	1.00
crabs	0.87	weapons	0.66	peristomal	0.73
oysters	0.87	race	0.57	stalk	0.62
crab	0.67	countries	0.40	trochal	0.51
lobsters	0.66	rights	0.37	vorticella	0.35
shrimp	0.62	india	0.27	campanella	0.32
hermit	0.50	russia	0.26	hairlike	0.17
mussels	0.27	philippines	0.26	swimmers	0.15
lice	0.23	brazil	0.25	epistylis	0.12
scallops	0.20	waste	0.22	telotroch	0.11

Table 2: Sample clusters from the TASA, NIPS and enTenTen collections. Shown are the clusters' top ten words, ranked by their normalized hub-score within the community. Note the differences in hub-score distributions between clusters.

Figure 4 shows the average coherence for our method, compared to that of a 20-topic latent Dirichlet allocation (LDA) model fit to the same corpora. Results from an LDA model fit to our corpora, as well as from a sample of published topics, are provided as a baseline to calibrate readers' intuitions about coherence[2]. Although topics from LDA do not necessarily consist of specialist terms those in the current model, the expectation of coherence remains: probable or central words should comprise a cohesive group. In every case, coherence is calculated over the top 10 words ranked using within-community hub-scores, for every community of 10 or more words. The results show that LDA provides relatively consistent coherence across collections, though with generally more variance than the communities of specialist terms. The term clusters are more coherent for the enTenTen collection than the others, which may

[2] Coherence was computed for the published results with C_{UMass} using TASA as the reference corpus.

be due to its larger size. This up-tick on the largest corpus may have to do with the proportional size of the monolithic community for the less structured documents in enTenTen. Figure 5 depicts how the proportional size of the core would effect the number and size of satellite clusters. It was found that the largest community (the core SCC) comprised 95% of TASA, 90% of NIPS and 97% of enTenTen. It may be that specialized language will have a proportionally smaller core and more satellite communities, whereas more general language will have a larger core and fewer satellites.

A critical question remains as to whether the method is actually observing the core-periphery structure of the lexicon or if it is an artifact. To test this, the frequencies of words in satellite communities were compared to those in the monolithic cases. If the monolithic community does indeed correspond to the core proposed in Figure 1, words in the satellites should have significantly lower frequencies. Indeed, the monolithic community in every corpus contained words that were significantly more frequent than those in the communities (Wilcoxon rank-sum test; Table 3). Taken with the coherence scores, these results show that there is coherent structure in the periphery of the lexicon, that can be extracted from unstructured text.

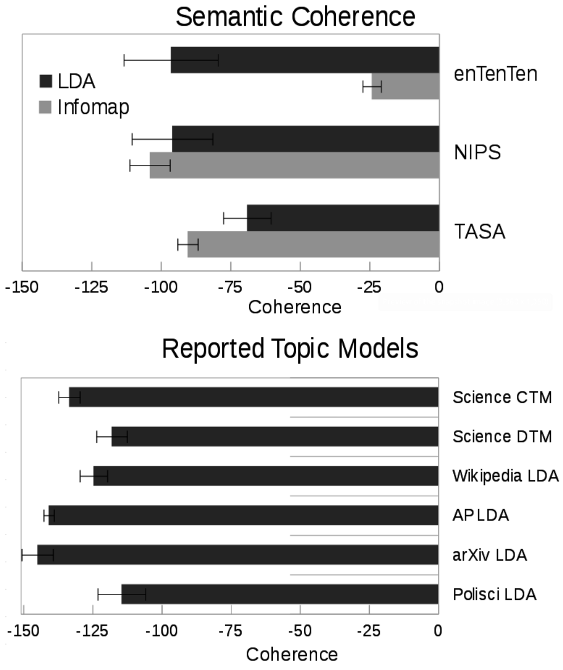

Figure 4: Mean coherence (C_{UMass}) for satellite clusters and topics from LDA on the TASA, NIPS and enTenTen collections (top). Also shown are the mean coherence of topics found in published models (LDA, a dynamic topic model, *DTM* and a correlated topic model, *CTM*; bottom). Error-bars are ±2 SE of the mean.

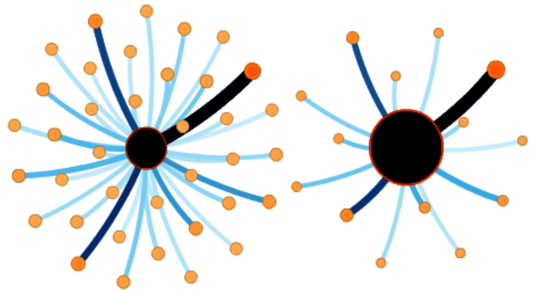

Figure 5: A proportionally larger core SCC (right) would force satellite communities to be smaller, less numerous and more isolated. Alternatively, with a small core (left), satellite communities would be more numerous and prominent.

Corpus	mean f_c	mean f_s	W	df
TASA	112.3	7.3	39985454	40895
NIPS	211.5	10.7	28342663	25077
enTenTen	365.1	15.9	246095083	72695

Table 3: Comparison of frequency for core words, f_c, found in the monolithic community and specialist terms, f_s, found in the satellite communities (Wilcoxon rank-sum test). All differences were significant at $p < 0.001$.

4 Discussion

The results of our method show that outlying structure in the lexicon can be extracted directly from large collections of unstructured text. The lexicon's topography, previously explored in dictionary studies, contains modular groups of satellite terms that are observable without appeal to external resources or document structure and with minimal normalization. The contribution of this method is two-fold: it confirms the structure of the *observed* lexicon is similar to that apparent in the organization of dictionaries (Picard et al., 2013). Second, it offers a tractable, reliable means of extracting and summarizing structure in the fringes of the lexicon.

The output of the model developed here is similar to topic models, but with some important differences. Topic models produce a probability distribution over words to define a topic, which can be summarized by the top 10 to 20 most likely words. Instead of probabilities, the within-community hub-scores were used to rank words in each cluster. This means that the actual structure of the community (to which topics have no analogue) is responsible for producing the scores that rate words' internal relevance. Another crucial difference is that topic size from a single sampling iteration tends to correlate with coherence (Mimno et al., 2011), but in the current method, there is no correlation between cluster size and coherence ($p = 0.98$). The other important difference is that whereas topic models produce a topic-document mixture that can be used for posterior inference, to perform such inference with our method, the output would have to be used indirectly.

One understated strength of the community detection method is the minimal required preprocessing. Whereas many solutions in NLP (including topic models) require document segmentation, lexical normalization and statistical normalizations on the co-occurrence matrix itself, the only variable in our method is the co-occurrence window size. However, lemmatization (or stemming) could help collapse morphosyntactic variation among terms in the results, but stop-word removal, sentence segmentation and TF-IDF weighting appear unnecessary. What might be most surprising given the examples in Table 2 is that word-document occurrence information is not used at all. This makes the the method particularly useful for large collections with little to no structure.

One question overlooked in our analysis concerns the effect the core has on the satellites. It could be that the proportional size of a collection's core is indicative of the degree of specialist terminology contained in the collection. Also, the raw number of satellite communities might indicate the level of diversity in a corpus. Addressing these questions could yield measures of previously vague and latent variables like specialty or topical diversity, without employing a direct semantic analysis. By measuring a collection's core size, relative to its satellites, one could also use measure changes in specialization. The Infomap algorithm could accommodate such an experiment: by varying the threshold of density that constitutes a community, the core could be made smaller, yielding more satellites, the coherence of which could be compared to those reported here. One could examine the position of individual words in the satellite(s) to explore what features signal important, emerging and dying terms or to track diachronic movement of terms like *computer* or *gene* from the specialized periphery to core of the lexicon.

At the level of inter-related term clusters, there are likely important or central groups that influence other satellites. There is no agreed upon measure of "community centrality" in a network sense (Eaton and Mansbach, 2012). One way to measure the importance of a community would be to use significance testing on the internal link mass compared to the external (Csardi and Nepusz, 2006). However, this approach discards some factors for which one might want to account, such as centrality in the network of communities and their composition. Future work could seek to combine graph-theoretic notions of centrality and intuitions about the defining features of term clusters. Another avenue for future research would be to use mixed membership community detection (Gopalan and Blei, 2013). Allowing terms to be represented in more than one community would accommodate words like *nuclear*, that might be found relating to weaponry, energy production and physics research at the same time. Using co-occurrence networks to extract clusters of specialist terms, though an important task, is perhaps only a starting point for exploring the observed lexicon. Network-based analysis of language offers a general and powerful potential to address a range of questions about the lexicon, other NLP tasks and language more generally.

Acknowledgments

Thanks to E. Duede and three reviewers for comments on earlier versions of this manuscript. This work was supported by a grant from the Templeton Foundation to the Metaknowledge Research Network and grant #1158803 from the National Science Foundation.

References

Andrea Baronchelli, Ramon Ferrer i Cancho, Romualdo Pastor-Satorras, Nick Chater, and Morten H. Christiansen. 2013. Networks in cognitive science. *Trends in cognitive sciences*, 17(7):348–360.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the German Society for Computational Linguistics & Language Technology*, pages 31–40.

Hiram Calvo, Alexander Gelbukh, and Adam Kilgarriff. 2005. Distributional thesaurus versus wordnet: A comparison of backoff techniques for unsupervised pp attachment. In *Computational Linguistics and Intelligent Text Processing*, pages 177–188. Springer.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, volume 22, pages 288–296.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Gabor Csardi and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Ido Dagan, Lillian Lee, and Fernando C.N. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.

Eric Eaton and Rachael Mansbach. 2012. A spin-glass model for semi-supervised community detection. In *Association for the Advancement of Artificial Intelligence*.

Ramon Ferror i Cancho and Richard V. Solé. 2001. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265.

Dipl-Math Bettina Friedl, Julia Heidemann, et al. 2010. A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2(6):371–385.

Prem K. Gopalan and David M. Blei. 2013. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 268–275.

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive topic modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 248–257.

Junko Itô and Armin Mester. 1995. The core-periphery structure of the lexicon and constraints on reranking. *University of Massachusetts occasional papers in linguistics*, 18:181–209.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

A. Blondin Massé, Guillaume Chicoisne, Yassine Gargouri, Stevan Harnad, Olivier Picard, and Odile Marcotte. 2008. How is meaning grounded in dictionary definitions? In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*, pages 17–24.

Rada Mihalcea and Dragomir Radev. 2011. *Graph-based natural language processing and information retrieval*. Cambridge University Press.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41.

Olivier Picard, Mélanie Lord, Alexandre Blondin-Massé, Odile Marcotte, Marcos Lopes, and Stevan Harnad. 2013. Hidden structure and function in the lexicon. *NLPCS 2013: 10th International Workshop on Natural Language Processing and Cognitive Science, Marseille, France.*

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57.

Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123.

Martin Rosvall and Carl T. Bergstrom. 2011. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

John Sinclair. 1996. The empty lexicon. *International Journal of Corpus Linguistics*, 1(1):99–119.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 638–647.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Xing Wei and Bruce Croft. 2006. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53–74.