Behavior Research Methods
2006, 38 (2), 262-279

Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping

ANDREW E. SMITH and MICHAEL S. HUMPHREYS
University of Queensland, Brisbane, Queensland, Australia

The Leximancer system is a relatively new method for transforming lexical co-occurrence information from natural language into semantic patterns in an unsupervised manner. It employs two stages of co-occurrence information extraction—*semantic* and *relational*—using a different algorithm for each stage. The algorithms used are statistical, but they employ nonlinear dynamics and machine learning. This article is an attempt to validate the output of Leximancer, using a set of evaluation criteria taken from content analysis that are appropriate for knowledge discovery tasks.

There are several reasons why one would want an automated system for content analysis of text. It is known that human decision makers are potentially subject to influences that they are unable to report (Nisbett & Wilson, 1977). Furthermore, the mitigation of subjectivity in human analysis requires extensive investment of time and money in the content analysis process. Code books or dictionaries must be validated, coders must be trained, and intercoder reliability must be tested (see, e.g., Weber, 1990). Increasing the automation of this process should reduce the cost and allow more rapid and frequent analysis and reanalysis of text. It is also hoped that such a system will be applicable to extremely large quantities of text where there is little possibility of intense human analysis. Text corpora of up to 300 Mb have been analyzed with Leximancer so far, but there is no theoretical limit, other than the utility of the results. When applied to larger quantities of text, this method of analysis can also be thought of as a form of text mining.

The form of semantic mapping evaluated in this article has been published elsewhere (Smith, 2000a, 2000b, 2003). The Leximancer system performs a style of automatic content analysis. The system goes beyond keyword searching by discovering and extracting thesaurus-based concepts from the text data, with no requirement for a prior dictionary, although one can be used if desired. These concepts are then coded into the text, using the thesaurus as a classifier. The resulting asymmetric concept co-occurrence information is then used to generate a concept map.

The key methods and their derivation will be described briefly below, but the essential features are as follows. A unified body of text is examined to select a ranked list of important lexical terms on the basis of word frequency and co-occurrence usage. These terms then seed a bootstrapping thesaurus builder, which learns a set of classifiers from the text by iteratively extending the seed word definitions. The resulting weighted term classifiers are then referred to as *concepts*. Next, the text is classified using these concepts at a high resolution, which is normally every three sentences. This produces a concept index for the text and a concept co-occurrence matrix. By calculating the relative co-occurrence frequencies of the concepts, an asymmetric co-occurrence matrix is obtained. This matrix is used to produce a two-dimensional concept map via a novel emergent clustering algorithm. The connectedness of each concept in this semantic network is employed to generate a third hierarchical dimension, which displays the more general parent concepts at the higher levels.

A major goal of the Leximancer system is to make the analyst aware of the global context and significance of concepts and to help avoid fixation on particular anecdotal evidence, which may be atypical or erroneous. We wish to evaluate the validity of this system. In particular, we will be examining the structure and concept names of the final concept map and, also, the nature of the weighted term sets that form the thesaurus.

FOUNDATIONS OF THE TECHNIQUE

The exploitation of information contained in the co-occurrence statistics of words within text has had a long history under the banner of corpus linguistics (Stubbs, 1996). In essence, a word can be defined by its context in usage. Beeferman and colleagues observed that words tend to correlate with other words over a certain range within the text stream (Beeferman, Berger, & Lafferty, 1997). Computational linguists have also exploited this aspect of language—for word sense disambiguation, as a particular example (Yarowsky, 1995). In the discipline

A.E.S. has inventor rights to the Leximancer intellectual property. Correspondence concerning this article should be addressed to A. E. Smith, ARC Key Centre for Human Factors and Applied Cognitive Psychology, University of Queensland, Brisbane, QLD 4072, Australia (e-mail: asmith@humanfactors.uq.edu.au).

of psychology, Burgess and Lund (1997) developed the hyperspace analogue to language (HAL). This system exploits lexical co-occurrence within a sliding window in the text to construct a matrix of representations of words in terms of other co-occurring words. These representations are then compared using similarity metrics, such as the standard cosine metric. The similarity measurements are used to demonstrate semantic and grammatical clustering, frequently by means of multidimensional scaling.

Landauer and his colleagues (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) developed latent semantic analysis (LSA), which exploits the occurrence of words in text segments. LSA uses single value decomposition to reduce the dimensionality of the word by text-segment matrix. The dimensional reduction again results in vectorial representations of words (and also of text segments) in terms of a kernel of vectors of reduced rank. The eigenvalues can be used to rank the contribution of each kernel vector. Again, similarity measurements between word representations allow the inference of indirect relationships between words that appear in similar contexts; this could also be described as approximating a transition from episodic to semantic linkage. Specifically, it is the reduction of the rank of the matrix and the corresponding information loss and abstraction of detail, which leads to the discovery of an indirect relationship. LSA has been shown to perform as well as humans in multichoice vocabulary tests, essay marking, and the acquisition of lexical knowledge. Specifically, Landauer and colleagues have shown that the induction of implicit relationships between contextually similar words can accelerate vocabulary growth from limited training examples to a degree similar to that observed in children.

The above-mentioned methods have demonstrated that there is considerable information contained in word co-occurrence statistics. In fact, we will normally refer to these induced word co-occurrence categories as *concepts*, since there is psychological evidence that correlates them with human learning and performance. However, we freely admit that they are still *textual concepts*, and any correlation with mental states is abductive. A discussion of the relationship between observable signs and their meanings is beyond the scope of this article, but a good discussion of the various taxonomies of meaning from the viewpoints of different disciplines can be found in chapter 1 of Osgood, Suci, and Tannenbaum (1957). In addition, there is no one method that can claim to optimally capture this information. The choice of text segment, of co-occurrence metric, and of the algorithm for inferring indirect relationships are design decisions. Indeed, as will be seen with Leximancer, it is possible to employ different metrics and inference algorithms in series to exploit different aspects of the co-occurrence information.

Leximancer employs two stages of extraction from episodic co-occurrence information, performed sequentially. These can be characterized as semantic extraction, followed by relational extraction. In each case, the data consist of actual episodic co-occurrence records. In the language of relational content analysis (see, e.g., Weber, 1990), a set of words that discriminate each category across the corpus of data is learned in the first phase; this can be considered as learning the categorical dictionary. These category classifiers are then used to code the text segments. Finally, the category frequency information and category co-occurrence information, which constitutes relational information, is analyzed. Equivalently, in the language of information systems (e.g., Sowa, 2000), attributes of entities or concepts are learned in the first phase, and relationships between entities and concepts are established in the second phase.[1]

This process of abstracting words for entities and primitive concepts *prior to* extracting the relationships between them is a very efficient way of controlling combinatoric explosion. For example, if a text collection has a vocabulary of 20,000 words, there are slightly fewer than 200 million possible pairwise relational combinations of words. Obviously, dimensional reduction must be employed to make the relational network easily comprehensible. If the vocabulary of the text can be grouped, say, into 100 concepts, each with 200 terms on average (neglecting repetitions), the maximum number of pairwise concept relationships is now 4,950, which is a reduction by a factor of 40,000. This process also allows retrieval of episodic text records, using semantic representations of cue words, even when the initial cue words are not present in the text records.

For the semantic extraction phase, there are several aims.

1. To construct classifiers for multiple concepts that can predict whether a small segment of text contains one or more of the concepts.

2. To provide a meaningful name for each concept as a signifier; this is done to support interpretation and visualization.

3. To allow the concept set to characterize the message conveyed by the text corpus.

4. To also allow manual customization of the concept set prior to learning of the representations. Even if it were possible to extract just one conceptual representation of the text that reflected its message, the infinite variability of the context and intent of the user means that modification of the conceptual view is essential to its usefulness. To support this, processes of concept *seeding* and *profiling* are desired. Seeding is a method whereby an incomplete but characteristic set of query terms can be expanded and refined by a machine-learning process into an effective lexical classifier. Profiling is a method for taking a prior set of concepts of interest and discovering a set of related concepts that depend either strongly or weakly, either directly or indirectly, on the prior concepts.

To achieve these criteria, a concept bootstrapping algorithm was developed from a word sense disambiguation algorithm (Yarowsky, 1995). The requirement that the resulting representation be capable of classifying small segments of text, with limited available evidence, led to the selection of a naive Bayesian co-occurrence metric

(Salton, 1989), which is known to perform well as a text classifier (Dumais, Platt, Heckerman, & Sahami, 1998). This metric, derived from Bayesian decision theory, takes into consideration not only how frequently two words co-occur, but also how often they occur apart; this is similar to a log odds, or two-way contingency statistic. This metric gives a tighter binding of relevant terms to concepts that is suitable for extracting discriminating attributes of entities or concepts. For example, consider a document in which the occupational hazards of postal workers are discussed. To characterize the identity of a concept such as *dog* in this text, terms such as *bark*, *kennel*, and *tail* may be diagnostic, in that those terms may appear frequently alongside *dog* and infrequently elsewhere. Note that in *other* documents, *bark* could be diagnostic of trees. However, the term *postman*, although it may appear in relational encounters with *dog*, will occur more often elsewhere in other relationships. Thus, it seems appropriate to consider *postman* and *dog* as separate categories in this text, with the category of *dog* being discriminated by such words as *bark*, *kennel*, and *tail*.

The second stage, relational extraction, begins with the classification, or coding, of text segments, using the learned semantic classifiers. This is an implementation of naive Bayesian accumulation of evidence, using the term weights. After this process, the following statistics are available: concept count, concept co-occurrence count and relative concept co-occurrence frequency, and word count within each text segment classified within a concept. In addition, the text episodes classified within each concept and each pair of concepts can be retrieved and inspected.

There are many forms of statistical, data mining, and network analyses that could be performed on the concept statistics. It must be noted that the concepts show an approximate power law distribution of decreasing frequency within most data sets. As a result, co-occurrence information will lead to asymmetric attachment between concepts if the frequency of each concept is considered. In concrete terms, the relative co-occurrence frequency between two concepts will change, in general, depending on which concept the frequency is relative to. The resulting information can be expressed as an asymmetric concept co-occurrence matrix containing relative co-occurrence frequencies. Equivalently, this can be viewed as a concept network with directed weighted arcs. Relative co-occurrence frequency can also be considered as a frequentist approximation to the conditional probability of finding a second concept, given the first.

The choice of relative co-occurrence frequency as the measure of concept co-occurrence was influenced by two factors. This measure is much less tightly binding than a two-way contingency measure, and this is desirable because we now want to measure incidental interactions between concepts, such as those between *dog* and *postman*. Second, it was felt that throwing away all the asymmetric attachment information, which is endemic to natural language (see, e.g., Nelson, McEvoy, & Pointer, 2003), was not justified. In very many instances in which word or document similarity measures are required, including many analyses of results from HAL and LSA, the vector cosine measure is used. However, vector cosine is a symmetric measure. Neither is it equivalent to symmetrizing the matrix by pairwise averaging of link values. Finally, it is noted that Nelson and colleagues have used the relative frequency of word free association to calculate their free association norms (e.g., Nelson et al., 2003).

As a result of this choice of a real-valued asymmetric measure, many analytical tools are not applicable. Multidimensional scaling (MDS), factor analysis, and the vast majority of social network and graph theory measures either do not incorporate both directions of an asymmetric link or do not deal with real-valued links. In addition, the Leximancer method seeks to discover implicit, indirect relationships between concepts. This facility can allow discovery of previously unknown relationships. As a result of these requirements, the techniques of complex systems simulations and emergent behavior were examined as approaches for calculating a concept map.

The Leximancer concept-mapping algorithm is based on a variant of the spring-force model for the many-body problem (e.g., Chalmers & Chitson, 1992). The method used in Leximancer simulates forces between the concepts. It is a highly dissipative iterative numerical model and comes under the definition of a complex network system. The map is an indicative visualization that presents concept frequency (brightness), total concept connectedness (hierarchical order of appearance), direct interconcept relative co-occurrence frequency (ray intensity), and total (direct and indirect) interconcept co-occurrence (proximity). The formation of groups of directly and indirectly related concepts displays emergent behavior—that is, exhibits information that was not apparent by inspection of the input concept co-occurrence matrix. For this reason, it is not appropriate to demand that the final concept map should explain as much of the initial variance as possible. If that were the case, concepts that were initially unrelated by the direct co-occurrence measure should be unrelated on the map, which in turn would not identify indirect relationships.

The emergent concept groups are normally referred to as *themes*. Identification of themes by the observer is greatly facilitated by employing the hierarchy of concept connectedness. Each highly connected concept is a parent of a thematic region and can be used to characterize that region. It is noted that the problem of matching structure between different concept networks is made much harder by the variation in names of equivalent concept nodes between the networks. The comparative maps of the Holy Bible in French and in English, which will be presented later, provide an extreme example of this; none of the concept names are identical.

It must be emphasized that as with most algorithms, there are parameters that must be set, and these choices will be expected to influence the results. The most critical parameter is the length of the text segment. This is selected by choosing the maximum number of sentences contained in each segment and whether or not the segment can cross a paragraph boundary. This setting will affect both the semantic and the relational extraction phases. The nature of this effect will be examined below in the Stability section.[2]

FORMULATION OF EVALUATION CRITERIA

The success or otherwise of a content-analytic method is often referred to as *validity*. The analysis of validity presented here will generally follow the typology presented by Krippendorff (2004, p. 319, Figure 13.1). This typology offers a promising framework for standardizing validation efforts not only in text content analysis, but also in knowledge discovery generally.

For reasons discussed in the sections below, we will combine some of Krippendorff's validation types: (1) face validity; (2) stability (including sampling validity of members); (3) reproducibility (including sampling validity of representatives and predictive validity), which also covers structural validity in the case of concept network comparisons; (4) correlative validity (also including semantic validity); and (5) functional validity. These categories will now be expanded upon.

Face Validity

Face validity is a measure of how plausible or defensible the Leximancer algorithms are. In more concrete terms, are the algorithms grounded in established practice?

The foundations of most of the Leximancer algorithms have been published elsewhere (Smith, 2000a, 2000b, 2003) and have been discussed above in the Foundations of the Technique section. In summary, Leximancer is founded in the observations of corpus linguistics, computational linguistics, and psycholinguistics that word co-occurrence statistics in natural language are a rich source of information that correlates with certain aspects of human language learning, comprehension, and performance. To achieve the design goals, two stages of co-occurrence information extraction are employed, using different statistical relevancy measures and different nonlinear clustering algorithms. The relevancy measures are grounded in Bayesian decision theory and word free association norms. The clustering algorithms are derived from computational linguistics and complex network simulation.

Discussion of the sensitivity of the algorithms also tends to confirm expectations about how information is structured within a text, as will be seen in the next section.

Stability

Stability is a measure of whether the same data produce the same results. Coder reliability is not an issue for Leximancer; text segments are always coded in the same way, given the same parameter settings. In addition, Leximancer can normally analyze the whole data set of interest. However, stability of the arrangement of the final concept cluster map must be tested, since that component is calculated using a stochastic algorithm. In fact, we have found that this stability is a good measure of contextual confusion in the data. Note that any instability in the map does not affect either the frequency statistics or the centrality rankings of concepts.

Of course, changing the parameter settings may change the results. There are many ways to formulate an analysis with Leximancer, depending on what sort of questions are being asked. The proportion of automatically selected concepts that are based on proper names can be controlled. The total number of automatically selected concepts can be increased in order to extract more specific concepts from the lower end of the ranking. Concepts can be hand-seeded and profiled to generate customized views. However, these formulations are generally determined by a deliberate analysis strategy that can be justified. More arbitrarily, the removal of stop-words (functional words with low semantic content, such as *and*, *is*, or *but*) may have an influence on the analysis. The presence of very frequent stop-words in the text can result in overgeneralizing in thesaurus learning. This has the most effect on other high-frequency words, since their occurrence statistics can align with very common stop-words, and so these frequent semantic words can abstract to the stop-word.

To demonstrate this sensitivity, we will analyze a body of text that can be obtained by other researchers for comparison. The text we have chosen is *The Personal Memoirs of U. S. Grant*, by Ulysses S. Grant (1885). This can be obtained from the Project Gutenberg repository. The complete text has been processed by Leximancer, using automatic concept selection but asking for the 100 top-ranked concepts.

Under the default Leximancer operating parameters, the map in Figure 1 was obtained. The thematic groups have been circled and labeled by hand.

In many of the example map figures that follow, large bold labels and circular groupings have been added by hand to clarify the structure. Some interpretation is required to place the borders and choose the names of these groupings, but the name of the group is commonly the name of a parent concept within the group. Recall that the measure of concept connectedness adds a hierarchy to the network. In addition, every attempt has been made to make the names of the constituent concepts legible for inspection by the reader. It should also be noted that rotation of some concept maps has been imposed in order to clarify the comparison of structure. Specifically, the structure of the map is correlated with the semantic relational structure, whereas rotational and reflective orientation is correlated with the relative emphasis given in the data to parts of the semantic structure.

After the stop-list was overridden so that the most frequent stop-word in this text, which was *the*, was preserved in the data, it was observed that several of the most frequent concepts (namely, *troops*, *time*, *fact*, *large*, *Union*, *Confederate*, *roads*, *morning*, *people*, and *North*) were subsumed under the concept *the*. The reason for this is

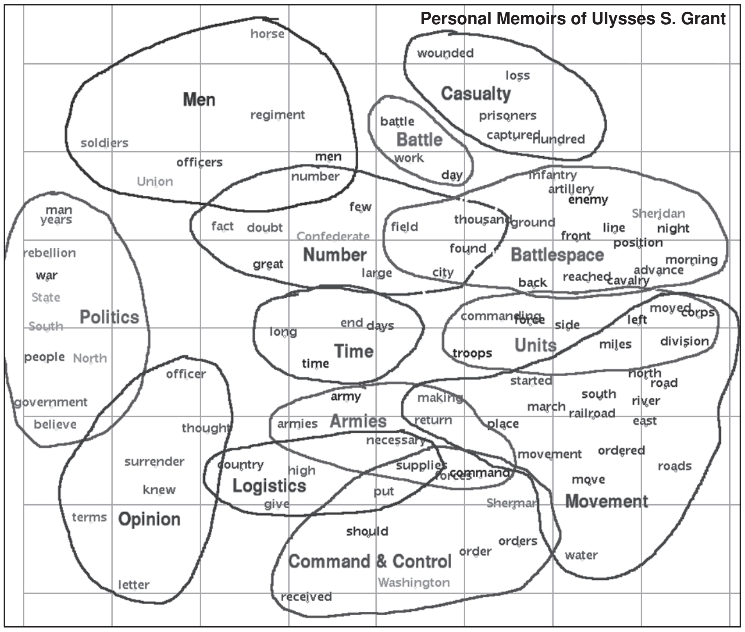

Figure 1. Standard map of Grant's *Memoirs* (top 100 concepts), with manual annotation.

that these words co-occur so often with the definite article and so infrequently appear apart that they are subsumed under that parent concept.

The addition or removal of a highly connected concept, such as a frequent stop-word, can strongly affect the concept map structure. This is analogous to removing or adding a hub in a network. This sensitivity disappears rapidly as the connectivity of the concept in question drops, so that the removal of one of the semantically meaningful content words does not normally perturb the network structure excessively. Of course, this does depend on the properties of the text; a document map that is dominated by one very central concept would, of course, be changed by the removal of that central concept. Figure 2 shows the map of the Grant text with the concept *the* added. Since this concept is extremely frequent and connected to most other concepts, the map is dramatically different.

Another setting that can affect the results of the analysis is the number of sentences per text segment. This can easily be varied with Leximancer to perform sensitivity analysis. The lists of words in Table 1 are the top-ranked words from trained thesaurus entries for the concept *city*, and those lists in Table 2 are for the concept *water*. Again, these were extracted from the Grant text. A thesaurus was extracted for a text segment of three sentences (with no paragraph crossing) and also for a segment of one sentence (with no paragraph crossing). The learning threshold was the same for both, at 1.8, and learning converged after six iterations for three-sentence segments and five iterations for one-sentence segments. The number following each list term is the weighting, which indicates how relevant each word is to the concept. The terms in double square brackets are proper names, which were automatically identified by Leximancer.[3] The lists have been truncated at around a relevancy score of 3, for brevity.

A comparison of the lists shows that when three-sentence segments are used, more words are learned, as might be expected. At the simplest level, if the algorithm is allowed to use a longer text segment around a seed word, a larger set of terms is likely to be measured. More interesting, in language usage, at least for English, the tendency is to avoid repeating a word in an adjacent sentence and to use a replacement term, such as a synonym (see, e.g., Beeferman et al., 1997). The relevancy scores of the words that are also seen in the one-sentence lists are, for the most part, higher in the three-sentence lists. Occasionally, a word will drop in relevancy when one goes from one- to three-sentence segments. It belongs to the character of the one-sentence lists to tend toward syntagmatic (within-sentence) associates of the seed word. For example, *yellow* is associated with *city* by means of *yellow fever* and also by the fact that Vicksburg was built on yellow clay. However, the iterative process of training certainly offers the opportunity for paradigmatic associates to be learned even with one-sentence segments. Most of the syntagmatic associates remain when three-sentence segments are employed, but other paradigmatic and indirect associates are also found, such as *plaza* and *house-tops*. The association of *city* with *flames* is domain specific. If thesaurus training is performed over text segments longer than three sentences or paragraph boundaries are ignored, the resulting thesaurus entries contain more noise terms, and convergence of the algorithm is slower.

Performing concept classification back on the text is also affected by the size of text segment. It can be seen that since relationships between concepts are measured

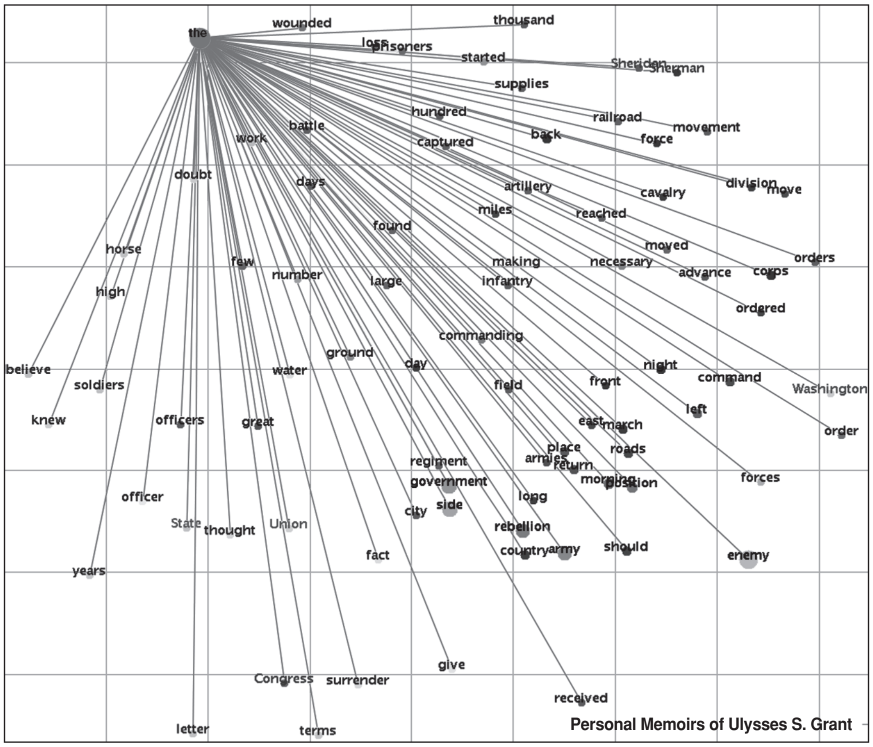

Figure 2. Map of Grant's *Memoirs* (top 100 concepts) plus the concept *the*.

by their co-occurrence within text segments, a shorter text segment would mean that a concept would tend to be related to fewer other concepts.

For example, the top relationships for the concept *water* for each of the three-sentence and one-sentence text segments are as follows.

Top concepts related to *water*:

> Three-sentence segments: river, 30.7%; troops, 21.2%; enemy, 18.8%; time, 17.3%; high, 16.5%; plus 85 more.
>
> One-sentence segments: river, 23.3%; troops, 15.3%; high, 13.3%; time, 12.6%; back, 9.3%; plus 66 more.

It can be seen that for one-sentence segments, there are fewer related concepts and that the relative co-occurrence frequency is less for matching concepts. Also, a strong relationship with the concept *enemy* has been reduced to a very weak one, which was actually measured at 4.6% for one-sentence segments. This results in some discourse relationships being neglected. Another effect of this reduction in concept co-occurrence is that the number of text segments that are indexed by only one concept rises. In this instance, the fraction of text segments classified with *water* but no other concept rises from 2.3% to 10.6% when one goes from three-sentence to one-sentence segments. This has the effect of reducing the completeness and effectiveness of the concept index into the text.

It could be expected that this parameter change would also alter the pattern of the concept map. However, this is not generally true for text segment sizes of less than four sentences. For example, Figure 3 shows the concept map for one-sentence text segments, and it is very similar to the map for three sentences (Figure 1). This is partially due to the ability of the mapping algorithm to "fill in" indirect relationships. However, if the text segment size is increased beyond four sentences, the relational interconnectivity rises strongly, as does the relational noise, because the further apart the conceptual evidences in the text, the less likely they are to be related. As a result, the concept map tends to become less differentiated and more unstable. It is occasionally useful when analyzing dialogue to employ longer text segments so that they can cross interspeaker boundaries, which, in turn, can take some relational account of consecutive speaker interaction. In this situation and in several other important situations in which short paragraphs are employed, it can be important to allow the text segments for the concept classification phase to cross paragraph boundaries. Examples of text styles where this may be appropriate are dialogue and novels that consist mainly of dialogue, electronic mail, press releases, and verse.[4]

Reproducibility

Reproducibility includes sampling validity of representatives and predictive validity (Krippendorff, 2004). We conflate the two in the following sense. Consider a theoretical population of data sets with known similar meanings (i.e., semantics) between all the constituent sets. Parts of the data set can be from different times or sources, with different surface representations (i.e., vocabulary, style,

Table 1
Truncated Thesaurus Lists for the Concept *City* Extracted From the Grant *Memoirs* Employing One- and Three-Sentence Text Segments

Three-Sentence Segments	One-Sentence Segments
city 8.5887	city 8.8482
plaza 4.773	gates 3.9236
[[black_fort]] 4.5742	villages 3.7299
[[walnut_springs]] 4.4533	yellow 3.7299
flames 4.4533	aqueducts 3.7299
house-tops 4.4533	[[san_juan_de_ulloa]] 3.7299
swept 4.3111	gun-shot 3.4584
arches 4.3111	
gates 4.138	
sand-bags 4.138	
villages 3.9161	
challenge 3.9161	
conducting 3.9161	
angles 3.9161	
extinguish 3.9161	
quit 3.9161	
yellow 3.9161	
denied 3.9161	
citadel 3.605	
windows 3.605	
moderate 3.605	
aqueducts 3.605	
[[west_forrest]] 3.605	
gun-shot 3.605	
outskirts 3.605	
expedient 3.605	
square 3.605	
[[bishop_s_palace]] 3.605	
[[san_juan_de_ulloa]] 3.605	
[[lake_chalco]] 3.605	
[[deep]] 3.605	
volley 3.605	

and even language). To show that Leximancer-induced patterns are reproducible, we need to show that similar patterns are found from each constituent data set. As a corollary, we also need to show that data sets with known different meanings result in different Leximancer patterns. Note that since we are measuring reproducibility by looking at similarity in concept network patterns, we are simultaneously testing structural validity, in Krippendorff's terminology.

Reproducibility of the thesaurus classifiers. Once the thesaurus network has been learned, the classification procedure is quite simple. This process is similar in conception to manual coding, or sense tagging, as performed in content analysis (Weber, 1990), where trained human coders attach conceptual tags to groups of sentences with reference to a code book.

The steps involved in the Leximancer classification algorithm are as follows.

1. Process the text sequentially in blocks of *n* sentences. It has been found that *n* should be similar to the number of sentences per block used during training, since this is the average length of text constrained by one instance of a concept.

2. Look up all the words from the text block in the thesaurus network and add their weightings in each concept represented.

3. Threshold the results to select the relevant concepts, with likelihood weightings.

Supervised training benchmarking. Within the supervised document classification community, a standard benchmark for testing the reproducibility of a classifier is to employ a set of human classified documents. This set is then split into a training set and a test set.

However, the standard Leximancer thesaurus learning algorithm is not operated as a supervised learner. Instead, it is designed to extend an incomplete definition—that is, a set of seed words. Nevertheless, a single iteration of the learning algorithm is equivalent to a naive Bayes supervised learner, so we undertook a standard benchmark of this, using the Reuters-21578 text categorization test collection, with the ModApte split (Apté, Damerau, & Weiss, 1994). This is a set of human classified media reports, and several standard training/test splits have been defined by the categorization community.

Unfortunately, the classification tags in Reuters-21578 are placed at the beginning of each article, because each article is generally fairly short and Reuters only had a need for whole-document classification. It takes much more effort and concentration for humans to tag reliably at a higher resolution, and the markup method would need to be changed to accommodate this.

If the human classification tags were placed within the contextual sentences that actually triggered the classification, Leximancer's term co-occurrence algorithm could then extract a lexical classifier, using the optimal text segment size. Essentially, the relevant contexts of the tags would be learned. It would be satisfactory if the tags were applied to every three sentences. Alternatively, the code book used by the Reuters classifiers would need to be available for use as a manual seed set, using the normal concept bootstrapping algorithm. As it stands, this benchmark does not allow the classifier to perform anywhere near optimally, particularly for more general categories such as *energy*, *housing*, *income*, *money-fx*, *instal-debt*, or *retail*. The relevant seed words can be guessed, but this is not appropriate for a benchmark. Given these caveats, it is still of some interest to reproduce the whole-document supervised classification benchmark by modifying the operation of Leximancer.

The Reuters-21578 ModApte standard benchmark for supervised training produced the following results with the Leximancer naive Bayes classifier after its operation was modified as alluded to above. In order to interpret the results, some definitions are required. *Precision* and *recall* are standard evaluation measures for supervised classification: Precision is the fraction of automatically classified documents that match the manual classification; recall is the fraction of manually classified documents that are successfully classified automatically. For evaluation with multiple classification categories, the results are often aggregated, using micro-averaging. This means that each classification category is considered in turn, and that, within each category, each document is simply designated as a hit (if automatic classification matches manual classification), a miss (if the manual classification has no

Table 2
Truncated Thesaurus Lists for the Concept *Water* Extracted From the Grant *Memoirs* Employing One- and Three-Sentence Text Segments

Three-Sentence Segments	One-Sentence Segments
water 7.871	water 7.8038
surface 4.2155	tide 3.7026
levees 4.0818	patient 3.5184
tide 3.9196	wade 3.5184
depth 3.9196	washed 3.5184
channels 3.7121	replenish 3.5184
angles 3.7121	oven 3.5184
pits 3.7121	ships 3.2607
patient 3.7121	undertaking 3.2607
bends 3.7121	regulated 3.2607
begins 3.7121	rainfall 3.2607
washed 3.7121	abound 3.2607
[[carthage]] 3.7121	waists 3.2607
oven 3.7121	bathing 3.2607
aqueducts 3.422	stones 3.2607
bucket 3.422	
recede 3.422	
undertaking 3.422	
rainfall 3.422	
bathing 3.422	
dam 3.422	
ships 3.422	
regulated 3.422	
abound 3.422	
stones 3.422	

matching automatic classification), or a false hit (if the automatic classification has no matching manual classification). The numbers of hits, misses, and false hits are added across all categories, and the final micro-averaged precision and recall are calculated by the following equations: precision = hits/(hits + false hits) and recall = hits/(hits + misses). Furthermore, there is almost always at least one classification threshold parameter that can be tuned, and adjusting this parameter almost always trades off precision against recall. Rigorously, the resulting precision versus recall curve can be plotted, but a simple figure of merit is the break-even point, where precision equals recall. Finally, some of the categories used in the Reuters-21578 set are much less frequent than others. This has implications for the relative effectiveness with which the different classifiers are learned. For this reason, precision and recall results are often calculated across the 10 most frequent categories, as well as across all of them (see, e.g., Dumais et al., 1998).

For the Leximancer naive Bayes classifier and the Reuters-21578 ModApte split, the micro-averaged break-even precision and recall were found to be 81.2% for the 10 most frequent classifications and 75.5% across all classifications. These results agree well with the best results for naive Bayes classifiers quoted by Dumais et al. (1998).

Thesaurus discovery benchmarking. A more appropriate cross-classification evaluation for the Leximancer thesaurus builder was then sought. It is of interest to know how useful a given thesaurus is at classifying text that is from the same domain but is different from the training set. This evaluation will test the limitations of reuse of the thesaurus for classifying new text.

To examine some aspects of thesaurus reproducibility, a 13.6-Mb set of data from the Internet news group sci .environment was obtained and split into two. The data set was ordered by article number and split into two contiguous halves. In other words, the parts were not interleaved in the original, and the second set of articles came after the first in terms of time. It is to be expected that the content of news group discussions should evolve over time.

Each half was used to learn a thesaurus entry for the concept *energy*, using just the seed word *energy*. The second half was then classified with its own thesaurus to produce a classification called *self*. The second half was then cross-classified, using the thesaurus learned from the first half, to produce a classification called *cross*. These two classifications of the same text, *self* and *cross*, were then compared. These classifications were carried out under the most standard operational settings: three sentences per block, a training threshold of 1.8 relevancy units with no paragraph crossing, and a classification threshold of 7.5 relevancy units total sum per block with no paragraph crossing. These are the default Leximancer parameter settings.

When classification instances of the concept *energy* were considered, it was found that when one went from self- to cross-classification, 3.6% of the classifications by weight (6.0% by number) disappeared, and 4.0% new classifications by weight (8.6% by number) were created. By number, 94% of the classifications were common to both, but their total weighting dropped by 35% when one went from self- to cross-classification.

The fraction of blocks that were classified as *energy* but did not contain the keyword *energy* was 7% for self-classification and 10% for cross-classification.

This performance is satisfactory, but it indicates that self-classification should be used where possible. The fact that 94% of the text blocks were allocated to the same class with both methods is pleasing. It is apparent from these results that the vocabulary shifted to some extent between the two sets.

As a further comparison, the same procedure was followed again, but this time the learning process was allowed to cross paragraph boundaries, and the learning threshold was increased to 2.0 relevancy units to compensate. It would be expected that allowing text segments to cross paragraph boundaries would add noise to the thesaurus and the relational network, since authors are more likely to change topic at a new paragraph.

In this case, it was found that, when one went from self- to cross-classification, 12.0% of the classifications by weight (15.7% by number) disappeared, and 10.4% new classifications by weight (5.2% by number) were created. By number, 84% of the classifications were common to both, but their total weighting dropped by 30% when one went from self- to cross-classification.

The fraction of blocks that were classified as *energy* but did not contain the keyword *energy* was 16% for self-classification and 12% for cross-classification.

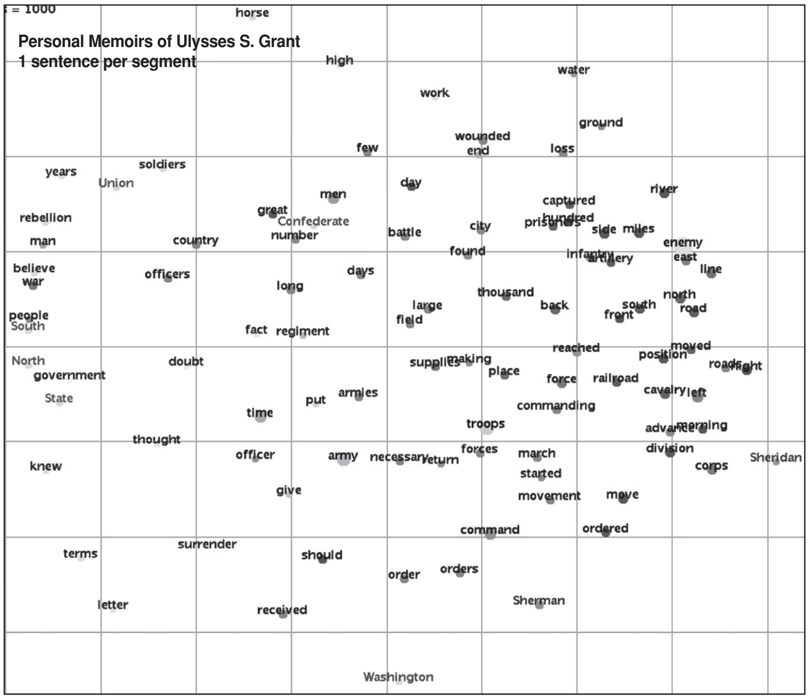

Figure 3. Map of Grant's *Memoirs* (top 100 concepts) with one-sentence segments.

The total quantity of self-classifications increased by 8% weight (11% by number) when the learning process was changed from no paragraph crossing to paragraph crossing.

These results indicate that allowing the learning process to use blocks of text crossing paragraph boundaries creates a more diverse but also a noisier thesaurus entry. In our experience, paragraph crossing is recommended only when the text uses short paragraphs, such as e-mail or dialogue.

Impact of sampling during learning on classification. It is important to know what impact the use of sampling during learning has on the performance of the thesaurus classifier. Sampling, as employed here, involves a uniform process of using only every *n*th text segment for learning the thesaurus. The reason for wanting to do this is to accelerate the iterative learning algorithm.

To examine this, the same 13.6-Mb set of data from the Internet news group sci.environment was used as that in the previous section.

The full data set was used to learn a thesaurus entry for the concept *energy*, using just the seed word *energy*. In one case, a classification of the data was produced using no sampling. In the other case, a classification was produced using a thesaurus trained with a sampling of two, or every second text block. These two classifications of the same text, *full* and *sampled*, were then compared. These classifications were carried out under the most standard operational settings: three sentences per block, a training threshold of 1.8 relevancy units with no paragraph crossing, and a classification threshold of 7.5 relevancy units total sum per block with no paragraph crossing.

It was found that when one went from full to sampled classification, 1.8% of the classifications by weight (3.0% by number) disappeared, and 1.6% new classifications by weight (2.7% by number) were created. By number, 97% of the classifications were common to both, and their total weighting dropped by 7.2% when one went from full to sampled learning.

These results are pleasing and show that sampling is a very viable way of accelerating the learning process. Contrasting these results with the comparable results from the previous section shows that sampling is much more reliable than learning and classifying on disjoint contiguous data sets.

Impact of learning stability on classification. It is interesting to evaluate the stability of the convergence point for the training algorithm and its effect on classification. The iterative concept-learning algorithm employed essentially searches for a local attractor in the lexical co-occurrence space. If the starting point of the trajectory is significantly perturbed away from the attractor, one can establish whether the learning system returns to the attractor or diverges away. If the system is overly sensitive to its starting location or, worse, exhibits chaotic trajectories, this will limit the reproducibility of identified conceptual patterns.

To examine this, the same 13.6-Mb set of data from the Internet news group sci.environment was used as that in the previous sections.

The full data set was used to learn a thesaurus entry for the concept *energy*, using just the seed word *energy*. This was then used to generate a classification called *full*. The final training network was then taken from this, and the highest ranked term was removed from the concept *energy*. In this case, the word was *energy*—the initial seed word, in fact—with a relevance of 8.38 relevancy units. The learning algorithm was then restarted using this perturbed network as the starting point. This system converged after eight iterations, and now the highest ranked term in the concept *energy* was again the word *energy*, with a relevance of 5.70. This new thesaurus was then used to produce a classification called *shifted*. These two classifications of the same text, *full* and *shifted*, were then compared. These classifications were carried out under the most standard operational settings: three sentences per block, a training threshold of 1.8 relevancy units with no paragraph crossing, and a classification threshold of 7.5 relevancy units total sum per block with no paragraph crossing.

It was found that when one went from full to shifted classification, 6.7% of the classifications by weight (14.5% by number) disappeared, and 0.7% new classifications by weight (0.8% by number) were created. By number, 85.5% of the classifications were common to both, and their total weighting dropped by 14.5% when one went from full to shifted learning.

Since the initial seed word *energy* reappeared as the highest ranked term, it appears that the learning algorithm is quite stable once converged. It was noted that the word *energy* had a weighting above the classification threshold (7.5) in the first case, but below in the second. This means that for the shifted classification, the word *energy* by itself is insufficient to trigger a classification. And yet, 85.5% of the classification instances remained in common, and 90.8% of the shifted classification instances contained the word *energy*.

This indicates that the concept of *energy*, as learned by this algorithm, is a stable maximum in the concept space. Also, and more importantly, this demonstrates that the cumulative support of all the words in the thesaurus is important to classification weighting. It is not just a keyword search engine.

Reproducibility of the concept maps. We now wish to examine the reproducibility of the Leximancer concept maps over different data sets. The aim of these analyses is to establish whether text sets with similar semantics produce similar concept maps and, conversely, whether text sets with different semantics produce different concept maps. The issue of how we "know" that the text sets have similar or different semantics is important; this problem leads to some overlap with the measure of correlative validity (see below). The overlap occurs because we must have some other method for establishing this similarity. The best that can be said is that we choose data sets where the similarity or difference is fairly obvious, usually due to the circumstances of creation of the data.

1. The first example shows maps of two translations of the Bible, the English King James version and the French Louis Segond version (see Figure 4). For these maps, automatic concept detection was used, and the top 100 concepts for each were asked for. Due to the size of these data sets and their high relational interconnectedness, the low-resolution indexing setting was set to three text blocks per low-resolution bucket.[5]

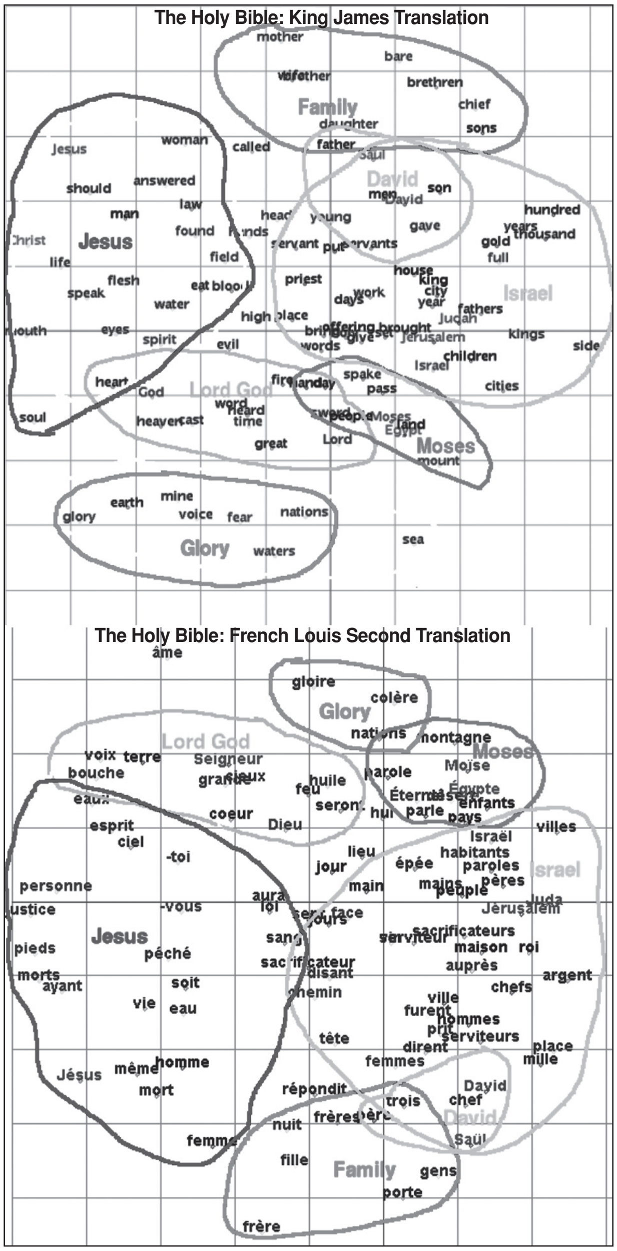

Figure 4. Concept maps of English and French translations of the Holy Bible.

Because these were translations of the same source, it was felt that the semantics should be similar. If one allows for a reflection about a horizontal axis, these maps do indeed show a similar structure.

2. Leximancer concept maps of the rules of baseball (Major League Baseball, 1999) and cricket (Marylebone Cricket Club, 2003) were compared (see Figure 5). For these maps, automatic concept selection was employed, and the top 80 concepts were asked for. It should be noted that for these maps, it was necessary to tune the thesaurus-

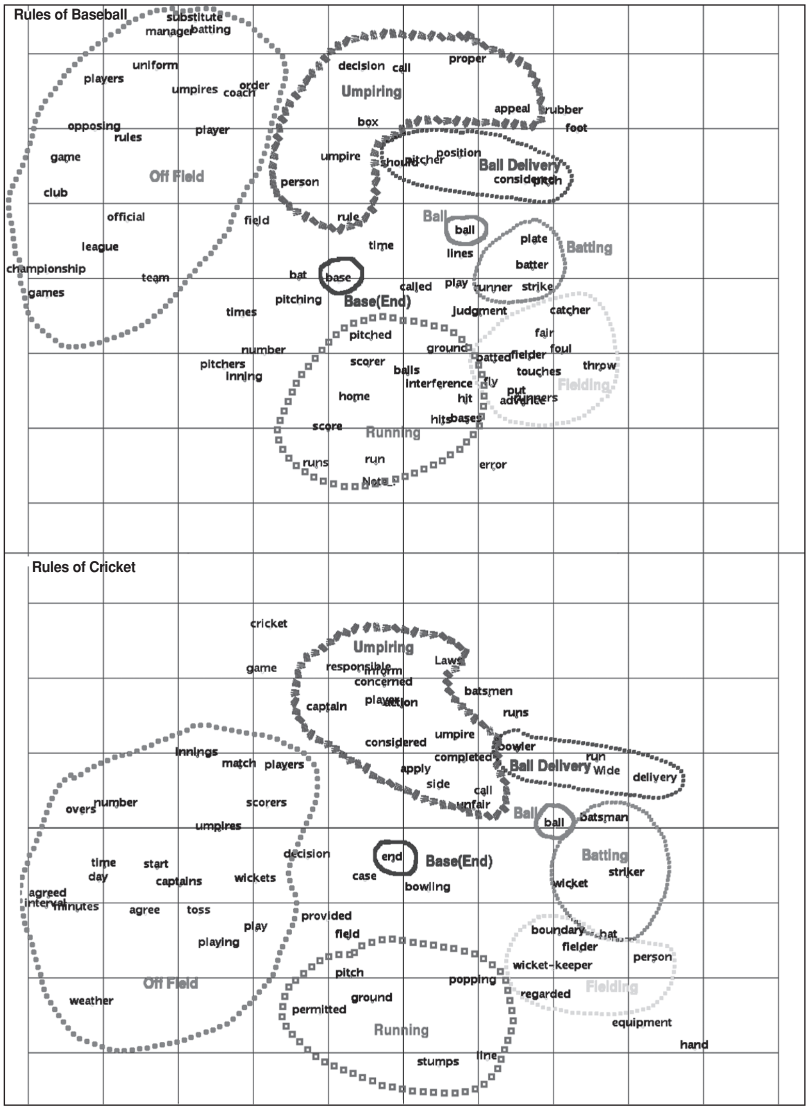

Figure 5. Concept maps of the rules of baseball and cricket.

learning threshold so that this learning converged after around the same number of iterations for each map. The number of iterations was six or seven.

As bat-and-ball games with players in similar roles, it was felt that the rules would show similar semantics. The origins of both games are uncertain; however, it is likely that both derive from late medieval French and English village games, with a game called *stool ball* being a possible common ancestor (Bahr & Johnston, 1992, Vol. 3, p. 660). This is not to say that the strategy and tactics are similar between baseball and cricket, but the rules do not normally work at that level of meaning. Comparison reveals similar arrangements in terms of the roles of key concepts in the structure of the games; alignment is seen between pitcher and bowler, batter and batsman, fielder and fielder, base and end, wicket and plate, wicket-keeper and catcher, match and game, ground and field, and so forth.

3. It is also important to show that maps of text data sets that are known not to have similar semantics are, in fact, dissimilar. The rules of American football (North American Football League, 2003) and rugby union (International Rugby Board, 2003) were compared (see Figure 6). American football originated directly from rugby union but underwent significant modifications in the 20th century (Bahr & Johnston, 1992, Vol. 10, p. 163). We anticipated showing that maps of these sports are moderately similar to each other but very different from the maps of cricket and baseball. For these maps, automatic concept selection was employed, and the top 80 concepts were asked for. If one allows for reflection around vertical axis, these maps show a structure that is moderately similar between the two and that is quite different from the baseball–cricket structure.

4. Leximancer concept maps (Figure 7) were created from the book *On War* by Carl von Clausewitz (1832/1873), and

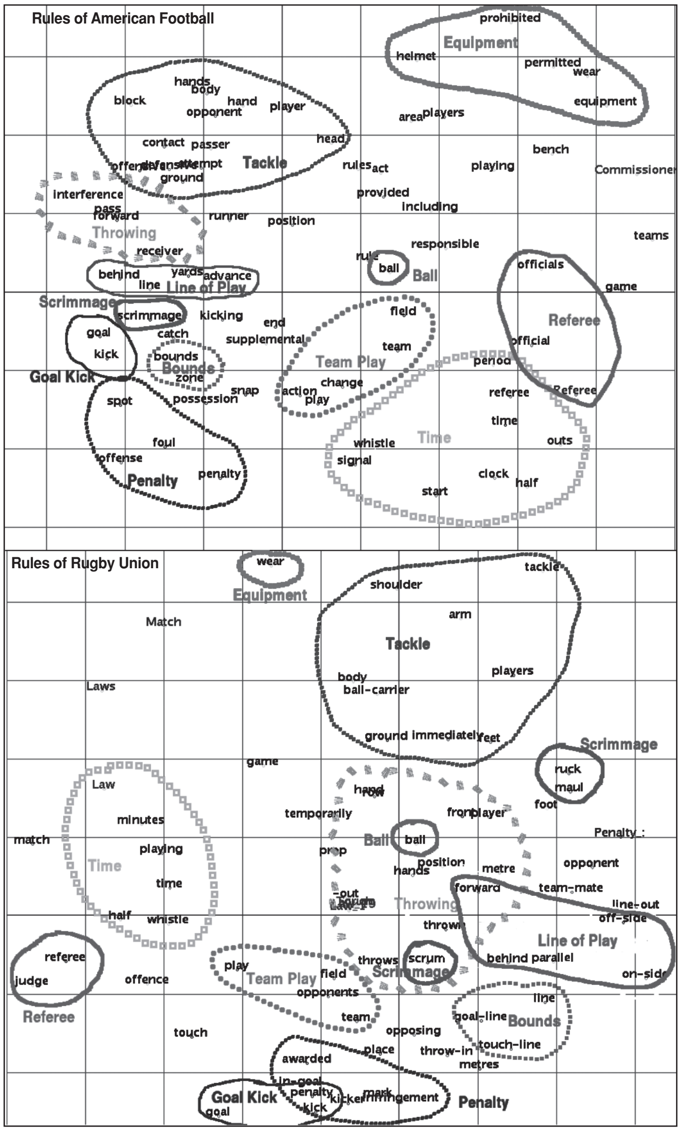

Figure 6. Concept maps of the rules of American football and rugby union.

from the Capstone doctrinal publications of the U.S. Marine Corps (1997). For these maps, automatic concept selection was employed, and the top 80 concepts were asked for. Striking similarities in the thematic structure can be seen, allowing for reflection around the vertical axis. Subsequent to this analysis, a private communication (Bassford, personal communication, March 21, 2004) revealed that, in fact, the author of much of the USMC Capstone doctrine is one of the leading authorities on *On War* (e.g., Bassford, 1994) and was not surprised by the similarity.

5. Leximancer concept maps were made from newspaper articles containing some mention of Iraq, from 5 weeks before to 3 weeks after the U.S.-led invasion of that country in March and April 2004 (see Figures 8 and 9). The data were obtained by using a text retrieval engine to find all articles containing the word *iraq* printed during the relevant weeks in *The Australian*, the major Australian national daily. The technique for map construction is to hand-seed the concept *Iraq* with the seed terms (*Iraq*, *iraq*, *Iraqi*, *iraqi*, *Iraqis*, *iraqis*) and ask the thesaurus learner to learn this concept and then to discover 100 associated concepts that will profile the concept of *Iraq* in the data. This method avoids totally unrelated content in the newspaper articles. A completely separate map was constructed for each week of data, and the set of eight maps were compared in an attempt to see some predictive validity of pattern matching or change. There are similar patterns between the weeks leading up to the conflict. However, the maps change dramatically at the point of the invasion, as would be expected.

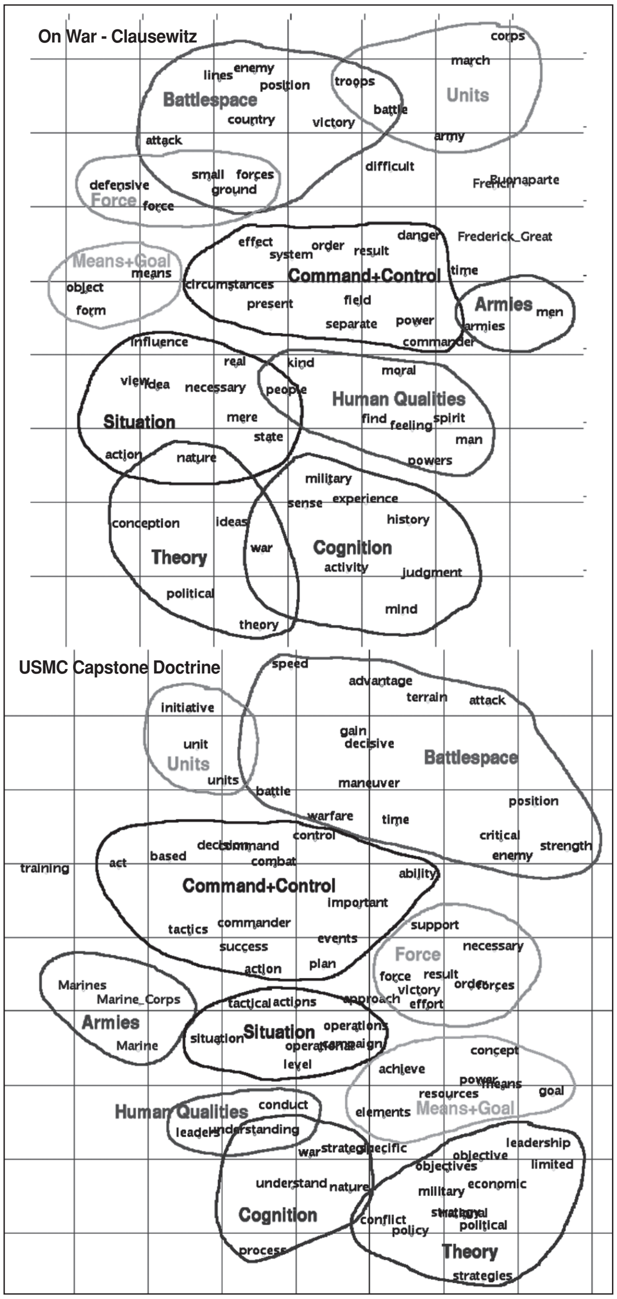

Figure 7. Concept maps of Clausewitz and the U.S. Marine Corps Capstone doctrine.

Correlative Validity

Correlative validity is established by showing that patterns measured by different methods that are considered valid correlate with patterns found by Leximancer. The corollary is that patterns known by other methods to be absent are not found by Leximancer. We believe that this includes semantic validity, for the reason that some other method must be used to establish the semantics of the situation. Hence, it is a type of correlative validity.

Correlative validity is interpreted here as strict comparison of the output of Leximancer with some valid independent analysis of the *same data*. This is viewed as distinct from evaluating the success of an analytic process that includes Leximancer analysis of *partial data* as part of the system. A key difference here is the exploitation of background knowledge by the analyst that is not available to Leximancer. Validation of the realistic analytic process as a whole, where the analyst only partially relies on Leximancer, will be discussed in the Functional Validity section.

There is some overlap between reproducibility and correlative validity. Whereas reproducibility focuses on comparisons between different Leximancer analyses and correlative validity focuses on comparing with other analysis methods, the overlap occurs because reproducibility still requires some other method to identify data sets that should be similar. This was referred to in the previous section.

It has been the policy since the creation of Leximancer to seek out data sets for which a domain expert can be identified to conduct informal parallel trials. The experts have been authors of the material, experienced analysts who have analyzed the material, or researchers who are very familiar with the content and the domain. Very many of these informal evaluations have been performed over the last 3 years, and feedback from the experts has guided development of the system. Although most of this evidence is informal and anecdotal, the work published on maritime accident reports (Grech, Horberry, & Smith, 2002) includes assessment of correlative validity by two domain experts both on the exploratory Leximancer maps and on a set of predefined variables.

We intend to develop a more rigorous method to enable publication of the results on correlative validity. However, there are methodological problems to address.

First, a suitable data set with known valid measurements must be found as a benchmark. The benchmark results must satisfy the same validity tests as those being examined here; it is not enough to assume that a human judgment is the "gold standard." The human is simply another text analysis machine for these purposes, with its own strengths, weaknesses, and biases. Such validated human analysis may be approximated for smaller text sets, but even then, most strict validation of human content analysis is conducted on confirmatory (or deductive) research, where there is a predefined set of concept variables. In exploratory mode, Leximancer induces the concept variables from the text. To compare this with human analysis, each person must also develop his or her own set of concept variables from the data, in the manner of grounded theory. Now, when larger text sets are considered, where Leximancer is believed to be of most utility, validated human inductive analysis is extremely hard to find.

Second, these valid measurements must be comparable with Leximancer results, without too much interpretation subverting the comparison. The issue here is that human analysts may possess background information, influences, and intentions that Leximancer does not possess. In addition, the human analysts may be subject to influences that

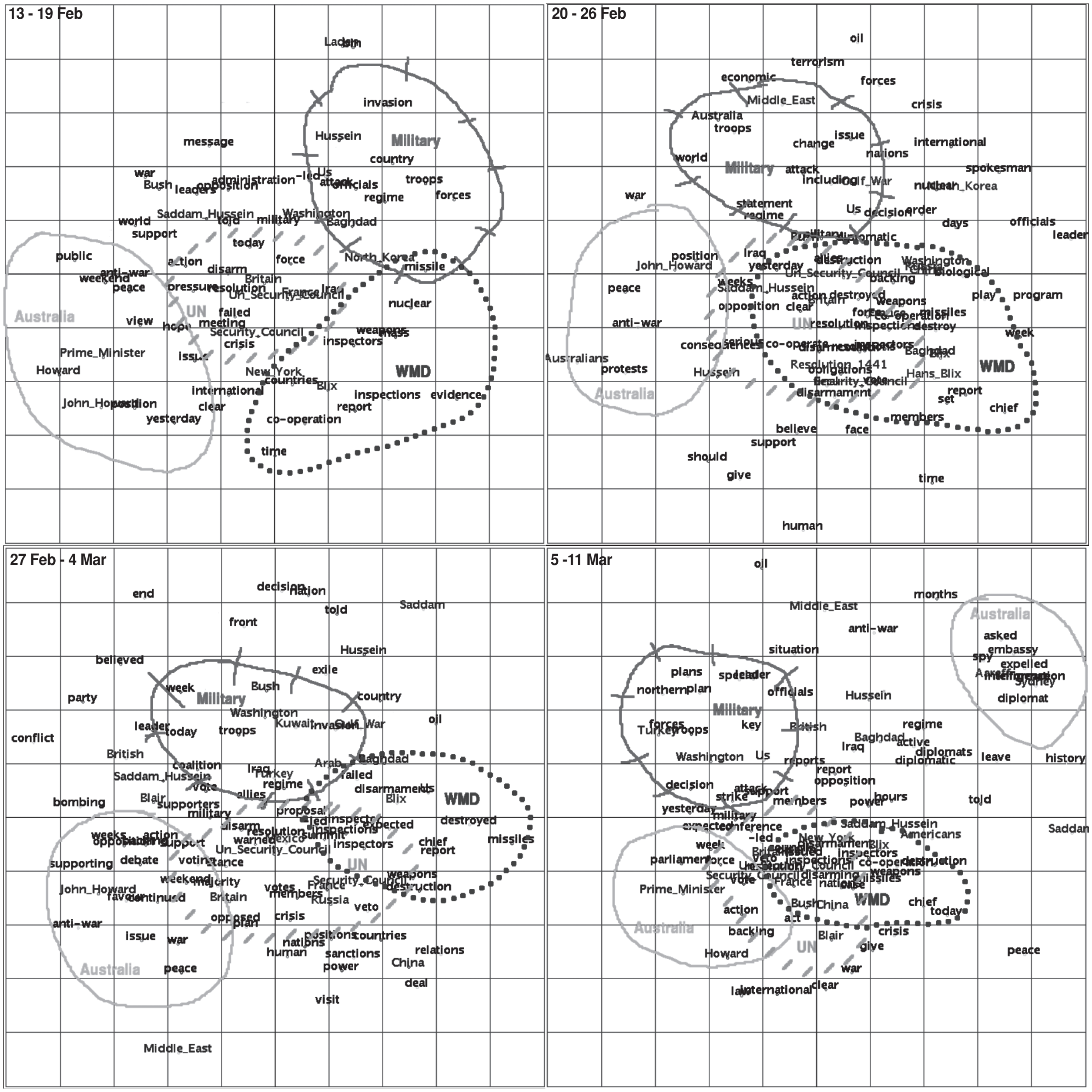

Figure 8. Concept maps of newspaper articles from weeks before invasion of Iraq.

they are unable to report (Nisbett & Wilson, 1977). As was mentioned above, the aim in this section is to set up an artificial situation in which the human analysts consider only data that are also available to Leximancer. The more realistic situation is considered under functional validity.

The use of background knowledge and accumulated experience by text analysts is endemic. For many real text data sets, the semantics contained within the text are partial. Successful interpretation relies on other semantics common to the author and the reader. This is particularly true in intelligence analysis, as is stated in Lefebvre (2004), who quotes from Katter, Montgomery, and Thompson (1979): "In other words, analysts never have a perfect information situation and 'information from memory provides the sole basis for hypothesizing relationships among data available for interpretation and for classifying various data as relevant, redundant, present, absent, or crucial for the interpretative task'" (p. 241). Somehow, either the human experts who create the benchmark must be isolated from other influences, including a lifetime of experience, or Leximancer must be provided with sufficient background material so that the results from both methods are based on the same inputs. Both options are difficult, but a semantic mapping system can allow a much larger contextual corpus

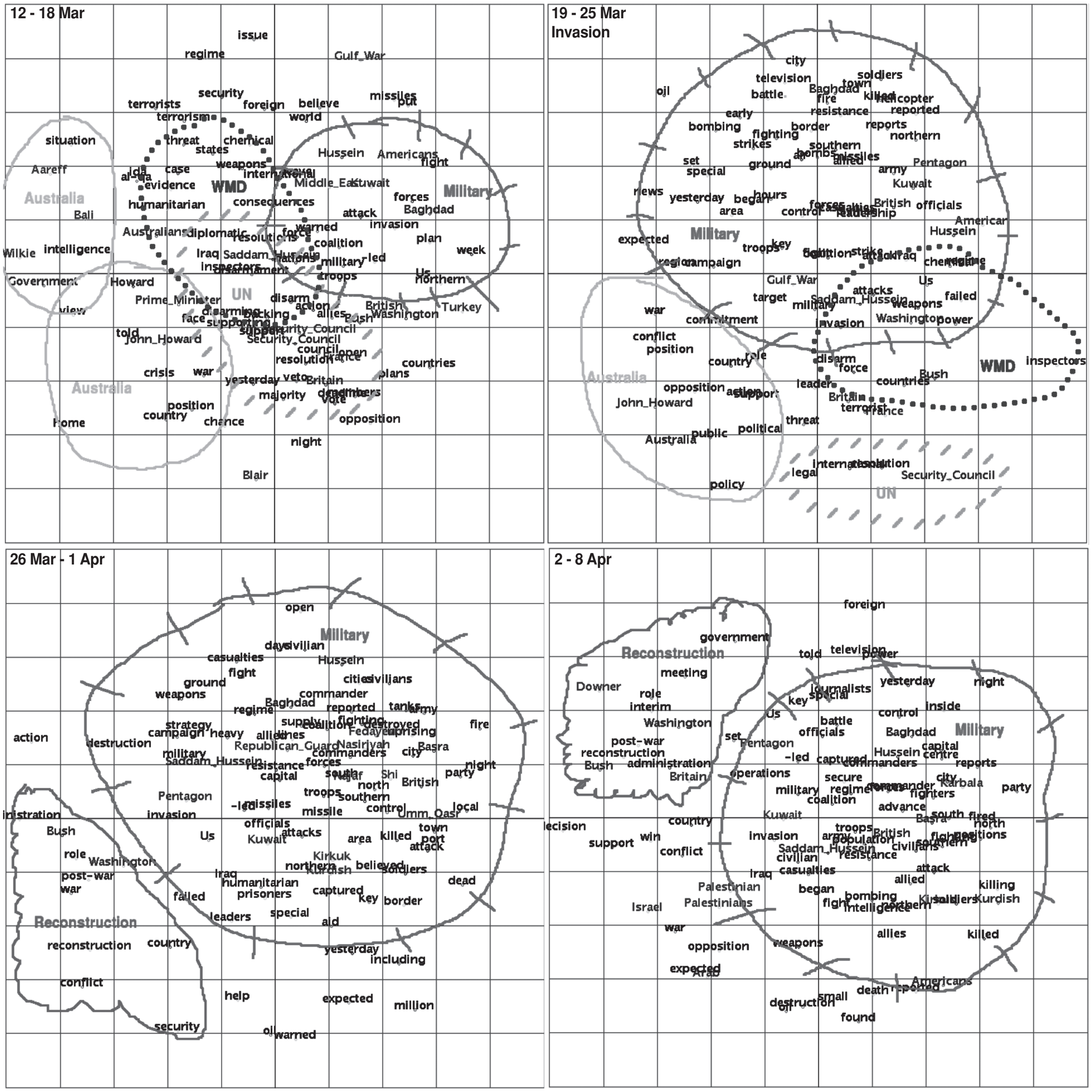

Figure 9. Concept maps of newspaper articles from weeks during invasion of Iraq.

of material to be mapped, which may make explicit some of the implicit background semantics. Alternatively, careful psychological experiment design with subjects who are unfamiliar with the text data in question may reduce the effect of background knowledge.

In summary, the rigorous evaluation of correlative validity for this system becomes one of bootstrapping: If all new techniques must replicate existing deficient means of exploratory text analysis, how can things ever improve?

Validity by inspection of thesaurus. It is normal practice to inspect the learned thesaurus weighted term sets to see whether they match expectations about word usage within each concept. Unfortunately, expectation can be wrong, and work must often be undertaken exploring the data to understand why terms in a set tend to travel together. Of course, domain expertise is valuable for this task.

Nevertheless, some thesaurus sets are compelling, such as the two given in the Appendix, which were learned from a set of maritime accident reports using the two seed words *engine* and *fire*. Terms in double square brackets are proper names that Leximancer has identified, those in single square brackets are tentative acronyms generated by Leximancer, and the numerical value indicates how relevant the term is to the concept in units of relevancy metric. The lists have been truncated.

Other thesaurus concepts are not so easy to understand without detailed inspection of the text data, but the algorithm is the same, and so some confidence can be developed in the method.

Functional Validity

The functional validity measure is, of course, the subject of most of the anecdotal feedback we have received, and of our professional services experiences in providing consulting text analysis services. However, if we define a functional goal of Leximancer mapping as the enhancement of learning and recall of text by people, this can be objectively measured. We are currently preparing to perform a set of psychological experiments on human subjects wherein study material is presented either in traditional paper form or as a Leximancer map. The subjects will then be tested at a later time for comprehension and recall.

CONCLUSION

In conclusion, we have addressed several forms of validity for Leximancer thesauri and concept maps, including face validity, stability (sampling of members), and reproducibility (including structural validity, sampling of representatives, and predictive validity). We have described research in progress for testing functional validity and have outlined some issues and progress in the area of correlative validity.

It will be interesting to see how much more information can be extracted from lexical co-occurrence, using combinations of different measurement formulae and nonlinear learning algorithms. Of course, much detailed grammatical information cannot be obtained using methods that discard word ordering within sentences, but it is apparent that there is an abundance of rich and complex information that *can* be extracted by means such as Leximancer. For rapid human appreciation of the information contained within nontrivial amounts of natural language, perhaps the challenge is to choose what level of detail to abstract.

REFERENCES

Apté, C., Damerau, F., & Weiss, S. M. (1994). Towards language independent automated learning of text categorization models. In W. B. Croft & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 23-30). New York: Springer.

Bahr, L. S., & Johnston, B. (Eds.) (1992). *Collier's encyclopedia*. New York: Macmillan Educational.

Bassford, C. (1994). *Clausewitz in English: The reception of Clausewitz in Britain and America, 1815–1945*. New York: Oxford University Press. (See also www.clausewitz.com)

Beeferman, D., Berger, A., & Lafferty, J. (1997). A model of lexical attraction and repulsion. In P. R. Cohen & W. Wahlster (Eds.), *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 373-380). Madrid: Association for Computational Linguistics.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language & Cognitive Processes*, **12**, 177-210.

Chalmers, M., & Chitson, P. (1992). Bead: Explorations in information visualisation. In N. J. Belkin, P. Ingwersen, & A. M. Pejtersen (Eds.), published as a special issue of sigir forum, *Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 330-337). New York: ACM Press.

Clausewitz, C. von (1873). *On war* (J. J. Graham, Ed. and Trans.). London: Trübner. (Original work published 1832) (This text obtained from: www.clausewitz.com)

Dumais, S. T., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In G. Gardarin, J. C. French, N. Pissinou, K. Makki, & L. Bougamin (Eds.), *CIKM '98*: *Proceedings of the 7th International Conference on Information and Knowledge Management* (pp. 148-155). New York: ACM Press.

Grant, U. S. (1885). *The personal memoirs of U. S. Grant*. Retrieved from Project Gutenberg, www.gutenberg.org, September 2004.

Grech, M. R., Horberry, T., & Smith, A. (2002). Human error in maritime operations: Analyses of accident reports using the leximancer tool. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*. Baltimore: Human Factors and Ergonomics Society.

International Rugby Board (2003). *The laws of the game of rugby union: 2003 edition*. Available at www.irb.com.

Katter, R. V., Montgomery, C. A., & Thompson, J. R. (1979). *Human processes in intelligence analysis: Phase I overview* (Research Rep. 1237). Woodland Hills, CA: Operating Systems, Inc.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Newbury Park, CA: Sage.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240.

Landauer, T., Foltz, P., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, **25**, 259-284.

Lefebvre, S. (2004). A look at intelligence analysis. *International Journal of Intelligence & Counterintelligence*, **17**, 231-264.

Major League Baseball (1999). *Official baseball rules: 1999 edition*. Available at www.amherst.edu/~baseball/rules.html.

Marylebone Cricket Club (2003). *The laws of cricket* (2000 Code 2nd edition). Available at www.lords.org.

Nelson, D., McEvoy, C. L., & Pointer, L. (2003). Spreading activation or spooky action at a distance? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 42-52.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, **84**, 231-259.

North American Football League (2003). *2003 playing rules of the NAFL*. Available at www.nafl.org.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Smith, A. E. (2000a). Machine learning of well-defined thesaurus concepts. In A.-H. Tan & P. S. Yu (Eds.), *Proceedings of the International Workshop on Text and Web Mining (PRICAI 2000)* (pp. 72-79). Melbourne.

Smith, A. E. (2000b). Machine mapping of document collections: The leximancer system. In *Proceedings of the Fifth Australasian Document Computing Symposium*. Sunshine Coast, Australia: DSTC.

Smith, A. E. (2003). Automatic extraction of semantic networks from text using Leximancer. In *HLT-NAACL 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Companion volume* (pp. Demo23-Demo24). Edmonton: ACL.

Sowa, J. F. (2000). *Knowledge representation: Logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks Cole.

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell.

U.S. Marine Corps (1997). *Marine corps doctrinal publications: Capstone publications* (MCDP Nos. 1, 1-1, 1-2, 1-3). Washington, DC: United States Government. (Available at www.doctrine.usmc.mil)

Weber, R. (1990). *Basic content analysis*. Newbury Park, CA: Sage.

Yarowsky, D. (1995). Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)* (pp. 189-196). Morristown, NJ: Association for Computational Linguistics.

NOTES

1. The process of concept extraction starts with preliminary definitions of categories, called *seed sets*, which can be either imposed by the investigator or selected by Leximancer.

2. The Leximancer Web site at www.leximancer.com includes the product manual, which describes the various parameters of the system.

3. Automatic multiword proper name extraction is performed simply from examination of character capitalization, when available. Of course, this will not be possible for some languages and for some transcribed speech. Performance is not perfect for words at the start of a sentence, but this effect is usually not statistically significant.

4. This evaluation work has been performed on at least moderately well-structured written language, partly since it is easier to find parallel semantic corpora in that scope. Further investigation is needed of informal and spoken language.

5. Low-resolution indexing is a type of discourse filter for removing concept tags that are not strongly represented over a set of consecutive text segments. It resembles a windowing noise filter algorithm and acts to remove incidental relationships. Concept classification weights are summed over multiple consecutive text segments, and any that do not make a certain threshold are deleted from all the segments. This has the important effect of enhancing relational signal-to-noise.

APPENDIX

Thesaurus Examples Extracted from Maritime Accident Reports

Engine:	Fire:
engine → 7.0192	fire → 8.5145
driving → 6.4023	fighting → 5.754
telegraph → 6.3363	detectors → 5.5341
compressor → 6.2133	brigade → 5.4339
crankcase → 5.6739	fight → 4.8983
extinguisher → 5.6482	extinguishing → 4.7603
room → 5.6299	extinguish → 4.5403
restarted → 5.5057	stair → 4.5403
[[mcr]] → 5.4409	retardant → 4.367
logger → 5.3706	[[nsw_fire_brigades]] → 4.2986
turbochargers → 5.3706	[n.f.b] → 4.2986
charger → 5.3706	[nfb] → 4.2986
[[electrician]] → 5.2522	erupted → 4.223
turbo → 5.2522	firemain → 4.223
[[ums]] → 5.2522	lethal → 4.223
lub → 5.2086	[[first_engineer]] → 4.223
restart → 5.1136	fighters → 4.1385
fans → 5.1136	plenums → 4.1385
governor → 5.0616	[[mvz]] → 4.1385
rung → 5.0616	outbreak → 4.0429
turbocharger → 5.0061	blaze → 4.0429
tachometer → 5.0061	igniting → 4.0429
lagging → 4.9464	signaling → 4.0429
transformer → 4.9464	[[total_endeavour]] → 4.0429
[[sulzer]] → 4.9464	lived → 3.9325
charged → 4.9464	monoxide → 3.9325
sump → 4.882	unburned → 3.9325
mist → 4.882	fought → 3.9325
injectors → 4.812	[[ccf]] → 3.9325
hunting → 4.812	hydrant → 3.9325
[[jcw]] → 4.812	fiercely → 3.802
starter → 4.7353	sengers → 3.802
rectifier → 4.7353	[[junior_first_mate]] → 3.802
jury → 4.6506	[[boundary]] → 3.802
fortunate → 4.6506	[j.f.m] → 3.802
strainers → 4.6506	imaging → 3.802
daywork → 4.6506	[jfm] → 3.802
builders → 4.6506	matic → 3.802
[[caterpillar_d399]] → 4.6506	[[korimul]] → 3.802
economizer → 4.5559	[[flames]] → 3.802
[[d399]] → 4.5559	extinguished → 3.7115
gearbox → 4.5559	smoldering → 3.6423
[[ecr]] → 4.5559	descending → 3.6423
cooler → 4.5559	sparks → 3.6423
ventilated → 4.5559	fusible → 3.6423
powered → 4.5474	addressable → 3.6423
lubricating → 4.5034	[[thevenard_island]] → 3.6423
atomised → 4.4487	avenues → 3.6423
[[b&w]] → 4.4487	marshal → 3.6423
[[bar_channel]] → 4.4487	[[coastal_tug]] → 3.6423
booster → 4.4487	withdraw → 3.6423
[[crt]] → 4.3249	accelerants → 3.6423
[[scr_60]] → 4.3249	galleys → 3.6423
[[robert_ h]] → 4.3249	[[hot]] → 3.6423
ancillary → 4.3249	arson → 3.6423
axial → 4.3249	[[leaving]] → 3.6423

(Manuscript received October 20, 2004;
revision accepted for publication April 1, 2005.)