Gregory Grefenstette,
Rank Xerox Research Centre, Meylan, France

Corpus–Derived First, Second and Third–Order Word Affinities

Abstract

A number of corpus–based extraction techniques have been successfully implemented which derive lists of similar words, based on some definition of the context in which they are found, from a corpus. We present here the results of affining such a list in order to extract semantic axes expressing nuances of a word's meaning. These semantic axes represent corpus–based meaning distinctions that are based on the word's usage in the corpus.

1. Introduction

A number of semantic–free or knowledge–poor techniques have been developed for extracting lists of semantically similar words from a large corpus of text. The commonality of these knowledge–poor techniques has been their limitation of knowledge of a word to its strict minimum, usually its part of speech. This limitation means that these techniques can treat texts from any domain, including domains for which no special purpose dictionaries exist. These techniques, to be described below, differ in what is considered as the context of a word and how this context is used. One purpose of these techniques is the discovery of word affinities, i.e., how words can be grouped together. First–order techniques examine the local context of a word attempting to discover what can co–occur with that word within that context. Second–order techniques derive a context for each term and compare these contexts to discover similar words or terms. Third–order techniques compare lists of similar words or terms and group them along semantic axes.

After a brief discussion of first and second–order techniques, we will present our third–order technique for deriving semantic axes.

2. Deriving affinities

2.1 First–order affinities

First–order term affinities describe what other words are likely to be found in the immediate vicinity of a given word. Different affinities can be extracted depending upon what is defined as context.

a) **Significant word associations**: Early work by Church and Hanks (1990) showed that calculating the mutual information [1] between words by passing a fixed–length window over a large corpus and noting when words co–occurred within the window, allowed them to recognize affinities between pairs such as *doctor ... nurse* and *save ... from.*

b) **Collocations:** Smadja (1993) presents a technique for recognizing fixed expressions, fixed patterns, and collocations by examining grammatically tagged windows of words around a given word. In order to find collocations or fixed expressions involving a given word, his technique will build a table listing all the words appearing within five words of the given word, as well as their positions with respect to this word, over a large corpus. Words which appear often and mostly in a fixed position are considered as possible collocates. For example, in a large corpus of newspaper text *hostile ... takeover* often appeared as a collocation with a fixed positioning between the two words whereas *federal ... takeover*, though appearing often within the same window of text, did not pass positioning filters as it is not considered as a collocate. Smadja's system, Xtract, also recognizes rigid noun phrases (Choueka 1988) as well phrasal templates such as *The Dow Jones average of 30 industrials fell *NUMBER* points to *NUMBER*.*

c) **Subcategorizations:** Manning (1993) following work by Brent(1991) derives subcategorization frames from free text through stochastic tagging, robust parsing, and statistical evaluation of the phrases appearing around a given verb.

d) **Morphological variants:** Grefenstette (1993) derives domain–dependent families of words by examining the context defined by a document and using weak string matching clues.

These knowledge–poor techniques provide a partial answer to the question: What kinds of other words will appear around a given word in a corpus?

2.2 Second–order affinities

Second–order affinities show which words share the same environments. Words sharing second–order affinities need never appear together themselves, but their environments are similar. A trivial example of such words is *tumor* and *tumour.* In an English language medical corpus one would expect that the environments of each form of this word would be the same whilst the two words would never appear together. More interesting is the discovery of near synonyms and semantically related words which share second–order affinities.

Techniques which use second–order relations have produced interesting results. Brown et al. (1992) show that a string matching technique over a window around each word provides enough information to cluster words into similar syntactic and semantic classes. A window of 1000 words

excluding the 5 words directly around each word was used to calculate mutual information. A clustering technique, that was aimed at maximizing the average mutual information within clusters, was then iteratively applied to pairs of clusters to create a specified number of general semantic classes such as {*tie, jacket, suit*}, {*morning, noon, evening, night, nights, midnight, bed*}, or { *problems, problem, solution, solve, analyzed, solved, solving*}, from a corpus of 365 million words from a variety of sources. This knowledge–poor technique, based solely on counting strings, provides interesting results, though it is computationally expensive, $O(N^3)$ with a large coefficient and where N is the number of distinct strings.

Deerwester et al. (1990) used document co–occurrence to build up a data matrix where each row represents a word and each column represents a document from some corpus of documents. The entry in each matrix position corresponds to the presence of that word in each document. They then used singular value decomposition to reduce the matrix to its principal axes. This has the effect of reducing the space described by all the words to a smaller space of semantic axes, reducing the problem from thousands of dimensions to hundreds. Each word can then be thought of as a point in this reduced space, specified by its value along each dimension's axis. By considering the distance between words in this space, semantically related words appear closer together. The composition of all the words appearing in the query on the corpus also defines a point in this reduced space, and documents found near that point are chosen in response to a query. Deerwester et al. have shown that this semantic space reduction can improve information retrieval. This technique suffers from the drawbacks of (i) computational complexity since matrix reduction is $O(N^3)$ where N is the smaller of the number of terms and the number of documents, and (ii) attacking only one part of the language variability problem, that of many terms concerning the same concept. Indeed, the other aspect of language variability, that one word can mean many things, introduces noise into the calculations of the semantic axes. Schutze (1992) uses a related technique called canonical discriminant analysis to create semantic axes, using co–occurrence of terms within windows of 1000 characters, which suffers from the same computational complexity.

Using more focussed information, Hindle (1990) reports on semantic extraction work using noun–verb combinations. He processed 6 million words of 1987 AP news with robust deterministic parsers (Hindle 1989) to extract large numbers of Subject–Verb–Object triples. He then calculated the mutual information between verb–noun pairs. For example, the nouns with the highest associations as objects of the verb *drink* were *bunch–beer, tea, Pepsi, champagne, liquid, beer, wine, water*. As a second order calculation using this mutual information association, he then calculated the similarity between nouns by considering how much mutual information they shared over all of the verbs in the corpus. He was able to produce intuitively pleasing results such as the result that the words most similar to *boat* were *ship, plane,*

bus, jet, vessel, truck, car, helicopter, ferry, man. Pereira and Tishby (1992) use just verb–object pairs and a dissimilarity measure to cluster words that are have little dissimilarity.

2.3 Third–order affinities

Many of the techniques mentioned in the last section, particularly the matrix reduction techniques, are prone to generating noise in the case of polysemous words which share sense–dependent environments.[2]

A third–order technique takes the lists of similar words produced by a second–order technique and reworks the context of these added words in order to derive subgroupings of similarity. Schutze and Pedersen (1993) extracts the right–hand contexts of words and compares them to produce similarity lists for a given word, and then examines the left–hand contexts of these similar words to subgroup them.

Here we will present a technique with which we have been experimenting for further exploiting the context of similar words in order to tease out subgroupings along axes of meaning.

3. Overview of derivation of similarity lists

We have developed a system called SEXTANT (Grefenstette 1994) that analyzes the lexical syntactic usage of a word over a corpus and calculates the similarity of words using this syntactic context. Briefly, the system first tokenizes the input text, performs limited morphological analysis of the input tokens assigning to each token possible parts of speech and source words; these parts of speech are resolved to one grammatical tag per token using a stochastic method (Cutting et al. 1992; de Marcken 1990). The unambiguously tagged text is robustly parsed using a technique of filters inspired by Debili (1982) and similar to those found in Constraint Grammars (Voutilainen et al. 1992), i.e., the input sentence need not be fully analyzed, in order to extract low–level syntactic components such as subject–verb, verb–objects, adjective–noun, and noun–noun relations. The output of this stage for each word can be compared to that produced by windowing techniques (Phillips 1985) with the advantage that cleaner context is produced since most spurious relations caused by positional contiguity are avoided. See Figure 1.

At this stage, the context of each word is the list of words that are found in local syntactic relations with it. The type of syntactic relation is retained for noun—verb relations, where SUBJ, DOBJ, and IOBJ stand for subject, direct object, and a generalized indirect object relation which covers both arguments and adjuncts. All these relations are derived from the robust syntactic parsers used in SEXTANT. See Grefenstette (1994) for further details.

With the arrival of Europeans in 1788 , many Aboriginal societies , caught within the coils of expanding white settlement , were gradually destroyed .

Contexts of nouns extracted after syntactic analysis

arrival european	society aboriginal	society destroy-DOBJ
society catch-SUBJ	coil catch-IOBJ	settlement white
settlement expand-DOBJ		

Some contexts extracted with 10 full-word window

arrival aboriginal	arrival society	arrival catch
arrival coil	arrival expand	arrival white
arrival settlement	arrival destroy	european arrival
european aboriginal	european society	european catch
european coil	european expand	european white
european settlement	european destroy	society arrival
society european	society aboriginal	society catch
society coil	society expand	society white
society settlement	society destroy	...

Figure 1: Comparison of extracted contexts using syntactic and non–syntactic techniques.

		Some similarity lists
	freq	
acquisition	1504	purchase bid transaction offer sale merger investment plan deal agreement
agreement	1738	plan offer transaction bid acquisition deal proposal price month investment
bid	2503	offer proposal plan acquisition transaction agreement purchase share year month
buy-out	673	transaction takeover deal purchase merger acquisition investment equity move buy-outs
offer	2744	bid proposal plan transaction agreement acquisition year sale stock board
plan	2022	proposal offer bid agreement transaction price year sale month acquisition
revenue	467	profit earning income gain net sale loss result growth number
shareholder	995	holder board share stock investor offer management director acquisition stake
unit	2625	company subsidiary group sale operation concern share year bank acquisition

Figure 2: Sample of similarity lists, calculated using each word's syntactic attributes, extracted from 6 Mbyte corpus of articles on Mergers from the *Wall Street Journal.*

The contexts for each noun in the corpus are compared to the contexts for every other noun using a weighted Jaccard [3] similarity measure (Romesburg 1990). This step provides a ranked list for each word of words used in the most similar way throughout the corpus being treated, as in Figure 2. Before describing our own third–order technique, exploiting these ranked lists, for deriving semantic axes, we will motivate this research with a brief discussion.

4. Background on semantic axes

The idea of a semantic field which is divided into overlapping areas by a set of words, such as the color words, or words expressing emotional state, has a long history in linguistics (Trier 1931) and is related to the dual view, much more popular, of considering a word as covering segments of a number of semantic features. If each semantic feature, such as *animate, concrete, human* is considered as an axis in the space of meaning, then, the thinking goes, each meaning of a word describes some region in this hyperspace, specified by the word's extension along each axis. Much work in computational semantics presupposes that each word, or word sense, in the lexicon is described as a set of semantic features. In the simplest case the features are either present or absent; in more complicated cases they are described as attribute–value pairs, which again can be visualized as describing a segment of the axis defined by that semantic feature.

The problem with this approach is that the set of semantic features must be defined beforehand by the linguist constructing the lexicon. The choice of these axes, although intellectually stimulating, is not practically simple (Eco 1984:46–68). Although the usually expressed hope is that a finite number of features may suffice, no proof of this has ever been offered, despite claims.[4] Another approach is to abandon the division of the space of meaning into predetermined axes and rather to induce the axes from some corpus. It is this approach that we will explore here.

In order to induce an axis, we shall start from some word and connect this word to a similar word. This connection defines one axis in some space of meaning. We then use this axis to place other words along this axis. Such a technique will be described in the next section.

5. Technique

Two words may be recognized as similar to a third word for different reasons, along different axes of that third word's nuances of meaning. For example, in a medical corpus, *administration* can relate to the organization of a hospital or to the injection of the drug. The straight list of similarly used words produced by a second–order technique does not bring out these semantic differences.

In order to derive semantic axes, we extend a concept introduced in Hindle (1990) and define words as being "reciprocally near neighbors" if the words appear on each other's similarity lists[5] within the closest N words (we use $N=10$ throughout). These words can serve as seeds for axis definition in the following way. We will consider the following case. Let us suppose that a word A was found close to B, C, D, E, and F, and suppose that B was reciprocally near to A; that is we will suppose that A was also one of the closest words to B. Due to this reciprocity, we can be confident that A–B forms a semantic axis. Now we wish to see if we should attach the other words

C, D, E, and F to this axis. One way to do this is to include any of these words which is also a near neighbor to B, independent of A. This will define a set of words which are (i) close to A, (ii) near neighbors to B, and (iii) close to this axis, supposing that A–B is a semantic axis.

Semantic Axis	*words closest to axis*
acquisition *as an* agreement	bid offer plan
acquisition *as a* bid	offer sale
acquisition *as a* deal	transaction merger investment agreement
acquisition *as a* merger	transaction
acquisition *as a* plan	bid offer sale
acquisition *as a* purchase	bid transaction sale
acquisition *as a* sale	offer
acquisition *as a* transaction	bid offer sale plan
agency *as a* firm	bank concern
agency *as a* united-states	thrift
agreement *as an* acquisition	plan offer bid
agreement *as a* bid	offer
agreement *as a* deal	transaction acquisition investment
agreement *as a* plan	offer bid price
agreement *as a* proposal	plan offer bid transaction
.....	...
approval *as an* action	decision
approval *as an* authority	review rule
approval *as a* clearance	review authority
approval *as a* review	clearance authority
approval *as a* vote	step
.....	...
bid *as an* acquisition	offer plan agreement sale
bid *as an* agreement	offer plan
bid *as a* plan	offer sale
bid *as a* proposal	offer plan transaction agreement
.....	...
buy-out *as a* deal	transaction takeover merger acquisition investment
buy-out *as a* merger	transaction acquisition
buy-out *as a* purchase	transaction acquisition
buy-out *as a* takeover	merger investment
buy-out *as a* transaction	acquisition offer
.....	...
transaction *as an* acquisition	offer plan sale bid
transaction *as a* bid	offer sale
transaction *as a* buy-out	purchase
transaction *as a* deal	buy-out merger acquisition
transaction *as a* merger	buy-out acquisition
transaction *as a* plan	offer sale bid
transaction *as a* proposal	offer plan bid
transaction *as a* purchase	acquisition sale bid
transaction *as a* sale	offer
value *as an* amount	cash number debt
value *as a* cash	profit debt
value *as an* earning	profit
value *as a* financing	cash
value *as an* interest	price
value *as a* price	year
value *as a* profit	price

Figure 3: Semantic clusters from a WSJ MERGERS corpus.

Semantic Axis	*words closest to axis*
a-crystallin *as a* dna	protein
ability *as a* capacity	production function
ability *as a* inability	capacity
abnormality *as a* anomaly	atresia
abnormality *as a* impairment	disorder disturbance
abnormality *as a* nature	manifestation
absence *as a* sibling	family
absorption *as a* exchange	transport
absorption *as a* na	exchange
absorption *as a* po	tension
accumulation *as a* extent	jaundice
acid *as a* dna	protein
acid *as a* fraction	protein
acidosis *as a* insufficiency	hypertrophy
act *as a* prolongation	deficiency
activity *as a* amount	concentration level number
addition *as a* absence	presence
adenocarcinoma *as a* carcinoma	tumor
adenoma *as a* hyperplasia	hypertrophy
adjunct *as a* chemotherapy	therapy
administration *as a* dose	injection
administration *as a* infusion	dose
administration *as a* secretion	deficiency
administration *as a* therapy	treatment
administration *as a* treatment	response
.....	...
tumor *as a* cancer	lesion tissue
tumor *as a* carcinoma	cancer disease
tumor *as a* growth	tissue effect
tumor *as a* lesion	cancer disease
tumor *as a* tissue	disease
.....	...

Figure 4: Semantic clusters from the MED corpus.

When this grouping technique is applied to the most frequent words from the corpus composed of 6MBytes of Wall Street Journal articles on mergers, we develop clusters such as those presented in Figure 3. Words are included in a cluster if they are nearly as frequent as, or more frequent than, the second word defining the axis; taking into account frequency in this way is an attempt to generalize from more specific to more general words. For example, in Figure 3, we see that *agreement* is a reciprocal near neighbor to *acquisition* in the MERGER corpus, so we take *acquisition–agreement* to be one semantic axis of the word *acquisition* for this corpus. Then, comparing the similarity lists of *agreement* to the other words closest to *acquisition*, we

discover that *bid*, *offer*, and *plan* are more general words (appearing more often) than *agreement* and are reciprocally close to it. This group seems to define a sense of *acquisition* having to do with the negotiation process involved in acquiring some company.

breast	compare-DOBJ	develop-DOBJ	disseminate-DOBJ	human
lung	mammary	metastatic	neck	present-DOBJ
primary	pulmonary			

Figure 5: Syntactic attributes shared by "tumor," "cancer," and "carcinoma" in the Medical corpus.

brain	discuss-DOBJ	human	lung	make-SUBJ
mammary	mouse	normal	patient	pituitary
rat	renal			

Figure 6: Syntactic attributes shared by "tumor" "growth," and "tissue" in the Medical corpus.

When the same clustering technique is applied to the corpus of medical abstracts, we get the clusters appearing in Figure 4. Again, the technique of using reciprocal near neighbors creates axes which are able to group similar words, although the non–medical person must resort to a medical dictionary to recognize the relations. For example, *a–crystallin* and *dna* are both examples of *proteins*. *Atresia* is a "congenital absence or closing of a normal body opening," and should be close to the axis *abnormality—anomaly*. It is not clear what relation, if any, exists between *acid*, *fraction* and *protein*. *Acidosis*, though, is an *insufficiency* resulting from renal *hypertrophy* which is captured in the *acidosis—insufficiency—hypertrophy* axis in Figure 4.

Another interesting result is the way the word *tumor* is divided along malignant and non–malignant axes. One axis is *tumor–growth*, which attracts the words *tissue* and *effect*, while the axes *tumor–cancer*, *tumor–carcinoma*, and *tumor–lesion* bring in each other to their axes. The axes that are derived here are the result of a cascade of knowledge–poor processes: tokenization, tagging, parsing, extracting lexical–syntactic contexts, automatically weighting the contexts using their frequency in the corpus, similarity comparison using these contexts as attributes, crossing similarity lists to find reciprocally nearest neighbors, using these neighbors with their corpus frequency to derive the lists. The reasons why words are grouped together in these third–order affinities depend on each step, yet an idea of why words are grouped can be intuited from the lexical–syntactic attributes that the words along an axis share, even though other words may have played a part in the calculation of the reciprocally nearest neighbors. Figure 5 shows which

contexts (without their frequencies) are shared by the words *tumor*, *cancer*, and *carcinoma*. The tags SUBJ, DOBJ, and IOBJ indicate a noun–verb context and untagged words indicate a noun–noun or noun–adjective context. Each lexical–syntactic attribute adds a small piece into the calculation of similarity and ultimately semantic axes. Figure 6 shows those contexts shared by the words *tumor*, *growth*, and *tissue*.

Going back to Figure 3, we can see similar sense differentiations in the *approval–action–decision* as opposed to the *approval–authority–review–rule* which distinguish the act of approving from the right of approving. On the other hand, we also see correspondences unlike ones that a human would draw such as between *agency–United States–thrift*. Although these words are all connected through the "Resolution Trust Corporation," it is difficult to see a clear semantic axis here. Looking at the attributes that all three terms have in common, shown in Figure 7, does not clarify the situation further, since each attribute seems to be adding one little piece of meaning to the composite judgment made by SEXTANT that the words are similar.

acquire-DOBJ	agree-DOBJ	base-DOBJ	big	continue-SUBJ
create-DOBJ	expect-DOBJ	hold-SUBJ	lose-SUBJ	make-DOBJ
make-IOBJ	medium-size-DOBJ	new	own-SUBJ	plan-DOBJ
regional	report-SUBJ	say-DOBJ	say-SUBJ	second-large
sell-DOBJ	small	think-DOBJ	time	top year

Figure 7: Attributes shared by "agency," "United States," and "thrift" in the MERGERS corpus.

6. Conclusion

We have presented a technique for deriving third–order affinities from a corpus of text. First–order affinities describe collocates of words, second–order affinities show similarly used words, and third–order affinities create semantic groupings among similar words. Rather than predefining the semantic axes, pairs of close words were used to define an axis and similar words were placed along these axes. A problem with the clustering method as it exists is that sometimes the distinctions seem to be too fine. For example, it might be perfectly satisfactory to group *acquisition–sale–purchase–transaction–merger* into one large group rather than its many small subsets as they appear in Figure 3. This level of distinction or grouping of course depends on the use to which these lists are to be put. If the use is for human consumption, such as an expansion proposing interface to a retrieval system, then larger groups would be all right, since the user could quickly pare down the list. If it is for an automatic expansion system, then smaller lists might be preferable (Sparck Jones 1971). Another problem, apparent in any knowledge–poor corpus–based technique, is that results contain noise that would have to be manually weeded out. One further problem is that words

appearing infrequently in the corpus do not possess enough context there to be included in the treatment and are mostly ignored. Still it is to be hoped that larger corpora, and further overlaying of knowledge–poor techniques will allow greater and finer exploitation of such words.

Notes

1 The formula for mutual information is $I(x\,y) = log(P(x\,y)/(P(x)\,P(y)))$ *where* $P(x\,y)$ is the joint probability of the events x and y and $P(x)$ and $P(y)$ are the probabilities of each individual event. The value reaches a maximum when x and y co–occur and are both rare.

2 Work has been done in assigning words to thesaurus headings (Yarowsky 1992) as a means for reducing polysemy, but this only captures a small portion of grossly polysemous words and is limited to words appearing in the thesaurus, and to the sense distinctions of the thesaurus.

3 The Jaccard measure is defined as the number of attributes shared by two objects divided by the total number of unique attributes possessed by both objects. If A is the number of attributes shared by two objects, and B is the number of attributes only appearing with the first object, and C is the number of attributes appearing only with the second object, then the Jaccard measure of similarity between the two objects is $A/(A+B+C)$. This similarity measure yields a value between 0 and 1. The attributes are weighted by taking the *log* of the attribute frequency with the object multiplied by an inverse entropy measure of the attribute over the corpus. For example, common adjectives have a high entropy and thus lower weights. See (Grefenstette 1994). Here the objects being compared are two nouns, and their attributes are the words found in lexical–syntactic relations to these nouns.

4 In (Wierzbicka.1990), Anna Wierzbicka proposes a number of definitions of hard–to–define words such as *bachelor, bird, game, lie* using her "proposed list of universal semantic primitives," (p. 359) but these definitions include computationally unusable primitives such as "thought as" as in the definition "bachelor — an unmarried man thought as someone who could marry.'

5 These similarity lists were derived as described above by comparing the syntactically derived contexts for each noun throughout the entire corpus. Note that, as with any similarity comparison method, similarity lists are not symmetric. For example, if you compare the similarity of a group containing two elephants, a dog, a rat, and ten mice, the animal most similar to the rat will be a mouse, while a mouse will be most similar to another mouse.

References

Brent, M.R. 1991. "Automatic acquisition of subcategorization frames from untagged, free–text corpora". In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.*

Brown, P.F., Della Pietra, V.J., deSouza, P.V., Lai, J.C., and Mercer, R.L. 1992. "Class–based n–gram models of natural language". *Computational Linguistics*, 184:467–479.

Choueka, Y. 1988. "Looking for a needle in a haystack, or locating interesting collocational expressions in large textual databases". In *RIAO'88 Conference Proceedings*, pages 609–623, MIT,Cambridge,Mass.

Church, K.W. and Hanks, P. 1990. "Word association norms, mutual information, and lexicography". *Computational Linguistics*, 161:22–29.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. 1992. "A practical part–of–speech tagger". *Proceedings of the Third Conference on Applied Natural Language Processing.*

de Marcken, C.G. 1990. "Parsing the LOB corpus". In *28th Annual Meeting of the Association for Computational Linguistics*, pages 243–251, Pittsburgh, PA. ACL.

Debili, F. 1982. *Analyse Syntaxico–Sémantique Fondée sur une Acquisition Automatique de Relations Lexicales–Sémantiques.* PhD thesis, University of Paris XI, France.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. 1990. "Indexing by latent semantic indexing". *Journal of the American Society for Information Science*, 416:391–407.

Eco, U. 1984. *Semiotics and the Philosophy of Language.* Indiana University Press, Bloomington.

Grefenstette, G. 1993. "Automatic thesaurus generation from raw text using knowledge–poor techniques". In *Making Sense of Words.* Ninth Annual Conference of the UW Centre for the New OED and text Research.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Press, Boston.

Hindle, D. 1989. "Acquiring disambiguation rules from text". In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 118–125. ACL.

Hindle, D. 1990. "Noun classification from predicate–argument structures". In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh. ACL.

Manning, C.D. 1993. "Automatic acquisition of a large subcategorization dictionary from corpora". In *31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242, Columbus, OH. ACL.

Pereira, F. and Tishby, N. 1992. "Distributional similarity, phase transitions and hierarchical clustering". In Goldman, R. (Ed.), *Fall Symposium on Probability and Natural Language.* AAAI, Cambridge, Mass.

Phillips, M. 1985. *Aspects of Text Structure: An investigation of the lexical organization of text.* Elsevier, Amsterdam.

Romesburg, H.C. 1990. *Cluster Analysis for Researchers.* Krieger Publishing Company, Malabar, Florida.

Schutze, H. 1992. "Context space". In *Fall Symposium on Probability and Natural Language*, Cambridge, Mass. AAAI.

Smadja, F. 1993. "Retrieving collocations from text: Xtract". *Computational Linguistics*, 191:143–178.

Sparck Jones, K. 1971. *Automatic Keyword Classification and Information Retrieval.* Butterworths, London.

Trier, J. 1931. *Der deutsche Wortschatz im Sinnbezirk des Verstandes: Die Geschichte eines sprachlichen Feldes. Band I.* Heidelberg.

Voutilainen, A., Heikkila, J., and Anttila, A. 1992. A lexicon and constraint grammar of English. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France. COLING'92.

Wierzbicka, A. 1990. "Protypes save". In Tsohatzidis, S.L. (Ed.), *Meanings and Prototypes*, chapter 17. Routledge, London.

Yarowsky, D. 1992. "Word–sense disambiguation using statistical models of Roget's thesaurus categories trained on a large corpus". In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France. COLING'92.