Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches

Gregory Grefenstette

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
grefen@cs.pitt.edu

Abstract

As large on-line corpora become more prevalent, a number of attempts have been made to automatically extract thesaurus-like relations directly from text using knowledge poor methods. In the absence of any specific application, comparing the results of these attempts is difficult. Here we propose an evaluation method using gold standards, i.e., pre-existing hand-compiled resources, as a means of comparing extraction techniques. Using this evaluation method, we compare two semantic extraction techniques which produce similar word lists, one using syntactic context of words , and the other using windows of heuristically tagged words. The two techniques are very similar except that in one case selective natural language processing, a partial syntactic analysis, is performed. On a 4 megabyte corpus, syntactic contexts produce significantly better results against the gold standards for the most characteristic words in the corpus, while windows produce better results for rare words.

1 Introduction

As more text becomes available electronically, it is tempting to imagine the development of automatic filters able to screen these tremendous flows of text extracting useful bits of information. In order to properly filter, it is useful to know when two words are similar in a corpus. Knowing this would alleviate part of the *term variability* problem of natural language discussed in Furnas et al. (1987) . Individuals will choose a variety of words to name the same object or operation, with little overlap between people's choices. This variability in naming was cited as the principal reason for large numbers of missed citations in a large-scale evaluation of an information retrieval system [Blair and Maron, 1985]. A proper filter must be able to access information in the text using any word of a set of similar words. A number of knowledge-rich [Jacobs and Rau, 1990, Calzolari and Bindi, 1990, Mauldin, 1991] and knowledge-poor [Brown *et al.*, 1992, Hindle, 1990, Ruge, 1991, Grefenstette, 1992] methods have been proposed for recognizing when words are similar. The knowledge-rich approaches require either a conceptual dependency representation, or semantic tagging of the words, while the knowledge-poor approaches require no previously encoded semantic information, and depend on frequency of co-occurrence of word contexts to determine similarity. Evaluations of results produced by the above systems are often been limited to visual verification by a human subject or left to the human reader.

In this paper, we propose gold standard evaluation techniques, allowing us to objectively evaluate and to compare two knowledge-poor approaches for extracting word similarity relations from large text corpora. In order to evaluate the relations extracted, we measure the overlap of the results of each technique against existing hand-created

repositories of semantic information such as thesauri and dictionaries. We describe below how such resources can be used as evaluation tools, and apply them to two knowledge-poor approaches.

One of the tested semantic extraction approaches uses selective natural language processing, in this case the lexical-syntactic relations that can be extracted for each word in a corpus by robust parsers [Hindle, 1983, Grefenstette, 1993]. The other approach uses a variation on a classic windowing technique around each word such as was used in [Phillips, 1985]. Both techniques are applied to the same 4 megabyte corpus. We evaluate the results of both techniques using our gold standard evaluations over thesauri and dictionaries and compare the results obtained by the syntactic based method to those obtained by the windowing method. The syntax-based method provides a better overlap with the manually defined thesaurus classes for the 600 most frequently appearing words in the corpus, while for rare words the windowing method performs slightly better for rare words.

2 Gold Standards Evaluation

2.1 Thesauri

Roget's Thesaurus is readily available via anonymous ftp[1]. In it are collected more than 30,000 unique words arranged in a shallow hierarchy under 1000 topic numbers such as Existence (Topic Number 1), Inexistence (2), Substantiality (3), Unsubstantiality (4), ..., Rite (998), Canonicals (999), and Temple (1000). Although this is far from the total number of semantic axes of which one could think, it does provide a wide swath of commonly accepted associations of English language words. We would expect that any system claiming to extract semantics from text should find some of the relations contained in this resource.

By transforming the online source of such a thesaurus, we use it as a gold standard by which to measure the results of different similarity extraction techniques. This measurement is done by checking whether the 'similar words' discovered by each technique are placed under the same heading in this thesaurus.

In order to create this evaluation tool, we extracted a list consisting of all single-word entries from our thesauri with their topic number or numbers. A portion of the extracted *Roget* list in Figure 1 shows that ***abatement*** appears under two topics: Nonincrease (36) and Discount (813). ***Abbe*** and ***abbess*** both belong under the same topic heading 996 (Clergy). The extracted *Roget's* list has 60,071 words (an average of 60 words for each of the 1000 topics). Of these 32,000 are unique (an average of two occurrence for each word). If we assume for simplicity that each word appears under exactly 2 of the 1000 topics, and that the words are uniformly distributed, the chance that two words w_1 and w_2 occur under the same topic is

$$P_{Roget} = 2 * (2/1000),$$

since w_1 is under 2 topic headings and since the chance that w_2 is under any specific topic heading is 2/1000. The probability of finding two randomly chosen words together under the same heading, then, is 0.4%.

Our measurement of a similarity extraction technique using this gold standard is performed as follows.

[1] For example, in March 1993 it was available via anonymous ftp at the Internet site *world.std.com* in the directory /obi/obi2/Gutenberg/etext91, as well at over 30 other sites.

Roget's	
entry	*Topic*
...	
abatement	36
abatement	813
abatis	717
abatjour	260
abattis	717
abattoir	361
abba	166
abbacy	995
abbatial	995
abbatical	995
abbatis	717
abbe	996
abbess	996
...	

Macquarie	
entry	*subheading*
...	
disesteem	036406
disesteem	063701
diseur	022701
disfavour	003901
disfavour	056601
disfavour	063701
disfeature	018212
disfeaturement	018201
disfigure	006804
disfigure	018212
disfigure	020103
disfigured	006803
disfigured	020102
...	

Figure 1: Samples from One Word Entries in Both Thesauri

> Given a corpus, use the similarity extraction method to derive similarity judgements between the words appearing in the corpus. For each word, take the word appearing as most similar. Examine the human compiled thesaurus to see if that pair of words appears under the same topic number. If it does, count this as a hit.

This procedure was followed on the 4 megabyte corpus described below to test two semantic extraction techniques, one using syntactically derived contexts to judge similarity and one using window-based contexts. The results of these evaluations are also given below.

2.2 Dictionary

We also use an online dictionary as a gold standard following a slightly different procedure. Many researchers have drawn on online dictionaries in attempts to do semantic discovery [Sparck Jones, 1986, Vossen *et al.*, 1989, Wilks *et al.*, 1989], whereas we use it here only as a tool for evaluating extraction techniques from unstructured text. We have an online version of *Webster's 7th* available, and we use it in evaluating discovered similarity pairs. This evaluation is based on the assumption that similar words will share some overlap in their dictionary definitions. In order to determine overlap, each the entire literal definition is broken into a list of individual words. This list of tokens contains all the words in the dictionary entry, including dictionary-related markings and abbreviations. In order to clean this list of non-information-bearing words, we automatically removed any word or token

1. of fewer than 4 characters,
2. among the most common 50 words of 4 or more letters in the Brown corpus,
3. among the most common 50 words of 4 or more letters appearing in the definitions of *Webster's 7th*,

ad-min-is-tra-tion n. 1. the act or process of administering 2. performance of executive duties :: c<MANAGEMENT> 3. the execution of public affairs as distinguished from policy making 4. a) a body of persons who administer b) i<cap> :: a group constituting the political executive in a presidential government c) a governmental agency or board 5. the term of office of an administrative officer. or body.

administer, administering, administrative, affairs, agency, board, constituting, distinguished, duties, execution, executive, government, governmental, making, management, office, officer, performance, persons, policy, political, presidential, public, term

Figure 2: *Webster* definition of "administration," and resulting definition list after filtering through stoplist.

4. listed as a preposition, quantifier, or determiner in our lexicon,
5. of 4 or more letters from a common information retrieval stoplist,
6. among the dictionary-related set: *slang, attrib, kind, word, brit, ness, tion, ment.*

These conditions generated a list of 434 stopwords of 4 or more characters which are retracted from any dictionary definition, The remaining words are sorted into a list. For example, the list produced for the definition of the word *administration* is given in Figure 2. For simplicity no morphological analysis or any other modifications were performed on the tokens in these lists.

To compare two words using these lists, the intersection of each word's filtered definition list is performed. For example, the intersection between the lists derived from the dictionary entries of *diamond* and *ruby* is (*precious, stone*); between *right* and *freedom* it is (*acting, condition, political, power, privilege, right*). In order to use these dictionary-derived lists as an evaluation tool, we perform the following experiment on a corpus.

> **Given a corpus, take the similarity pairs derived by the semantic extraction technique in order of decreasing frequency of the first term. Perform the intersection of their respective two dictionary definitions as described above. If this intersection contains two or more elements, count this as a hit.**

This evaluation method was also performed on the results of both semantic extraction techniques applied to the corpus described in the next section.

3 Corpus

The corpus used for the evaluating the two techniques was extracted from *Grolier's Encyclopedia* for other experiments in semantic extraction. In order to generate a relatively coherent corpus, the corpus was created by extracting only those those sentences which contained the word *Harvard* or one of the thirty hyponyms found under the word **institution** in *WordNet*[2] [Miller *et al.*, 1990], viz. *institution, establishment, charity, religion, ..., settlement.* This produced a corpus of 3.9 megabytes of text.

[2] WordNet was not used itself as a gold standard since its hierarchy is very deep and its inherent notion of semantic classes is not as clearly defined as in *Roget.*

4 Semantic Extraction Techniques

We will use these gold standard evaluation techniques to compare two techniques for extracting similarity lists from raw text.

The first technique [Grefenstette, 1992] extracts the syntactic context of each word throughout the corpus. The corpus is divided into lexical units via a regular grammar, each lexical unit is assigned a list of context-free syntactic categories, and a normalized form. Then a time linear stochastic grammar similar to the one described in [de Marcken, 1990] selects a most probable category for each word. A syntactic analyzer described in [Grefenstette, 1993] chunks nouns and verb phrases and create relations within chunks and between chunks. A noun's context becomes all the other adjectives, nouns, and verbs that enter into syntactic relations with it.

As a second technique, more similar to classical knowledge-poor techniques [Phillips, 1985] for judging word similarity, we do not perform syntactic disambiguation and analysis, but simply consider some window of words around a given word as forming the context of that word. We suppose that we have a lexicon, which we do, that gives all the possible parts of speech for a word. Each word in the corpus is looked up in this lexicon as in the first technique, in order to normalize the word and know its possible parts of speech [Evans *et al.*, 1991]. A noun's context will be all the words that can be nouns, adjectives, or verbs within a certain window around the noun. The window that was used was all nouns, adjectives, or verbs on either side of the noun within ten and within the same sentence.

In both cases we will compare nouns to each other, using their contexts. In the first case, the disambiguator determines whether a given ambiguous word is a noun or not. In the second case, we will simply decide that if a word can be at once a noun or verb, or a noun or adjective, that it is a noun. This distinction between the two techniques of using a cursory syntactic analysis or not allows us to evaluate what is gained by the addition of this processing step.

Figure 3 below shows the types of contexts extracted by the selective syntactic technique and by the windowing technique for a sentence from the corpus.

Once context is extracted for each noun, the contexts are compared for similarity using a weighted Jaccard measure [Grefenstette, 1993]. In order to reduce run time for the similarity comparison, only those nouns appearing more than 10 times in the corpus were retained. 2661 unique nouns appear 10 times or more. For the windowing technique 33,283 unique attributes with which to judge the words are extracted. The similarity judging run takes 4 full days on a DEC 5000, compared to 3 and 1/2 hours for the similarity calculation using data from the syntactic technique, due to greatly increased number of attributes for each word. For each noun, we retain the noun rated as most similar by the Jaccard similarity measure. Figure 4 shows some examples of words found most similar by both techniques.

5 Results

The first table, in Figure 5, compares the hits produced by the two techniques over *Roget's* and over another online thesaurus, *Macquarie's*, that we had available in the Laboratory for Computational Linguistics at Carnegie Mellon University. This table compares the results obtained from the windowing technique described in preceding paragraphs to those

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 3: Comparison of Extracted Contexts using Syntactic and Non-Syntactic Techniques

	Technique used	
Corpus word	*Syntax*	*Window*
formation	creation	system
work	school	religious
foundation	institution	system
government	constitution	state
education	training	public
religious	religion	century
university	institution	institution
group	institution	member
establishment	creation	government
power	authority	government
creation	establishment	state
state	law	government
program	institution	education
law	constitution	public
year	century	government
center	development	city
art	architecture	science
form	group	life
century	year	religious
member	group	group
part	center	government

Figure 4: Sample of words found to be most similar, by the syntactic based technique, and by the window technique, to some frequently occurring words in the corpus

results over corpus using Window vs Syntactic Contexts						
	ROGET		MACQUARIE		WEBSTER	
RANK	WINDOW	SYNTAX	WINDOW	SYNTAX	WINDOW	SYNTAX
1-20	25%	50%	15%	40%	55%	50%
21-40	10%	30%	20%	45%	40%	60%
41-60	25%	30%	30%	35%	55%	70%
61-80	15%	30%	20%	30%	45%	65%
81-100	15%	40%	15%	35%	35%	55%
101-200	14%	31%	19%	34%	34%	55%
201-300	21%	29%	20%	30%	29%	34%
301-400	13%	17%	12%	18%	25%	29%
401-500	15%	16%	12%	13%	24%	26%
501-600	13%	11%	10%	15%	19%	16%
601-700	8%	11%	11%	14%	20%	14%
701-800	11%	9%	9%	9%	17%	17%
801-900	17%	6%	13%	7%	25%	12%
901-1000	8%	10%	9%	9%	29%	12%
1001-2000	10.2%	4.9%	11.8%	5.3%	19.2%	6.9%
2001-3000	7.9%	2.4%	7.9%	2.1%	15.2%	5.2%

Figure 5: Windowing vs Syntactic Percentage of Hits for words from most frequent to least

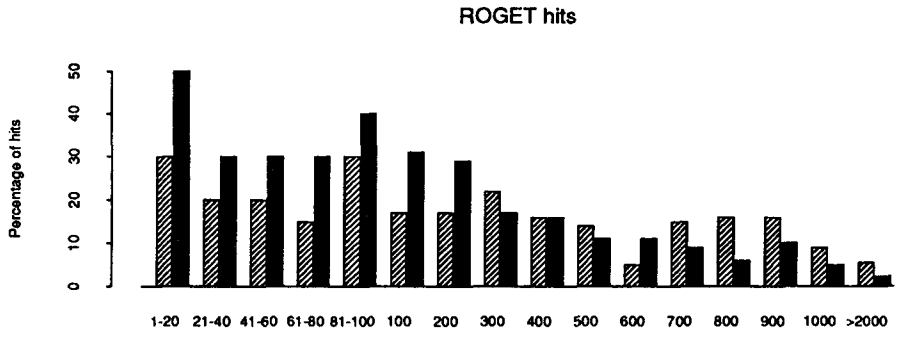

Figure 6: Comparison of hit percentage in *Roget's* using simple 10-word windowing technique(clear) vs syntactic technique(black). The y-axis gives the percentage of hits for each group of frequency-ranked terms.

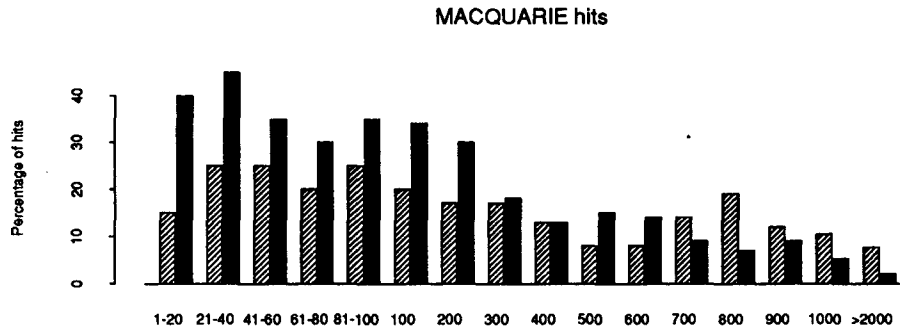

Figure 7: Comparison of hits in *Macquarie's* using simple 10-word windowing technique(clear) vs syntactic technique(black). The y-axis gives the percentage of hits for each group of frequency-ranked terms.

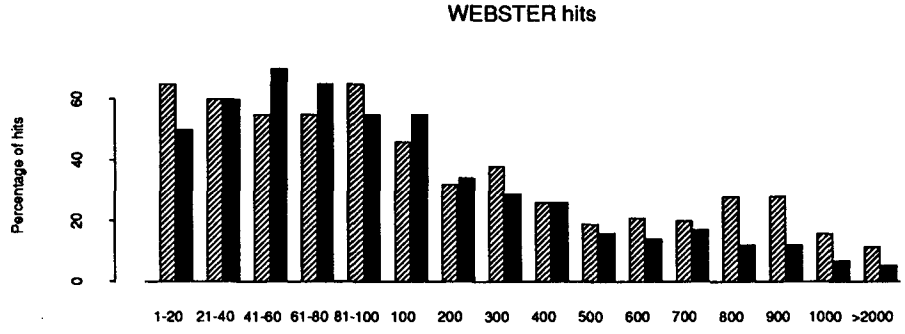

Figure 8: Comparison of hit percentage in *Webster's* using simple 10-word windowing technique (hashed bars) vs syntactic technique (solid bars). The y-axis gives the percentage of hits for each group of frequency-ranked terms.

Roget *First 600*	SYNTACTIC HITS	SYNTACTIC MISS
WINDOW HITS	48	60
WINDOW MISS	91	401

$\chi^2 = 6.4$
$p < .025$

Macquarie *First 600*	SYNTACTIC HITS	SYNTACTIC MISS
WINDOW HITS	42	54
WINDOW MISS	103	401

$\chi^2 = 15.3$
$p < .005$

Roget *Last 600*	SYNTACTIC HITS	SYNTACTIC MISS
WINDOW HITS	2	28
WINDOW MISS	14	556

$\chi^2 = 4.6$
$p < .05$

Macquarie *Last 600*	SYNTACTIC HITS	SYNTACTIC MISS
WINDOW HITS	4	40
WINDOW MISS	14	542

$\chi^2 = 12.5$
$p < .0005$

Figure 9: χ^2 results comparing Syntactic and windowing hits in man-made thesauri

obtained from the syntactic technique, retaining only words for which similarity judgements were made by both techniques.

It can be seen in Figure 5 that simple technique of moving a window over a large corpus, counting co-occurrences of words, and eliminating empty words, provides a good hit ratio for frequently appearing words, since about 1 out of 5 of the 100 most frequent words are found similar to words appearing in the same heading in a hand-built thesaurus.

It can also be seen that the performance of the partial syntactic analysis based technique is better for the 600 most frequently appearing nouns, which may be considered as the characteristic vocabulary of the corpus. The difference in performance between the two techniques is statistically significant (p ¡ 0.05). The results of a χ^2 test are given in Figure 9. Figures 6 and 7 show the same results as histograms. In these histograms it becomes more evident that the window co-occurrence techniques give more hits for less frequently occurring words, after the 600th most frequent word. One reason for this can be seen by examining the 900th most frequent word, *employment*. Since the windowing technique extracts up to 20 non-stopwords from either side, there are still 537 context words attached to this word, while the syntactically-based technique, which examines finer-grained contexts, only provides 32 attributes.

Figure 8 shows the results of applying the less focused dictionary gold standard experiment to the similarities obtained from the corpus by each technique. For this experiment, both techniques provide about the same overlap for frequent words, and the same significantly stronger showing for the rare words for the windowing technique.

6 Conclusion

In this paper we presented a general method for comparing the results of two similarity extraction techniques via gold standards. This method can be used when no application-specific evaluation technique exists and provides a relative measurement of techniques against human-generated standard semantic resources. We showed how these gold standards could be processed to produce a tool for measuring overlap between their contents and the results of a semantic extraction method. We applied these gold standard evaluations to two different semantic extraction techniques passed over the same 4 megabyte corpus. The syntactic-based technique produced greater overlap with the gold standards derived from thesauri for the characteristic vocabulary of the corpus, while the window-based technique provided relatively better results for rare words.

This dichotomous result suggests that no one statistical technique is adapted to all ranges of frequencies of words from a corpus. Everyday experience suggests that frequently occurring events can be more finely analyzed than rarer ones. In the domain of corpus linguistics, the same reasoning can be applied. For frequent words, finer grained context such as that provided by even rough syntactic analysis, is rich enough to judge similarity. For less frequent words, reaping more though less exact information such as that given by windows of N words provides more information about each word. For rare words, the context may have to be extended beyond a window, to the paragraph, or section, or entire document level, as Crouch (1990) did for rarely appearing words.

Acknowledgements. This research was performed under the auspices of the Laboratory for Computational Linguistics (Carnegie Mellon University) directed by Professor David A. Evans.

References

[Blair and Maron, 1985] D.C. Blair and M.E. Maron. An evaluation of retrieval effectiveness. *Communications of the ACM*, 28:289–299, 1985.

[Brown *et al.*, 1992] Peter F. Brown, Vincent J. Della Pietra, Petere V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[Calzolari and Bindi, 1990] Nicoletta Calzolari and Remo Bindi. Acquisition of lexical information from a large textual italian corpus. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki, 1990.

[Crouch, 1990] C. J. Crouch. An approach to the automatic construction of global thesauri. *Information Processing and Management*, 26(5):629–640, 1990.

[de Marcken, 1990] Carl G. de Marcken. Parsing the LOB corpus. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 243–251, Pittsburgh, PA, June 6–9 1990. ACL.

[Evans *et al.*, 1991] David A. Evans, Steve K. Handerson, Robert G. Lefferts, and Ira A. Monarch. A summary of the CLARIT project. Technical Report CMU-LCL-91-2, Laboratory for Computational Linguistics, Carnegie-Mellon University, November 1991.

[Furnas *et al.*, 1987] George W. Furnas, Tomas K. Landauer, L.M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, November 1987.

[Grefenstette, 1992] G. Grefenstette. Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis. In *30th Annual Meeting of the Association for Computational Linguistics*, Newark, Delaware, 28 June – 2 July 1992. ACL'92.

[Grefenstette, 1993] Gregory Grefenstette. Extracting semantics from raw text, implementation details. *Heuristics: the Journal of Knowledge Engineering*, 1993. To Appear in the Special Issue on Knowledge Extraction from Text, Available as TR CS92-05, from the University of Pittsburgh, CS Dept.

[Hindle, 1983] Donald Hindle. User manual for Fidditch. Technical Report 7590–142, Naval Research Laboratory, 1983.

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

[Jacobs and Rau, 1990] Paul Jacobs and Lisa Rau. SCISOR: Extracting information from on-line news. *Communications of the ACM*, 33(11):88–97, 1990.

[Mauldin, 1991] M. L. Mauldin. *Conceptual Information Retrieval: A case study in adaptive parsing.* Kluwer, Norwell, MA, 1991.

[Miller *et al.*, 1990] George A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244, 1990.

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

[Ruge, 1991] Gerda Ruge. Experiments on linguistically based term associations. In *RIAO'91*, pages 528–545, Barcelona, April 2–5 1991. CID, Paris.

[Sparck Jones, 1986] Karen Sparck Jones. *Synonymy and Semantic Classification.* Edinburgh University Press, Edinburgh, 1986. PhD thesis delivered by University of Cambridge in 1964.

[Vossen *et al.*, 1989] P. Vossen, W. Meijs, and M. den Broeder. Meaning and structure in dictionary definitions. In Bran Boguraev and Ted Briscoe, editors, *Computational Lexicography for Natural Language Processing*, pages 171–190. Longman Group UK Limited, London, 1989.

[Wilks *et al.*, 1989] Yorick Wilks, D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. A tractable machine dictionary as a resource for computational semantics. In Bran Boguraev and Ted Briscoe, editors, *Computational Lexicography for Natural Language Processing*, pages 193–228. Longman Group UK Limited, London, 1989.