Shipping Departments vs. Shipping Pacemakers: Using Thematic Analysis to Improve Tagging Accuracy

Uri Zernik
General Electric - Research and Development Center
PO Box 8, Schenectady, NY 12301

Abstract

Thematic analysis is best manifested by contrasting collocations[1] such as "shipping pacemakers" vs. "shipping departments". While in the first pair, the pacemakers are being shipped, in the second one, the departments are probably engaged in some shipping activity, but are not being shipped.

Text pre-processors, intended to inject corpus-based intuition into the parsing process, must adequately distinguish between such cases. Although statistical tagging [Church *et al.*, 1989; Meteer *et al.*, 1991; Brill, 1992; Cutting *et al.*, 1992] has attained impressive results overall, the analysis of multiple-content-word strings (i.e., collocations) has presented a weakness, and caused accuracy degradation.

To provide acceptable coverage (i.e., 90% of collocations), a tagger must have accessible a large database (i.e., 250,000 pairs) of individually analyzed collocations. Consequently, training must be based on a corpus ranging well over 50 million words. Since such a large corpus does not exist in a tagged form, training must be from raw corpus.

In this paper we present an algorithm for text tagging based on thematic analysis. The algorithm yields high-accuracy results. We provide empirical results: The program NLcp (NL corpus processing) acquired a 250,000 thematic-relation database through the 85-million word Wall-Street Journal Corpus. It was tested over the Tipster 66,000-word Joint-Venture corpus. [2] [3]

Pre-processing: The Big Picture

Sentences in a typical newspaper story include idioms, ellipses, and ungrammatical constructs. Since authentic language defies textbook grammar, we must rethink our basic parsing paradigm, and tune it to the nature of the text under analysis.

Hypothetically, parsing could be performed by one huge unification mechanism [Kay, 1985; Shieber, 1986; Tomita, 1986] which would receive its tokens in the form of words, characters, or morphemes, negotiate all given constraints, and produce a full chart with all possible interpretations.

However, when tested on a real corpus (i.e., Wall Street Journal (WSJ) news stories), this mechanism collapses. For a typical well-behaved 33-word sentence it produces hundreds of candidate interpretations.

To alleviate problems associated with processing real text, a new strategy has emerged. A pre-processor, capitalizing on statistical data [Church *et al.*, 1989; Zernik and Jacobs, 1990; Dagan *et al.*, 1991], and trained to exploit properties of the corpus itself, could highlight regularities, identify thematic relations, and in general, feed digested text into the unification parser.

In this paper we investigate how a parser can be aided in the analysis of multiple content-word strings, which are problematic since they do not include syntax "sugar" in the form of function words.

What is Pre-Processing Up Against?

The Linguistic Phenomenon

Consider the following Wall Street Journal (WSJ), (August 19, 1987) paragraph processed by the NLcp pre-processor [Zernik *et al.*, 1991].

> Separately, Kaneb Services **spokesman/nn said/vb holders/nn** of its Class A **preferred/jj stock/nn failed/vb** to elect two directors to the **company/nn board/nn** when the **annual/jj meeting/nn resumed/vb** Tuesday because there are questions as to the validity of the **proxies/nn submitted/vb** for review by the group.
>
> The **company/nn adjourned/vb** its **annual/jj meeting/nn** May 12 to **allow/vb time/nn** for ne-

[1] In this discussion, a collocation is defined as a pair of cooccurring content words

[2] This research was sponsored (in part) by the Defense Advanced Research Project Agency (DOD) and other government agencies. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Project Agency or the US Government.

[3] We thank ACL/DCI (Data Collection Initiative), the Collins publishing company, the Wall Street Journal, for providing invaluable on-line data, and the TreeBank project for providing tagged corpus for reference.

gotiations and **expressed/vb concern/nn** about **future/jj actions/nn** by **preferred/vb holders/nn.**

The task under investigation is the classification of content-word pairs into one of three categories.

1. and **expressed/VB concern/NN** about
2. Services **spokesman/NN said/VB** holders
3. class A **preferred/JJ stock/NN** *comma*

The constructs *expressed concern* and *spokesman said* must be tagged verb-object and subject-verb respectively. *Preferred stock*, on the other hand, must be identified and tagged as a fixed adjective-noun construct.

An Architecture for Text-Processing

Text processing proceeds through the following stages [Zernik and Krupka, submitted 1992]:

Training-Time Thematic Analysis: High-frequency collocations are collected from a large corpus. A thematic-relation database (250,000 items) is constructed, based on the diversity of each collocation in the corpus.

Processing-Time Tagging:

- Perform lexical analysis based on Collins on-line dictionary.
- Perform initial tagging based on a fixed set of knowledge-base rules. Difficult cases such as content-word strings are left untagged.
- Based on the thematic-relation database, tag the collocations. Leave untagged cases not covered by the database.

Processing-Time Parsing: Perform syntactic analysis of the tagged text by a unification parser [Tomita, 1986].

The training and the consequent tagging of collocations are addressed in this paper.

The Input: Ambiguous Lexical Tags

The complex scope of the pre-processing task is best illustrated by the input to the pre-processor shown in Figure 1. This lexical analysis of the sentence is based on the Collins on-line dictionary (about 49,000 lexical entries extracted by NLcp) plus morphology. Each word is associated with *candidate* parts of speech, and almost all words are ambiguous. The tagger's task is to resolve the ambiguity.

Ambiguous words such as *services, preferred,* and *expressed,* should be resolved as noun (*nn*), adjective (*jj*), and verb (*vb*), respectively. While some pairs (e.g., *annual meeting*) can be resolved easily, other pairs (e.g., *preferred stock* and *expressed concerns*) are more difficult, and require statistical training.

Part-Of-Speech Resolution

A program can bring to bear 3 types of clues in resolving part-of-speech ambiguity:

Local context: Consider the following 2 cases where local context dominates:

1. **the** preferred stock raised
2. **he** expressed concern about

The words *the* and *he* dictate that *preferred* and *expressed* are adjective and verb respectively. This kind of inference, due to its local nature, is captured and propagated by the pre-processor.

Global context: Global-sentence constraints are shown by the following two examples:

1. and preferred stock sold yesterday **was** ...
2. and expressed concern about ... ***period***

In case 1, a main verb is found (i.e., *was*), and *preferred* is taken as an adjective; in case 2, a main verb is not found, and therefore *expressed* itself is taken as the main verb. This kind of ambiguity requires full-fledged unification, and it is not handled by the pre-processor. Fortunately, only a small percent of the cases (in newspaper stories) depend on global reading.

Thematic Analysis: Corpus analysis provides certain preferences [Beckwith *et al.*, 1991]

collocation	**total**	**vb-nn**	**jj-nn**
preferred stock	2314	100	0
expressed concern	318	1	99

The construct *expressed concern*, which appears 318 times in the corpus, is almost always (99 times out of 100 counted cases) a verb-noun construct; on the other hand, *preferred stock*, which appears in the corpus 2314 times, is 100 times out of 100 an adjective-noun, construct.

Figure 2 which illustrates the use of a fixed and a variable collocation in context, motivates the need for thematic analysis. In this small sample, 8 out of 35 cases (the ones marked "-") cannot be resolved reliably by using local context only. Without using thematic analysis, a tagger will produce arbitrary tags for *taking* and *operating*.

Indeed, existing statistical taggers [Church *et al.*, 1989; Meteer *et al.*, 1991; Brill, 1992; Cutting *et al.*, 1992] which rely on bigrams or trigrams, but do not employ thematic analysis of individual collocations fare poorly on this linguistic aspect. [4]

Learning from Raw Corpus

A database of collocations must be put in place in order to perform educated thematic analysis as shown above.

[4] The univariate-analysis strategy [Brill, 1992] of using default single-word probability, is not successful in this case. All cases of *operating* would by default be tagged incorrectly as verb since the noun/verb ratio for *operating* is 454/331 in the 2-million word portion of WSJ manually tagged by the TreeBank project [Santorini, 1990]).

Kaneb	NM	Services	NN VB	spokesman	NN
said	JJ VB	holders	NN	of	PP
its	DT	Class	JJ NN	A	DT JJ
preferred	JJ VB	stock	NN VB	failed	AD VB
to	PP	elect	VB	two	JJ NN
directors	NN	to	PP	the	DT
company	NN	board	NN VB	when	CC
annual	JJ	meeting	NN VB	resumed	JJ VB
tuesday	NM	questions	NN VB	validity	NN
proxies	NN	submitted	JJ VB	group	NN VB

Figure 1: Lexical Analysis of Sentence: Words plus Parts of Speech

```
e latest version of the UNIX V  operating  system software and some  -
th Microsoft 's MS *slash* DOS  operating  system *period* Microsoft -
ties obtained licenses for the  operating  system *period* With the  +
nths before IBM can provide an  operating  system that taps its mach +
  *comma* much as Microsoft 's  operating  system software is now th +
r *colon* eta systems inc. its  operating  system has not been debug +
cyber uses an unusual internal  operating  system *s-colon* to sell  +
*hyphen* Telegraph Co. 's UNIX  operating  system *comma* fast becom -
willing to suffer with a crude  operating  system *period*           +
at someday the Macintosh II 's  operating  system would be enhanced  +
phen* compatible computers and  operating  systems has created an op -

 allow the equity investors to  take  advantage of federal tax benef +
spect that some countries will  take  advantage of the option to pay +
 *comma* probably will want to  take  advantage of an option such as +
 scheduling *comma* some might  take  advantage of the opportunity t +
ed that rotated 360 degrees to  take  advantage of the view *period* +
th cheap local deposits and by  taking  advantage of its low overhea -
ins by nimbly trading zeros to  take  advantage of short *hyphen* te +
dexes and futures contracts to  take  advantage of various differenc +
itional financing *s-colon* to  take  advantage of future business o +
pendent publishers *comma* and  take  advantage of our considerable  +
olon* but if brazil decides to  take  advantage of any price rally * +
that some practical jokers had  taken  advantage of the offer *dash* +
onent systems on time *period*  Taking  advantage of changing demogr -
ravel plans by a few months to  take  advantage of the low fares *pe +
tic producers can successfully  take  advantage of the tax to eke ou +
ing lobbyists and scurrying to  take  advantage of the current hosti +
  homeowners 's refinancing to  take  advantage of lower interest ra +
g complete pc systems *period*  Taking  advantage of their lower *hy -
 rally came from investors who  took  advantage of rising stock pric +
n part by investors rushing to  take  advantage of britain 's high c +
 *comma* stayed long enough to  take  advantage of the amenities tha +
ber of institutional investors  took  advantage of the rally to roll +
mma* mo *period* Companies are  taking  advantage of that to rebuild +
for example *dash* *dash* have  taken  advantage of the strong yen t +
```

Figure 2: KWIC Table for operating-system and take-advantage. Note the diverse inflections of *take advantage* compared with the fixed nature of *operating systems*. The sentences marked "+" can be tagged appropriately using local context. The sentences marked "-" cannot be tagged without thematic analysis. Unless the tagger is familiar with the appropriate phrases, it cannot determine whether the combination is verb-noun or adjective-noun.

Verb-Noun Relations

2	produced-car	387	expressed-concern	72	taken-advantage
9	produced-cars	25	expressed-concerns	22	takes-advantage
5	produces-cars	10	expresses-concern	995	take-advantage
4	produce-car	31	expressing-concern	2	take-advantages
13	produce-cars	3	expressing-concerns	260	taking-advantage
17	producing-cars	33	express-concern	159	took-advantage
2	production-cars				

Noun-Verb Relations

947	companies-said	118	analysts-note	51	spokesman-acknowledged
242	companies-say	192	analysts-noted	8	spokesman-acknowledges
13	companies-saying	192	analysts-noted	2	spokesman-acknowledging
135	companies-says	13	analysts-noting		
14146	company-said	79	analyst-noted		
43	company-say	6	analyst-notes		
20	company-saying	6	analyst-notes		
698	company-says	6	analyst-notes		
		9	analyst-noting		

Adjective-Noun Constructs

3491	joint-venture	3558	preferred-stock	2	operates-systems
807	joint-ventures	11	preferred-stocks	627	operating-system
2	joint-venturing			86	operating-systems
				2	operational-systems
				2	operates-system

Figure 3: Fixed and variable collocations. Fixed phrases (e.g., *preferred stocks*) allow only a narrow variance. Full-fledged thematic relations (i.e., *produced cars*) appear in a wide variety of forms.

Where Is the Evidence?

Ideally, the database could be acquired by counting frequencies over a tagged corpus. However, a sufficiently large tagged corpus is not available.

Other statistical taggers have required much smaller training texts: A database of univariate (i.e., single word) statistics can be collected from a 1-million word corpus [Brill, 1992]; a database of state-transitions for part-of-speech tagging can also be collected from a 1-million corpus ([Church *et al.*, 1989]), or even from a smaller 60,000-word corpus ([Meteer *et al.*, 1991]).

However, to acquire an adequate database of collocations, we needed the full 85-million WSJ corpus. As shown by [Church *et al.*, 1991], the events we are looking for, i.e., word cooccurrence, are much sparser than the events required for state-transition, or for univariate-statistics.

In conclusion, since no apriori tagged training corpus exists, there is no direct evidence regarding part-of-speech. All we get from the corpus are numbers that indicate frequency and mutual information score (MIS) [Church *et al.*, 1991] of collocations. It is necessary to infer the nature of combinations from indirect corpus-based statistics as shown by the rest of this paper.

Identifying Collocation Variability

The basic linguistic intuition of our analysis is given in KWIC tables such as Figure 2. In this table we compare the cooccurence of the pairs *operating-system* and *take-advantage*. The verb-noun collocation shows a diverse distribution while the adjective-noun collocation is quite unchanged.

A deeper analysis of variation analysis is presented in figure 3, which provides the frequencies found for each variant in the WSJ corpus. For example, *joint venture* takes 3 variants totaling 4300 instances, out of which 4288 are concentrated in 2 patterns, which in effect (stripping the plural S suffix) are a single pattern. For *produce car* no single pattern holds more than 21% of the cases. Thus, when more than 90% of the phrases are concentrated in a single pattern we classify it as a fixed adjective-noun (or noun-noun) phrase. Otherwise, it is classified as a noun-verb (or verb-noun) thematic relation.

Training-Time Thematic Analysis

Training over the corpus requires inflectional morphology. For each collocation *P* it P's Variability Factor *VF(P)* is calculated according to the following formula:

$$VF(P) = \frac{fW(plural(P)) + fW(singular(P))}{fR(stemmed(P))}$$

Where *fW(plural(P))* means the word frequency of the plural form of the collocation; *fW(singular(P))* means the frequency of the singular form of the collocation; *fR(stemmed(P))* means the frequency of the stemmed collocation. The *VF* for *produced cars* is given as an example:

$$\begin{aligned} VF(produced - cars) &= \\ \frac{fW(produced - cars) + fW(produced - car)}{fR(produce - car)} &= \\ \frac{2+9}{2+9+5+4+13+17+2} &= \\ \frac{11}{52} &= 0.21 \end{aligned}$$

Accordingly, $VF(producing - car) = VF(producing - cars) = 0.32$; and VF(produce-car) is (by coincidence) 0.32. In contrast, VF(joint-venture) is 1.00. A list of the first 38 content-word pairs encountered in the the Joint-Venture corpus is shown in Figure 4. The figure illustrates the frequency of each collocation P in the corpus relative to its stem frequency. The ratio, called VF, is given in the first column. The second and third columns present the collocation and its frequency. The fourth and fifth column present the stemmed collocation and its frequency. The sixth column presents the mutual information score.

Notice that fixed collocations are easily distinguishable from thematic relations. The smallest *VF* of a fixed collocation has a VF of 0.86 (finance specialist); the largest *VF* of a thematic relation is 0.56 (produce concrete). Thus, a threshold, say 0.75, can effectively be established.

Processing-Time Tagging

Relative to a database such as in Figure 4, the tagging algorithm proceeds as follows, as the text is read word by word:

1. Use local-context rules to tag words. When no rule applies for tagging a word, **then** tag the word "??" ("untagged").
2. **If** the last word pair is a collocation (e.g., *holding companies*), **and**
   one of the two words is tagged "??",
   then generate the S-stripped version (i.e., *holding company*), and the affix-stripped version (i.e., *hold company*).
3. Look up database.
   (a) **If** neither collocation is found, **then** do nothing;
   (b) **if** only affix-stripped collocation is found, or
       if VF (variability factor) is smaller than threshold, **then**
       tag first word a verb and the second word a noun;
   (c) **If** VF is larger than threshold, **then** tag adjective-noun or noun-noun (depending on lexical properties of word, i.e., running vs. meeting).

Checking for the noun-verb case is symmetrical (in step 2.b). The threshold is different for each suffix and should be determined experimentally (initial threshold can be taken as 0.75).

VF(P)	*P*	*fW(P)*	*stemmed(P)*	*fR(st'd(P))*	*MIS(P)*
1.00	business-brief	10083	business-brief	10083	9.95
1.00	joint-ventures	4298	joint-venture	4300	12.11
1.00	aggregates-operation	9	aggregate-operation	9	5.84
0.56	produce-concrete	5	produce-concrete	9	4.59
1.00	crushed-stones	12	crush-stone	12	11.08
0.00	forming-ventures	0	form-venture	44	5.50
0.00	leases-equipment	0	lease-equipment	12	4.35
1.00	composite-trading	10629	composite-trade	10629	9.41
1.00	related-equipment	65	relate-equipment	65	5.28
0.17	taking-advantage	260	take-advantage	1510	9.25
0.99	electronics-concern	482	electronic-concern	485	6.87
1.00	work-force	2014	work-force	2014	7.79
0.00	beginning-operation	0	begin-operation	160	4.11
1.00	makes-additives	5	make-additive	5	4.39
1.00	lubricating-additive	4	lubricate-additive	4	14.66
0.18	showed-signs	62	show-sign	339	6.28
1.00	telephone-exchange	66	telephone-exchange	66	5.56
0.95	holding-company	7752	hold-company	8124	6.21
1.00	phone-equipment	51	phone-equipment	51	6.02
1.00	phone-companies	572	phone-company	572	5.56
0.93	venture-partner	140	venture-partner	150	6.17
0.26	report-net	283	report-net	1072	6.10
1.00	net-income	9759	net-income	9759	10.54
1.00	home-appliance	96	home-appliance	96	11.01
0.99	brand-name	683	brand-name	687	8.98
0.96	product-lines	965	product-line	1009	7.12
1.00	equity-stake	266	equity-stake	266	6.65
1.00	earning-asset	46	earn-asset	46	4.46
1.00	problem-loans	252	problem-loan	252	5.10
0.86	finance-specialists	30	finance-specialist	35	5.06
1.00	finished-products	93	finish-product	93	5.79
1.00	mining-ventures	18	mine-venture	18	5.03
1.00	gas-industry	154	gas-industry	154	5.05
0.18	began-talks	27	begin-talk	152	4.56
0.55	produce-electricity	27	produce-electricity	49	6.14
1.00	power-plants	1353	power-plant	1353	8.12
1.00	oil-heating	14	oil-heat	14	4.01
0.97	contract-dispute	187	contract-dispute	193	6.64

Figure 4: Thematic-Relations Database: Each collocation is associated with a Variability Factor (VF). A high VF indicates a fixed construct while a low VF (under 0.75) indicates a verb-noun thematic relation.

Notice that local-context rules override corpus preference. Thus, although ***preferred stocks*** is a fixed construct, in a case such as ***John preferred stocks***, the algorithm will identify ***preferred*** as a verb.

Evaluation

The database was generated over the WSJ corpus (85-million words). The database retained about 250,000 collocations (collocations below a certain MIS are dropped). The count was performed over the Tipster Joint-Venture 1988 corpus (66,186 words). In the evaluation, only content words (i.e., verbs, nouns, adverbs, and adjectives, totaling 36,231 words) are observed.

Out of 36,231 content words, 1,021 are left untagged by the tagger due to incomplete coverage.

12,719 of the words in the text fall into collocations (of 2 or more content words). 6,801 of these words are resolved by local context rules. 4,652 of these words are resolved by thematic analysis. The remaining 1266 are untagged.

Part-of-speech accuracy is 97%, estimated by checking 1000 collocations. A mismatch between adjective and noun was not counted as an error.

Problematic Cases

Our algorithm yields incorrect results in two problematic cases.

Ambiguous Thematic Relations: Collocations that entertain both subject-verb and verb-object relations, i.e., *selling-companies* (as in "the company sold its subsidiary ..." and "he sold companies ...").

Interference: Coinciding collocations such as: *market-experience* and *marketing-experience*, or *ship-agent* and *shipping-agent*.

Fortunately, these cases are very infrequent.

Limitations

Adjectives and nouns are difficult to distinguish in raw corpus (unless they are marked as such lexically). For example, since the lexicon marks ***light*** as both adjective and noun, there is no visible difference in the corpus between ***light/JJ beer*** and ***light/NN bulb***. Our algorithm tags both ***light*** cases as a noun.

Corpus Size and Database Size

Two parameters are frequently confused when assessing tagging effectiveness: training-time corpus size and run-time database size.

A larger training corpus improves both coverage (the number of cases that are tagged) because more collocations have been encountered. It also improves precision (the number of cases that are tagged correctly) since for each collocation, more variations have been analyzed.

In order to acommodate the tagger to a specific architecture (20Mbyte SPARC, in our case), the program might be linked with only a partial database (low frequency collocations are removed). Cutting down on run-time database does not reduce precision. In the configuration evaluated above, the run-time tagger used only the most frequent 200,000 collocations out of the entire collection of 250,000.

Conclusions

We have presented a mechanism for injecting corpus-based preference in the form of thematic relations into syntactic text parsing. Thematic analysis (1) is crucial for semantic parsing accuracy, and (2) presents the weakest link of existing statistical taggers.

The algorithm presented in this paper capitalizes on the fact that text writers draw fixed phrases, such as ***cash flow***, ***joint venture***, and ***preferred stock***, from a limited vocabulary of collocations which can be captured in a database. Human readers, as well as computer programs, are successful in interpreting the text because they have previously encountered and acquired the embedded collocations.

Although the algorithm identifies fixed collocations as such, it allows local-context rules to override those corpus-based preferences. As a result, exceptional cases such as *he is operating systems*, or *he preferred stocks* are handled appropriately. It turns out that writers of a highly edited text such as WSJ know how to avoid potential false readings by making sure that exceptions are marked by local context "sugar".

Our general line of thinking follows [Church *et al.*, 1991; Beckwith *et al.*, 1991; Dagan *et al.*, 1991; Zernik and Jacobs, 1990; Smadja, 1991]: in order for a program to interpret natural language text, it must train on and exploit word connections in the text under interpretation.

References

R. Beckwith, C. Fellbaum, D. Gross, and G. Miller. Wordnet: A lexical database organized on psycholinguistic principles. In U. Zernik, editor, *Lexical Acquisition: Exploiting On-Line Dictionary to Build a Lexicon.* Lawrence Erlbaum Assoc., Hissdale, NJ, 1991.

Eric Brill. A simple rule-based part of speech taggers. In *Proceedings of Third Conference on Applied Natural Language Processing*, Morristown, NJ, 1992. Association for Computational Linguistics.

K. Church, W. Gale, P. Hanks, and D. Hindle. Parsing, word associations, and predicate-argument relations. In *Proceedings of the International Workshop on Parsing Technologies*, Carnegie Mellon University, 1989.

K. Church, W. Gale, P. Hanks, and D. Hindle. Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition: Using On-Line Resources to Build a*

Lexicon. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

D. Cutting, J. Kupiee, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proceedings of Third Conference on Applied Natural Language Processing*, Morristown, NJ, 1992. Association for Computational Linguistics.

I. Dagan, A. Itai, and U. Schwall. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, 1991.

M. Kay. Parsing in Functional Unification Grammar. In D. Dowty, L. Kartunnen, and A. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives.* Cambridge University Press, Cambridge, England, 1985.

M. Meteer, R. Schwartz, and R. Weischedel. Post: Using probabilities in language processing. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, 1991.

B. Santorini. Annotation manual for the pen treebank project. Technical report, University of Pennsylvania, Computer and Information Science, Philadelphia, PA, 1990.

S. Shieber. *An Introduction to Unification-based Approaches to Grammar.* Center for the Study of Language and Information, Palo Alto, California, 1986.

F. Smadja. Macrocoding the lexicon with cooccurrence knowledge. In U. Zernik, editor, *Lexical Acquisition: Using On-Line Resources to Build a Lexicon.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

M. Tomita. *Efficient Parsing for Natural Language.* Kluwer Academic Publishers, Hingham, Massachusetts, 1986.

U. Zernik and P. Jacobs. Tagging for learning. In *COLING 1990*, Helsinki, Finland, 1990.

U. Zernik and G. Krupka. Pre-processing for parsing: Is 95% accuracy good enough? submitted 1992.

U. Zernik, A. Dietsch, and M. Charbonneau. Imtoolset programmer's manual. Ge-crd technical report, Artificial Intelligence Laboratory, Schenectady, NY, 1991.