How Verb Subcategorization Frequencies Are Affected By Corpus Choice

Douglas Roland
University of Colorado
Department of Linguistics
Boulder, CO 80309-0295
Douglas.Roland@colorado.edu

Daniel Jurafsky
University of Colorado
Dept. of Linguistics & Inst. of Cognitive Science
Boulder, CO 80309-0295
jurafsky@colorado.edu

Abstract

The probabilistic relation between verbs and their arguments plays an important role in modern statistical parsers and supertaggers, and in psychological theories of language processing. But these probabilities are computed in very different ways by the two sets of researchers. Computational linguists compute verb subcategorization probabilities from large corpora while psycholinguists compute them from psychological studies (sentence production and completion tasks). Recent studies have found differences between corpus frequencies and psycholinguistic measures. We analyze subcategorization frequencies from four different corpora: psychological sentence production data (Connine et al. 1984), written text (Brown and WSJ), and telephone conversation data (Switchboard). We find two different sources for the differences. *Discourse influence* is a result of how verb use is affected by different discourse types such as narrative, connected discourse, and single sentence productions. *Semantic influence* is a result of different corpora using different senses of verbs, which have different subcategorization frequencies. We conclude that verb sense and discourse type play an important role in the frequencies observed in different experimental and corpus based sources of verb subcategorization frequencies.

1 Introduction

The probabilistic relation between verbs and their arguments plays an important role in modern statistical parsers and supertaggers (Charniak 1995, Collins 1996/1997, Joshi and Srinivas 1994, Kim, Srinivas, and Trueswell 1997, Stolcke et al. 1997), and in psychological theories of language processing (Clifton et al. 1984, Ferreira & McClure 1997, Garnsey et al. 1997, Jurafsky 1996, MacDonald 1994, Mitchell & Holmes 1985, Tanenhaus et al. 1990, Trueswell et al. 1993).

These probabilities are computed in very different ways by the two sets of researchers. Psychological studies use methods such as sentence completion and sentence production for collecting verb argument structure probabilities. In sentence completion, subjects are asked to complete a sentence fragment. Garnsey at al. (1997) used a proper name followed by a verb, such as "Debbie remembered _____." In sentence subjects are asked to write any sentence containing a given verb. An example of this type of study is Connine et al. (1984).

An alternative to these psychological methods is to use corpus data. This can be done automatically with unparsed corpora (Briscoe and Carroll 1997, Manning 1993, Ushioda et al. 1993), from parsed corpora such as Marcus et al.'s (1993) Treebank (Merlo 1994, Framis 1994) or manually as was done for COMLEX (Macleod and Grishman 1994). The advantage of any of these corpus methods is the much greater amount of data that can be used, and the much more natural contexts. This seems to make it preferable to data generated in psychological studies.

Recent studies (Merlo 1994, Gibson et al. 1996) have found differences between corpus frequencies and experimental measures. This suggests that corpus-based frequencies and experiment-based frequencies may not be interchangeable. To clarify the nature of the differences between various corpora and to find the causes of these differences, we analyzed

psychological sentence production data (Connine et al. 1984), written discourse (Brown and WSJ from Penn Treebank - Marcus et al. 1993), and conversational data (Switchboard - Godfrey et al. 1992). We found that the subcategorization frequencies in each of these sources are different. We performed three experiments to (1) find the causes of general differences between corpora, (2) measure the size of these differences, and (3) find verb specific differences. The rest of this paper describes our methodology and the two sources of subcategorization probability differences: *discourse influence* and *semantic influence.*

2 Methodology

For the sentence production data, we used the numbers published in the original Connine et al. paper as well as the original data, which we were able to review thanks to the generosity of Charles Clifton. The Connine data (CFJCF) consists of examples of 127 verbs, each classified as belonging to one of 15 subcategorization frames. We added a 16th category for direct quotations (which appeared in the corpus data but not the Connine data). Examples of these categories, taken from the Brown Corpus, appear in figure 1 below. There are approximately 14,000 verb tokens in the CFJCF data set.

For the BC, WSJ, and SWBD data, we counted subcategorizations using tgrep scripts based on the Penn Treebank. We automatically extracted and categorized all examples of the 127 verbs used in the Connine study. We used the same verb subcategorization categories as the Connine study. There were approximately 21,000 relevant verb tokens in the Brown Corpus, 25,000 relevant verb

1	[O]	Barbara **asked**, as they heard the front door close.
2	[PP]	Guerrillas were **racing** [toward him].
3	[inf-S]	Hank thanked them and **promised** [to observe the rules].
4	[inf-S]/PP/	Labor **fights** [to change its collar from blue to white].
5	[wh-S]	I **know** now [why the students insisted that I go to Hiroshima even when I told them I didn't want to].
6	[that-S]	She **promised** [that she would soon take a few day's leave and visit the uncle she had never seen, on the island of Oyajima --which was not very far from Yokosuka].
7	[verb-ing]	But I couldn't **help** [thinking that Nadine and Wally were getting just what they deserved].
8	[perception complement.]	Far off, in the dusk, he **heard** [voices singing, muffled but strong].
9	[NP]	The turtle immediately withdrew into its private council room to **study** [the phenomenon].
10	[NP][NP]	The mayor of the town **taught** [them] [English and French].
11	[NP][PP]	They **bought** [rustled cattle] [from the outlaw], kept him supplied with guns and ammunition, harbored his men in their houses.
12	[NP][inf-S]	She had assumed before then that one day he would **ask** [her] [to marry him].
13	[NP][wh-S]	I **asked** [Wisman] [what would happen if he broke out the go codes and tried to start transmitting one].
14	[NP][that-S]	But, in departing, Lewis **begged** [Breasted] [that there be no liquor in the apartment at the Grosvenor on his return], and he took with him the first thirty galleys of Elmer Gantry.
15	[passive]	A cold supper **was ordered** and a bottle of port.
16	Quotes	He **writes** ["Confucius held that in times of stress, one should take short views – only up to lunchtime."]

Figure 1 - examples of each subcategorization frame from Brown Corpus

tokens in the Wall Street Journal Corpus, and 10,000 in Switchboard. Unlike the Connine data, where all verbs were equally represented, the frequencies of each verb in the corpora varied. For each calculation where individual verb frequency could affect the outcome, we normalized for frequency, and eliminated verbs with less than 50 examples. This left 77 out of 127 verbs in the Brown Corpus, 74 in the Wall Street Journal, and only 30 verbs in Switchboard. This was not a problem with the Connine data where most verbs had approximately 100 tokens.

3 Experiment 1

The purpose of the first experiment is to analyze the general (non-verb-specific) differences between argument structure frequencies in the data sources. In order to do this, the data for each verb in the corpus was normalized to remove the effects of verb frequency. The average frequency of each subcategorization frame was calculated for each corpus. The average frequencies for each of the data sources were then compared.

3.1 Results

We found that the three corpora consisting of connected discourse (BC, WSJ, SWBD) shared a common set of differences when compared to the CFJCF sentence production data. There were three general categories of differences between the corpora, and all can be related to discourse type. These categories are:

(1) passive sentences
(2) zero anaphora
(3) quotations

3.1.1 Passive Sentences

The CFJCF single sentence productions had the smallest number of passive sentences. The connected spoken discourse in Switchboard had more passives, followed by the written discourse in the Wall Street Journal and the Brown Corpus.

Data Source	% passive sentences
CFJCF	0.6%
Switchboard	2.2%
Wall Street Journal	6.7%
Brown Corpus	7.8%

Passive is generally used in English to emphasize the undergoer (to keep the topic in subject position) and/or to de-emphasize the identity of the agent (Thompson 1987). Both of these reasons are affected by the type of discourse. If there is no preceding discourse, then there is no pre-existing topic to keep in subject position. In addition, with no context for the sentence, there is less likely to be a reason to de-emphasize the agent of the sentence.

3.1.2 Zero Anaphora

The increase in zero anaphora (not overtly mentioning understood arguments) is caused by two factors. Generally, as the amount of surrounding context increases (going from single sentence to connected discourse) the need to overtly express all of the arguments with a verb decreases.

Data Source	% [0] subcat frame
CFJCF	7%
Wall Street Journal	8%
Brown	13%
Switchboard	18%

Verbs that can describe actions (agree, disappear, escape, follow, leave, sing, wait) were typically used with some form of argument in single sentences, such as:

"I had a test that day, so I really wanted to escape from school." (CFJCF data).

Such verbs were more likely to be used without any arguments in connected discourse as in:

"She escaped , crawled through the usual mine fields, under barbed wire, was shot at, swam a river, and we finally picked her up in Linz." (Brown Corpus)

In this case, the argument of "escaped", ("imprisonment") was understood from the previous sentence. Verbs of propositional attitude (agree, guess, know, see, understand) are typically used transitively in written corpora and single-sentence production:

"I guessed the right answer on the quiz." (CFJCF).

In spoken discourse, these verbs are more likely to be used metalinguistically, with the previous

discourse contribution understood as the argument of the verb:

"I see." (Switchboard)
"I guess." (Switchboard)

3.1.3 Quotations

Quotations are usually used in narrative, which is more likely in connected discourse than in an isolated sentence. This difference mainly effects verbs of communication (e.g. answer, ask, call, describe, read, say, write).

Data Source	Percent Direct Quotation
CFJCF	0%
Switchboard	0%
Brown	4%
Wall Street Journal	6%

These verbs are used in corpora to discuss details of the contents of communication:

"Turning to the reporters, she asked, `Did you hear her?'"(Brown)

In single sentence production, they are used to describe the (new) act of communication itself :

"He asked a lot of questions at school." (CFJCF)

We are currently working on systematically identifying indirect quotes in the corpora and the CFJCF data to analyze in more detail how they fit in to this picture.

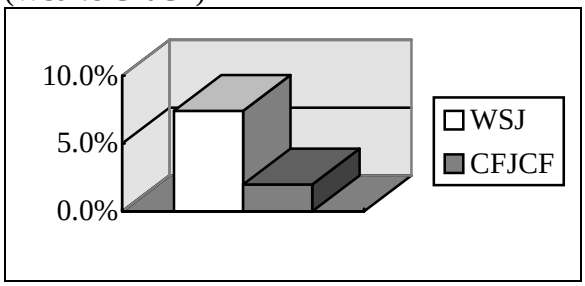
4 Experiment 2

Our first experiment suggested that discourse factors were the primary cause of subcategorization differences. One way to test this hypothesis is to eliminate discourse factors and see if this removes subcategorization differences.

We measure the difference between the way a verb is used in two different corpora by counting the number of sentences (per hundred) where a verb in one corpus would have to be used with a different subcategorization in order for the two corpora to yield the same subcategorization frequencies. This same number can also be calculated for the overall subcategorization frequencies of two corpora to show the overall difference between the two corpora.

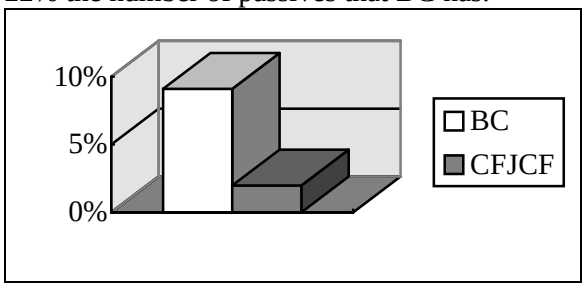
Our procedure for measuring the effect of discourse is as follows (illustrated using passive as an example):

1. Measure the difference between two corpora (WSJ vs CFJCF)



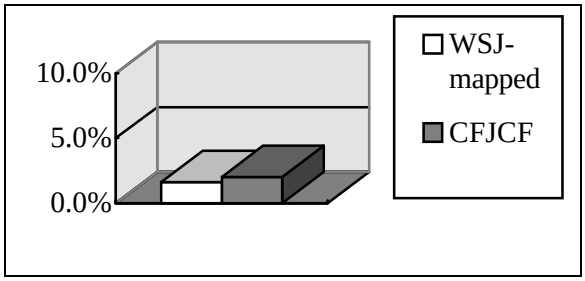
% Passive - WSJ vs CFJCF

2. Remove differences caused by discourse effects (based on BC vs CFJCF). CFJCF has 22% the number of passives that BC has.



% Passive - BC vs CFJCF

We then linearly scale the number of passives found in WSJ to reflect the difference found between BC and CFJCF.



% Passive - WSJ (adjusted) vs CFJCF

3. re-measure the difference between two corpora (WSJ vs CFJCF)
4. amount of improvement = size of discourse effect

This method was applied to the **passive**, **quote**, and **zero** subcat frames, since these are the ones that show discourse-based differences. Before

the mapping, WSJ has a difference of 17 frames/100 overall difference when compared with CFJCF. After the mapping, the difference is only 9.6 frames/100 overall difference. This indicates that 43% of the overall cross-verb differences between these two corpora are caused by discourse effects.

We use this mapping procedure to measure the size and consistency of the discourse effects. A more sophisticated mapping procedure would be appropriate for other purposes since the verbs with the best matches between corpora are actually made worse by this mapping procedure.

5 Experiment 3

Argument preference was also affected by verb semantics. To examine this effect, we took two sample ambiguous verbs, "charge" and "pass". We hand coded them for semantic senses in each of the corpora we used as follows:

Examples of 'charge' taken from BC.

accuse: "His petition charged mental cruelty."
attack: "When he charged Mickey was ready."
money: "... 20 per cent ... was all he charged the traders."

Examples of 'pass' taken from BC.

movement: "Blue Throat 's men spotted him ... as he passed."
law: "The President noted that Congress last year passed a law providing grants ..."
transfer: "He asked , when she passed him a glass."
test: "Those who T stayed had * to pass tests."

We then asked two questions:

1. Do different verb senses have different argument structure preferences?
2. Do different corpora have different verb sense preferences, and therefore potentially different argument structure preferences?

For both verbs examined (pass and charge) there was a significant effect of verb sense on argument structure probabilities (by X^2 p <.001 for 'charge' and p <.001 for 'pass'). The following chart shows a sample of this difference:

	that	NP	NP PP	passive
$Charge_{(accuse)}$	32	0	24	25
$Charge_{(money)}$	0	19	24	1

Sample Frames and Senses from WSJ

We then analyzed how often each sense was used in each of the corpora and found that there was again a significant difference (by X^2 p <.001 for 'charge' and p <.001 for 'pass').

	accuse	money	run/attack	other
BC	22	13	15	4
WSJ	88	69	1	7
SWBD	1	16	0	0

Senses of 'Charge' used in each corpus

	movement	law	transfer	test	other
BC	136	32	16	2	44
WSJ	11	76	31	8	22
SWBD	0	5	2	1	0

Senses of 'Pass' used in each corpus

This analysis shows that it is possible for shifts in the relative frequency of each of a verbs senses to influence the observed subcat frequencies.

We are currently extending our study to see if verb senses have constant subcategorization frequencies across corpora. This would be useful for word sense disambiguation and for parsing. If the verb sense is known, then a parser could use this information to help look for likely arguments. If the subcatagorization is known, then a disambiguator could use this information to find the sense of the verb. These could be used to bootstrap each other relying on the heuristic that only one sense is used within any discourse (Gale, Church, & Yarowsky 1992).

6 Evaluation

We had previously hoped to evaluate the accuracy of our treebank induduced subcategorization probabilities by comparing them with the COMLEX hand-coded probabilities (Macleod and

Grishman 1994), but we used a different set of subcategorization frames than COMLEX. Instead, we hand checked a random sample of our data for errors.

The error rate in our data is between 3% and 7% for all verbs excluding 'say' type verbs such as 'answer', 'ask', 'call', 'read', 'say', and 'write'. The error rate is given as a range due to the subjectivity of some types of errors. The errors can be divided into two classes; errors which are due to mis-parsed sentences in Treebank[1], and errors which are due to the inadequacy of our search strings in indentifying certain syntactic patterns.

Treebank-based errors	
PP attachment	1%
verb+particle vs verb+PP	2%
NP/adverbial distinction	2%
misc. miss-parsed sentences	1%
Errors based on our search strings	
missed traces and displaced arguments	1%
"say" verbs missing quotes	6%

Error rate by category

In trying to estimate the maximum amount of error in our data, we found cases where it was possible to disagree with the parses/tags given in Treebank. Treebank examples given below include prepositional attachment (1), the verb-particle/preposition distinction (2), and the NP/adverbial distinction (3).

1. "Sam, I thought you [knew [everything]$_{NP}$ [about Tokyo]$_{PP}$]" (BC)
2. "...who has since moved [on to other methods]$_{PP?}$" (BC)
3. "Gross stopped [briefly]$_{NP?}$**,** then went on." (BC)

Missed traces and displaced argument errors were a result of the difficulty in writing search strings to find arguments that were located to the left of the verb. This is because arbitrary amounts of structure can intervene, expecially in the case of traces.

Six percent of the data (overall) was improperly classified due to the failure of our search patterns to identify all of the quote-type arguments which occur in 'say' type verbs. The identification of these elements is particularly problematic due to the asyntactic nature of these arguments, ranging from a sound (He said 'Argh!') to complex sentences. The presence or absense of quotation marks was not a completely reliable indicator of these arguments. This type of error affects only a small subset of the total number of verbs. 27% of the examples of these verbs were mis-classified, always by failing to find a quote-type argument of the verb. Using separate search strings for these verbs would greatly improve the accuracy of these searches.

Our eventual goal is to develop a set of regular expressions that work on flat tagged corpora instead of TreeBank parsed structures to allow us to gather information from larger corpora than have been done by the TreeBank project (see Manning 1993 and Gahl 1998).

7 Conclusion

We find that there are significant differences between the verb subcategorization frequencies generated through experimental methods and corpus methods, and between the frequencies found in different corpora. We have identified two distinct sources for these differences. *Discourse influences* are caused by the changes in the ways language is used in different discourse types and are to some extent predictable from the discourse type of the corpus in question. *Semantic influences* are based on the semantic context of the discourse. These differences may be predictable from the relative frequencies of each of the possible senses of the verbs in the corpus. An extensive analysis of the frame and sense frequencies of different verbs across different corpora is needed to verify this. This work is presently being carried out by us and others (Baker, Fillmore, & Lowe 1998). It is certain, however, that verb sense and

[1] All of our search patterns are based only on the information available in the Treebank 1 coding system, since the Brown Corpus is only available in this scheme. The error rate for corpora available in Treebank 2 form would have been lower had we used all available information.

discourse type play an important role in the frequencies observed in different experimental and corpus based sources of verb subcategorization frequencies

Acknowledgments

This project was supported by the generosity of the NSF via NSF IRI-9704046 and NSF IRI-9618838 and the Committee on Research and Creative Work at the graduate school of the University of Colorado, Boulder. Many thanks to Giulia Bencini, Charles Clifton, Charles Fillmore, Susanne Gahl, Michelle Gregory, Uli Heid, Paola Merlo, Bill Raymond, and Philip Resnik.

References

Baker, C. Fillmore, C., & Lowe, J.B. (1998) Framenet. *ACL 1998*

Biber, D. (1993) Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics, 19/2, pp. 219-241.*

Briscoe T. and Carrol J. (1997) Automatic Extraction of Subcategorization from Corpora.

Charniak, E. (1997) Statistical parsing with a context-free grammar and word statistics. *Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI Press, Menlo Park.*

Clifton, C., Frazier, L., & Connine, C. (1984) Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 23, 696-708.*

Collins, M. J. (1996) A new statistical parser based on bigram lexical dependencies. *In Proceedings of ACL-96, 184--191, Santa Cruz, CA.*

Collins, M. J. (1997) Three generative, lexicalised models for statistical parsing. *In Proceedings of ACL-97.*

Connine, Cynthia, Fernanda Ferreira, Charlie Jones, Charles Clifton and Lyn Frazier. (1984) Verb Frame Preference: Descriptive Norms. *Journal of Psycholinguistic Research 13, 307-319*

Ferreira, F., and McClure, K.K. (1997). Parsing of Garden-path Sentences with Reciprocal Verbs. *Language and Cognitive Processes, 12, 273-306.*

Framis, F.R. (1994). An experiment on learning appropriate selectional restrictions from a parsed corpus. Manuscript.

Gahl, S. (1998). Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus. *Proceedings of ACL-98, Montreal.*

Gale, W.A., Church, K.W., and Yarowsky, D. (1992). One Sense Per Discourse. *Darpa Speech and Natural Language Workshop.*

Garnsey, S. M., Pearlmutter, N. J., Myers, E. & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37, 58-93.*

Gibson, E., Schutze, C., & Salomon, A. (1996). The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research 25(1), 59-92.*

Godfrey, J., E. Holliman, J. McDaniel. (1992) SWITCHBOARD : Telephone speech corpus for research and development. *Proceedings of ICASSP-92, 517--520, San Francisco.*

Joshi, A. & B. Srinivas. (1994) Disambiguation of super parts of speech (or supertags): almost parsing. *Proceedings of COLING '94.*

Juliano, C., and Tanenhaus, M.K. Contingent frequency effects in syntactic ambiguity resolution. *In proceedings of the 15th annual conference of the cognitive science society, LEA: Hillsdale, NJ.*

Jurafsky, D. (1996) A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20, 137-194.*

Lafferty, J., D. Sleator, and D. Temperley. (1992) Grammatical trigrams: A probabilistic model of link grammar. *In Proceedings of the 1992 AAAI Fall Symposium on Probabilistic Approaches to Natural Language.*

MacDonald, M. C. (1994) Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes 9.157--201.*

MacDonald, M. C., Pearlmutter, N. J. & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101, 676-703.*

Macleod, C. & Grishman, R. (1994) COMLEX Syntax Reference Manual Version 1.2. *Linguistic Data Consortium, University of Pennsylvania.*

Manning, C. D. (1993) Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. *Proceedings of ACL-93, 235-242.*

Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A.. (1993) Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics 19.2:313-330.*

Marcus, M. P., Kim, G. Marcinkiewicz, M.A., MacIntyre, R., Ann Bies, Ferguson, M., Katz, K., and Schasberger, B.. (1994) The Penn Treebank: Annotating predicate argument structure. *ARPA Human Language Technology Workshop, Plainsboro, NJ, 114-119.*

Meyers, A., Macleod, C., and Grishman, R.. (1995) Comlex Syntax 2.0 manual for tagged entries.

Merlo, P. (1994). A Corpus-Based Analysis of Verb Continuation Frequencies for Syntactic Processing. *Journal of Pyscholinguistic Research 23.6:435-457.*

Mitchell, D. C. and V. M. Holmes. (1985) The role of specific information about the verb in parsing sentences with local structural ambiguity. *Journal of Memory and Language 24.542--559.*

Stolcke, A., C. Chelba, D. Engle, V. Jimenez, L. Mangu, H. Printz, E. Ristad, R. Rosenfeld, D. Wu, F. Jelinek and S. Khudanpur. (1997) Dependency Language Modeling. *Center for Language and Speech Processing Research Note No. 24. Johns Hopkins University, Baltimore.*

Thompson, S. A. (1987) The Passive in English: A Discourse Perspective. *In Channon, Robert & Shockey, Linda (Eds.) In Honor of Ilse Lehiste/Ilse Lehiste Puhendusteos. Dordrecht: Foris, 497-511.*

Trueswell, J., M. Tanenhaus and C. Kello. (1993) Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference from Garden-Paths. *Journal of Experimental Psychology: Learning, Memory and Cognition 19.3, 528-553*

Trueswell, J. & M. Tanenhaus. (1994) Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. *In C. Clifton, K. Rayner & L. Frazier (Eds.) Perspectives on Sentence Processing. Hillsdale, NJ: Erlbaum, 155-179.*

Ushioda, A., Evans, D., Gibson, T. & Waibel, A. (1993) The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. *In Boguraev, B. & Pustejovsky, J. eds. SIGLEX ACL Workshop of Acquisition of Lexical Knowledge from Text. Columbus, Ohio: 95-106*