# Extracting and Evaluating General World Knowledge from the Brown Corpus

**Lenhart Schubert**
University of Rochester
schubert@cs.rochester.edu

**Matthew Tong**
University of Rochester
mt004i@mail.rochester.edu

## Abstract

We have been developing techniques for extracting general world knowledge from miscellaneous texts by a process of approximate interpretation and abstraction, focusing initially on the Brown corpus. We apply interpretive rules to clausal patterns and patterns of modification, and concurrently abstract general "possibilistic" propositions from the resulting formulas. Two examples are "A person may believe a proposition", and "Children may live with relatives". Our methods currently yield over 117,000 such propositions (of variable quality) for the Brown corpus (more than 2 per sentence). We report here on our efforts to evaluate these results with a judging scheme aimed at determining how many of these propositions pass muster as "reasonable general claims" about the world in the opinion of human judges. We find that nearly 60% of the extracted propositions are favorably judged according to our scheme by any given judge. The percentage unanimously judged to be reasonable claims by multiple judges is lower, but still sufficiently high to suggest that our techniques may be of some use in tackling the long-standing "knowledge acquisition bottleneck" in AI.

## 1 Introduction: deriving general knowledge from texts

We have been exploring a new method of gaining general world knowledge from texts, including fiction. The method does not depend on full or exact interpretation, but rather tries to glean general facts from particulars by combined processes of compositional interpretation and abstraction. For example, consider a sentence such as the following from the Brown corpus (Kucera and Francis, 1967):

*Rilly or Glendora had entered her room while she slept, bringing back her washed clothes.*

From the clauses and patterns of modification of this sentence, we can glean that an individual may enter a room, a female individual may sleep, and clothes may be washed. In fact, given the following Treebank bracketing, our programs produce the output shown:

```
((S
  (NP
    (NP (NNP Rilly) )
    (CC or)
    (NP (NNP Glendora) ))
  (AUX (VBD had) )
  (VP (VBN entered)
    (NP (PRP\$ her) (NN room) ))
  (SBAR (IN while)
    (S
      (NP (PRP she) )
      (VP (VBD slept) )))
  (\, \,)
  (S
    (NP (\-NONE\- \*) )
    (VP (VBG bringing)
      (PRT (RB back) )
      (NP (PRP\$ her) (JJ washed) (NNS clothes) ))))
 (\. \.) )

 A NAMED-ENTITY MAY ENTER A ROOM.
 A FEMALE-INDIVIDUAL MAY HAVE A ROOM.
 A FEMALE-INDIVIDUAL MAY SLEEP.
 A FEMALE-INDIVIDUAL MAY HAVE CLOTHES.
 CLOTHES CAN BE WASHED.

((:I (:Q DET NAMED-ENTITY) ENTER[V] (:Q THE ROOM[N]))
 (:I (:Q DET FEMALE-INDIVIDUAL) HAVE[V] (:Q DET ROOM[N]))
 (:I (:Q DET FEMALE-INDIVIDUAL) SLEEP[V])
 (:I (:Q DET FEMALE-INDIVIDUAL) HAVE[V]
                              (:Q DET (:F PLUR CLOTHE[N])))
 (:I (:Q DET (:F PLUR CLOTHE[N])) WASHED[A]))
```

The results are produced as logical forms (the last five lines above – see Schubert, 2002, for some details), from which the English glosses are generated automatically. Our work so far has focused on data in the Penn Treebank (Marcus et al., 1993), particularly the Brown corpus and some examples from the Wall Street Journal corpus. The advantage is that Treebank annotations allow us to postpone the challenges of reasonably accurate parsing, though we will soon be experimenting with "industrial strength" parsers on unannotated texts.

We reported some specifics of our approach and some preliminary results in (Schubert, 2002). Since then we have refined our extraction methods to the point where we can reliably apply them the Treebank corpora, on average extracting more than 2 generalized propositions per sentence. Applying these methods to the Brown corpus, we have extracted 137,510 propositions, of which 117,326 are distinct. Some additional miscellaneous examples are "A PERSON MAY BELIEVE A PROPOSITION", "BILLS MAY BE APPROVED BY COMMITTEES", "A US-STATE MAY HAVE HIGH SCHOOLS", "CHILDREN MAY LIVE WITH RELATIVES", "A COMEDY MAY BE DELIGHTFUL", "A BOOK MAY BE WRITE-ED (i.e., written) BY AN AGENT", "A FEMALE-INDIVIDUAL MAY HAVE A SPOUSE", "AN ARTERY CAN BE THICKENED", "A HOUSE MAY HAVE WINDOWS", etc.

The programs that produce these results consist of (1) a Treebank preprocessor that makes various modifications to Treebank trees so as to facilitate the extraction of semantic information (for instance, differentiating different kinds of "SBAR", such as S-THAT and S-ALTHOUGH, and identifying certain noun phrases and prepositional phrases, such as "next Friday", as temporal); (2) a pattern matcher that uses a type of regular-expression language to identify particular kinds of phrase structure patterns (e.g., verb + complement patterns, with possible inserted adverbials or other material); (3) a semantic pattern extraction routine that associates particular semantic patterns with particular phrase structure patterns and recursively instantiates and collects such patterns for the preprocessed tree, in bottom-up fashion; (4) abstraction routines that abstract away modifiers and other "type-preserving operators", before semantic patterns are constructed at the next-higher level in the tree (for instance, stripping the interpreted modifier "washed" from the interpreted noun phrase "her washed clothes"); (5) routines for deriving propositional patterns from the resulting miscellaneous semantic patterns, and rendering them in a simple, approximate English form; and (6) heuristic routines for filtering out many ill-formed or vacuous propositions. In addition, semantic interpretation of individual words involves some simple morphological analysis, for instance to allow the interpretation of (VBD SLEPT) in terms of a predicate SLEEP[V].

In (Schubert, 2002) we made some comparisons between our project and earlier work in knowledge extraction (e.g., (muc, 1993; muc, 1995; muc, 1998; Berland and Charniak, 1999; Clark and Weir, 1999; Hearst, 1998; Riloff and Jones, 1999)) and in discovery of selectional preferences (e.g., (Agirre and Martinez, 2001; Grishman and Sterling, 1992; Resnik, 1992; Resnik, 1993; Zernik, 1992; Zernik and Jacobs, 1990)). Reiterating briefly, we note that knowledge extraction work has generally employed carefully tuned extraction patterns to locate and extract some predetermined, specific kinds of facts; our goal, instead, is to process every phrase and sentence that is encountered, abstracting from it miscellaneous general knowledge whenever possible. Methods for discovering selectional preferences do seek out conventional patterns of verb-argument combination, but tend to "lose the connection" between argument types (e.g., that a road may carry traffic, a newspaper may carry a story, but a road is unlikely to carry a story); in any event, they have not led so far to amassment of data interpretable as general world knowledge.

Our concern in this paper is with the evaluation of the results we currently obtain for the Brown corpus. The overall goal of this evaluation is to gain some idea of what proportion of the extracted propositions are likely to be credible as world knowledge. The ultimate test of this will of course be systems (e.g., QA systems) that use such extracted propositions as part of their knowledge base, but such a test is not immediately feasible. In the meantime it certainly seems worthwhile to evaluate the outputs subjectively with multiple judges, to determine if this approach holds any promise at all as a knowledge acquisition technique.

In the following sections we describe the judging method we have developed, and two experiments based on this method, one aimed at determining whether "literary style makes a difference" to the quality of outputs obtained, and one aimed at assessing the overall success rate of the extraction method, in the estimation of several judges.

## 2 Judging the output propositions

We have created judging software that can be used by the researchers and other judges to assess the quality and correctness of the extracted information. The current scheme evolved from a series of trial versions, starting initially with a 3-tiered judging scheme, but this turned out to be difficult to use, and yielded poor inter-judge agreement. We ultimately converged on a simplified scheme, for which ease of use and inter-judge agreement are significantly better.

The following are the instructions to a judge using the judger program in its current form:

> Welcome to the sentence evaluator for the KNEXT knowledge extraction program. Thank you for your participation. You will be asked to evaluate a series of sentences based on such criteria as comprehensibility and truth. Do your best to give accurate responses. The judgement categories are selected to try to ensure that each sentence fits best in one and only one category. Help is available for each menu item, along with example sentences, by selecting 'h'; PLEASE consult this if this is your first time using this program even if you feel confident of your choice. There is also a tutorial available, which should also be done if this is your first time. If you find it hard to make a choice for a particular sentence even after carefully considering the alternatives, you should probably choose 6 (HARD TO JUDGE)! But if you strongly feel none of the choices fit a sentence, even after consulting the help file, please notify Matthew Tong (mtong@cs.rochester.edu) to allow necessary modifications to the menus or available help infor-

mation to occur. You may quit at any time by typing 'q'; if you quit partway through the judgement of a sentence, that partial judgement will be discarded, so the best time to quit is right after being presented with a new sentence.

<*here the first sentence to be judged is presented*>

---

1. SEEMS LIKE A REASONABLE GENERAL CLAIM (Of course. Yes.)
   A grand-jury may say a proposition. A report can be favorable.
2. SEEMS REASONABLE BUT EXTREMELY SPECIFIC OR OBSCURE (I suppose so)
   A surgeon may carry a cage. Gladiator pecs can be Reeves-type.
3. SEEMS VACUOUS (That's not saying anything)
   A thing can be a hen. A skiff can be nearest.
4. SEEMS FALSE (No. I don't think so. Hardly)
   A square can be round. Individual -s may have a world.
5. SOMETHING IS OBVIOUSLY MISSING (Give me a complete sentence)
   A person may ask. A male-individual may attach an importance.
6. HARD TO JUDGE (Huh?? How do you mean that? I don't know.)
   A female-individual can be psychic. Supervision can be with a company.

Based on this judging scheme, we performed two types of experiments: an experiment to determine whether literary style significantly impacts the percentage of propositions judged favorably; and experiments to assess overall success rate, in the judgement of multiple judges. We obtained clear evidence that literary style matters, and achieved a moderately high success rate – but certainly sufficiently high to assure us that large numbers of potentially useful propositions are extracted by our methods. The judging consistency remains rather low, but this does not invalidate our approach. In the worst case, hand-screening of output propositions by multiple judges could be used to reject propositions of doubtful validity or value. But of course we are very interested in developing less labor-intensive alternatives. The following subsections provide some details.

## 2.1 Dependence of extracted propositions on literary style

The question this experiment addressed was whether different literary styles correlated with different degrees of success in extracting intuitively reasonable propositions. The experiment was carried out twice, first by one of the authors (who was unaware of the contents of the files being sampled) and the second time by an outside recruit. While further experimentation is desirable, we believe that the evidence from two judges that literary style correlates with substantial differences in the perceived quality of extracted propositions demonstrates that future work on larger corpora should control the materials used for literary style.

Judgements were based on 4 Brown files (ck01, ck13, cd02, cd01). The 4 files were chosen by one of us on purely subjective grounds. Each contains about 2,200 words of text. (Our extraction methods yield about 1 proposition for every 8 words of text. So each file yields about 250-300 propositions.) The first two, ck01 and ck13, are straightforward, realistic narratives in plain, unadorned English, while cd01 and cd02 are philosophical and theological essays employing much abstract and figurative language. The expectation was that the first two texts would yield significantly more propositions judged to be reasonable general claims about the world than the latter two. To give some sense of the contents, the first few sentences from each of the texts are extracted here:

**Initial segments of each of the four texts**

**ck01:** Scotty did not go back to school. His parents talked seriously and lengthily to their own doctor and to a specialist at the University Hospital– Mr. McKinley was entitled to a discount for members of his family– and it was decided it would be best for him to take the remainder of the term off, spend a lot of time in bed and, for the rest, do pretty much as he chose– provided, of course, he chose to do nothing too exciting or too debilitating. His teacher and his school principal were conferred with and everyone agreed that, if he kept up with a certain amount of work at home, there was little danger of his losing a term.

**ck13:** In the dim underwater light they dressed and straightened up the room, and then they went across the hall to the kitchen. She was intimidated by the stove. He found the pilot light and turned on one of the burners for her. The gas flamed up two inches high. They found the teakettle. And put water on to boil and then searched through the icebox.

**cd01:** As a result, although we still make use of this distinction, there is much confusion as to the meaning of the basic terms employed. Just what is meant by "spirit" and by "matter"?? The terms are generally taken for granted as though they referred to direct and axiomatic elements in the common experience of all. Yet in the contemporary context this is precisely what one must not do. For in the modern world neither "spirit" nor "matter" refer to any generally agreed-upon elements of experience...

**cd02:** If the content of faith is to be presented today in a form that can be "understanded of the people"– and this, it must not be forgotten, is one of the goals of the perennial theological task– there is no other choice but to abandon completely a mythological manner of representation. This does not mean that mythological language as such can no longer be used in theology and preaching. The absurd notion that demythologization entails the expurgation of all mythological concepts completely misrepresents Bultmann's intention.

Extracted propositions were uniformly sampled from the 4 files, for a total count of 400, and the number of judgements in each judgement category were then separated out for the four files. In a preliminary version of this experiment, the judgement categories were still the 3-level hierarchical ones we eventually dropped in favor of a 6-alternatives scheme. Still, the results clearly indicated that the "plain" texts yielded significantly more propositions judged to be reasonable claims than the more abstract texts. Two repetitions of the experiment (with newly sampled propositions from the 4 files) using the 6-category judging scheme, and the heuristic postprocessing and filtering routines, yielded the following unequivocal results. (The exact sizes of the samples from files ck01, ck13, cd01, and cd02 in both repetitions were 120, 98, 85, and 97 respectively, where the relatively high count for ck01 reflects the relatively high count of extracted propositions for that text.)

- For ck01 and ck13 around 73% of the propositions (159/218 for judge 1 and 162/218 for judge 2) were judged to be in the "reasonable general claim" category; for cd01 and cd02, the figures were much lower, at 41% (35/85 for judge 1 and 40/85 for judge 2) and less than 55% (53/97 for judge 1 and 47/97 for judge 2) respectively.

- For ck01 and ck13 the counts in the "hard to judge" category were 12.5-15% (15-18/120) and 7.1-8.2% (6-7/85) respectively, while for cd01 and cd02 the figures were substantially higher, viz., 25.9-28.2% (22-24/85) and 19.6-23% (19-34/97) respectively.

Thus, as one would expect, simple narrative texts yield more propositions recognized as reasonable claims about the world (nearly 3 out of 4) than abstruse analytical materials (around 1 out of 2). The question then is then how to control for style when we turn our methods to larger corpora. One obvious answer is to hand-select texts in relevant categories, such as literature for young readers, or from authors whose writings are realistic and stylistically simple (e.g., Hemingway). However, this could be quite laborious since large literary collections available online (such as the works in Project Gutenberg, `http://promo.net/pg/`, `http://www.thalasson.com/gtn/`, with expired copyrights) are not sorted by style. Thus we expect to use automated style analysis methods, taking account of such factors as vocabulary (checking for esoteric vocabulary and vocabulary indicative of fairy tales and other fanciful fiction), tense (analytical material is often in present tense), etc. We may also turn our knowledge extraction methods themselves to the task: if, for instance, we find propositions about animals talking, it may be best to skip the text source altogether.

### 2.2 Overall quality of extracted propositions

To assess the quality of extracted propositions over a wide variety of Brown corpus texts, with judgements made by multiple judges, the authors and three other individuals made judgements on the same set of 250 extracted propositions. The propositions were extracted from the third of the Brown corpus (186 files) that had been annotated with WordNet senses in the SEMCOR project (Landes et al., 1998) (chiefly because those were the files at hand when we started the experiment – but they do represent a broad cross-section of the Brown Corpus materials). We excluded the cj-files, which contain highly technical material.

Table 1 shows the judgements of the 5 judges (as percentages of counts out of 250) in each of the six judgement categories. The category descriptions have been mnemonically abbreviated at the top of the table. Judge 1 appears twice, and this represents a repetition, as a test of self-consistency, of judgements on the same data presented in different randomized orderings.

| | reasonable | obscure | vacuous | false | incomplete | hard |
|---|---|---|---|---|---|---|
| judge 1 | 60.0 | 9.6 | 9.6 | 0.4 | 7.6 | 12.8 |
| judge 1 | 61.6 | 9.6 | 9.2 | 0.4 | 7.2 | 11.6 |
| judge 2 | 58.4 | 4.4 | 0.8 | 2.8 | 10.0 | 23.2 |
| judge 3 | 54.8 | 10.4 | 14.8 | 5.6 | 8.8 | 5.2 |
| judge 4 | 64.0 | 6.4 | 3.2 | 7.6 | 10.0 | 8.4 |
| judge 5 | 49.0 | 8.4 | 22.5 | 4.8 | 2.8 | 12.4 |

**Table 1. Judgements (in %) for 250 randomly sampled propositions**

As can be seen from the first column, the judges placed about 49-64% of the propositions in the "reasonable general claim" category. This result is consistent with the results of the style-dependency study described above, i.e., the average lies between the ones for "straightforward" narratives (which was nearly 3 out of 4) and the ones for abstruse texts (which was around 1 out of 2). This is an encouraging result, suggesting that mining general world knowledge from texts can indeed be productive.

One point to note is that the second and third judgement categories need not be taken as an indictment of the propositions falling under them – while we wanted to distinguish overly specific, obscure, or vacuous propositions from ones that seem potentially useful, such propositions would not corrupt a knowledge base in the way the other categories would (false, incomplete, or incoherent propositions). Therefore, we have also collapsed our data into three more inclusive categories, namely "true" (collapsing the first 3 categories), "false" (same as the original "false" category), and "undecidable" (collapsing the last two categories). The corresponding variant of Table 1 would thus be obtained by summing the first 3 and last 2 columns. We won't do so explicitly, but it is easy to verify that the proportion of "true" judgements comprise about three out of four judgements, when averaged over the 5 judges.

We now turn to the extent of agreement among the judgements of the five judges (and judge 1 with himself on the same data). The overall pairwise agreement results for classification into six judgement catagories are shown in Table 2.

| | judge 1 | judge 2 | judge 3 | judge 4 |
|---|---|---|---|---|
| judge 1 | 90.1 | | | |
| judge 2 | 60.1 | | | |
| judge 3 | 56.9 | 10.4 | | |
| judge 4 | 61.7 | 62.4 | 54.5 | |
| judge 5 | 58.5 | 57.3 | 56.0 | 49.3 |

**Table 2.** **Overall % agreement among judges for 250 propositions**

A commonly used metric for evaluating interrater reliability in categorization of data is the kappa statistic (Carletta, 1996). As a concession to the popularity of that statistic, we compute it in a few different ways here, though – as we will explain – we do not consider it particularly appropriate. For 6 judgement categories, kappa computed in the conventional way for pairs of judges ranges from .195 to .367, averaging .306. For 3 (more inclusive) judgement categories, the pairwise kappa scores range from .303 to .462, with an average of .375.

These scores, though certainly indicating a positive correlation between the assessments of multiple judges, are well below the lower threshold of .67 often employed in deciding whether judgements are sufficiently consistent across judges to be useful. However, to see that there is a problem with applying the conventional statistic here, imagine that we could improve our extraction methods to the point where 99% of extracted propositions are judged by miscellaneous judges to be reasonable general claims. This would be success beyond our wildest dreams – yet the kappa statistic might well be 0 (the worst possible score), if the judges generally reject a *different* one out of every one hundred propositions!

One somewhat open-ended aspect of the kappa statistic is the way "expected" agreement is calculated. In the conventional calculation (employed above), this is based on the observed average frequency in each judgement category. This leads to low scores when one category is overwhelmingly favored by all judges, but the exceptions to the favored judgement vary randomly among judges (as in the hypothetical situation just described). A possible way to remedy this problem is to use a uniform distribution over judgement categories to compute expected agreement. Under such an assumption, our kappa scores are significantly better: for 6 categories, they range from .366 to .549, averaging .482; for 3 categories, they range from .556 to .730, averaging .645. This approaches, and for several pairs of judges exceeds, the minimum threshold for significance of the judgements.[1]

Since the ideal result, as implied above, would be agreement by multiple judges on the "reasonableness" or truth of a large proportion of extracted propositions, it seems worthwhile to measure the extent of such agreement as well. Therefore we have also computed the "survival rates" of extracted propositions, when we reject those not judged to be reasonable general claims by $n$ judges (or, in the case of 3 categories, not judged to be true by $n$ judges). Figure 1 shows the results, where the survival rate for $n$ judges is averaged over all subsets of size $n$ of the 5 available judges.

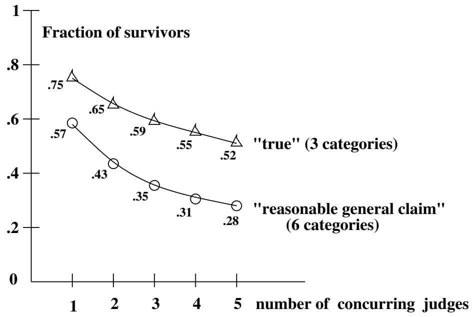


**Figure 1.** **Fraction of propositions placed in best category by multiple judges**

Thus we find that the survival rate for "reasonable general claims" starts off at 57%, drops to 43% and then 35% for 2 and 3 judges, and drops further to 31% and 28% for 4 and 5 judges. It appears as if an asymptotic level above 20% might be reached. But this may be an unrealistic extrapolation, since virtually *any* proposition, no matter how impeccable from a knowledge engineering perspective, might eventually be relegated to one of the other 5 categories by *some* uninvolved judge. The survival rates based on 2 or 3 judges seem to us more indicative of the likely proportion of (eventually) useful propositions than an extrapolation to infinitely many judges. For the 3-way judgements, we see that 75% of extracted propositions are judged "true" by individual judges (as noted earlier), and this drops to 65% and then 59% for 2 and 3 judges. Though again sufficiently many judges may eventually bring this down to 40% or less, the survival rate is certainly high enough to support the claim that our method of deriving propositions from texts can potentially deliver very large amounts of world knowledge.

[1] The fact that for some pairs of judges the kappa-agreement (with this version of kappa) exceeds 0.7 indicates that with more careful training of judges significant levels of agreement could be reached consistently.

## 3 Conclusions and further work

We now know that large numbers of intuitively reasonable general propositions can be extracted from a corpus that has been bracketed in the manner of the Penn Treebank. The number of "surviving" propositions for the Brown corpus, based on the judgements of multiple judges, is certainly in the tens of thousands, and the duplication rate is a rather small fraction of the overall number (about 15%).

Of course, there is the problem of screening out, as far as possible, the not-so-reasonable propositions. One step strongly indicated by our experiment on the effect of style is to restrict extraction to the kinds of texts that yield higher success rates – namely those written in straightforward, unadorned language. As we indicated, both style analysis techniques and our own proposition extraction methods could be used to select stylistically suitable materials from large online corpora.

Even so, a significant residual error rate will remain. There are two remedies – a short-term, brute-force remedy, and a longer-term computational remedy. The brute-force remedy would be to hand-select acceptable propositions. This would be tedious work, but it would still be far less arduous than "dreaming up" such propositions; besides, most of the propositions are of a sort one would not readily come up with spontaneously ("A person may paint a porch", "A person may plan an attack", "A house may have a slate roof", "Superstition may blend with fact", "Evidence of shenanigans may be gathered by a tape recorder", etc.)

The longer-term computational remedy is to use a well-founded parser and grammar, providing syntactic analyses better suited to semantic interpretation than Treebank trees. Our original motivation for using the Penn Treebank, apart from the fact that it instantly provides a large number of parsed sentences from miscellaneous genres, was to determine how readily such parses might permit semantic interpretation. The Penn Treebank pays little heed to many of the structural principles and features that have preoccupied linguists for decades. Would these turn out to be largely irrelevant to semantics? We were actually rather pessimistic about this, since the Treebank data tacitly posit tens of thousands of phrase structure rules with inflated, heterogeneous right-hand sides, and phrase classifications are very coarse (notably, with no distinctions between adjuncts and complements, and with many clause-like constructs, whether infinitives, subordinate clauses, clausal adverbials, nominalized questions, etc., lumped together as "SBAR" – and these are surely semantically crucial distinctions). So we are actually surprised at our degree of success in extracting sensible general propositions on the basis of such rough-and-ready syntactic annotations.

Nonetheless, our extracted propositions in the "something missing" and "hard to judge" categories do quite often reflect the limitations of the Treebank analyses. For example, the incompleteness of the proposition "A male-individual may attach an importance" seen above as an illustration of judgement category 5 can be attributed to the lack of any indication that the PP[to] constituent of the verb phrase in the source sentence is a verb complement rather than an adjunct. Though our heuristics try to sort out complements from adjuncts, they cannot fully make up for the shortcomings of the Treebank annotations. It therefore seems clear that we will ultimately need to base knowledge extraction on more adequate syntactic analyses than those provided by the Brown annotations.

Another general conclusion concerns the ease or difficulty of broad-coverage semantic interpretation. Even though our interpretive goals up to this point have been rather modest, our success in providing rough semantic rules for much of the Brown corpus suggests to us that full, broad-coverage semantic interpretation is not very far out of reach. The reason for optimism lies in the "systematicity" of interpretation. There is no need to hand-construct semantic rules for each and every phrase structure rule. We were able provide reasonably comprehensive semantic coverage of the many thousands of distinct phrase types in Brown with just 80 regular-expression patterns (each aimed at a class of related phrase types) and corresponding semantic rules. Although our semantic rules do omit some constituents (such as prenominal participles, non-initial conjuncts in coordination, adverbials injected into the complement structure of a verb, etc.) and gloss over subtleties involving gaps (traces), comparatives, ellipsis, presupposition, etc., they are not radical simplifications of what would be required for full interpretation. The simplicity of our outputs is due not so much to oversimplification of the semantic rules, as to the deliberate abstraction and culling of information that we perform in extracting general propositions from a specific sentence. Of course, what we mean here by semantic interpretation is just a mapping to logical form. Our project sheds no light on the larger issues in text understanding such as referent determination, temporal analysis, inference of causes, intentions and rhetorical relations, and so on. It was the relative independence of the kind of knowledge we are extracting of these issues that made our project attractive and feasible in the first place.

Among the miscellaneous improvements under consideration are the use of lexical distinctions and WordNet abstraction to arrive at more reliable interpretations; the use of modules to determine the types of neuter pronouns and of traces (e.g., in "She looked in the cookie jar, but it was empty", we should be able to abstract the proposition that a cookie jar may be empty, using the referent of "it"); and extracting properties of events by making use of information in adverbials (e.g.,, from "He slept soundly" we should be able to abstract the proposition that sleep may be sound; also many *causal* propositions can be inferred from adverbial constructions). We also hope to demonstrate extraction results through knowledge elicitaton questions (e.g.,"What do you know about books?", etc.)

## 4 Acknowledgements

The authors are grateful to David Ahn for contributing ideas and for extensive help in preparing and processing Brown corpus files, conducting some of the reported experiments, and performing some differential analyses of results. We also benefited from the discussions and ideas contributed by Greg Carlson and Henry Kyburg in the context of our "Knowledge Mining" group, and we appreciate the participation of group members and outside recruits in the judging experiments. As well, we thank Peter Clark and Phil Harrison (at Boeing Company) for their interest and suggestions. This work was supported by the National Science Foundation under Grant No. IIS-0082928.

## References

Eneko Agirre and David Martinez. 2001. Learning class-to-class selectional preferences. In *Proc. of the 5th Workshop on Computational Language Learning (CoNLL-2001)*, Toulouse, France, July 6-7,.

Matthew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proc. of the 37th Ann. Meet. of the Assoc. for Computational Linguistics (ACL-99)*, Univ. of Maryland, June 22 - 27,.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Stephen Clark and David Weir. 1999. An iterative approach to estimating frequencies over a semantic hierarchy. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Also available at http://www.cogs.susx.ac.uk/users/davidw/research/papers.html.

Ralph Grishman and John Sterling. 1992. Acquisition of selectional patterns. In *Proc. of COLING-92*, pages 658–664, Nantes, France.

Marti A. Hearst. 1998. Automated discovery of WordNet relations. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 131–153? MIT Press.

H. Kucera and W.N. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Shari Landes, Claudia Leacock, and Randee I. Tengi. 1998. Building semantic concordances. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages chapter 8, 199–216. MIT Press, Cambridge, MA.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June.

1993. *Proc. of the 5th Message Understanding Conference (MUC-5)*. Morgan Kaufmann, Los Altos, CA.

1995. *Proc. of the 6th Message Understanding Conference (MUC-6)*. Morgan Kaufmann, Los Altos, CA.

1998. *Proc. of the 7th Message Understanding Conference (MUC-7)*. Morgan Kaufmann, Los Altos, CA, April 29 – May 1, Virginia.

P. Resnik. 1992. A class-based approach to lexical discovery. In *Proc. of the 30th Ann. Meet. of the Assoc. for Computational Linguistics (ACL-92)*, pages 327–329, Newark, DE.

P. Resnik. 1993. Semantic classes and syntactic ambiguity. In *Proc. of ARPA Workshop on Human Language Technology*, Plainsboro, NJ.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proc. of the 16th Nat. Conf. on Artificial Intelligence (AAAI-99)*.

Lenhart K. Schubert. 2002. Can we derive general world knowledge from texts? In *Proc. of 2nd Int. Conf. on Human Language Technology Research (HLT 2002)*, pages 94–97, San Diego, CA, March 24-27.

Uri Zernik and Paul Jacobs. 1990. Tagging for learning: Collecting thematic relations from corpus. In *Proc. of the 13th Int. Conf. on Computational Linguistics (COLING-90)*, pages 34–39, Helsinki.

Uri Zernik. 1992. Closed yesterday and closed minds: Asking the right questions of the corpus to distinguish thematic from sentential relations. In *Proc. of COLING-92*, pages 1304–1311, Nantes, France, Aug. 23-28,.