# Experience in WordNet Sense Tagging in the Wall Street Journal

**Janyce Wiebe†, Julie Maples†, Lei Duan†, and Rebecca Bruce‡**
†Dept. of Computer Science and the Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003
‡Dept. of Computer Science and Engineering
Southern Methodist University
Dallas, TX 75275-0112
wiebe,jmaples,lduan@crl.nmsu.edu, rbruce@seas.smu.edu
*Proc. ANLP-97 Workshop, Tagging Text with Lexical Semantics: Why, What, How?*
Association for Computational Linguistics SIGLEX, Washington, D.C., April 1997, pp. 8-11

**Abstract**

This paper reports on our experience hand tagging the senses of 25 of the most frequent verbs in 12,925 sentences of the Wall Street Journal Treebank corpus (Marcus et al. 1993). The verbs are tagged with respect to senses in WordNet (Miller 1990). Some of the annotated verbs can function as both main and auxiliary verbs, and some are often used in idioms. This paper suggests consistently representing these as separate subclasses. Strategies described in the coding instruction for recognizing idioms are described, as well as some challenging ambiguities found in the data.

## 1 Introduction

This paper reports on our experience hand tagging the senses of 25 of the most frequent verbs in 12,925 sentences of the Wall Street Journal Treebank corpus (Marcus et al. 1993). The purpose of this work is to support related work in automatic word-sense disambiguation.

The verbs are tagged with respect to senses in WordNet (Miller 1990), which has become widely used, for example in corpus-annotation projects (Miller et al. 1994, Ng & Hian 1996, and Grishman et al. 1994) and for performing disambiguation (Resnik 1995 and Leacock et al. 1993).

The verbs to tag were chosen on the basis of how frequently they occur in the text, how wide their range of senses, and how distinguishable the senses are from one another.

In related work, we have begun to tag nouns and adjectives as well. These are being chosen additionally on the basis of co-occurrence with the verbs already tagged, to support approaches such as (Hirst 1987), in which word-sense ambiguities are resolved with respect to one another.

Some of the chosen verbs can function as both main and auxiliary verbs, and some are often used in idioms. In this paper, we suggest consistently representing these as separate subclasses.

We apply a preprocessor to the data, which automatically identifies some classes of verb occurrence with good accuracy. This facilitates manual annotation, because it is easier to fix a moderate number of errors than to tag the verbs completely from scratch. The preprocessor performs other miscellaneous tasks to aide in the tagging task, such as separating out punctuation marks and contractions.

At the end of the paper, we share some strategies from our coding instructions for recognizing idioms, and show some challenging ambiguities we found in the data.

## 2 The Verbs and the Basic Tag Format

The following are the verbs that were tagged. The total number of occurrences is 6,197.

| VERB | NUMBER | VERB | NUMBER |
|---|---|---|---|
| have | 2740 | make | 473 |
| take | 316 | get | 231 |
| add | 118 | pay | 189 |
| see | 159 | call | 151 |
| decline | 84 | hold | 127 |
| come | 191 | give | 168 |
| keep | 101 | know | 87 |
| find | 130 | lose | 82 |
| believe | 103 | raise | 124 |
| drop | 61 | lead | 105 |
| work | 101 | leave | 81 |
| run | 105 | look | 95 |
| meet | 75 | | |

The basic tags have the following form. Extensions will be given below.

```
word_<lemma, WordNet POS, WordNet sense number>
```

For example:

> The Sacramento-based S&L had_⟨have verb 4⟩ assets of $2.4 billion at the end of September.

That is, 'had' is a form of the main verb 'have' occurring as WordNet sense number 4.

## 3 Refinements

We consistently break out certain uses of verbs to a greater extent than WordNet does, in particular, idioms and verbs of intermediate (and auxiliary) function. There are several reasons for doing so.

The primary reason is to perform more accurate tagging. Not all such uses are covered by WordNet entries.

A second reason is to support identifying better features for automatic tagging. Some of these special-case uses can be identified with good accuracy with simple grammars, while the more semantically weighty uses of the same verb generally cannot be. Thus, different features will be appropriate for the special-case versus other uses. By tagging them as separate categories, one can search for separate features characterizing each class.

Finally, it is helpful to the human tagger for the preprocessor to target these distinguished classes, for which relatively high-accuracy automatic solutions are possible.

### 3.1 Auxiliary Verbs

WordNet does not provide sense information for auxiliary uses of verbs. SEMCOR (Miller et al. 1994) leaves these uses untagged. Among the verbs tagged in our corpus, only 'have' has an auxiliary use, which we tag as follows, with the string "aux" replacing the sense number:

> South Korea has_⟨have verb_aux⟩ recorded a trade surplus of 71 million dollars so far this year.

As they can be recognized automatically with high accuracy, auxiliaries are automatically annotated by the preprocessor.

### 3.2 Intermediate Verbs

"Intermediate verb" is a term used in Quirk et al. (1985; pp. 138-148), defined as an occurrence "whose status is in some degree intermediate between auxiliaries and main verbs." Quirk et al. arrange verbs on a scale ranging from modal auxiliaries to main verbs, and "many of the intermediate verbs, particularly those at the higher end of the scale, have meanings associated with aspect, tense, and modality: meanings which are primarily expressed through auxiliary verb constructions."

Among the verbs tagged in our corpus, 'had', 'get', and 'keep' are used with intermediate function in the following constructions: 'had better' (or 'had best') and 'have got to' (called "modal idioms" by Quirk et al.), 'have to' (called a "semi-auxiliary"), 'get' + -ed participle, and 'keep' + -ing participle (which are given the title "catenatives").

Some but not all of these are represented by senses in WordNet (and none are identified as having this special function). Since WordNet senses cannot be consistently assigned to these occurrences, we use a new tag, "int", in place of a sense number (or in addition to one, when there is an appropriate sense), creating a new category, as we did with the auxiliaries.

An example of an intermediate verb occurrence is the following. Note that sense 5 of 'have' is an appropriate WordNet sense for this occurrence:

> Apple II owners, for example, had_⟨have_to verb_int 5⟩ to use their television sets as screens and stored data on audiocassettes.

These intermediate occurrences can also be recognized with good accuracy, and so are also added to the corpus by the preprocessor.

The auxiliary and intermediate uses of 'have' together represent well over half of the occurrences, so breaking these out as separate categories enables the preprocessor to assist the tagger greatly. In addition, it would allow for separate evaluation of an automatic classifier tagging 'have'.

## 4 Verb Idioms

### 4.1 Manual Annotation

The occurrence of a variety of verb idioms—semantic units consisting of a verb followed by a particle or other modifying word—accounted for a recognizable segment—about 6%– of the tagged data. For example:

> The borrowing to raise these funds would be paid_⟨pay_off verb 1⟩ off as assets of sick thrifts are sold.

WordNet does not provide entries for all idioms, and the entries it does provide do not always include a sense for the occurrences observed in the corpus.

It is important to recognize idioms, because interpreting their constituent words separately would often change the meaning of the sentence (cf., e.g., Wilks 1977 and Wilensky & Arens 1980). Our coding instructions specify that the tagger should attempt to identify idioms even if WordNet does not provide an entry for it. The preprocessor assists in this task, by identifying potential idioms.

The following are strategies we found useful in dealing with the difficult problem of manually identifying idioms.

1. Does the word following the verb cease to have any of its usual or literal meanings as supplied by WordNet when used with that verb?

   > If America can keep_⟨keep_up verb 1⟩ up the present situation ... the economies of these countries would be totally restructured to be able to almost sustain growth by themselves.

   The 'situation' here does not need to be kept in a lofty position, but rather maintained. The use of 'up' as a particle takes away its literal, physical meaning, and attaches it semantically to 'keep', making an idiom definition necessary.

2. Could the idiom be replaced with a single verb which has the same meaning?

   For example:

   > But the New York Stock Exchange chairman said he doesn't support reinstating a "collar" on program trading, arguing that firms could get_⟨get_around verb 2⟩ around such a limit.

   WordNet's entry for this sense of "get around" includes as synonyms "avoid" and "bypass", which, if used in place of the idiom, do not change the meaning of the sentence.

3. Would the particle be mistaken for a preposition beginning a prepositional phrase–and thereby changing the meaning of the sentence–if viewed as separate from the main verb?

   Consider this example:

   > Coleco failed to come_⟨come_up_with verb 1⟩ up with another winner and filed for bankruptcy-law protection ...

   This example actually meets all three criteria. 'Come up with' must be considered a single idiom partly to avoid a literal interpretation that would change the meaning of the sentence, as described in criterion (1), and it also has the meaning "locate", which further qualifies this sentence as an idiom according to criterion (2). If this sentence were given a literal reading, perhaps by an automatic tagger, 'with another winner' might be identified as an acceptable prepositional phrase.

### 4.2 A Flexible Tag Format

For the purposes of the larger project of which this annotation project is a part, the words are annotated with information in addition to the WordNet sense tags. A simple example is the richer part-of-speech tags produced by Brill's tagger (1992). We note here a problem that we encountered using SEMCOR's tag format for idioms: SEMCOR merges the component words of the idiom into one annotation, thereby making it impossible to unambiguously represent information about the individual words. Representing split idioms is also a problem with this scheme.

To maintain separate annotations and also tie the constituents of an idiom together, we suggest the format below (or an analogous one), which is generated by the preprocessor. The annotations for the individual words are delimited by "⟨wf" and "⟨/wf ⟩". The human annotator's tags are included in the individual word annotations. For example, below the annotator tagged "take" with the first WordNet entry for 'take place'. When there is an appropriate WordNet entry for the idiom as a whole, we store that entry with the first word of the idiom (but the entry could be stored with both). Appropriate WordNet entries for the individual words can also be stored in the individual word annotations. The Brill part-of-speech tags illustrate other information we would like to retain for the individual words.

```
<wf BrillPOS=VBD  idiom=take_place-1
   wnentry=_<take_place verb 1>>took</wf>
<wf pos=NN idiom=take_place-2>place</wf>
```

The first two lines contain the annotation for the first word in the idiom. It contains a Brill POS tag for 'take' and a WordNet entry for 'take place'. The string 'take-place-1' encodes the fact that this is the first word of a 'take place' idiom.

The third line represents the second word in the idiom ('take-place-2'), which is a noun ('NN').

An intervening adverb, for example, would simply be represented with its own annotation placed between the annotations for the words in the idiom.

## 5 Challenging Ambiguities

There are some instances in the corpus that we found to be truly ambiguous. These instances support two completely different interpretations even with the help of the context. For example:

The group has_⟨have verb 1?aux⟩ forecast 1989 revenue of 56.9 billion francs.

In this sentence, two interpretations of the verb 'have' are equally possible, even when the sentence is viewed in context: 'Have' can be seen as an auxiliary, meaning that the group have themselves done the forecasting, or as WordNet sense 1 (in which case 'forecast' is an adjective), implying that someone else has given them an amount, 56.9 billion francs, that represents their expected revenue.

A problem found several times in the corpus occurred when a single verb is used in a sentence that has two objects, and each object suggests a different sense of the verb. In the sentence below, for example, two senses of the main verb 'have' are represented simultaneously in the sentence. Sense 4 carries the idea of ownership, which should be applied to the object 'papers', while sense 3 has the meaning "to experience or receive", which should be applied to the object 'sales'.

PAPERS: Backe Group Inc. agreed to acquire Atlantic Publications Inc., which has_⟨have verb 4|14⟩ 30 community papers and annual sales of $7 million.

Such cases are borderline, hovering in between two distinct meanings.

## 6 Conclusion

Data manually annotated with lexical semantics clearly has many applications in NLP. This paper shared our experience in manual annotation of WordNet senses in the Wall Street Journal Treebank corpus. WordNet proved to be a valuable and useful tool. Its wide range of senses made possible a highly specific level of tagging. WordNet's structure, with the alignment of hierarchical information and the addition of synsets and sample sentences, was especially helpful. We have made some suggestions for consistently identifying certain uses of verbs and for representing tags, and have shared some guidelines from our annotation instructions for identifying idioms in the corpus.

## 7 Acknowledgements

This research was supported in part by the Office of Naval Research under grant number N00014-95-1-0776.

## References

Brill, E. 1992. A simple rule-based part of speech tagger. In *Proc. of the Third Conference on Applied Natural Language Processing*, pp. 152–155.

Grishman, R., Macleod, C., & Meyers, A. (1994). COMPLEX syntax: building a computational lexicon. In *Proc. 15th International Conference of Computational Linguistics (COLING 94)*.

Hirst, G. 1987. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge: Cambridge University Press.

Leacock, C., Towell, G., Voorhees, E. 1993. In *Proc. of the ARPA Human Language Technology Workshop.*

Marcus, M., Santorini, B., & Marcinkiewicz, M. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics 19* (2): 313–330.

Miller, G. 1990. Special Issue, WordNet: An on-line lexical database. *International Journal of Lexicography 3* (4).

Miller, G., Chodorow, M., Landes, S., Leacock, C., & Thomas, R. 1994. Using a semantic concordance for sense identification. In *Proc. ARPA Human Language Technology Workshop.*

Ng, H.T. & Hian, B.L. 1996. Integrating multiple knowledge sources to disambiguate word senses: an exemplar-based approach. In *Proc. 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, pp. 40–47.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language.* (New York: Longman).

Resnik, P. 1995. Disambiguating noun groupings with respect to wordnet senses. In *Proc. of the Third Workshop on Very Large Corpora*, pp. 54–68.

Wilensky, R. & Arens, Y. 1980. PHRAN - a knowledge based natural language understander. In *Proc. 18th Annual Meeting of the Association for Computational Linguistics (ACL-80)*, pp. 117–121.

Wilks, Y. 1977. Making preferences more active. *Artificial Intelligence 8*, pp. 75–97.