Mapping Lexical Entries in a Verbs Database to WordNet Senses*

Rebecca Green, Lisa Pearl, Bonnie J. Dorr, and **Philip Resnik**
Department of Computer Science
University of Maryland

Abstract

This paper describes automatic techniques for mapping 9611 entries in a database of English verbs to WordNet senses. The verbs were initially grouped into 491 classes based on syntactic categories. Mapping these classified verbs into WordNet senses provides a resource that may be used for disambiguation in multilingual applications such as machine translation and cross-language information retrieval. Our techniques make use of (1) a training set of 1791 disambiguated entries, representing 1442 verb entries from 167 of the categories; (2) word sense probabilities based on frequency counts in a previously tagged corpus; (3) semantic similarity of WordNet senses for verbs within the same class; (4) probabilistic correlations between WordNet data and attributes of the verb classes. The best results achieved 72% precision and 58% recall, versus a lower bound of 62% precision and 38% recall for assigning the most frequently occurring WordNet sense, and an upper bound of 87% precision and 75% recall for human judgment.

1 Introduction

Our goal is to map entries in a lexical database of 4076 English verbs automatically to WordNet senses (Miller and Fellbaum, 1991), (Fellbaum, 1998) to better support applications such as machine translation and cross-language information retrieval. For example, the verb *drop* is multiply ambiguous and, thus, has many potential translations in Spanish: *bajar*, *caerse*, *dejar*, *caer*, *derribar*, *disminuir*, *echar*, *hundir*, *soltar*, etc. The lexical database specifies a set interpretations for the verb *drop*, depending on its context in the source-language (SL). Inclusion of WordNet senses into the lexical database enables the selection of an appropriate verb in the target language (TL). Final selection is based on a frequency count of WordNet senses across all semantic classes to which the verb belongs—e.g., *disminuir* is selected in the case where the WordNet sense corresponds to the meaning of *drop* in *Prices dropped*.

Our mapping approach is related to, but distinct from, the "lexical-sample" approach of SENSEVAL (Kilgarriff and Rosenzweig, 2000), where only a specific set of words from a body of text are to be disambiguated. This makes it possible to take advantage of detailed knowledge of specific senses of those words in performing the disambiguation. Given that the words occur in sentential contexts, it is assumed in SENSEVAL that only one word sense is accurate for each token. In contrast, our task takes an "all-words" approach, in which all words in the body of text (i.e., all entries in the lexical database) are to be disambiguated.

As we lack contextual data for the disambiguation task, we use instead information about verb senses encoded in terms of thematic grids and lexical-semantic representations from (Olsen et al., 1997). Given the lack of context, several WordNet senses may be equally appropriate for a database entry. Indeed, since distinctions between WordNet senses are often fine-grained (Palmer, 2000), it may be unclear, even in context, which of several senses is invoked or even if only one sense is invoked. When words occur out of context, it is that much more likely that more than one WordNet sense will be applicable.

The verb database contains mostly syntactic

[0]This report supersedes Rebecca Green, Lisa Pearl, and Bonnie J. Dorr, Mapping WordNet Senses to a Lexical Database of Verbs, CS-TR-4206, UMIACS-TR-2001-02, LAMP TR-061, University of Maryland, College Park, MD.

information about its entries, with much of that information applying at the level of the classes used within the database. WordNet, on the other hand, is a significant source for information about semantic relationships, with much of that information applying at the "synset" level ("synsets" are WordNet's groupings of synonymous word senses). Thus, by mapping entries in the database to their corresponding WordNet senses, the semantic potential of the verb database is extended significantly.

2 Lexical Resources

We use an existing classification of 4076 English verbs, based initially on *English Verbs Classes and Alternations* (Levin, 1993) and extended through the splitting of some classes into subclasses and the addition of new classes. The resulting 491 classes (e.g., "Roll Verbs, Group I", which includes *drift, drop, glide, roll, swing*) are referred to here as *Levin+ classes*. As verbs may be assigned to multiple Levin+ classes, the number of entries in the database is rather larger, viz., 9611.

Following the model of (Dorr and Olsen, 1997), each Levin+ class is associated with a *thematic grid* (henceforth abbreviated θ-grid) which summarizes a verb's syntactic behavior through specifying its predicate argument structure. For example, the Levin+ class "Roll Verbs, Group I" is associated with the θ-grid `[th goal]`, in which a theme and a goal are used (e.g., *The ball dropped to the ground*).[1] Each θ-grid specification corresponds to a *Grid class*. There are 48 Grid classes, with a one-to-many relationship between Grid and Levin+ classes.

WordNet, the lexical resource to which we are mapping entries from the lexical database, groups synonymous word senses into "synsets" and structures the synsets into part-of-speech hierarchies. Our mapping operation uses several other data elements pertaining to WordNet: semantic relationships between synsets, frequency data, and syntactic information.

Seven semantic relationship types exist between synsets, including, for example, antonymy, hyperonymy, and entailment. Synsets are often related to a half dozen or more other synsets; they may be related to multiple synsets through a single relationship or may be related to a single synset through multiple relationship types.

Our frequency data for WordNet senses is derived from SEMCOR—a semantic concordance incorporating tagging of the Brown corpus with WordNet senses.[2]

Syntactic patterns ("frames") are associated with each synset, e.g., Somebody ----s something; Something ----s; Somebody ----s somebody into V-ing something. There are 35 such verb frames in WordNet and a synset may have only one or as many as a half dozen or so frames assigned to it.

Our mapping of verbs in Levin+ classes to WordNet senses relies in part on the relation between thematic roles in Levin+ and verb frames in WordNet. Both reflect how many and what kinds of arguments a verb may take. However, constructing a direct mapping between θ-grids and WordNet frames is not possible, since the underlying classifications differ in significant ways. The correlations between the two sets of data are better viewed probabilistically—as will be described in Section 3.

Table 1 illustrates the relation between each of the resources above for the verb *drop*. In our multilingual applications (e.g., lexical selection in machine translation), the Grid information provides a context-based means of associating a verb with a Levin+ class according to its usage in the SL sentence. The WordNet sense possibilities are thus pared down during SL analysis, but not sufficiently for the final selection of a TL verb. For example, Levin+ class 9.4 has three possible WordNet senses for *drop*. However, the WordNet sense 8 is not associated with any of the other classes; thus, it is considered to have a higher "information content" than the others. The upshot is that the lexical-selection routine prefers *dejar caer* over other translations such as *derribar* and *bajar*.[3] The other classes are

[1] There is also a Levin+ class "Roll Verbs, Group II" which is associated with the θ-grid [th particle(down)], in which a theme and a particle 'down' are used (e.g., *The ball dropped down*).

[2] For further information see the WordNet manuals, section 7, SEMCOR at http://www.cogsci.princeton.edu.

[3] This lexical-selection approach is an adaptation of the notion of *reduction in entropy*, measured by information gain (Mitchell, 1997). Using information content to quantify the "value" of a class in the WordNet hierarchy has also been used for measuring semantic similarity

Levin+	Grid/Example	WN Sense	Spanish Verb(s)
9.4 Directional Put	[ag th mod-loc src goal] *I dropped the stone*	1. move, displace 2. descend, fall, go down 8. drop set down, put down	1. derribar, echar 2. bajar, caerse 8. dejar caer, echar, soltar
45.6 Calibratable Change of State	[th] *Prices dropped*	1. move, displace 3. decline, go down, wane	1. derribar, echar 3. disminuir
47.7 Meander	[th src goal] *The river dropped from the lake to the sea*	2. descend, fall, go down 4. sink, drop, drop down	2. bajar, caerse 4. hundir, caer
51.3.1 Roll I	[th goal] *The ball dropped to the ground*	2. descend, fall, go down	2. bajar, caerse
51.3.1 Roll II	[th particle(down)] *The ball dropped down*	2. descend, fall, go down	2. bajar, caerse

Table 1: Relation Between Levin+ and WN Senses for 'drop'

similarly associated with appropriate TL verbs during lexical selection: *disminuir* (class 45.6), *hundir* (class 47.7), and *bajar* (class 51.3.1).[4]

3 Training Data

As a starting point, we used the lexical database of (Dorr and Jones, 1996), which contains a significant number of WordNet-tagged verb entries. Some of the assignments were in doubt, since class splitting had occurred subsequent to those assignments, with all old WordNet senses having been carried over to new subclasses. New classes had also been added since the manual tagging. It was determined that the tagging for only 1791 entries—including 1442 verbs in 167 classes—could be considered stable; for these entries, 2756 assignments of WordNet senses had been made. Data for these entries, taken from both WordNet and the verb lexicon, constitute the training data for this study.

The following probabilities were generated from the training data:

- **Grid probability**$_x$ $= \frac{|\{r_x \ \& \ \text{Grid}_1 = \text{Grid}_2\}|}{|\{r_x\}|}$, where each occurrence of r_x involves the relating of one synset s_1 through relationship type x to another synset s_2, and where s_1 is mapped to by a verb in Grid class Grid$_1$ and s_2 is mapped to by a verb in Grid class Grid$_2$. This is the probability that if one synset is related to another through a particular relationship type, then a verb mapped to the first synset will belong to the same Grid class as a verb mapped to the second synset. Computed values generally range between .3 and .35.

- **Levin + probability**$_x$ $= \frac{|\{r_x \ \& \ \text{Levin+}_1 = \text{Levin+}_2\}|}{|\{r_x\}|}$, where r_x is as above, and where s_1 is mapped to by a verb in Levin+ class Levin+$_1$ and s_2 is mapped to by a verb in Levin+ class Levin+$_2$. This is the probability that if one synset is related to another through a particular relationship type, then a verb mapped to the first synset will belong to the same Levin+ class as a verb mapped to the second synset. Computed values generally range between .25 and .3.

- **Combo frame probability**$_{i,j}$ $= \frac{|\{\theta_{i,v} \ \& \ \text{cf}_{j,v}\}|}{|\{\theta_{i,v}\}|}$, where $\theta_{i,v}$ is the occurrence of θ-grid i for verb entry v and cf$_{j,v}$ is the occurrence of the entire frame sequence j for a WordNet sense to which verb entry v is mapped. This is the probability that a verb in a Levin+ class is mapped to a WordNet verb sense with some specific combination of frames. Values average only .11, but in some cases the probability is 1.0.

in a taxonomy (Resnik, 1999b). More recently, context-based models of disambiguation have been shown to represent significant improvements over the baseline (Bangalore and Rambow, 2000), (Ratnaparkhi, 2000).

[4] The full set of Spanish translations is selected from WordNet associations developed in the EuroWordNet effort (Dorr et al., 1997).

- **Indv frame probability**$_{i,j} = \frac{|\{\theta_{i,v} \ \& \ \mathrm{cf}_{j,v}\}|}{|\{\theta_{i,v}\}|}$, where $\theta_{i,v}$ is the occurrence of θ-grid component i for verb entry v and $\mathrm{cf}_{j,v}$ is the occurrence of the single frame j for a WordNet sense to which verb entry v is mapped. This is the probability that a verb in a Levin+ class with a particular θ-grid component (possibly among others) is mapped to a WordNet verb sense assigned a specific frame (possibly among others). Values average .20, but in some cases the probability is 1.0.

- **Prior WN probability**$_\mathbf{s} = \frac{|\{t_s\}|}{|\{t_v\}|}$, where t_s is an occurrence of tag s (for a particular synset) in SEMCOR and t_v is an occurrence of any of a set of tags for verb v in SEMCOR, with s being one of the senses possible for verb v. This probability is the prior probability of specific WordNet verb senses. Values average .11, but in some cases the probability is 1.0.

In addition to the foregoing data elements, based on the training set, we also made use of a semantic similarity measure, which reflects the confidence with which a verb, given the total set of verbs assigned to its Levin+ class, is mapped to a specific WordNet sense. This represents an implementation of a class disambiguation algorithm (Resnik, 1999a), modified to run against the WordNet verb hierarchy.[5]

We also made a rather powerful assumption (referred to hereafter as the "same-synset assumption"): If (1) two verbs are assigned to the same Levin+ class, (2) one of the verbs v_1 has been mapped to a specific WordNet sense s_1, and (3) the other verb v_2 has a WordNet sense s_2 that is synonymous with s_1, then v_2 should be mapped to s_2. Since WordNet groups synonymous word senses into "synsets," s_1 and s_2 would correspond to the same synset. Moreover, Levin+ verbs are mapped to WordNet senses via their corresponding synset identifiers. Thus, when the set of conditions enumerated above are met, the two verb entries should be mapped to the same WordNet synset.

As an example, the two verbs *tag* and *mark* have been assigned to the same Levin+ class. In WordNet, each occurs in five synsets, only one in which they both occur. If *tag* has a WordNet synset assigned to it for the Levin+ class it shares with *mark*, and it is the synset that covers senses of both *tag* and *mark*, we can safely assume that that synset is also appropriate for *mark*, since in that context, the two verb senses are synonymous.

4 Evaluation

Subsequent to the culling of the training set, several processes were undertaken that resulted in full mapping of entries in the lexical database to WordNet senses. Much, but not all, of this mapping was accomplished manually.

Each entry whose WordNet senses were assigned manually was considered by at least two coders, one coder who was involved in the entire manual assignment process and the other drawn from a handful of coders who worked independently on different subsets of the verb lexicon. In the manual tagging, if a WordNet sense was considered appropriate for a lexical entry by any one of the coders, it was assigned. Overall, 13452 WordNet sense assignments were made. Of these, 51% were agreed upon by multiple coders. The kappa coefficient (K) of intercoder agreement was .47 for a first round of manual tagging and (only) .24 for a second round of more problematic cases.[6]

While the full tagging of the lexical database may make the automatic tagging task appear superfluous, the low rate of agreement between coders and the automatic nature of some of the tagging suggest that there is still room for adjustment of WordNet sense assignments in the lexical database. On the one hand, even the higher of the kappa coefficients mentioned above is significantly lower than the standard suggested for good reliability ($K > .8$) or even the

[5]The assumption underlying this measure is that the appropriate word senses for a group of semantically related words should themselves be semantically related. Given WordNet's hierarchical structure, the semantic similarity between two WordNet senses corresponds to the degree of informativeness of the most specific concept that subsumes them both.

[6]The kappa statistic measures the degree to which pairwise agreement of coders on a classification task surpasses what would be expected by chance; the standard definition of this coefficient is: $K = (P(A) - P(E))/(1 - P(E))$, where $P(A)$ is the actual percentage of agreement and $P(E)$ is the expected percentage of agreement, averaged over all pairs of assignments. Several adjustments in the computation of the kappa coefficient were made necessary by the possible assignment of multiple senses for each verb in a Levin+ class, since without prior knowledge of how many senses are to be assigned, there is no basis on which to compute $P(E)$.

level where tentative conclusions may be drawn ($.67 < K < .8$) (Carletta, 1996), (Krippendorff, 1980). On the other hand, if the automatic assignments agree with human coding at levels comparable to the degree of agreement among humans, it may be used to identify assignments that should be reviewed and to suggest other assignments for further consideration.

In addition, there are consistency checks that can be made much more easily by the automatic process than can be made manually. For example, the same-synset assumption is much more easily enforced automatically than manually. When such WordNet sense assignments are made automatically on the basis of the 2756 senses in the training set, another 967 sense assignments are generated, only 131 of which were actually assigned manually. Similarly, when such a premise is enforced on the entirety of the lexical database of 13452 assignments, another 5059 sense assignments are generated. If the premise is valid and if the senses assigned in the database are accurate, then the human tagging has a recall of no more than 73%.

Because a word sense was assigned even if only one coder judged it to apply, human coding has been treated as having a precision of 100%. However, some of the solo judgments were likely to have been idiosyncratic. To determine what proportion of such judgments were in reality precision errors, a random sample of 50 WordNet senses supported by only one of the two original coders was investigated further by a team of three judges. In this round, judges rated the WordNet senses assigned to the verb entries as falling into one of three categories: definitely correct, definitely incorrect, and arguable whether correct. As it turned out, if any one of the judges rated a sense definitely correct, another judge independently judged it definitely correct; this accounts for 31 instances. In 13 instances the assignments were judged definitely incorrect by at least two of the judges. No consensus was reached on the remaining 6 instances. Extrapolating from this sample to the full set of judgments in the database supported by only one coder leads to an estimate that approximately 1725 (26% of 6636 solo judgments) of those senses are incorrect. This suggests that the precision of the human coding is approximately 87%.

The upper bound for this task, as set by human performance, is thus 73% recall and 87% precision. The lower bound, based on assigning the WordNet sense with the greatest prior probability, is 38% recall and 62% precision.

5 Mapping Strategies

Recent work (Van Halteren et al., 1998) has demonstrated improvement in part-of-speech tagging when the outputs of multiple taggers are combined. When the errors of multiple classifiers are not significantly correlated, the result of combining votes from a set of individual classifiers often outperforms the best result from any single classifier. Using a voting strategy seems especially appropriate here: Most of the data available for picking out WordNet senses for entries in the lexical database function as only weak indicators of correct senses; on average, they identify correct senses from the training data about 40% of the time. At the same time, there is significant variation in which senses they pick out.

The investigations undertaken here used both *simple* and *aggregate* voters, combined using various voting strategies. The simple voters were the 7 measures introduced above in the Training Data section.[7] In addition, three aggregate voters were generated: (1) the product of the seven simple measures (smoothed so that zero values wouldn't offset all other measures); (2) the weighted sum of the seven simple measures, with weights representing the percentage of the training set assignments correctly identified by the highest score of the simple probabilities; and (3) the maximum score of the seven simple measures.

Using these data, two different sorts of voting schemes were investigated. These schemes differ most significantly on the circumstances under which a voter casts its vote for a WordNet sense, the size of the vote cast by each voter, and the circumstances under which a WordNet sense was selected. We will refer to these two schemes as *Majority Voting Scheme* and *Threshold Voting Scheme*.

[7] There are actually 6 measures in this previous section (including the semantic similarity measure), but Indv frame probability is used in two different ways.

5.1 Majority Voting Scheme

Although we do not know in advance how many WordNet senses should be assigned to an entry in the lexical database, we assume that, in general, there is at least one. In line with this intuition, one strategy we investigated was to have both simple and aggregate measures cast a vote for whichever sense(s) of a verb in a semantic class received the highest (non-zero) value for that measure. Ten variations are given here:

- **PriorProb:** Prior Probability of WordNet senses
- **SemSim:** Semantic Similarity
- **SimpleProd:** Product of all simple measures
- **SimpleWtdSum:** Weighted sum of all simple measures
- **MajSimpleSgl:** Majority vote of all (7) simple voters
- **MajSimplePair:** Majority vote of all (21) pairs of simple voters[8]
- **MajAggr:** Majority vote of SimpleProd and SimpleWtdSum
- **Maj3Best:** Majority vote of SemSim, SimpleProd, and SimpleWtdSum
- **MajSgl+Aggr:** Majority vote of MajSimpleSgl and MajAggr
- **MajPair+Aggr:** Majority vote of MajSimplePair and MajAggr

Table 2 gives recall and precision measures for all variations of this voting scheme, both with and without enforcement of the same-synset assumption. The best voting scheme is MajAggr, based on the product and weighted-sum aggregate voters, with 58% recall and 72% precision without enforcement of the same-synset assumption. Note that if the same-synset assumption is correct, the drop in precision with its enforcement mostly reflects inconsistencies in human judgments in the training set; the true precision value for MajAggr is probably close to 67%.

[8]A pair cast a vote for a sense if, among all the senses of a verb, a specific sense had the highest value for both measures.

Variation	W/O SS		W/ SS	
	R	P	R	P
PriorProb	38%	62%	45%	46%
SemSim	56%	71%	60%	55%
SimpleProd	51%	74%	57%	55%
SimpleWtdSum	53%	77%	58%	56%
MajSimpleSgl	23%	71%	30%	48%
MajSimplePair	38%	60%	45%	43%
MajAggr	58%	72%	63%	53%
Maj3Best	52%	78%	57%	57%
MajSgl+Aggr	44%	74%	50%	54%
MajPair+Aggr	49%	77%	55%	57%

Table 2: Recall (R) and Precision (P) for Majority Voting Scheme, With and Without the Same-Synset assumption

Variation	R	P
AutoMap+	61%	54%
AutoMap-	61%	54%
Triples	63%	52%
Combo	53%	44%
Combo&Auto	59%	45%

Table 3: Recall (R) and Precision (P) for Threshold Voting Scheme

5.2 Threshold Voting Scheme

The second voting strategy commenced by identifying, for each simple and aggregate measure, the threshold value at which the product of recall and precision scores in the training set has the highest value if that threshold is used to select WordNet senses. During the voting, if a WordNet sense has a higher score for a measure than its threshold, the measure votes for the sense; otherwise, it votes against it. The weight of the measure's vote is the precision-recall product at the threshold. This voting strategy has the advantage of taking into account each individual attribute's strength of prediction.

Five variations on this basic voting scheme were investigated. In each, senses were selected if their vote total exceeded a variation-specific threshold. Table 3 summarizes recall and precision for these variations at their optimal vote thresholds.

The first variation, **AutoMap+**, implements the same-synset assumption taking Grid probability and Levin+ probability into account. The second variation, **AutoMap-**, differs in that it disregards the Grid and Levin+ probabilities. The **Triples** variation places the simple and composite measures into three groups, the three with the highest weights, the three with the lowest weights, and the middle or remaining three. Voting first occurred within the group, and the group's vote was brought forward with a weight equaling the sum of the group members' weights. This variation also added to the vote total if the sense had been assigned in the training data. The **Combo** variation is like Triples, but rather than using the weights and thresholds calculated for the single measures from the training data, this variation calculated weights and thresholds for combinations of two, three, four, five, six, and, seven measures. Finally, the **Combo&Auto** variation adds the same-synset assumption to the previous variation.

6 Conclusions and Future Work

The voting schemes still leave room for improvement, as the best results (58% recall and 72% precision, or, optimistically, 63% recall and 67% precision) fall shy of the upper bound of 73% recall and 87% precision for human coding.[9] At the same time, these results are far better than the lower bound of 38% recall and 62% precision for the most frequent WordNet sense.

As has been true in many other evaluation studies, the best results come from combining classifiers (MajAggr): not only does this variation use a majority voting scheme, but more importantly, the two voters take into account all of the simple voters, in different ways. The next-best results come from Maj3Best, in which the three best single measures vote. We should note, however, that the single best measure, the semantic similarity measure from SemSim, lags only slightly behind the two best voting schemes.

This research demonstrates that credible all-words sense disambiguation results can be achieved without recourse to contextual data. Lexical resources enriched with, for example, syntactic information, in which some portion of the resource is hand-mapped to another lexical resource may be rich enough to support such a task. The degree of success achieved here also owes much to the confluence of WordNet's hierachical structure and SEMCOR tagging, as used in the computation of the semantic similarity measure, on the one hand, and the classified structure of the verb lexicon, which provided the underlying groupings used in that measure, on the other hand. Even where one measure yields good results, several data sources needed to be combined to enable its success.

[9] The criteria for the majority voting schemes preclude their assigning more than 2 senses. Controlled relaxation of these criteria may achieve somewhat better results.

References

Srinivas Bangalore and Owen Rambow. 2000. Corpus-Based Lexical Choice in Natural Language Generation. In *Proceedings of the ACL*, Hong Kong.

Olivier Bodenreider and Carol A. Bean. 2001. Relationships among Knowledge Structures: Vocabulary Integration within a Subject Domain. In C.A. Bean and R. Green, editors, *Relationships in the Organization of Knowledge*, pages 81–98. Kluwer, Dordrecht.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Lingustics*, 22(2):249–254, June.

Bonnie J. Dorr and Douglas Jones. 1996. Robust Lexical Acquisition: Word Sense Disambiguation to Increase Recall and Precision. Technical report, University of Maryland, College Park, MD.

Bonnie J. Dorr and Mari Broman Olsen. 1997. Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 151–158, Madrid, Spain, July 7-12.

Bonnie J. Dorr, M. Antonia Martí, and Irene Castellón. 1997. Spanish EuroWordNet and LCS-Based Interlingual MT. In *Proceedings of the Workshop on Interlinguas in MT, MT Summit, New Mexico State University Technical Report MCCS-97-314*, pages 19–32, San Diego, CA, October.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Eduard Hovy. In press. Comparing Sets of Semantic Relations in Ontologies. In R. Green, C.A. Bean, and S. Myaeng, editors, *The Semantics of Relationships: An Interdisciplinary Perspective*. Book manuscript submitted for review.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and Results for English SENSEVAL. *Computers and the Humanities*, 34:15–48.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage, Beverly Hills.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

George A. Miller and Christiane Fellbaum. 1991. Semantic Networks of English. In Beth Levin and Steven Pinker, editors, *Lexical and Conceptual Semantics*, pages 197–229. Elsevier Science Publishers, B.V., Amsterdam, The Netherlands.

Tom Mitchell. 1997. *Machine Learning*. McGraw Hill.

Mari Broman Olsen, Bonnie J. Dorr, and David J. Clark. 1997. Using WordNet to Posit Hierarchical Structure in Levin's Verb Classes. In *Proceedings of the Workshop on Interlinguas in MT, MT Summit, New Mexico State University Technical Report MCCS-97-314*, pages 99–110, San Diego, CA, October.

Martha Palmer. 2000. Consistent Criteria for Sense Distinctions. *Computers and the Humanities*, 34:217–222.

Adwait Ratnaparkhi. 2000. Trainable methods for surface natural language generation. In *Proceedings of the ANLP-NAACL*, Seattle, WA.

Philip Resnik. 1999a. Disambiguating noun groupings with respect to wordnet senses. In S. Armstrong, K. Church, P. Isabelle, E. Tzoukermann S. Manzi, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 77–98. Kluwer Academic, Dordrecht.

Philip Resnik. 1999b. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. In *Journal of Artificial Intelligence Research*, number 11, pages 95–130.

Hans Van Halteren, Jakub Zavrel, and Walter Daelemans. 1998. Improving data-driven wordclass tagging by system combination. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 491–497.