# Automatic Extraction of Semantic Classes from Syntactic Information in Online Resources

**Bonnie J. Dorr**
Department of Computer Science and
Institute for Advanced Computer Studies
A. V. Williams Building
College Park, MD 20742
bonnie@cs.umd.edu

**Doug Jones**
Institute for Advanced
Computer Studies
A. V. Williams Building
College Park, MD 20742
jones@umiacs.umd.edu

## Abstract

This paper addresses the issue of word-sense ambiguity in extraction from machine-readable resources for the construction of large-scale knowledge sources. We describe two experiments: one which took word-sense distinctions into account, resulting in 97.9% accuracy for semantic classification of verbs based on (Levin, 1993); and one which ignored word-sense distinctions, resulting in 6.3% accuracy. These experiments were dual purpose: (1) to validate the central thesis of the work of (Levin, 1993), i.e., that verb semantics and syntactic behavior are predictably related; (2) to demonstrate that a 20-fold improvement can be achieved in deriving semantic information from syntactic cues if we first divide the syntactic cues into distinct groupings that correlate with different word senses. Finally, we show that we can provide effective acquisition techniques for novel word senses using a combination of online sources.

## 1 Introduction

This paper addresses the issue of word-sense ambiguity in extraction from machine-readable resources for the construction of large-scale knowledge sources. We describe two experiments: one which took word-sense distinctions into account, resulting in 97.9% accuracy for semantic classification of verbs based on (Levin, 1993); and one which ignored word-sense distinctions, resulting in 6.3% accuracy. These experiments were dual purpose: (1) to validate the central thesis of the work of (Levin, 1993), i.e., that verb semantics and syntactic behavior are predictably related; (2) to demonstrate that a 20-fold improvement can be achieved in deriving semantic information from syntactic cues if we first divide the syntactic cues into distinct groupings that correlate with different word senses. Finally, we show that we can provide effective acquisition techniques for novel word senses using a combination of online sources, in particular, Longman's Dictionary of Contemporary English (LDOCE) (Procter, 1978), Levin's verb classification scheme (Levin, 1993), and WordNet (Miller, 1985). We have used these techniques to build a database of 10,000 English verb entries containing semantic information that we are currently porting into languages such as Arabic, Spanish, and Korean for multilingual NLP tasks such as foreign language tutoring and machine translation.

## 2 Automatic Lexical Acquisition for NLP Tasks

As machine-readable resources (i.e., online dictionaries, thesauri, and other knowledge sources) become readily available to NLP researchers, automated acquisition has become increasingly more attractive. Several researchers have noted that the average time needed to construct a lexical entry can be as much as 30 minutes (see, e.g., (Neff and McCord, 1990; Copestake et al., 1995; Walker and Amsler, 1986)). Given that we are aiming for large-scale lexicons of 20-60,000 words, automation of the acquisition process has become a necessity.

Previous research in automatic acquisition focuses primarily on the use of statistical techniques, such as bilingual alignment (Church and Hanks, 1990; Klavans and Tzoukermann, 1996; Wu and Xia, 1995), or extraction of syntactic constructions from online dictionaries and corpora (Brent, 1993; Dorr et al., 1995). In such cases, the objective is typically to build a large set of translation equivalences between words and phrases, e.g., for transfer MT. Others who have taken a more knowledge-based (interlingual) approach (Lonsdale et al., 1996) do not provide a means for systematically deriving the relation between surface syntactic structures and their underlying semantic representations. Such approaches tend to ignore the wide range argument structures (beyond intransitive and transitive) that could potentially be associated with verbs. Those who have taken more sophisticated argument structures into account, e.g., (Copestake et al., 1995), do not take full advantage of the systematic relation between syntax and semantics during the lexical acquisition stage. Our own approach exploits certain linguistic constraints that govern the relation between syntactic structure and word meaning. We demonstrate that verb meaning can be systematically derived from information about syntactic realizations; these meaning components are used to build verb entries which are then ported into different languages.

## 3 Syntax-Semantics Relation: Verb Classification Based on Syntactic Behavior

The central thesis of (Levin, 1993) is that the semantics of a verb and its syntactic behavior are predictably related. As a demonstration that such predictable relationships are not confined to an insignificant portion of the vocabulary, Levin surveys 4183 verbs, grouped into 191 semantic classes in Part Two of her book. The syntactic behavior of these classes is illustrated with 1668 example sentences, an average of 8 sentences per class.

Given the scope of Levin's work, it is not easy to verify the central thesis. To this end, we created a database of Levin's verb classes and example sentences from each class, and wrote a parser to extract basic syntactic patterns from the sentences.[1] We then characterized each semantic class by a set of syntactic patterns, which we call a *syntactic signature*, and used the resulting database as the basis of two experiments, both designed to to discover whether the syntactic signatures tell us anything about the meaning of the verbs.[2] The first experiment, which we label *Class-Based*, implicitly takes word-sense distinctions into account by considering each occurrence of a verb individually and assigning it a single syntactic signature according to class membership. The second experiment, which we label *Verb-Based*, ignores word-sense distinctions by assigning one syntactic signature to each verb, regardless of whether it occurred in multiple classes.

The remainder of this section describes the assignment of signatures to semantic classes and the two experiments for determining the relation of syntactic information to semantic classes. We will see that our classification technique shows a 20-fold improvement in the experiment where we implicitly account for word-sense distinctions.

### 3.1 Assignment of Signatures to Semantic Classes

In order to assign signatures to semantic classes, we first needed to decide what syntactic information to extract. It turns out that a very simple strategy works very well, namely, flat parses that contain lists of the major categories in the sentence, the verb, and a handful of other elements. The "parse", then, for the sentence `Tony broke the crystal vase` is simply the syntactic pattern `[np,v,np]`. For `Tony broke the vase to pieces` we get `[np,v,np,pp(to)]`. Notice that the `pp` nodes is marked with its head preposition. Figure 1 shows an example class, the *break* subclass of the Change of State verbs (45.1), along with example sentences and the derived syntactic signature based on sentence patterns. Positive example sentences are denoted by the number 1 in the sentence patterns and negative example sentences are denoted by the number 0 (corresponding to sentences marked with a `*`).

**Verbs:** `break, chip, crack, crash, crush, fracture, rip, shatter, smash, snap, splinter, split, tear`

**Example Sentences:**
```
Crystal vases break easily.
The hammer broke the window.
The window broke.
Tony broke her arm.
Tony broke his finger.
Tony broke the crystal vase.
Tony broke the cup against the wall.
Tony broke the glass to pieces.
Tony broke the piggy bank open.
Tony broke the window with a hammer.
Tony broke the window.
* Tony broke at the window.
* Tony broke herself on the arm.
* Tony broke himself.
* Tony broke the wall with the cup.
A break.
```

**Derived Syntactic Signature:**
```
1-[np,v] 1-[np,v,np]
1-[np,v,np,adjective]
1-[np,v,np,pp(against)]
1-[np,v,np,pp(to)]
1-[np,v,np,pp(with)] 1-[np,v,poss,np]
1-[np,v,adv(easily)] 1-[n]
0-[np,v,np,pp(with)] 0-[np,v,self]
0-[np,v,self,pp(on)] 0-[np,v,pp(at)]
```

Figure 1: Syntactic Signature for Change of State – *break* subclass

### 3.2 Experiment 1: Class-based Approach

In the first experiment, we attempt to discover whether each syntactic signature uniquely identifies

[1] Both the database and the parser are encoded in Quintus Prolog.

[2] The design of this experiment is inspired by the work of (Dubois and Saint-Dizier., 1995). In particular, we depart from the alternation-based data in (Levin, 1993), which is primarily binary in that sentences are presented in pairs which constitute an alternation. Following Saint-Dizier's work, we construct N-ary syntactic characterizations. The choice is of no empirical consequence, but it simplifies the experiment by eliminating the problem of naming the syntactic patterns.

a single semantic class. The outline for this class-based experiment is as follows:

1. Automatically extract syntactic information from the example sentences to yield the syntactic signature for the class.
2. Discover which semantic classes have uniquely-identifying syntactic signatures.

When we parsed the 1668 example sentences in Part Two of Levin's book (including the negative examples), these sentences reduce to 282 unique patterns. The 191 sets of sentences listed with each of the 191 semantic classes in turn reduces to 189 unique syntactic signatures. 187 of them uniquely identify a semantic class, meaning that 97.9% of the classes have uniquely identifying syntactic signatures. As it turns out, only two classes do not have enough syntactic information to distinguish them uniquely.

Because we were interested in the role of prepositions in the signatures, we also ran the experiment with two different parse types: ones that ignored the actual prepositions in the `pp`'s, and parses that threw away all information except for the values of the prepositions. Interestingly, we still got useful results with these impoverished parses, although fewer semantic classes had uniquely-identifying syntactic signatures under these conditions. These results are shown in Figure 2.

We note that the use of negative examples, i.e., plausible uses of the verb in contexts which are disallowed, was a key component of this experiment. There are 1082 positive examples and 586 negative examples. Although this evidence is useful, it is not available in dictionaries, corpora, or other convenient resources that could be used to extend Levin's classification. Thus, to extend our approach to novel word senses (i.e., words not occurring in Levin), we would not be able to use negative evidence. For this reason, we felt it necessary to determine the importance of negative evidence for building uniquely identifying syntactic signatures. As one might expect, throwing out the negative evidence degrades the usefulness of the signatures across the board. The best result, using only the positive evidence to identify semantic classes, gives 88.0% of the semantic classes uniquely identifying syntactic signatures. See Figure 2 for the full results.

**Disambiguated**

| | Overlap | With Negative Evidence | No Negative Evidence |
|---|---|---|---|
| Marked Prepositions | Median | 1.00 | 1.00 |
| | Mean | 0.99 | 0.93 |
| | Perfect | 97.9% | 88.0% |
| Ignored Prepositions | Median | 1.00 | 1.00 |
| | Mean | 0.96 | 0.69 |
| | Perfect | 87.4% | 52.4% |
| Only Prepositions | Median | 1.00 | 0.54 |
| | Mean | 0.82 | 0.57 |
| | Perfect | 66.5% | 42.9% |

**Not Disambiguated**

| | Overlap | With Negative Evidence | No Negative Evidence |
|---|---|---|---|
| Marked Prepositions | Median | 0.10 | 0.09 |
| | Mean | 0.17 | 0.17 |
| | Perfect | 6.3% | 5.2% |
| Ignored Prepositions | Median | 0.10 | 0.09 |
| | Mean | 0.17 | 0.16 |
| | Perfect | 6.3% | 4.2% |
| Only Prepositions | Median | 0.10 | 0.09 |
| | Mean | 0.16 | 0.15 |
| | Perfect | 3.1% | 3.1% |

Figure 2: Overall Results

### 3.3 Experiment 2: Verb-based Approach

In this experiment, we abstracted away from word sense distinctions and considered each verb only once, regardless of whether it occurred in multiple classes. In fact, 46% appear more than once. In some cases, the verb appears to have a related sense even though it appears in different classes. For example, the verb *roll* appears in two subclasses of Manner of Motion Verbs that are distinguished on the basis of whether the grammatical subject is animate or inanimate. In other cases, the verb may have (largely) unrelated senses. For example, the verb *move* is both a Manner of Motion verb and verb of Psychological State.

The composition of a syntactic signature is different for this experiment. Here, we collect all of the syntactic patterns associated with every class a particular verb appears in, regardless of whether that verb is semantically related in the different classes. Now a syntactic signature is the union of the frames extracted from every example sentence for each verb. The outline of the verb-based experiment is as follows:

1. Automatically extract syntactic information from the example sentences.
2. Group the verbs according to their syntactic signature.
3. See where the two ways of grouping verbs overlap:
   (a) the semantic classification given by Levin.
   (b) the syntactic classification based on the derived syntactic signatures.

To return to the Change of State verbs, we now consider the syntactic signature of the verb *break*, rather than the signature of the semantic class as a unit. The verb *break* belongs not only to the Change of State class, but also four other classes: 10.6 *Cheat*, 23.2 *Split*, 40.8.3 *Hurt*, and 48.1.1 *Appear*. Each of these classes is characterized syntactically with a set of sentences. The union of the syntactic patterns corresponding to these sentences forms the syntactic signature for the verb. So although the signature for the Change of State class had 13 frames, the verb *break* has 39 frames from the other classes it appears in.

One way to view the difference between this experiment and the previous one is the difference between the *intension* of a function versus its *extension*. In this case, we are interested in the functions that group the verbs syntactically and semantically. Intensionally speaking, the definition of the function that groups verbs semantically would have something to do with the actual meaning of the verbs.[3] Likewise, the intension of the function that groups verbs syntactically would be defined in terms of something strictly syntactic, such as subcategorization frames. But the intensions of these functions are matters of significant theoretical investigation, and although much has been accomplished in this area, the question of mapping syntax to semantics and vice versa is an open research topic. Therefore, we can turn to the *extensions* of the functions: the actual groupings of verbs, based on these two separate criteria. The semantic extensions are sets of verb tokens, and likewise, the syntactic extensions are sets of verb tokens. To the extent that these functions map between syntax and semantics intensionally, they will pick out the same verbs extensionally.

So for the verb-based experiment, we need a different methodology to establish relatedness between the syntactic signatures and the semantic classes, since the signatures are now mediated by the verbs themselves. A direct method is to compare the two orthogonal groupings of the inventory of verbs: the semantic classes defined by Levin and the sets of verbs that correspond to each of the derived syntactic signatures. When these two groupings overlap, we have discovered a mapping from the syntax of the verbs to their semantics. More specifically, let us define the overlap index as the number of overlapping verbs divided by the average of the number of verbs in the semantic class and the number of verbs in the syntactic signature. Thus an overlap index of 1.00 is a complete overlap and an overlap of 0 is completely disjoint. In this experiment, the sets of verbs with a high overlap index are of interest.

If we use the class-based syntactic signatures containing preposition-marked `pp`'s and both positive and negative evidence, the 1668 example sentences reduce to 282 syntactic patterns, just as before. But now there are 748 verb-based syntactic signatures, as compared with 189 class-based signatures from before. Since there are far more syntactic signatures than the 191 semantic classes, it is clear that the mapping between signatures and semantic classes is not direct. Only 12 mappings have complete overlaps. That means 6.3% of the 191 semantic classes have a complete overlap with a syntactic signature.

[3] An example of the intensional characterization of the Levin classes are the definitions of Lexical Conceptual Structures which correspond to each of Levin's semantic classes. See (Dorr and Voss, to appear).

## 4 The Role of Word-Sense Disambiguation

In the class-based experiment, we counted the percentage of semantic classes that had uniquely identifying signatures. In the verb-based experiment, we counted the number of perfect overlaps (i.e., index of 1.00) between the verbs as grouped in the semantic classes and grouped by syntactic signature. The overall results of the suite of experiments, illustrating the role of disambiguation, negative evidence, and prepositions, is shown in Figure 2. There were three ways of treating prepositions: (i) mark the `pp` with the preposition, (ii) ignore the preposition, and (iii) keep only the prepositions. For these different strategies, we see the percentage of perfect overlaps, as well as both the median and mean overlap ratios for each experiment. These data show that the most important factor in the experiments is word-sense disambiguation.

## 5 Semantic Classification of Novel Words

As we saw above, word sense disambiguation is critical to the success of any lexical acquisition algorithm. The Levin-based verbs are already disambiguated by virtue of their membership in different classes. The difficulty, then, is to disambiguate and classify verbs that do not occur in Levin. Our current direction is to make use of the results of the first two experiments, i.e., the relation between syntactic patterns and semantic classes, but to use two additional techniques for disambiguation and classification of non-Levin verbs: (1) extraction of synonym sets provided in WordNet (Miller, 1985), an online lexical database containing thesaurus-like relations such as synonymy; and (2) selection of appropriate synonyms based on correlations between syntactic information in Longman's Dictionary of Contemporary English (LDOCE) (Procter, 1978) and semantic classes in Levin. The basic idea is to first determine the most likely candidates for semantic classification of a verb by examining the verb's synonym sets, many of which intersect directly with the verbs classified by Levin. The "closest" synonyms are then selected from these sets by comparing the LDOCE grammar codes of the unknown word with those associated with each synonym candidate. The use of LDOCE as a syntactic filter on the semantics derived from WordNet is the key to resolving word-sense ambiguity during the acquisition process. The full acquisition algorithm is given in figure 3.

Given a verb, check Levin class.

1. If in Levin, classify directly.
2. If not in Levin, find synonym set from WordNet.
   (a) If synonym in Levin, select the class that has the closest match with canonical LDOCE codes.
   (b) If no synonyms in Levin or canonical LDOCE codes are completely mismatched, hypothesize new class.

Figure 3: Algorithm for Semantic Classification of Novel Words

Note that this algorithm assumes that there is a "canonical" set of LDOCE codes for each of Levin's semantic classes. Figure 4 describes the significance of a subset of the syntactic codes in LDOCE. (The total number of codes is 174.) We have developed a relation between LDOCE codes and Levin classes, in much the same way that we associated syntactic signatures with the semantic classes in the earlier experiments. These canonical codes are for syntactic filtering (checking for the closest match) in the classification algorithm.

As an example of how the word-sense disambiguation process and classification , consider the non-Levin verb *attempt*. The LDOCE specification for this verb is: T1 T3 T4 WV5 N. Using the synonymy feature of WordNet, the algorithm automatically extracts five candidate classes associated with the synonyms of this word: (1) Class 29.6 "Masquerade Verbs" (*act*), (2) Class 29.8 "Captain Verbs" (*pioneer*), (3) Class 31.1 "Amuse Verbs" (*try*), (4) Class 35.6 "Ferret Verbs" (*seek*), and (5) Class 55.2 "Complete Verbs" (*initiate*). The synonyms for each of these classes have the following LDOCE encodings, respectively: (1) I I-FOR I-ON I-UPON L1 L9 T1 N; (2) L9 T1 N; (3) I T1 T3 T4 WV4 N; (4) I I-AFTER I-FOR T1 T3; and (5) T1 T1-INTO N. The largest intersection with the syntactic codes for *attempt* occurs with the verb *try* (T1 T3 T4 N). However, Levin's class 31.1 is not the correct class for *attempt* since this sense of *try* has a "negative amuse" meaning (e.g., *John's behavior tried my patience*. In fact, the codes T1 T3 T4 are not part of the canonical class-code mapping associated with class 31.1. Thus, *attempt* falls under case 2(b) of the algorithm, and a new class is hypothesized. This is a case where word-sense disambiguation has allowed us to classify a new word *and* to enhance Levin's verb classification by adding a new class to the word *try* as well. In our experiments, our algorithm found several additional non-Levin verbs that fell into this newly hypothesized class, including *aspire*, *attempt*, *dare*, *decide*, *desire*, *elect*, *need*, and *swear*.

We have automatically classified 10,000 "unknown" verbs, i.e., those not occurring in the Levin classification, using this technique. These verbs are taken from English "glosses" (i.e., translations) provided in bilingual dictionaries for Spanish and Arabic.[4] As a preliminary measure of success, we picked out 84 LDOCE *control vocabulary* verbs, (i.e., primitive words used for defining dictionary entries) and hand-checked our results. We found that 69 verbs were classified correctly, i.e., 82% accuracy.

## 6 Summary

We have conducted two experiments with the intent of addressing the issue of word-sense ambiguity in extraction from machine-readable resources for the construction of large-scale knowledge sources. The first experiment attempted to determine a relationship between a semantic class and the syntactic information associated with each class. Not surprisingly, but not insignificantly, this relationship was very clear, since this experiment avoided the problem of word sense ambiguity. In the second experiment, verbs that appeared in different classes collected the syntactic information from each class it appeared in. Therefore, the syntactic signature was composed from all of the example sentences from every class the verb appeared in. In some cases, the verbs were semantically unrelated and consequently the mapping from syntax to semantics was muddied. These experiments served to validate Levin's claim that verb semantics and syntactic behavior are predictably related and also demonstrated that a significant component of any lexical acquisition program is the ability to perform word-sense disambiguation.

We have used the results of our first two experiments to help in constructing and augmenting online dictionaries for novel verb senses. We have used the same syntactic signatures to categorize new verbs into Levin's classes on the basis of WordNet and LDOCE. We are currently porting these results to new languages using online bilingual lexicons.

## Acknowledgements

The research reported herein was supported, in part, by Army Research Office contract DAAL03-91-C-0034 through Battelle Corporation, NSF NYI IRI-9357731, Alfred P. Sloan Research Fellow Award BR3336, and a General Research Board Semester Award.

[4] The Spanish-English dictionary was built at the University of Maryland; The Arabic-English dictionary was produced by Alpnet, a company in Utah that develops translation aids. We are also in the process of developing bilingual dictionaries for Korean and French, and we will be porting our LCS acquisition technology to these languages in the near future.

| LDOCE Code | Arguments | Adjuncts | Example |
|---|---|---|---|
| I | — | — | Olivier is **acting** tonight |
| I-AFTER | — | PP[after] | She **sought** after the truth |
| I-FOR | — | PP[for] | They **sought** for the right one |
| I-ON | — | PP[on] | He **acted** on our suggestion |
| I-UPON | — | PP[upon] | The drug **acted** upon the pain |
| L1 | NP | — | He **acts** the experienced man |
| L9 | ADV/PP | — | The play **acts** well |
| T1 | NP | — | I **pioneered** the new land |
| T1-INTO | NP | PP[into] | We **initiated** him into the group |
| T3 | VP[to+inf] | — | He **tried** to do it |
| T4 | VP[+prog] | — | She **tried** eating the new food |
| WV4 | -ing adjectival | — | I've had a **trying** day |
| WV5 | -ed adjectival | — | He was convicted for **attempted** murder |
| N (denominal verb) | — | — | **pioneer** (noun) |

Figure 4: Sample Syntactic Codes used in LDOCE

## Note

This report is published as a technical report jointly at UMIACS and the CS Department at the University of Maryland:

- UMIACS-TR-95-65
- CS-TR-3481

## References

M. Brent. 1993. Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19:243–262.

K. Church and P. Hanks. 1990. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16:22–29.

A. Copestake, T. Briscoe, P. Vossen, A. Ageno, I. Castellon, F. Ribas, G. Rigau, H. Rodríguez, and A. Samiotou. 1995. Acquisition of Lexical Translation Relations from MRDS. *Machine Translation*, 9.

B. Dorr and C. Voss. to appear. A Multi-Level Approach to Interlingual MT: Defining the Interface between Representational Languages. *International Journal of Expert Systems.*

B. Dorr, J. Garman, and A. Weinberg. 1995. From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT. *Machine Translation*, 9.

D. Dubois and P. Saint-Dizier. 1995. Construction et représentation de classes sémantiques de verbes: une coopération entre syntaxe et cognition. manuscript, IRIT- CNRS, Toulouse, France.

J.L. Klavans and E. Tzoukermann. 1996. Dictionaries and Corpora: Combining Corpus and Machine-readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation*, 10.

B. Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago, IL.

D. Lonsdale, T. Mitamura, and E. Nyberg. 1996. Acquisition of Large Lexicons for Practical Knowledge-Based MT. *Machine Translation*, 9.

G. Miller. 1985. WORDNET: A Dictionary Browser. In *Proceedings of the First International Conference on Information in Data*, University of Waterloo Centre for the New OED, Waterloo, Ontario.

M. Neff and M. McCord. 1990. Acquiring Lexical Data from Machine-Readable Dictionary Resources for Machine Translation. In *Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI-90)*, Austin, Texas.

P. Procter. 1978. *Longman Dictionary of Contemporary English.* Longman, London.

D. Walker and R. Amsler. 1986. The Use of Machine-readable Dictionaries in Sublanguage Analysis. In R. Grishman and R. Kittredge, editors, *Analyzing Language in Restricted Domains*, pages 69–83. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

D. Wu and X. Xia. 1995. Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon. *Machine Translation*, 9.