Sense Tagging the Penn Treebank

Martha Palmer, Hoa Trang Dang, Joseph Rosenzweig

University of Pennsylvania
200 South 33rd Street, Philadelphia, PA, USA
{mpalmer, htd, josephr}@linc.cis.upenn.edu

Abstract

This paper describes the methodology that is being used to augment the Penn Treebank annotation with sense tags and other types of semantic information. Inspired by the results of SENSEVAL, and the high inter-annotator agreement that was achieved there, similar methods were used for a pilot study of 5000 words of running text from the Penn Treebank. Using the same techniques of allowing the annotators to discuss difficult tagging cases and to revise WordNet entries if necessary, comparable inter-annotator rates have been achieved. The criteria for determining appropriate revisions and ensuring clear sense distinctions are described. We are also using hand correction of automatic predicate argument structure information to provide additional thematic role labeling.

1. Introduction

The success of recent applications of machine learning techniques to tasks such as part-of-speech tagging and parsing has kindled the hope that these same techniques might have equal or greater success in other areas such as lexical semantics. Advances in automated and semi-automated methods of acquiring lexical semantics in particular would release the field from its dependence on well-defined subdomains with small vocabularies and enable broad-coverage natural language processing. However, supervised machine learning requires large amounts of publicly available training data, and a prerequisite for this training data is general agreement on which elements should be tagged and with what tags. With respect to lexical semantics, this type of general agreement has been strikingly elusive. The field has yet to develop a clear consensus on guidelines for a computational lexicon that could provide a springboard for training data. This is in spite of much effort being devoted to individual approaches to lexicon development such as Meaning Text Theory, (Mel'cuk, 1988), the Generative Lexicon, (Pustejovsky, 1991), Sensus (Hovy, 1993), MikroKosmos, (Nirenburg et al., 1992), Acquilex, (Copestake and Sanfilippo, 1993), Framenet, (Lowe et al., 1997), Lexical Conceptual Structures, (Dorr, 1997), and WordNet,(Miller et al., 1990). Each takes a very different approach and makes reference to different underlying theories of semantics.

One of the most controversial areas in semantic representation has to do with polysemy. What constitutes a clear separation into senses for any one word, and how can these senses be computationally characterized and distinguished, (Palmer, 1999)? The answer to this question is the key to breaking the bottleneck of semantic representation that is currently the single greatest limitation on the general application of natural language processing techniques.

SIGLEX98-SENSEVAL [1] shed light on the task of sense tagging, and whether or not sufficient training data could be consistently tagged with a set of pre-existing sense distinctions.[2] Assuming appropriate training data could be provided, an exercise was set up to evaluate different systems on the word sense disambiguation task. [3] This data was prepared using a set of senses from the Hector project (Atkins, 1993), and the results of the exercise were very encouraging. By allowing for revision of senses that caused disagreements among annotators during a training period, inter-annotator agreement for the words that were tagged was well over 90%, and the best supervised systems achieved precision and recall scores in the 80's.

All of the participants in SIGLEX98-SENSEVAL agreed that they would prefer evaluations based on running text rather than corpus instances, but this is only feasible if the Gold Standard sense inventory being used for tagging can be appropriately mapped onto several different lexical resources. Inspired by the SENSEVAL results, we used similar methods for a pilot study of sense tagging 5000 words of running text from the Penn Treebank. Using the same techniques of allowing the annotators to discuss difficult tagging cases and to revise WordNet entries if necessary, comparable inter-annotator rates have been achieved.

2. Semantically tagged running text

As a spin-off from SENSEVAL98, for SIGLEX99 we tagged running text with WordNet senses, as well as other senses. This tagged text was made available on the web prior to the workshop, and participants seemed satisfied with the quality of the sense-tagged running text. It was agreed that the next SENSEVAL should include text of this type, with two qualifications. It must be possible to revise WordNet senses along the lines of the Hector revisions when clear sense distinctions could not be made, and the running text should be augmented with sufficient corpus instances of the words chosen for training and testing.

2.1. Sense tagged text

We sense-tagged a 5000-word corpus of running text using WordNet1.6 senses. The most obvious example of

[1] http://www.itri.brighton.ac.uk/events/senseval

[2] There were concerns about inter-annotator agreement for SemCor.

[3] The exercise was also supported by Euralex, Elsenet, ECRAN and SPARKLE. A special issue of Computers and the Humanities that includes a detailed overview of the exercise and reports of individual systems will appear shortly, (Kilgarriff and Palmer, 2000).

```
<wf lemma=Donald_Trump wnsn=person>Donald Trump</wf> ,
<wf lemma=who wnsn=person:DT>who</wf>
<wf cmd=arb lemma=face wnsn=?>faced</wf> rising
<wf cmd=done lemma=doubt wnsn=1>doubt</wf> about his
<wf cmd=done lemma=bid wnsn=2>bid</wf> for American Airlines parent
<wf lemma=AMR_Corp. wnsn=company>AMR Corp.</wf> even before a United Airlines
<wf cmd=done lemma=buy-out wnsn=1>buy-out</wf>
<wf cmd=done lemma=come_apart wnsn=1?>came apart</wf>
<wf cmd=done lemma=Friday wnsn=1>Friday</wf> ,
<wf cmd=arb lemma=withdraw wnsn=5?>withdrew</wf> his $ 7.54 billion
<wf cmd=done lemma=offer wnsn=2>offer</wf> .

Donald Trump <person>, who faced <face7?> rising doubt <doubt1> about
his bid <bid2> for American Airlines parent AMR Corp. <company>, even
before a United Airlines buy-out <buy-out1> came apart <come_apart1>
Friday, withdrew <withdraw5?> his $7.54 billion offer <offer2>.
```

Figure 1: Sample sense-tagged text

an existing corpus of semantically tagged running text is SemCor, in which WordNet1.5 senses were used to tag approximately 200,000 words from the Brown corpus. Unlike SemCor, which assigns WordNet senses to all the verbs, nouns, adjectives, and adverbs, we chose to sense-tag only the verbs and headwords of their noun arguments and adjuncts. In addition, proper nouns which were not in WordNet were tagged as either person, company, date, or name (indicating none-of-the-above) and, wherever possible, pronouns were tagged with the sense of their antecedents.

Corpus Selection Our corpus sample comprised five texts from the Penn Treebank II Wall Street Journal corpus. We chose to work on the WSJ corpus because it was already manually annotated for part of speech and bracketed for syntactic structure, and because we have developed tools to automatically extract predicate-arguments from bracketed text. The particular WSJ articles that we chose contained interesting verbs that had been discussed in previous SIGLEX meetings, and covered a range of topics, including an account of an earthquake that occurred during court proceedings, a report about insurance and claims adjusting following an earthquake, a letter to the editor about the need to fight the Columbian drug mafia, a report about a bid by Donald Trump to buy an airline, and a description of a case in which government can seize the assets of criminal defendants, including lawyers fees.

Annotation Process Our test corpus was annotated once by a linguistics graduate student, and then checked by a computer science graduate student. The first annotator made one pass through the corpus and assigned initial sense tags for each of the relevant words, making note of problematic cases. Then the second annotator independently chose a sense for each word, checked it against the initial tag, and then assigned the "final" tag. Instances with differing initial and final tags were flagged with "arb", indicating that the second annotator could not agree with the decision of the first annotator even after carefully analyzing the disagreement. Tagging in the first pass was done serially and required approximately 40 hours to tag the 5000-word sample. The second pass required the same amount of time but included manual lemmatization of the tagged words (this will be done automatically in the future) so that all inflected forms of a word could be tagged at one time, allowing for more reliable results than serial annotation. We are currently developing a graphical interface that will allow more rapid and efficient annotation in the future, (see Section 4.).

Annotation Results Approximately 2100 words were sense-tagged, with an overall inter-annotator agreement rate of 89%, as measured by tokens not flagged with "arb". Of the tagged words, 700 were verbs, with an inter-annotator agreement rate of 81%. There were 350 different verb lemmas, 90 of which had at least one occurrence in the corpus in which the annotators disagreed on the correct sense. In many cases, this was because WordNet either did not have the correct sense or else did not adequately define how the different senses should be distinguished.

Figure 1 shows a sample of the tagged text, followed by a simplified form that is easier to read. Some words, such as *face*, did not have appropriate senses in WordNet, while others, such as *withdraw*, are marked with question marks because of inter-annotator disagreement. However, with our high agreement level we are confident that when we can allow for senses to be revised as was done for SENSEVAL we will also get inter-annotator agreement in the 94% range.

2.2. Automatic tagging of predicate-argument structure

We are also automatically annotating the same sentences with more explicit verbal predicate-argument structure that is closely linked to the Penn Treebank II bracketing (Marcus, 1994). Under this bracketing scheme, subscripts are appended to standard parse-tree nonterminals (such as NP, PP) to indicate approximately what semantic role a constituent plays in a sentence. For instance, the subscript SBJ is used to indicate that an NP is the subject of some verb, while subscript TMP may be used to indicate that a prepositional phrase specifies temporal information about an event that is being described. Additionally, indexed empty constituents are used to mark extraction and movement phenomena in cases when the locus of semantic interpretation differs from the position where a constituent appears in a tree.

To induce a useful predicate-argument annotation from this parse-tree representation, however, it is necessary to perform additional analysis on the trees. For instance, when an NP is marked as a subject, there is no explicit indication of what verb it is the subject of. To link verbs and their subjects, it is necessary to determine what the semantic heads of phrases are. Morphological information is not given in the Treebank II annotation scheme, so this also must be added, and phrasal lexical entries must also be identified when they form complex semantic predicates comprising several nodes in the parse tree.

We have developed a module to perform this additional analysis on the Treebank II materials, outputting an SGML representation of all predicate-argument relations detected. The module consults a lexical semantic knowledge base including information about verb subcategorization, the ontology of noun-phrase referents, and complex lexical items comprised of more than one word of text.

The module is also able to infer some of the information represented by the Treebank II formalism even if this information has not been explicitly coded in its input. For instance, it can still analyze parse trees produced by statistical parsers such as the one developed by Collins (Collins, 1997), the output of which lacks some of the semantic cues that have been added to the Treebank II manually by annotators.

To evaluate the accuracy of the automatic predicate-argument analyzer, we examined 65 sentences from our sense-tagged running corpus containing 162 automatically annotated verb predicate-argument structures, and found 132 to be correct (81% precision). We are currently hand correcting the entire set of predicate-argument structures.

3. Defining criteria for sense distinctions

We have demonstrated that, allowing for revisions, WordNet senses can be just as consistently tagged as Hector senses. However, this does not solve the problem of using training data tagged with one set in order to evaluate a system that uses a different set. To illustrate how pervasive mismatches between lexical resources can be, here are some of the discrepancies between the Hector *shake* and WordNet 1.6 *shake.*

3.1. *Shake* mismatches

WordNet 1.6 had 8 senses for *shake*, with an additional 5 senses for *shake up* and 2 for *shake off*, 15 all together, see Fig 2). Hector also has 8 main senses for *shake*, with the first one having 3 additional subsenses and the second one having 2. It has 3 senses for *shake up* instead of 5, and 2 for *shake off*. In addition, it has 3 for *shake down* and 2 for *shake out*, (see Fig 3). On the surface this seems fairly compatible, with WordNet just missing a few verb particle constructions. However, looking in detail at the content of the senses reveals a more fundamental mismatch. Hector distinguishes between *shaking hands with someone*, and *shaking one's fist* and *shaking one's head*. This is quite legitimate, since although these are all similar in that they are communicative acts, they communicate quite different things. Hector also distinguishes between the intransitive TREMBLE sense, *My hands were shaking from the cold*, and the more proactive MOVE sense, *He shook the bag violently*, where someone intentionally moves something back and forth. WordNet collects these together, along with *She shook her cousin's hands*, as WN1, and instead makes distinctions with respect to the type of action: WN2, gentle tremors, WN3, rapid vibrations, or WN4, swaying. So 3 Hector senses map onto WN1, and Hector 1 maps equally onto WN1,WN2, WN3 and WN4, (see Fig 4) . Hector also includes *shake out* and *shake off* as examples of 1.1, the CLEAN subsense of 1, *Richard removed her socks and shoes and shook all the gravel out of them, The sand gets shaken off them at the knockout*, then goes on to have separate entries for both of these for their more abstract variations, such as, *A jittery stock market has shaken out more shareholders in United Scientific Holdings, ... unable to shake off the memories of the trenches.* One could argue that shaking out shareholders is not exactly the same thing as shaking apples out of a tree, but that does not mean they are completely unconnected, either.

3.2. Concrete criteria for sense distinctions

In parallel with our sense tagging work we have been developing a lexical resource for verbs that uses Levin classes, (Levin, 1993), as the basis of an hierarchical organization allowing for inheritance, and as a source of cross-linguistic semantic features (Dang et al., 1998). To make our resource more usable, we are mapping our verb entries to the relevant WordNet senses. In addition to class membership information, we are also augmenting the WordNet entries with explicit syntactic information such as that found in COMLEX or the XTAG Syntactic Database. Where applicable we include semantic class constraints on verb arguments and information about resulting states, and will be using corpus-based techniques and machine learning to enrich these categorizations.

The verbs covered by the Levin classes are a subset of what is covered by WordNet, and many of the senses in WordNet entries are not addressed. However, in discussing the different entries for *shake*, and in going back and forth between Hector and WordNet, we find that we often make reference to the inclusion of specific lexical items, different syntactic frames, different semantic class constraints on verb arguments, or differences in outcome to distinguish the senses. The simplest and most obvious distinguishers are prepositions. *Shake down* is clearly marked as being different in meaning from *shake up*. Having access to even rudimentary syntactic structure makes it quite straightforward to tease apart all of the verb particle structures and idiomatic expressions by virtue of the presence of specific lexical items. Transitive and intransitive usages are also fairly easy to distinguish, but unfortunately they only too often cross sense boundaries.

Semantic class constraints are more subtle and more difficult to capture. However, even semantic preferences can help to distinguish senses. Verb class membership itself,

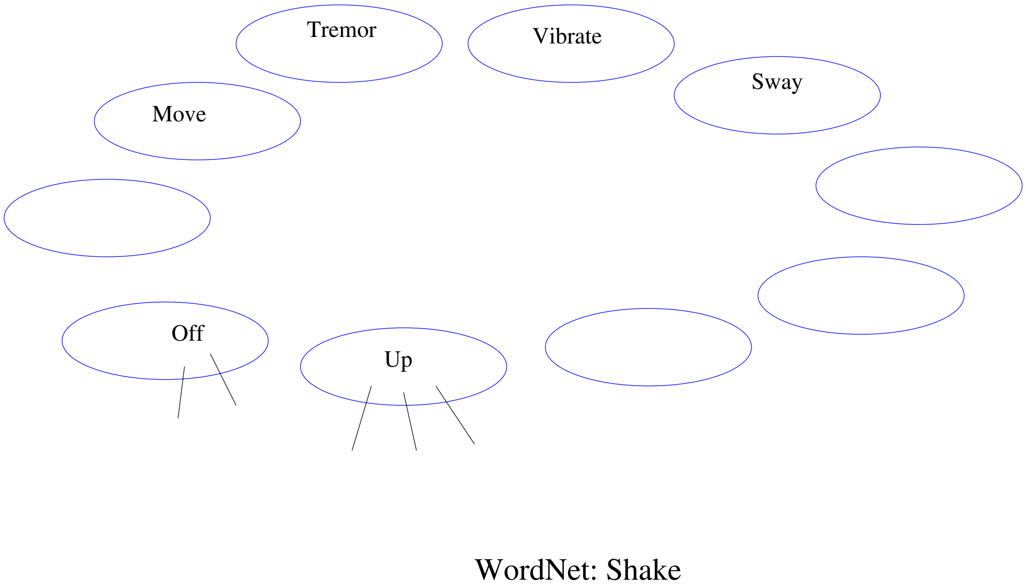

Figure 2: WordNet sense distinctions for *Shake*

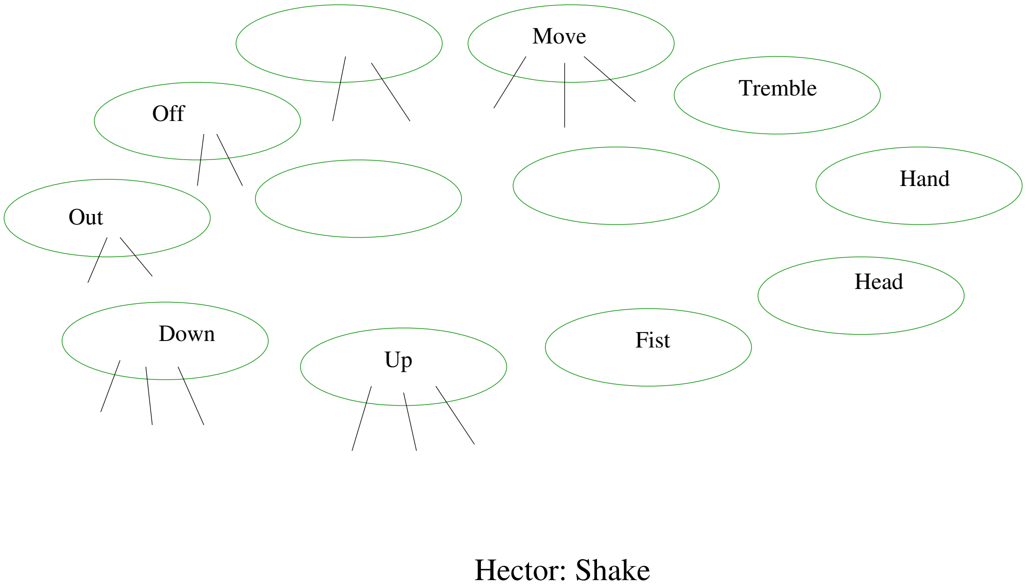

Figure 3: Hector sense distinctions for *Shake*

if it can be determined, can play a central role, and can indicate either homonyms or polysemes that are produced through regular extensions of meaning that can apply uniformly to entire sets of verbs. Since these regular extensions, such as resultatives, are often produced by adjunctions that can be seen as extending the subcategorization frame, this highlights the fundamental role argument structure plays in distinguishing senses. For instance, the 27 Hector *shake* senses and the 15 WordNet *shake* senses can all be partitioned into the same five major divisions that are illustrated in Figure 5, each one of which corresponds to a different Levin class. Idioms are handled separately.

The basic sense is the Externally controlled *shaking* motion which results when a person or an earthquake or some other major force shakes an object. This same motion can be further amplified with directional information specifying a result such as *off, down, up, out* or *away*. If a path prepositional phrase is actually specified, such as *shook the apples out of the tree* or *shook water from the umbrella*, then a change of location (CH-LOC) occurs, and these usages are now classed as Funnel verbs. The same back and forth motion can occur during Body-Internal states such as *shaking from cold or fear*, i.e., TREMBLING. If a particular BODY-PART is shaken in a stereotypical way, such as *shaking hands or fists or fingers* then a communicative act takes place and these are Crane verbs (as in *craning one's neck* or *blinking ones eyes.*) Then there are the abstract usages, which are all classified as Psych verbs, such as *shaken by the news, or the attack, or his father's death, etc.*. The Crane verbs and the Psych verbs are distinguished syntactically from the others in that they cannot occur in the intransitive, or even the passive.[4] Finally we are left with the

[4]With the exception of his hand was shaken, which being symmetrical is something of an outlier anyway.

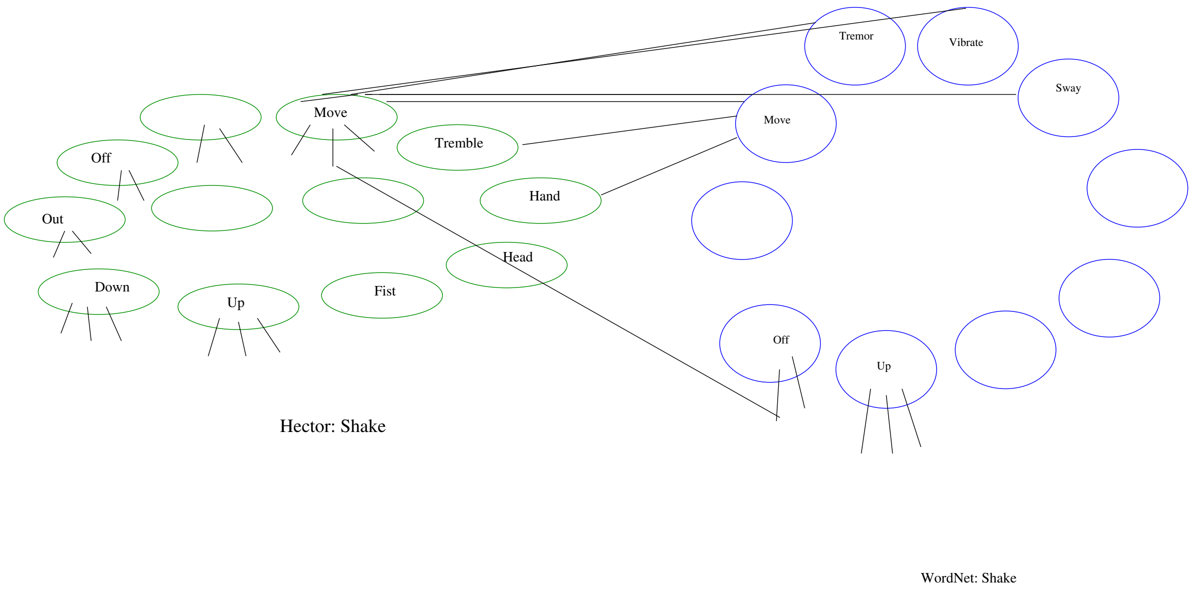

Figure 4: Mismatches between Hector and WordNet: *Shake*

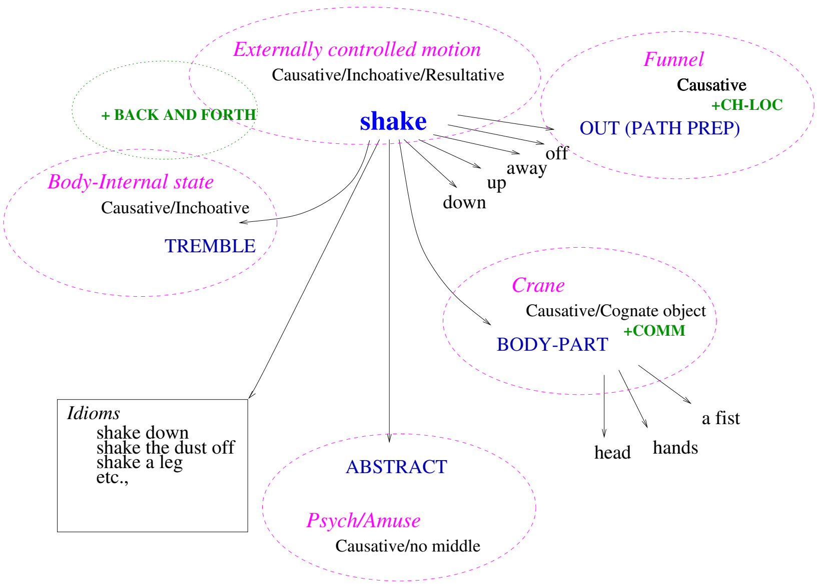

Figure 5: High level organization for *Shake*

idioms, which of course have to be listed individually. All of the Hector and WordNet senses can be categorized under one of these major divisions, providing more fine-grained distinctions if needed, (see Fig 6).

4. A Methodology for the tagging process

To summarize, we have thus far presented the approach used to provide semantic tags for 5000 words of running text from the Penn Treebank. We used WordNet senses, but where we found difficulty in getting inter-annotator agreement, allowed for the revision of the WordNet senses. In determining these revisions we make recourse to additional corpus examples as well as concrete criteria for sense distinctions such as those discussed in Section 3.2.. We also tagged the same data with predicate-argument structures marked with Agent and Patient thematic roles. This was done semi-automatically, by first running an algorithm for performing the annotation and then hand correcting the twenty percent with errors.

The Penn Treebank seems the ideal corpus for a major semantic tagging effort. It already has syntactic bracketing which can be extremely helpful, as demonstrated by the thematic role labeling, and it is widely available, and widely used.

We will continue to tag running text, but will do it more slowly, so that we can tag the additional corpus instances of each word as we come to it. This is more efficient, and more importantly, gives us sufficient data to determine if the WordNet senses need to be revised. If the Treebank itself has only a few instances of a particular word, these can always be augmented. We can automatically separate out all of the verb particle constructions and idiomatic usages which are clearly lexically marked as distinct senses. We can also automatically recognize simple

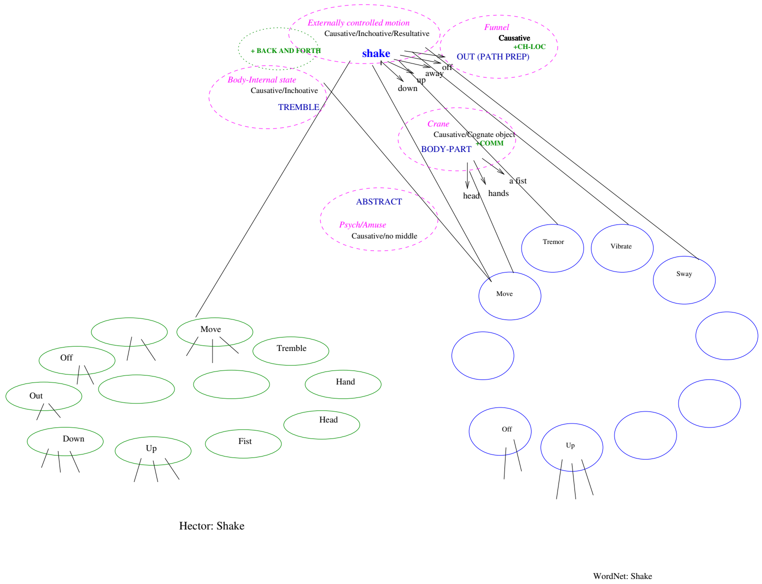

Figure 6: Mapping from VerbNet to Hector and WordNet: *Shake*

predicate-argument structure, another useful indicator.

We will also be using an improved interface that simplifies the tagging itself, and displays both the sentence to be tagged and its predicate-argument structure. The interface can also be used to retrieve surrounding text, sometimes necessary in making a subtle sense distinction.

5. References

Atkins, Sue, 1993. Tools for computer-aided corpus lexicography: The hector project. *Acta Linguistica Hungarica*, 41:5–72.

Collins, M., 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain.

Copestake, Ann and Antonio Sanfilippo, 1993. Multilingual lexical representation. In *Proceedings of the AAAI Spring Symposium: Building Lexicons for Machine Translation*. Stanford, California.

Dang, Hoa Trang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig, 1998. Investigating regular sense extensions based on intersective levin classes. In *Proceedings of Coling-ACL98*. Montreal, CA.

Dorr, Bonnie J., 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12:1–55.

Hovy, E., 1993. Ontologies for mt. Presented at the *Japan - US MT Workshop*, Washington, D.C.

Kilgarriff, Adam and Martha Palmer, 2000. Senseval: Evaluating word sense disambiguation systems. *Computers and the Humanities*, 34(1-2).

Levin, Beth, 1993. *English Verb Classes and Alternations A Preliminary Investigation*.

Lowe, J.B., C.F. Baker, and C.J. Fillmore, 1997. A frame-semantic approach to semantic annotation. In *Proceedings 1997 Siglex Workshop/ANLP97*. Washington, D.C.

Marcus, Mitch, 1994. The penn treebank: A revised corpus design for extracting predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*. Princeton, NJ.

Mel'cuk, I. A, 1988. Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, I:3:165–188.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, 1990. Five papers on wordnet. Technical Report 43, Cognitive Science Laboratory, Princeton University.

Nirenburg, S., J. Carbonell, M. Tomita, and K. Goodman, 1992. *Machine translation: a knowledge-based approach*. San Mateo, California, USA: Morgan Kaufmann.

Palmer, Martha, 1999. Consistent criteria for sense distinctions. *Computers and the Humanities*. Special Issue on SENSEVAL.

Pustejovsky, James, 1991. The generative lexicon. *Computational Linguistics*, 17(4).