Computational Lexical Semantics and its Constraints

Martha Palmer and Szu-ting Yi

Department of Computer & Information Science

University of Pennsylvania

Page 1 ~ 20

Proceedings of Research on Computational Linguistics

Conference XIII (ROCLING XIII)

Taipei, Taiwan

2000-08-24/2000-08-25

Computational Lexical Semantics and its Constraints

Martha Palmer and Szu-ting Yi

Department of Computer & Information Science

University of Pennsylvania

Philadelphia, PA 19104-6389

mpalmer@linc.cis.upenn.edu

1 Introduction

In the last decade of research and development in Natural Language Technology, basic tools have been put in place and are already impacting daily life. Speech recognition saves the phone company millions of dollars. Text to speech synthesis aids the blind. Massive resources for training and analysis are available in the form of annotated and analyzed corpora for spoken and written language. This explosion in applications has been largely due to new algorithms using statistical techniques and above all to the huge increase in power per dollar in computing machinery.

Yet the goals of accurate information extraction, focused information retrieval and fluent machine translation still remain tantalizingly out of reach. Our next quantum leap in language processing capabilities has to come from a closer integration of syntax and lexical semantics. However, the difficulty of achieving adequate hand-crafted semantic representations has limited the field of natural language processing to applications that can be contained within well-defined subdomains. The only escape from this limitation will be through the application of robust, statistical techniques to the task of semantic representation. 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. In spite of several different research efforts into lexicon development [13], [14], [17], [15], [3], [10], [6], the field has yet to develop a clear consensus on guidelines for a computational lexicon that could provide a springboard for such methods. The most encouraging development is recent work in linguistics on word classifications that have a more semantic orientation, such as WordNet, [14], [26], [27], and Levin's verb classes, [9]. These classes, and refinements on them [5], provide the key to developing concrete criteria for sense distinctions that will enable consistent semantic annotation of corpora [30].

The pressing need for a consensus on a task-oriented level of semantic representation is the driving force behind our current research efforts. The objective is not to prove the truth or correctness of any particular theoretical approach, but rather to provide a useful level of representation that is as theory-neutral as possible. We believe that a shallow dependency structure, i.e., predicate-argument structure, for verbs, participial modifiers and nominalizations provides a feasible level of annotation that would be of great benefit. This annotation, coupled with minimal co-reference links, could provide a major advance in our ability to train our new statistical parsers, [2], [11], [18], [1], [20], to recognize not just syntactic structures, but structures that are rich with semantic content as well. In the same way that the existence of the Penn Treebank enabled the development of extremely powerful new syntactic analysis methods, a correspondingly richer level of annotation that includes word senses, predicate argument structure, noun-phrase semantic categories and co-reference will facilitate our ability to extract salient relationships from text. Tightly interwoven syntactic and semantic processing can provide the levels of accuracy that are required to support discourse analysis and inference and reasoning; the foundation of any natural language processing application. This will in turn enable breakthroughs in message understanding, machine translation, fact retrieval, and perhaps even information retrieval.

In this paper we describe our use of linguistic theories in determining the sense distinctions and predicate-argument structures that we are currently using for the semantic annotation of the Penn Treebank.

2. Regular Sense Extensions

The most difficult questions in lexical semantics involve distinguishing between different senses and between arguments and adjuncts. The answers to these two questions can often be intrinsically related, as it may be the addition of a new argument that forms the basis for distinguishing one sense from another. For instance, the verbs in the sentences

(1) The bottle floated into the cave.

and

(2) The train roared through the station.

appear with a path prepositional phrase, *into the cave, through the station,* that forms an integral part of the meaning of the sentence. The events depicted can be described as change of location events that involve an object in motion along a path. The same verbs can also occur without the path prepositional phrase, in simple intransitive frames where they do not seem to make reference to events involving a path:

(1') The bottle floated.

and

(2') The train roared.

This suggests that the semantic predicates associated with the intransitive verbs in (1) and (2) do not contain placeholder argument positions for paths, so the path has to be added as an adjunct rather than an argument. The only way to consider the path as an intrinsic argument would be to postulate a separate sense with a binary predicate argument structure. We see the adjunction of the path prepositional phrase as explicitly adding a change-of-location semantic component to the basic representation of the event type corresponding to the intransitive usage, and this is the same whether the intransitive verb depicts a manner of motion event or a sound emission event. In this case the distinction between monosemy and polysemy, [24], [25], for these verbs can be viewed in computational terms as a choice between a dynamic interpretation of a monosemous meaning that can be used to produce two interpretations during the analysis process versus a static interpretation that is compiled out into two separate senses before processing begins. In other words, the monosemous position and the polysemous position can be seen as representationally equivalent from a computational perspective. We will apply this notion to an examination of the sense distinctions we will be using for our semantic tagging project. However, we must first look more closely at WordNet, [14], [26], [27], the on-line lexical database of English that is our primary source of sense distinctions, and also the association between syntax and semantics that is evidenced in Levin's English Verb Classification, [9].

2.1 Verb Classes

WordNet currently contains approximately 120,000 sets of noun, verb, adjective, and adverb synonyms, each representing a lexicalized concept. A synset (synonym set) contains, besides all the word forms that can refer to a given concept, a definitional gloss and - in most cases - an example sentence. Words and synsets are interrelated by means of lexical and semantic-conceptual links, respectively. Antonymy or semantic opposition links individual words, while the super-/subordinate relation links entire synsets. Although it was not the intention of the original designers, aving access to a public domain on-line lexical resource has proven extremely beneficial to the NLP community, and has enabled experimentation with dozens of new techniques for applications, [36]. It has not been straightforward to integrate WordNet with existing parsers, partly because it was designed principally as a semantic network, and contains little syntactic information. This may also have impeded its ability to give rise to high inter-annotator agreement scores for sense-tagging, (78%), since the taggers are provided with few explicit cues for distinguishing between senses.

In contrast, Levin verb classes are based on the ability of a verb to occur or not occur in pairs of syntactic frames that are in some sense meaning preserving (diathesis alternations) [9]. The distribution of syntactic frames in which a verb can appear is closely correlated with its class membership under the fundamental assumption that the syntactic frames are a direct reflection of the underlying semantics. Levin classes are supposed to provide specific sets of syntactic frames that are associated with the individual classes. These sets are not intended to be arbitrary, and they are supposed to reflect underlying semantic components that constrain allowable arguments. For example, *break* verbs and *cut* verbs are similar in that they can all participate in the transitive and in the middle construction, *John broke the window, Glass breaks easily, John cut the bread, This loaf cuts easily*. However, only *break* verbs can also occur in the simple intransitive, *The window broke, *The bread cut*. In addition, *cut* verbs can occur in the conative, *John valiantly cut/hacked at the frozen loaf, but his knife was too dull to make a dent in it*, whereas *break* verbs cannot, **John broke at the window*. The explanation given is that *cut* describes a series of recognizable actions directed at achieving the goal of separating some object into pieces. These actions consist of grasping an instrument with a sharp edge such as a knife, and applying it in a cutting fashion to the object. It is possible for these actions to be performed and still be recognizable without the end result of a separation into pieces of the object being achieved. Where *break* is concerned, the only thing specified is the resulting change-of-state where the object becomes separated into pieces. If the result is not achieved, there are no separate, recognizable *breaking* actions. For the *cut* class of verbs, when there is an *at* in between the verb and its direct object, it qualifies the assumption of the goal state being achieved. The *at* has the same effect on the *grab*, *hit, push/pull, swat* and *poke* classes, although it is not commonly found otherwise.

Ambiguities in Levin classes It is not clear how much WordNet synsets should be expected to overlap with Levin classes, and preliminary indications are that there is a wide discrepancy [28, 29, 6]. However, it would be useful for the WordNet senses to have access to the detailed syntactic information that the Levin classes contain, and it would be equally useful to have more guidance as to the semantic components the members of a Levin class share. Identification of these components is critical to the use of classes and their semantic features for translation purposes, whether transfer-based or interlingua based. Although Levin classes group together verbs with similar argument structures, the meanings of the verbs are not necessarily synonymous. Some classes such as *break (break, chip, crack, crash, crush, fracture, rip, shatter, smash, snap, splinter*, tear) and *cut (chip, clip, cut, hack, hew, saw, scrape, scratch, slash, snip)* contain verbs that are quite synononomous, but others, such as *braid (bob, braid, brush, clip, coldcream, comb, condition, crimp, crop, curl, etc.)* are clearly not intended to be synonymous.

The association of sets of syntactic frames with individual verbs in each class is also not as straightforward as one might suppose. The listing of a verb in more than one class (many verbs are in three or even four classes), with sometimes conflicting sets of alternations, is left open to interpretation in Levin. Does it reflect systematic meaning relationships or is it merely idiosyncratic, i.e., homonyms? For example, the verb *draw* is listed as a *remove* verb, as a *scribble* verb and as a *performance* verb. While the latter two senses seem systematically related (both seem to be involved, for example, in a usage like *draw a portrait*), the *remove* sense (as in *draw water from the well*) is clearly distinct.

A very different example of multiple membership is *seize*, which is in both the *obtain* class as in *He seized his watch from his dresser and dashed out the door*, and also the *possessional deprivation - steal* class, as in *He seized the woman's purse and dashed through the crowd.* Are these two separate senses of *seize*, or just one? And in fact, what are the differences in alternations between these two classes that distinguish them? These classes, which have a large overlap, have only one tangible syntactic difference. A few of the *obtain* verbs can take the Sum of Money Alternation: *$50 will purchase a dress at Sears*., but most of them do not. Both sets of verbs are distinguished more by alternations they do not take rather than by alternations they do take. The grounds for deciding that a verb belongs in a particular class because of the alternations that it does not take are elusive at best.

Intersective classes. We have augmented the existing database of Levin semantic classes with 129 *intersective* classes, which were created by grouping together subsets of existing classes with overlapping members. All subsets were included which shared a minimum of three members. If only one or two verbs were shared between two classes, we assumed this might be due to homophony, an idiosyncrasy involving individual verbs, rather than a systematic relationship involving coherent sets of verbs. This filter allowed us to reject potential intersective class with only one member, *draw*, that would have resulted from combining the *remove* verbs with the *scribble* verbs. On the other hand, the *scribble* verbs do form an intersective class with the performance verbs, since *paint* and *write* are also in both classes, in addition to *draw*. The algorithm we used is discussed in [5].

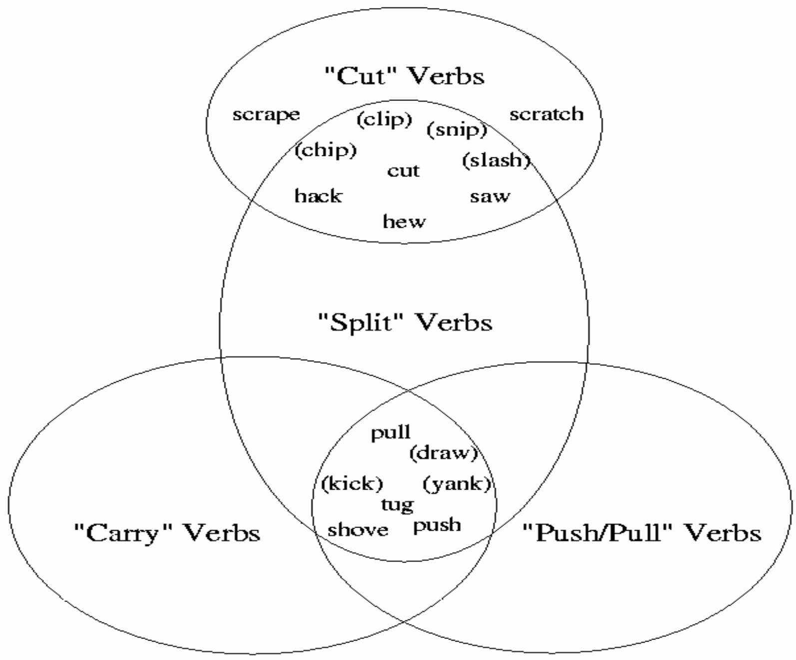

Figure 1. The Intersective class formed from Levin *carry*, *push/pull* and *split* verbs[1]

The intersection between the *push/pull* verbs of exerting force, the *carry* verbs and the *split* verb illustrates how the basic sense of a verb can be extended in several different directions. Depending on the particular syntactic frame in which they appear, members of this intersective class (*pull, push, shove, tug, kick, draw, yank* can have their basic sense extended to include either change-of-state, accompanied motion, or the absence of a change-of-location. The change-of-state comes with *split* class membership, when and *apart* has been adjoined on to the transitive form of the verb. Accompanied motion is added by the adjunction of a path prepositional phrase, and the conative denies any change-of-location of the object. The apparent conflict comes because *carry* verbs are not supposed to take the conative, *Jane

[1] verbs in () are not listed by Levin in all the intersecting classes but participate in all the alternations

carried at the baby, even though the entire subset of *push/pull* verbs demonstrably does. As a verb of exerting force, in the conative *push* indicates a recognizable *pushing* action that fails to result in a change-of-location; with a path prepositional phrase there is a definite change-of-location and it earns membership in the *carry* class. The critical point is that, while the verb's meaning can be extended to either ``attempted'' action or accompanied motion, these two extensions cannot occur simultaneously, as illustrated by the following examples.

Nora pushed the package.

(verb of exerting force; change-of-location of object is possible)

Nora pushed at/against the package.

(verb of exerting force, no change of location

Nora pushed the package to Pamela.

(*carry* sense of push implies causation of accompanied motion)

Nora pushed the branches apart.

(*split* sense implies change-of-state of the branches)

**Nora pushed at the package to Pamela.*

(adjoining both the conative and a path-pp at the same time is disallowed.)

Manner of motion verbs Similar types of regular sense extensions can be illustrated with the manner of motion verbs. Figure 2 shows intersective classes involving two classes of verbs of manner of motion (*run* and *roll* verbs) and a class of verbs of existence (*meander* verbs). *Roll* and *run* verbs have semantic components describing a manner of motion that typically, though not necessarily, involves change-of-location. In the absence of a goal or path adjunct they do not specify any change of location, and in some cases (e.g., *float, bounce*) require the adjunct to explicitly specify any displacement at all. The two classes differ in that *roll* verbs relate to manners of motion characteristic of inanimate entities, while run verbs describe manners in which animate entities can move. Some manner of motion verbs allow a transitive alternation in addition to the basic intransitive. When a *roll* verb occurs in the transitive, (*Bill moved the box across the room*), the subject physically causes the object to move, whereas the subject of a transitive *run* verb merely induces the object to move (*the coach ran the athlete around the track*). Some verbs can be used to describe motion of both animate and inanimate objects, and thus appear in both *roll* and *run* verb classes. The *slide* class partitions this *roll/run* intersection into verbs that can take the transitive alternation and verbs that cannot (*drift* and *glide* cannot be causative, because they are not typically externally controllable). Verbs in the *slide/roll/run* intersection are also allowed to appear in the dative alternation, (*Carla slid the book to Dale, Carla slid Dale the book*), in which the sense of change-of-location is extended to change-of-possession.

When used intransitively with a path prepositional phrase, some of the manner of motion verbs can take on a sense of pseudo-motional existence, in which the subject does not actually move, but has a shape that could describe a path for the verb (e.g., *The stream twists through the valley*). These verbs are listed in the intersective classes with *meander* verbs of existence.

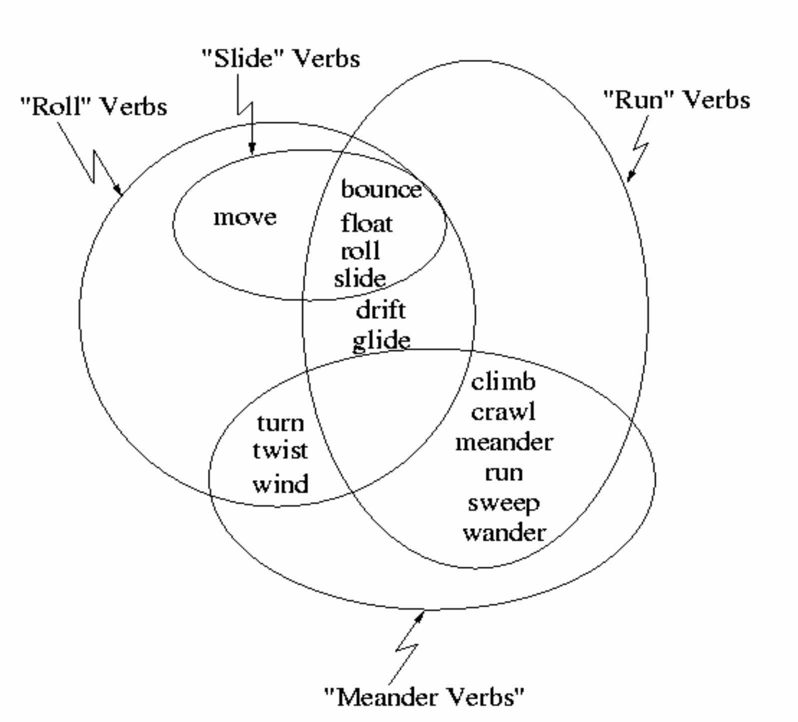

Figure 2. Intersections between *roll* and *run* verbs of motion and *meander* verbs of existence

Discussion We are currently constructing a lexical resource for verbs, VerbNet, that is intended to overcome some of the limitations of WordNet by addressing specifically the needs of natural language processing applications. This lexicon uses Levin classes as the basis of a hierarchical organization allowing for inheritance, and as a source of cross-linguistic semantic features such as the ones discussed here [5, 8, 4]. It also exploits the systematic link between syntax and semantics that motivates the Levin classes, and thus provides a clear and regular association between syntactic and semantic properties of verbs and verb classes. Specific sets of syntactic configurations and appropriate selectional restrictions on arguments are associated with individual senses. This lexicon gives us a first approximation of sense distinctions that are reflected in varying predicate argument structures. As described below, these entries provide a suitable foundation for directing consistent predicate-argument labeling of training data. To make our resource more usable, we are also mapping our verb entries to the relevant WordNet senses.

2.3 Possible Levin classes in Portuguese and Chinese

Preliminary investigations have indicated that a straightforward translation of Levin classes into other languages is not feasible [31], [32], [33]. However, we have found interesting parallels in how Portuguese and English treat regular sense extensions [5]. We examined a mapping between the English verbs discussed above and their Portuguese translations, most of which take the same alternations as in English and, by virtue of these alternations, achieve the same regular sense extensions. For instance, the Portuguese *carry* verbs naturally cluster into two different subclasses, based on their ability to take the conative, *(contra)*, and apart, *(separando)*, alternations as well as path prepositional phrases. The Portuguese motion verbs show more variance. Looking at the *slide/roll/run* class: *rebater (bounce), flutuar (float), rolar (roll) and deslizar (slide),* we see different alternations as well as different class memberships. In contrast with English, the resultative in Portuguese requires a gerund plus a reflexive, as in *A porta deslizou abrindo-se (The door slid opening itself),* and *flutuar* can also not be causativized without adding a light verb, *Maria fez o barco flutuar (Mary made the boat float).* This moves it out of the slide class and associates it with *derivar* (drift) and *planar* (glide) in the closely related *roll/run* class. As in English, *derivar* and *planar* are not externally controllable actions and thus don't take the causative/inchoative alternation common to other verbs in the *roll* class.

In this paper we make a preliminary examination of the Chinese translations of these same verb classes, and find that, although they differ more than the Portuguese verbs, there are still striking similarities.

Carry verbs

The Chinese translations of the *carry* verb class behave much more similarly to their English counterparts than we expected. Many of the verbs take part in the alternations that are direct translations of the English ones. However, there are some interesting differences:

Table 1. Chinese *carry* verbs with their alternations.[2]

English	Chinese	Conative*	Apart	Path
Carry	攜帶、運送	No	No	Yes
Heft	舉、舉起	No	No	Yes
Hoist	懸起、升	No	No	Yes
Tote	搬運	No	No	Yes
Tow	拖曳	No	No	Yes
Drag	拖、拖曳	No	Yes	Yes
Haul	拖、搬運	No	Yes	Yes
Draw	牽、引	No	Yes	Yes
Lug	搬、拖（重物）	No	Yes	Yes
Push	推	No	Yes	Yes
Pull	拉	No	Yes	Yes
Tug	用力拉	No	Yes	Yes
Yank	用力猛拉	No	Yes	Yes
Shove	撞、撞開	Yes	Yes	Yes
Kick	踢	Yes	Yes	Yes

Conative alternations:

English *carry* verbs do not allow conative alternation, however, from Table 1, one can see two of the Chinese verbs allow the conative alternation with a syntactic structure of "V 在 NP 上," e.g., 瑪莉撞在門上 *(Mary shoved at the door)*. The conative alternation in Chinese serves the same purpose as it does in English. They both insert a preposition between the verb and the object, and make it clear that the action is attempted but the object does not move. However, we have to add one more localizer after the object in order to generate a natural sentence in Chinese.

Apart alternations:

The *apart* alternation in Chinese, as shown in following example sentences, requires two changes in the syntactic frame of the base verb:

(1) 我 拉開 約翰 和 湯尼

I pull-apart John and Tony

[2] Chinese *carry* verbs naturally cluster into three different classes based on the conative, apart, and path prepositional phrases alternations, however, not like Portuguese, Chinese verb classes

(2) 我 把 約翰 和 湯尼 拉開

I ba John and Tony pull-apart

(3) 我 把 約翰 從 湯尼那邊 拉開

I ba John from Tony's side pull-apart

First, the Chinese verb combines with the verb 開, *apart*, and becomes a resultative compound. Second, in sentences (2) and (3), which are both a translation of the English sentence, *I pull John and Tony apart*, the Chinese word *ba* must be included.

Table 2. Chinese *slide/roll/run* and *roll/run* verbs with their alternations

	跳(回)、彈(回) (bounce)	浮、流動 (float)	滾動、轉動 (roll)	滑、溜 (slide)	漂流 (drift)	滑動、滑翔 (glide)
Dative	No	No	No	No	No	No
*conative	No	No	No	No	No	No
Caus./inch.	Yes	Yes	Yes	Yes	Yes	Yes
Middle	Yes	Yes	Yes	Yes	Yes	Yes
Resultative	Yes	Yes	Yes	Yes	Yes	Yes
Adj Part	Yes	Yes	Yes	Yes	Yes	Yes
Ind. Action	Yes	Yes	Yes	Yes	Yes	Yes
Locative Inversion	Yes	Yes	Yes	Yes	Yes	Yes
Measure	Yes	Yes	Yes	Yes	Yes	Yes
*Adj Perf	Yes ; N/A	Yes ; N/A	Yes ; N/A	Yes ; N/A	Yes ; N/A	Yes ; N/A
*cognate object	No	No	No	No	No	No
Zero nom.	Yes	Yes	Yes	Yes	Yes	Yes

***Manner of Motion* verbs**

We also investigated the Chinese translation of the *slide/roll/run*, and *roll/run* intersective classes.

Transitivity is not always preserved in the translations. For example, "滑 溜(*slide*)" does not take a direct object, so some of the alternations that are related to its transitive meaning are not present. For these verbs, we have to insert the marker "把-*ba*" into the sentence, and adjust the position of the verb, as in 約翰把書滑過桌面 (*John slid the books across the table*).

The resultative in Chinese is expressed using the verb-resultative compound; for example, "皮球滾開了(The ball slid away)"

---

do not map well to the intersective levin classes.

As can be seen in the table, the alternations for the Chinese translations of the verbs in this intersective class indicate that they share similar properties with the English verbs, except for dative alternation and adjectival perfect participle alternation. As shown in the following example sentences:

*Dative alternation:

(4) 我 把 球 滾 給 你

I ba- the ball roll to you

*(5) 我 滾 你 球

I rolled you the ball

Adjectival perfect participle alternation:[3]

(6) 已漂走的船

*drifted boat

The two Chinese verbs of the *roll/run* intersective class, drift and glide, behave in the same way as those four *slide/roll/run* verbs, it seems that these two subclasses should merge to be one class in Chinese.

3 Creating a corpus with Semantic Annotation

In this section we return to the task of data preparation, and directly address the question of how the regular sense extensions presented above can assist in revising WordNet entries to improve their suitability for semantic annotation. Our annotation efforts are aimed in part at coordinating with a new series of SENSEVAL workshops for testing Word Sense Disambiguation.

SIGLEX98-SENSEVAL[4] Assuming appropriate training data could be provided, an exercise was set up to evaluate different systems on the word sense disambiguation task.[5] A special issue of Computers and the Humanities that includes a detailed overview of the exercise and reports of individual systems will appear shortly [34]. This data was prepared using a set

[3] Whether there exists the adjectival perfect participle alternation in Chinese is still debating. If we treat that Chinese does not have this alternation and fill the table up with "N/A," we found that the behavior of the English and Chinese motion verbs are strikingly similar.
[4] http://www.itri.brighton.ac.uk/events/senseval 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.
[5] The exercise was also supported by Euralex, Elsenet, ECRAN and SPARKLE.

of senses from the Hector project [35], 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 there is a public-domain Gold Standard sense inventory being used for tagging that 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.[6] This includes traces for extraction and movement phenomena, and subscripts appended to standard parse-tree nonterminals (such as NP, PP) to indicate grammatical subject, (SBJ), and temporal and locative adjuncts, (TMP and LOC) with senses from WordNet 1.6. 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. By once again allowing the annotators to discuss difficult tagging cases and to revise WordNet entries if necessary, comparable inter-annotator rates were achieved [16]. Not surprisingly, the most problematic tags were for verbs. Out of 350 verbs in our corpus, 90 caused tagging difficulties. We also experimented with predicate-argument structure labeling and tagged the text automatically with Agent, Patient and Oblique roles labels and then hand corrected it. The automatic tagger correctly tagged 81% of the data [16].

3.1 Mismatches between lexical resources

Even if revised WordNet senses can be just as consistently tagged as Hector senses, this does not solve the problem of using testing data tagged with one lexical resource in order to evaluate a system based on a different one. 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* definitions.

[6] 100M words of Wall Street Journal text that has been hand annotated with part-of-speech tags and syntactic bracketing [12].

WordNet 1.6 had 8 senses for *shake*, with an additional 5 senses for *shake up* and 2 for *shake off*, 15 all together. Hector also has 8 main senses for *shake*, with the first one having 3 additional sub-senses 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*. 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 reasonable, since although these are all communicative acts, they involve different body parts and 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, since they all involve the same physical action, along with *She shook her cousin's hands*, as WN1. The following WN senses are distinguished 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. 3). Hector also includes *shake out* and *shake off* as examples of 1.1, the CLEAN sub-sense of 1, *Richard removed her socks and shoes and shook all the gravel out of them, The sand gets shaken off them*, then goes on to have separate entries for their more abstract variations as well, *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.

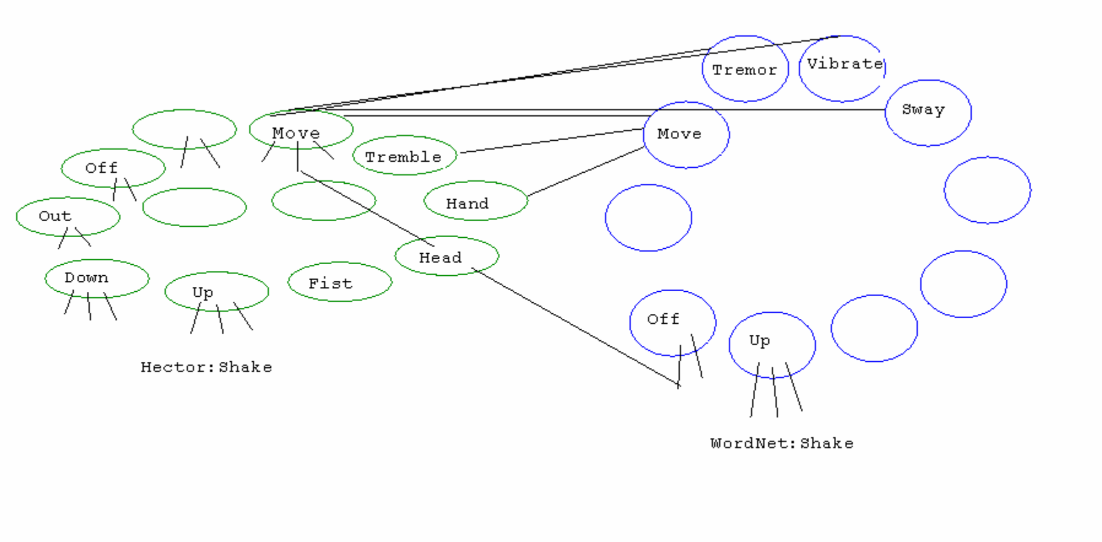

Figure 3. The discrepancies between the Hector *Shake* and WordNet 1.6 *Shake* definitions

Concrete criteria for *shake* sense distinctions 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 examining the mismatches between Hector and WordNet, we find that in order to distinguish the senses we often make reference to the very information that we are including in VerbNet: the inclusion of specific lexical items; different syntactic frames; different semantic class constraints on verb arguments; or differences in outcome. The simplest and most obvious distinguishers are lexical items such as the prepositions discussed earlier and verb particle constructions. *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 subtler and more difficult to capture. However, even semantic preferences can help to distinguish senses. Verb class membership itself, 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 4., 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 one's eyes*.) Then there are the abstract usages, which are 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, in common with the *Funnel* verbs, are distinguished syntactically from the others in that they cannot occur in the intransitive. Finally we are left with the idioms, which of course have to be listed individually. As illustrated by Figure 5, all of the Hector and WordNet senses can be categorized under one of these major divisions, thus providing more fine-grained distinctions if needed.

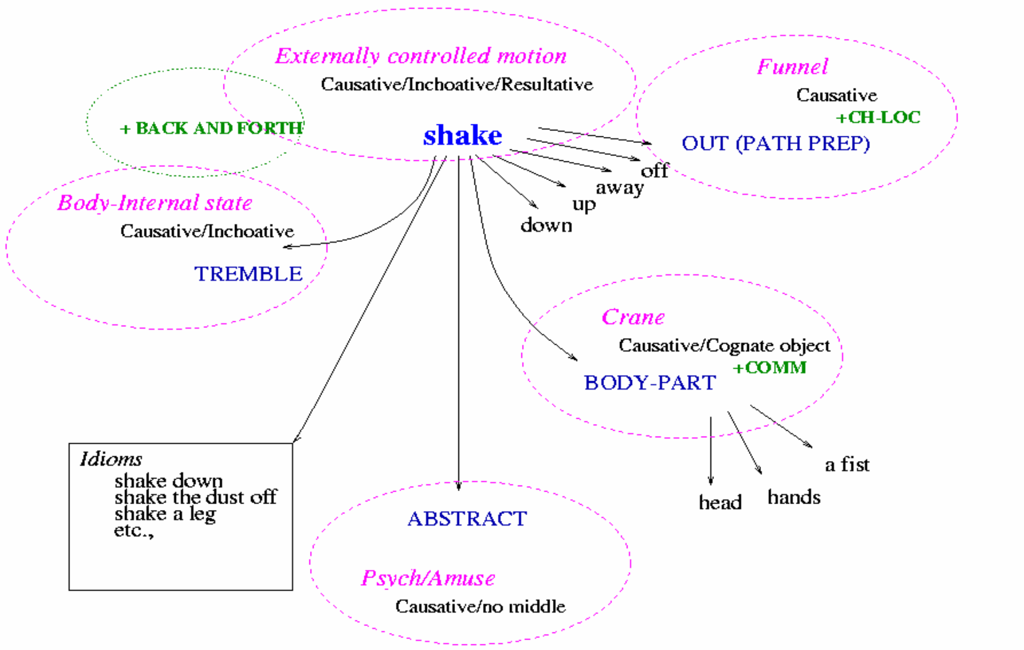

Figure 4. VerbNet representation for *Shake*

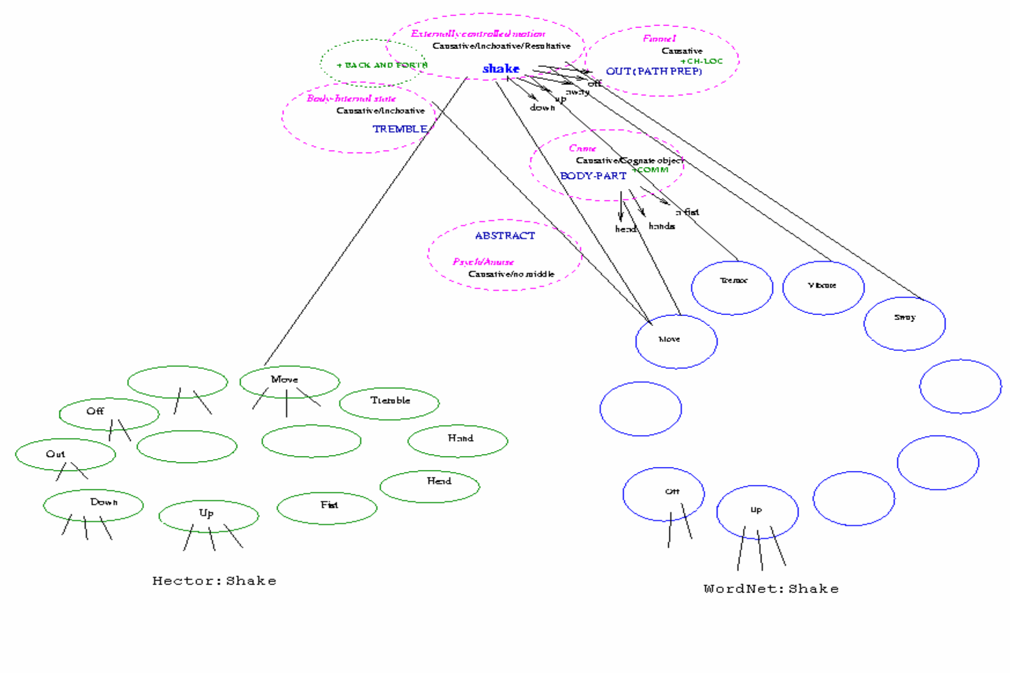

Figure 5. Mapping from VerbNet to Hector and WordNet: Shake

4 Conclusion

To summarize, we have presented results from our exploration of the close correlation between syntax and semantics that is evidenced by the Levin classes. The criteria for determining class membership is effective for clarifying distinctions between specific types of word senses, a major prerequisite for the semantic annotation of corpora. We have found evidence for similar criteria for classifications in other languages, tantalizing us with the promise of useful cross-linguistic generalizations about semantic components. This would prove invaluable to accurate multilingual information processing.

References

[1] Eugene Charniak. Parsing with Context-Free Grammars and Word Statistics. In Technical Report: CS-95-28, Brown University, 1995.

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

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

[4] Hoa Trang Dang, Karin Kipper, and Martha Palmer. Integrating compositional semantics into a verb lexicon. In Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000), Saarbr"ucken, Germany, July-August 2000.

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

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

[7] Fernando Pereira Ido Daggan, Lillian Lee. Similarity-based methods forword sense dismabiguation. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain, July 1997.

[8] Karin Kipper, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb

lexicon. In Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-2000), Austin, TX, July-August 2000.

[9] Beth Levin. English Verb Classes and Alternations A Preliminary Investigation. 1993.

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

[11] D. Magerman. Statistical decision-tree models for parsing. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 1995.

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

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

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

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

[16] Martha Palmer, Hoa Trang Dang, and Joseph Rosenzweig. Sense tagging the penn treebank. In Proceedings of the Second Language Resources and Evaluation Conference, Athens, Greece.

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

[18] Adwait Ratnaparkhi. A linear observed time statistical parser based on maximum entropy models. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, Providence, Rhode Island, August 1997.

[19] P. Resnik. Selection and Information: A Class-Based Approach to Lexical Relationships. PhD thesis, University of Pennsylvania Department of Computer and Information Sciences, 1993.

[20] B. Srinivas. Performance Evaluation of Supertagging for Partial Parsing. In Proceedings of Fifth International Workshop on Parsing Technology, Boston, USA, September 1997.

[21] Jiri Stetina and Makoto Nagao. Corpus Based PP Attachment Ambiguity Resolution with a Semantic Dictionary. In Jou Zhou and Kenneth Church, editors, Proceedings of the

Fifth Workshop on Very Large Corpora, pages 66-80, Beijing and Hong Kong, Aug. 18 - 20 1997.

[22] The XTAG-Group. A Lexicalized Tree Adjoining Grammar for English. Technical Report IRCS 95-03, University of Pennsylvania, 1995.

[23] D. Yarowsky. Three Machine Learning Algorithms for Lexical Ambiguity Resolution. PhD thesis, University of Pennsylvania Department of Computer and Information Sciences, 1995.

[24] Talmy, L. Path to realization: A typology of event conflation. Proceedings of the Berkeley Linguistics Society Parasession. Berkeley, CA, 1991.

[25] B. Levin and M. Rappaport-Hovav. Unaccusativity. At the syntax-semantics interface. MIT Press. 1995.

[26] Miller, George A. WordNet: An On-Line Lexical Database. International Journal of Lexicography, 3:235-312, 1990.

[27] Miller, George A. and Christiane Fellbaum. Semantic Networks of English. Lexical and Conceptual Semantics, Cognition Special Issue. Beth Levin and Steve Pinker. Elsevier Science Publishers, B. V. Amsterdam, The Netherlands, 1991. Pages, 197-229.

[28] Dorr, Bonnie J. and Doug Jones. Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision. Proceedings of SIGLEX. Santa Cruz, California. 1996.

[29] Doug Jones and Boyan Onyshkevych. Comparisons of Levin and WordNet. Presentation in working session of Semantic Tagging Workshop, ANLP-97. Washington, D.C., 1997.

[30] Martha Palmer. Consistent Criteria for Sense Distinctions, Special Issue of Computers and the Humanities, SENSEVAL98: Evaluating Word Sense Disambiguation Systems, Kluwer, 34: 1-2, 2000

[31] Jones, Douglas and Robert Berwick and Franklin Cho and Zeeshan Khan and Karen Kohl and Naouky Nomura and Anand Radhakrishnan Ulrich Sauerland and Bryan Ulicny, Verb Classes and Alternations in Bangla, German, English, and Korean, Massachussets Institute of Technology, 1994

[32] Nomura, Naoyuki and Douglas A. Jones and Robert C. Berwick, An Architecture for a Universal Lexicon: A Case Study on Shared Syntactic Information in Japanese, Hindi,

Bengali, Greek, and English, Proceedings of COLING, Santa Cruz, California, 1994, 243-249.

[33] Saint-Dizier, Patrick, Semantic Verb Classes Based on 'Alternations' and on WordNet-like Semantic Criteria: A Powerful Convergence, Proceedings the Workshop on Predicative Forms in Natural Language and Lexical Knowledge Bases, 1996, Toulouse, France.

[34] Kilgarriff99, Adam Kilgarriff and Martha Palmer, SENSEVAL: Evaluating Word Sense Disambiguation Systems, Computers and the Humanities, Adam Kilgarriff and Martha Palmer, 2000, 34, 1-2, Kluwer.

[35] Sue Atkins, Tools for computer-aided corpus lexicography: The Hector project, Acta Linguistica Hungarica, 1993, 41, 5-72

[36] Fellbaum, Christiane, ed. 1998. WordNet: An Electronic Lexical Database. MIT Press.