[*AMTA/SIG-IL First Workshop on Interlinguas*, San Diego, CA, October 28, 1997]

Associating semantic components with intersective Levin classes

Hoa Trang Dang, Joseph Rosenzweig, Martha Palmer
Department of Computer and Information Science
University of Pennsylvania
Philadelphia PA 19104-6389
htd/josephr/mpalmer@linc.cis.upenn.edu

1 Introduction

This paper examines the question of differences between a traditional interlingua approach and a transfer-based approach that uses cross-linguistic semantic features to generalize its transfer lexicon entries, and concludes that the two approaches share a common interest in lexical classifications that can be distinguished by cross-linguistic semantic features. The paper goes on to discuss current approaches to English classification, Levin classes [8] and WordNet [9]. We present a refinement of Levin classes - Intersective Classes - that shows interesting correlations to WordNet and that makes more explicit the semantic components that serve to distinguish different classes.

Tradition holds that an interlingua approach has a deeper analysis than a transfer approach, and that it can serve as a representation for many languages, thus providing for major gains in efficiency. Classic transfer approaches which are more syntactic are more amenable to statistical acquisition methods, but they fail to generalize their treatment of structural divergences. They also require that all possible language pairs be dealt with individually which, while an advantage for language-specific constructions such as idioms, is unnecessarily laborious for more frequently occurring items which exhibit regular syntactic behavior.

With the recent lexico-structural approach to transfer lexicons, [10, 11, 1], these approaches are no longer as distinct as traditionally viewed, and are not necessarily antithetical, in that they are both concerned with cross-linguistic semantic components. The lexico-structural approach gains efficiency by recognizing that structural correspondences hold for entire classes of lexical items. For example, a classic problem is the translation of motion verbs from English to French. In English the manner of motion can be incorporated into the matrix verb, with the direction of the motion being adjoined on by a prepositional phrase, as in *John swam across the lake.* In many cases, this is not allowed in French, where the direction becomes incorporated into the matrix verb, and the manner is adjoined on as an adverbial or a prepositional phrase, as in *Jean a traversé le lac à la nage,* (Jean crosses the lake by swimming). This type of structural correspondence has been typically handled best by interlingua approaches, since traditional transfer approaches required that every possible combination of manner of motion verb and path prepositional phrase be listed explicitly, and paired with its target language equivalent. The lexico-structural approach allows the entire class of English manner of motion verbs that have adjoined path prepositional phrases, to be associated in a single transfer lexicon entry with the class of French directed motion verbs with adjoined manners of motion. This is effected by treating manner of motion, path and directed motion as cross-linguistic semantic features that occur in both languages, and serve to anchor the correspondences [11]. These are the same basic components that Jackendoff ascribes to change-of-location verbs in his Lexical Conceptual Structures (LCS), GO, PATH and MANNER, [5]. A similar interlingua treatment, also based on LCS, would decompose the English phrase *swim across the lake* into the same three separate components which would constitute the predicates of the predicate-argument structure. This predicate argument structure, the LCS, then also serves as the representation for the French translation [2].

The new lexico-structural transfer approach is more similar to the interlingua approach in that they

both have a predicate-argument structure representation of the meaning of the sentence, which gives them roughly equivalent semantic depth. Another similarity is that the new transfer approach can also combine lexical items from several languages together into a single transfer lexicon entry, greatly simplifying the task of adding the mapping to a new language [10]. An important remaining difference is that the interlingua approach would claim that a single predicate-argument structure can serve as a common representation for many languages, whereas the transfer approach allows for language-specific predicate-argument structures..

A fundamental assumption of either approach, and the most important similarity, is that these classifications can be made based on distinguished semantic features, and that these semantic features will be relevant to classification schemes in other languages. Whether the classification schemes serve as a means of associating a single logical form composed of semantic primitives with many lexical items, as in the LCS approach, or as a means of enriching a set of logical forms with a collection of semantic features, the classifications still have to be determined, and the associations with semantic features have to be made. The rest of this paper discusses specific issues with respect to the association of semantic features with the classifications in English verbs.

2 Verb classes

Two current approaches to English verb classifications are WordNet synonym sets [9] and Levin classes [8]. WordNet is an on-line lexical database of English that currently contains about 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. WordNet was designed principally as a semantic network, and contains little syntactic information.

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, hence the term diathesis alternations [8]. The distribution of syntactic frames a verb can appear in determines its class membership. The fundamental assumption is that the syntactic frames are a direct reflection of the underlying semantics. Levin classes are supposed to provide very specific sets of syntactic frames that are associated with the individual classes.

The sets of syntactic frames associated with a particular Levin class 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 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 without the end result being achieved, but where the *cutting* manner can still be recognized, i.e., *John cut at the loaf.* 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 attempted *breaking* actions that can still be recognized. 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 *hit, push/pull, swat* and *poke* classes, although it is not commonly found otherwise.

2.1 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 [4], [6], [3]. However, it would be useful for the WordNet

synsets 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 when membership in a Levin class does in fact indicate shared semantic components. 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 synonymous, but others, such as *braid (bob, braid, brush, clip, coldcream, comb, condition, crimp, crop, curl, etc.)* do not, which at least partly explains the lack of overlap between Levin and WordNet.

The association of sets of syntactic frames with individual verbs in each class is not as straightforward as one might suppose. For instance, *carry* verbs are described as not taking the conative, **The mother carried at the baby,* and yet many of the verbs in the *carry* class *(push, pull, tug, heave, shove)* are also listed in the *push/pull* class, which does take the conative. This double listing of a verb in more than one class (many verbs are in three or even four classes) is open to interpretation. Does it indicate that more than one sense of the verb is involved, or is one sense primary, and the alternations for that class should take precedence over the alternations for the other classes the verb is listed in? Another example 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, definitely less tangible. Neither class can take the locative or the benefactive. The *steal* verbs also do not take the conative and the causative, and the *obtain* verbs do not take the *dative.* The grounds for deciding that a verb belongs in a particular class because of the alternations it does not take are elusive at best.

The confusion about what alternations actually apply to which verbs, and what the significance of a verb being in more than one class is, has hampered researchers' ability to reference Levin classes directly in applications. In the next section we describe an extension of the basic Levin verb classes, intersective Levin classes, that clarifies the issues around multiple listings and competing sets of alternations, as well as more precisely highlighting and isolating the meaning components of a verb class.

3 Intersective Levin classes

We began with the hypothesis that a verb that was double listed (or triple or quadruple listed) did not necessarily have more than one sense. Instead, we assumed that multiple listings simply indicated a further refinement of the semantic components associated with the verb that distinguished it subtly from other verbs in the same class. This refinement of semantic components would be more reliable and more consistent when several verbs participated in the multiple listing, i.e., when the multiple listing actually defined a well-formed subset of the original classes.

3.1 Construction of intersective classes

Many of the lexical items classified into Levin's verb classes are listed as members of more than one semantic class [8]. There are in fact 3104 verbs, but 4194 verb/class pairings, or *verb senses,* for an average of 1.35 senses per verb. Levin gives only a few informal indications about how to interpret a multiple listing for a verb. Sometimes the verb is listed in several classes because there is a systematic meaning relationship among them. Other times, the multiple categorization seems to be an idiosyncrasy involving two verbs that happen to have the same spelling, i.e., homonyms. For example, the verb *draw* is listed as a *remove* verb (class 10.1), as a *scribble* verb (class 25.2) and as a performance verb (class 26.7). 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.

To better understand the sense distinctions in the Levin database, we augmented her existing semantic classes with a set of *intersective* classes, which were created by grouping together sets of existing classes which shared some members. All sets 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 a coincidence, an idiosyncrasy involving individual verbs rather than a systematic relationship involving coherent sets of verbs. This filter allowed us to reject the potential intersective class that would have resulted from combining the *remove* verbs with the *scribble* verbs, for instance. The sole member of this intersection is the verb *draw.* 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 given in Figure 1.

1. Enumerate all sets $S = \{c_1,... ,c_n\}$ of semantic classes such that $|c_1 \cap ... \cap c_n| \geq \epsilon$, where ϵ is a **relevance cut-off**.
2. For each such S, define an *intersective class* I_S such that a verb $v \in I_S$ iff $v \in c_1 \cap ... \cap c_n$ $(S = \{c_1,..., c_n\})$, and there is no $S' = \{c'_1,..., c'_m\}$ such that $S \subset S'$ and $v \in c'_1 \cap ... \cap c'_m$ (= **subset criterion**).

Figure 1: Algorithm for identifying relevant semantic-class intersections

After filtering in this way, 129 intersective classes remained. We then reclassified the verbs in the database as follows. A verb was assigned membership in an intersective class if it was listed in each of the existing classes that were combined to form the new intersective class. Simultaneously, the verb was removed from the membership lists of those existing classes. For example, *draw* was added to the new class 25.2/26.7 (intersection of *scribble* and performance verbs), and removed from the old classes 25.2 and 26.7. As a result of this reclassification, the number of verb senses (verb/class pairings) in the database decreased by 500, to 3694 senses, and the average number of senses per verb dropped to 1.19. For example, the three senses of *draw* discussed above were reduced to two by combining *draw*/25.2 and *draw/26.7.* It was not always the case that the number of listings for a given verb decreased. Theoretically, the number of listings for a verb could have increased exponentially, since, if it belonged to n existing classes, it potentially could be reassigned to any of 2^n new intersective classes in the power set of the existing ones. To reduce this proliferation, a verb was not added to any intersective class if it was already a member of another larger intersective class containing all the elements of the first class, as illustrated in Figure 2 which indicates the formation of three new classes.

The remaining multiple listings seemed more uniformly to reflect truly idiosyncratic ambiguities. In addition, the resulting intersective classes suggested more precise classifications of their members' meanings, and also shed light on systematic meaning shifts in the English verb lexicon. For example, the systematic linking of *hit* verbs and verbs of sound emission gives insight into the detailed event semantics of a verb like *knock.* Another large intersective class is formed from the pairing of verbs of spatial configuration and verbs of assuming a position, making explicit the relationship between stative and eventive readings of verbs like *crouch* and *kneel.*

3.2 Comparisons to WordNet

Even though the Levin verb classes are defined by their syntactic behavior, they reflect semantic distinctions made by WordNet, a classification hierarchy defined in terms of purely semantic word relations (synonyms, hypernyms, etc.). When examining in detail the intersective classes just described, which emphasize not only the individual classes, but also their relation to other classes, we see a rich semantic lattice much like WordNet. This is exemplified by the Levin *cut* verbs and the intersective class formed by the *cut* verbs (class

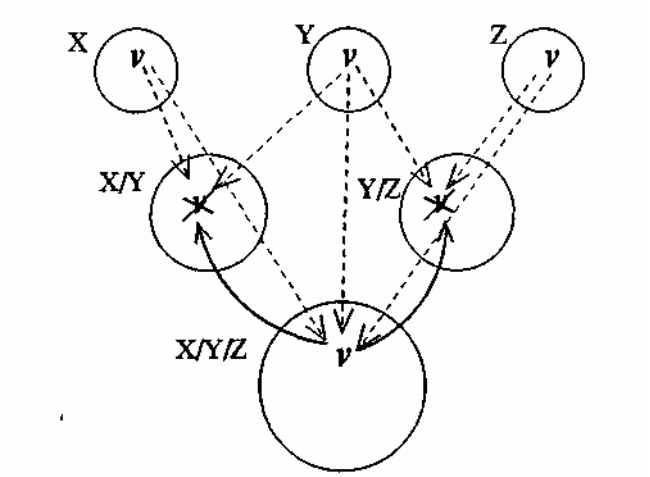

Figure 2: Filter for assigning membership to intersective classes

21.1) and *split* verbs (class 23.2) in Figure 3. The intersective class 21.1/23.2 *(cut, hack, hew, saw)* exhibits alternations of both parent classes, and has been augmented with *chip, clip, slash, snip* since these *cut* verbs also display the syntactic properties of *split* verbs. [1]

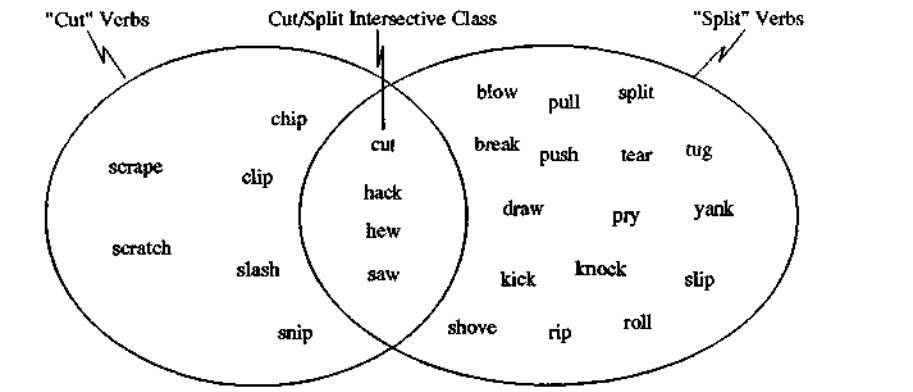

Figure 3: Intersective class formed from list of Levin *cut* verbs and *split* verbs

Figure 4 shows the augmented class membership for the Levin *cut* verbs, and their WordNet semantic classification. WordNet distinguishes two subclasses of *cut,* differentiated by the type of result:

1. Manner of cutting that results in separation into pieces *(chip, clip, cut, hack, hew, saw, slash, snip)*
2. Manner of cutting that doesn't separate completely *(scrape, scratch)*

This distinction appears in the second-order Levin classes as membership vs. nonmembership in the intersective class with *split.*

Levin verb classes are based on an underlying lattice of partial semantic descriptions, which are manifested indirectly in diathesis alternations. Whereas high level semantic relations (synonym, hypernym) are represented directly in WordNet, they can sometimes be inferred from the intersection between Levin verb classes.

[1] The list of members for each Levin verb class is not always complete, so to check if a particular verb belongs to a class it is better to check that the verb exhibits all the alternations that define the class. Since intersective classes were built using membership *lists* rather than the set of defining alternations, they were similarly incomplete. This is an obvious shortcoming of the current implementation of intersective classes, and might affect the choice of 3 as a threshold in later implementations.

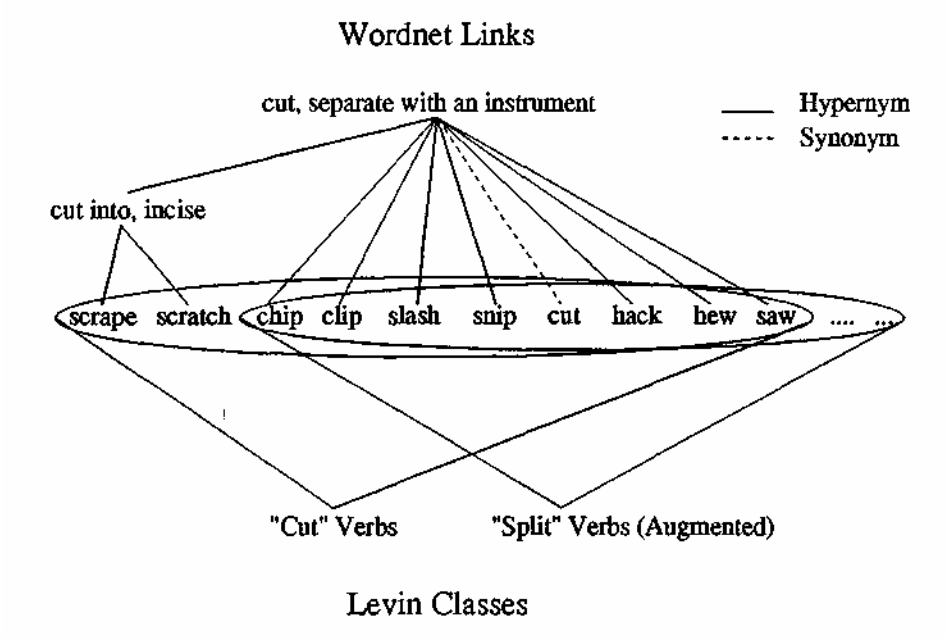

Figure 4: Levin *cut* verbs viewed as second-order classes reflect WordNet semantic relations

3.3 Using intersective Levin classes to isolate semantic components

In addition to Levin classes like *cut* whose members have core senses that are closely and systematically related in the WordNet hierarchy, other Levin classes are composed of verbs that exhibit a wider range of possible semantic components. The *split* verbs *(blow, break, cut, draw, hack, hew, kick, knock, pry, pull, push, rip, roll, saw, shove, slip, split, tear, tug, yank)* do not obviously form a tight semantic class. Instead, in their use as *split* verbs, each verb manifests an extended sense that can be paraphrased as "separate by V-ing," where "V" is the basic meaning of that verb [8]. Many of the verbs (e.g., *draw, pull, push, shove, tug, yank)* that do not have an inherent semantic component of "separating" belong to this class because of the component of *force* in their meaning. They are interpretable as verbs of splitting or separating only in particular syntactic frames. The adjunction of the *apart* adverb adds a *change of state* semantic component with respect to the object which is not present otherwise.

1. I pulled the twig and the branch apart.
2. I pulled the twig off (of) the branch.
3. *I pulled the twig and the branch.
   (on the interpretation of separating the twig and the branch)

These fringe *split* verbs appear in several other intersective classes that highlight the *force* aspect of their meaning. Figure 5 depicts the intersection of *split, carry* and *push/pull.*

The intersection between the *push/pull* verbs of exerting force (class 12), the *carry* verbs (class 11.4) and the *split* verbs (class 23.2) illustrates how the *force* semantic component of a verb can also be used to extend its meaning so that one can infer a: causation of accompanied motion. Depending on the particular syntactic frame in which they appear, members of this intersective class *(pull, push, shove, tug, kick, draw, yank* [2]) can be used to exemplify any one (or more) of the component Levin classes.

[2] Although *kick* is not listed *as* a verb of exerting force, it displays all the alternations that define this class. Similarly, *draw* and *yank* can be viewed as *carry* verbs although they are not listed as such.

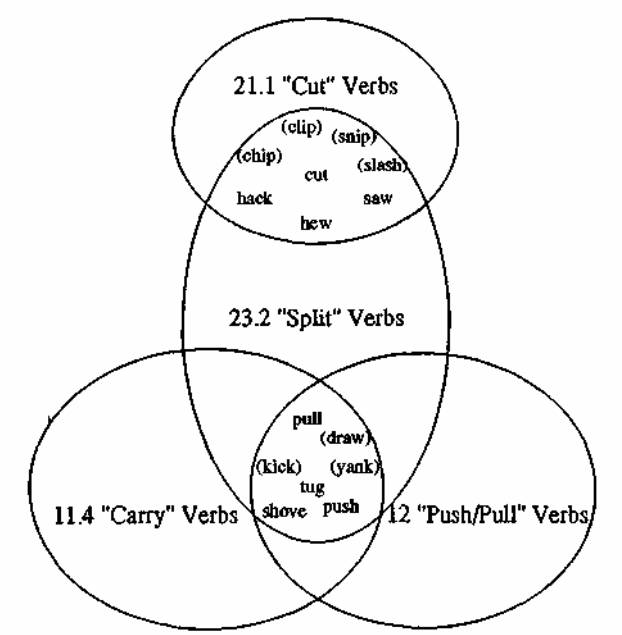

Figure 5: Intersective class formed from Levin *carry, push/pull* and *split* verbs - verbs in ( ) are not listed by Levin in all the intersecting classes but participate in all the alternations

1. Nora pushed the package to Pamela.
   (carry verb implies causation of accompanied motion, no separation)
2. Nora pushed at/against the package.
   (verb of exerting force, no separation or causation of accompanied motion implied)
3. Nora pushed the branches apart.
   (split verb implies separation, no causation of accompanied motion)
4. Nora pushed the package.
   (verb of exerting force; no separation implied, but causation of accompanied motion possible)
5. *Nora pushed at the package to Pamela.

Although the Levin classes that make up an intersective class may have conflicting alternations (e.g., verbs of exerting force can take the conative alternation, while *carry* verbs cannot), this does not invalidate the semantic regularity of the intersective class. As a verb of exerting force, *push* can appear in the conative alternation, which emphasizes its *force* semantic component and ability to express an "attempted" action where any result that might be associated with the verb (e.g., motion) is not necessarily achieved; as a *carry* verb (used with a goal or directional phrase), *push* cannot take the conative alternation, which would conflict with the core meaning of the *carry* verb class (i.e., causation of motion). The critical point is that, while the verb's meaning can be extended to either "attempted" action or directed motion, these two extensions cannot co-occur - they are mutually exclusive. However the simultaneous potential of mutually exclusive extensions is not a problem. It is exactly those verbs that are triple-listed in the *split/push/carry* intersective class (which have force exertion as a semantic component) that can take the conative. The *carry* verbs that are not in the intersective class *(carry, drag, haul, heft, hoist, lug, tote, tow)* are more "pure" examples of the *carry* class and always imply the achievement of causation of motion. Thus they cannot take the conative alternation.

4 Discussion

In this paper we have presented a more fine-grained analysis of the Levin classes which highlights the semantic components entailed by certain syntactic frames, and hence the semantic components of entire classes of verbs. We hypothesize that the semantic components we are identifying will be useful for cross-linguistic generalizations. An important avenue of future research which we intend to explore is the comparison of the translations of these classes to independently-defined classes in other languages, such as French verb classes [7] or European WordNet. [3]

These cross-linguistic generalizations will be equally valuable for both transfer-based and interlingua-based approaches to machine translation. Presumably both approaches need to be augmented with pragmatic information about tense and aspect and information structure, in particular coreference, in order to provide an adequate basis for translation in many circumstances. It could be argued that a language-specific predicate-argument structure will lend itself more readily to language-specific pragmatic annotation than a language-independent one, but it would still be necessary to ensure that the pragmatic annotation was meaningful in the target languages as well, i.e., cross-linguistic. The discovery of cross-linguistic pragmatic features is an equally important area for future research.

References

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

[2] B. Dorr. *Machine Translation: A View from the Lexicon.* MIT Press, Boston, Mass, 1993.

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

[4] Bonnie J. Dorr and Doug Jones. Acquisition of semantic lexicons: Using word sense disambiguation to improve precision. In *Proceedings of SIGLEX,* Santa Cruz, California, 1996.

[5] R. Jackendoff. *Semantic Structures.* MIT Press, Boston, Mass, 1990.

[6] Doug Jones and Boyan Onyshkevych. Comparisons of levin and wordnet. Presentation in working session of Semantic Tagging Workshop, ANLP-97, 1997.

[7] Christian Leclerc. Organisation du lexique-grammaire des verbes français. In *Langue Français: Dictionnaires Électroniques du Français.* Larousse, September 1990.

[8] B. Levin. *English Verb Classes and Alternations.* 1993.

[9] 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.

[10] Alexis Nasr, Owen Rambow, Martha Palmer, and Joseph Rosenzweig. Enriching lexical transfer with cross-linguistic semantic features. In *Proceedings of the Interlingua Workshop at the MT Summit,* San Diego, California, October 1997.

[11] Martha Palmer and Joseph Rosenzweig. Capturing motion verb generalizations with synchronous tags. In *Proceedings of AMTA-96,* Montreal, Quebec, October 1996.

[3] In the EuroWordNet project a multilingual database is developed with wordnets for four European Languages linked to the existing Princeton WordNet.