Investigating regular sense extensions based on intersective Levin classes

Hoa Trang Dang, Karin Kipper, Martha Palmer, Joseph Rosenzweig
Department of Computer and Information Sciences
and the Institute for Research in Cognitive Science
University of Pennsylvania
400A, 3401 Walnut Street/6228
Philadelphia, PA 19104, USA
htd/kipper/mpalmer/josephr@cis.upenn.edu

Abstract

In this paper we specifically address questions of polysemy with respect to verbs, and how regular extensions of meaning can be achieved through the adjunction of particular syntactic phrases. We see verb classes as the key to making generalizations about regular extensions of meaning. Current approaches to English classification, Levin classes and WordNet, have limitations in their applicability that impede their utility as general classification schemes. We present a refinement of Levin classes, intersective sets, which are a more fine-grained classification and have more coherent sets of syntactic frames and associated semantic components. We have preliminary indications that the membership of our intersective sets will be more compatible with WordNet than the original Levin classes. We also have begun to examine related classes in Portuguese, and find that these verbs demonstrate similarly coherent syntactic and semantic properties.

1 Introduction

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 use of automated or semi-automated methods of lexical acquisition. However, the field has yet to develop a clear consensus on guidelines for a computational lexicon that could provide a springboard for such methods, although attempts are being made (Pustejovsky, 1991), (Copestake and Sanfilippo, 1993), (Lowe et al., 1997), (Dorr, 1997). One of the most controversial areas has to do with polysemy. What constitutes a clear separation into senses for any one verb, and how can these senses be computationally characterized and distinguished? 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.

In this paper we specifically address questions of polysemy with respect to verbs, and how regular extensions of meaning can be achieved through the adjunction of particular syntactic phrases. We base these regular extensions on a fine-grained variation on Levin classes, intersective Levin classes, as a source of semantic components associated with specific adjuncts. We also examine similar classes in Portuguese, and the predictive powers of alternations in this language with respect to the same semantic components. The difficulty of determining a suitable lexical representation becomes multiplied when more than one language is involved and attempts are made to map between them. Preliminary investigations have indicated that a straightforward translation of Levin classes into other languages is not feasible (Jones et al., 1994), (Nomura et al., 1994), (Saint-Dizier, 1996). However, we have found interesting parallels in how Portuguese and English treat regular sense extensions.

2 Classifying verbs

Two current approaches to English verb classifications are WordNet (Miller et al., 1990) and Levin classes (Levin, 1993). WordNet is an on-line lexical database of English that currently contains approximately 120,000 sets of noun, verb, adjective, and adverb synonyms, each representing a lexicalized concept. A synset (syn-

The authors would like to acknowledge the support of DARPA grant N66001-94C-6043, ARO grant DAAH04-94G-0426, and CAPES grant 0914/95-2.

onym 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 (diathesis alternations) (Levin, 1993). The distribution of syntactic frames in which a verb can appear 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 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. 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.

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 (Dorr and Jones, 1996), (Jones and Onyshkevych, 1997), (Dorr, 1997). 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 when membership in a Levin class does in fact indicate shared semantic components. Of course, some Levin classes, such as *braid* (*bob, braid, brush, clip, coldcream, comb, condition, crimp, crop, curl*, etc.) are clearly not intended to be synonymous, 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, shove, kick)* are also listed in the *push/pull* class, which does take the conative. This listing of a verb in more than one class (many verbs are in three or even four classes) is left open to interpretation in Levin. 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 in which the verb is listed? 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.

3 Intersective Levin classes

We augmented the existing database of Levin semantic classes with a set of *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 the potential intersective class that would have resulted from combining the *remove* verbs with the *scribble* verbs, for example. 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 $S = \{c_1, \ldots, c_n\}$ of semantic classes such that $|c_1 \cap \ldots \cap c_n| \geq \epsilon$, where ϵ is a **relevance cut-off**.
2. For each such $S = \{c_1, \ldots, c_n\}$, define an *intersective class* I_S such that a verb $v \in I_S$ iff $v \in c_1 \cap \ldots \cap c_n$, and there is no $S' = \{c'_1, \ldots, c'_m\}$ such that $S \subset S'$ and $v \in c'_1 \cap \ldots \cap c'_m$ (**subset criterion**).

Figure 1: Algorithm for identifying relevant semantic-class intersections

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.

3.1 Using intersective Levin classes to isolate semantic components

Some of the large Levin classes comprise verbs that exhibit a wide range of possible semantic components, and could be divided into smaller subclasses. The *split* verbs (*cut, draw, kick, knock, push, rip, roll, shove, slip, split*, etc.) do not obviously form a homogeneous 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 (Levin, 1993). 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 (*I pulled the twig and the branch apart, I pulled the twig off (of) the branch*, but not **I pulled the twig and the branch*). The adjunction of the *apart* adverb adds a *change of state* semantic component with respect to the object which is not present otherwise. These fringe *split* verbs appear in several other intersective classes that highlight the *force* aspect of their meaning. Figure 2 depicts the intersection of *split, carry* and *push/pull*.

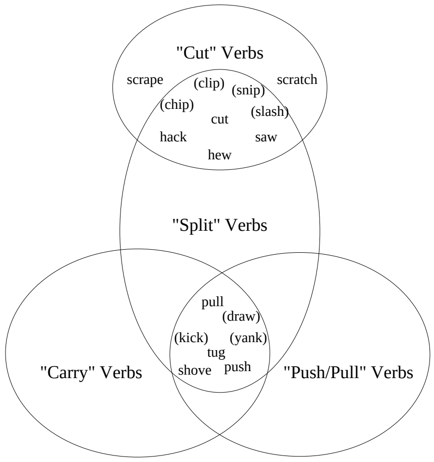

Figure 2: Intersective class formed from Levin *carry*, *push/pull* and *split* verbs - verbs in () are not listed by Levin in all the intersecting classes but participate in all the alternations

The intersection between the *push/pull* verbs of exerting force, the *carry* verbs and the *split* verbs 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*) * can be used to exemplify any one (or more) of the the component Levin classes.

1. Nora pushed the package to Pamela.
(*carry* verb implies causation of accompanied motion, no separation)
2. Nora pushed at/against the package.

*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. 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 relevance cut-off in later implementations.

(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.

3.2 Comparisons to WordNet

Even though the Levin verb classes are defined by their syntactic behavior, many 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 and *split* verbs. The original intersective class (*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.

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*), having *cut, separate with an instrument* as an immediate hypernym.

2. Manner of cutting that doesn't separate completely (*scrape, scratch*), having *cut into, incise* as an immediate hypernym, which in turn has *cut, separate with an instrument* as an immediate hypernym.

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, as with the *cut/split* class.

However, other intersective classes, such as the *split/push/carry* class, are no more consistent with WordNet than the original Levin classes. The most specific hypernym common to all the verbs in this intersective class is *move, displace*, which is also a hypernym for other *carry* verbs not in the intersection. In addition, only one verb (*pull*) has a WordNet sense corresponding to the *change of state - separation* semantic component associated with the *split* class. The fact that the *split* sense for these verbs does not appear explicitly in WordNet is not surprising since it is only an extended sense of the verbs, and separation is inferred only when the verb occurs with an appropriate adjunct, such as *apart*. However, *apart* can also be used with other classes of verbs, including many verbs of motion. To explicitly list separation as a possible sense for all these verbs would be extravagant when this sense can be generated from the combination of the adjunct with the *force* (potential cause of change of physical state) or motion (itself a special kind of change of state, i.e., of position) semantic component of the verb. WordNet does not currently provide a consistent treatment of regular sense extension (some are listed as separate senses, others are not mentioned at all). It would be straightforward to augment it with pointers indicating which senses are basic to a class of verbs and which can be generated automatically, and include corresponding syntactic information.

3.3 Sense extension for manner of motion

Figure 3 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 direction of motion, 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.

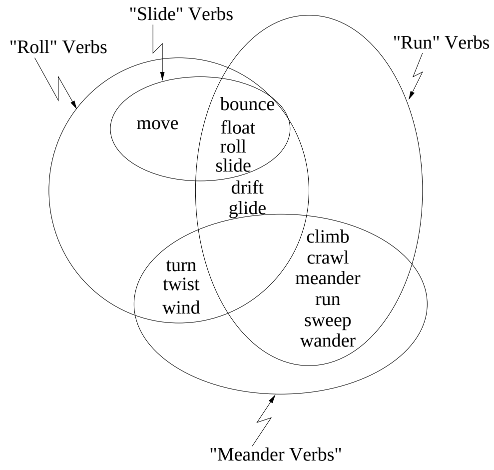

Figure 3: Intersections between *roll* and *run* verbs of motion and *meander* verbs of existence

4 Cross-linguistic verb classes

The Portuguese verbs we examined behaved much more similarly to their English counterparts than we expected. Many of the verbs participate in alternations that are direct translations of the English alternations. However, there are some interesting differences in which sense extensions are allowed.

4.1 Similar sense extensions

We have made a preliminary study of the Portuguese translation of the *carry* verb class. As in English, these verbs seem to take different alternations, and the ability of each to participate in an alternation is related to its semantic content. Table 1 shows how these Portuguese verbs naturally cluster into two different subclasses, based on their ability to take the conative and apart alternations as well as path prepositions. These subclasses correspond very well to the English subclasses created by the intersective class.

The conative alternation in Portuguese is mainly *contra* (against), and the apart alternation is mainly *separando* (separating). For example, *Eu puxei o ramo e o galho separando-os*

English	Portuguese	Conat.	Apart	Path
carry	levar	no	no	yes
drag	arrastar	no	yes	yes
haul	fretar	no	no	yes
heft	levantar com dificuldade	no	no	yes
hoist	icar	no	no	yes
lug	levar com dificuldade	no	no	yes
tote	levar facilmente	no	no	yes
tow	rebocar	no	no	yes
shove	empurrar com violencia	yes	yes	yes
push	empurrar	yes	yes	yes
draw	puxar	yes	yes	yes
pull	puxar	yes	yes	yes
kick	chutar	yes	yes	yes
tug	puxar com forca	yes	yes	yes
yank	arrancar	yes	yes	yes

Table 1: Portuguese *carry* verbs with their alternations

(*I pulled the twig and the branch apart*), and *Ele empurrou contra a parede* (*He pushed against the wall*).

4.2 Changing class membership

We also investigated the Portuguese translation of some intersective classes of motion verbs. We selected the *slide/roll/run*, *meander/roll* and *roll/run* intersective classes. Most verbs have more than one translation into Portuguese, so we chose the translation that best described the meaning or that had the same type of arguments as described in Levin's verb classes.

The elements of the *slide/roll/run* class are *rebater* (*bounce*), *flutuar* (*float*), *rolar* (*roll*) and *deslizar* (*slide*). The resultative in Portuguese cannot be expressed in the same way as in English. It takes a gerund plus a reflexive, as in *A porta deslizou abrindo-se* (*The door slid opening itself*). Transitivity is also not always preserved in the translations. For example, *flutuar* 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 the induced action alternation by using the light verb *fazer* (*make*) before the verb, as in *Maria fez o barco flutuar* (*Mary floated the boat*).

As can be seen in Table 2 the alternations for the Portuguese translations of the verbs in this intersective class indicate that they share similar properties with the English verbs, including the causative/inchoative. The exception to this, as just noted, is *flutuar* (*float*). The result of this is that *flutuar* should move out of the *slide* class, which puts 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. *Planar* doesn't take a direct object in Portuguese, and it shows the induced action alternation the same way as *flutuar* (by using the light verb *fazer*). *Derivar* is usually said as "estar a deriva" ("to be adrift"), showing its non-controllable action more explicitly.

5 Discussion

We have presented a refinement of Levin classes, intersective classes, and discussed the potential for mapping them to WordNet senses. Whereas each WordNet synset is hierarchicalized according to only one aspect (e.g., *Result*, in the case of *cut*), Levin recognizes that verbs in a class may share many different semantic features, without designating one as primary. Intersective Levin sets partition these classes according to more coherent subsets of features (*force, force+motion, force+separation*), in effect highlighting a lattice of semantic features that determine the sense of a verb. Given the incompleteness of the list of members of Levin classes, each verb must be examined to see whether it exhibits all the alternations of a class. This might be approximated by automatically extracting the syntactic frames in which the verb occurs in corpus data, rather than manual analysis of each verb, as was done in this study.

We have also examined a mapping between the English verbs that we have discussed and their Portuguese translations, which have several of the same properties as the corresponding verbs in English. Most of these verbs take the same alternations as in English and, by virtue of these alternations, achieve the same regular sense extensions.

There are still many questions that require further investigation. First, since our experiment was based on a translation from English to Portuguese, we can expect that other verbs in Portuguese would share the same alternations, so the classes in Portuguese should by no means be considered complete. We will be using resources such as dictionaries and on-line corpora to investigate potential additional members of our classes. Second, since the translation mappings may often be many-to-many, the alterna-

	rebater (bounce)	flutuar (float)	rolar (roll)	deslizar (slide)	derivar (drift)	planar (glide)
dative	yes		yes	yes		
*conative	no		no	no		
caus./inch.	yes		yes	yes		
middle	yes		yes	yes		
accept. coref.	yes		yes	yes		
caus./inch.	yes		yes	yes		
resultative	yes	yes	yes	yes	yes	yes
adject. part.	yes	yes	yes	yes	yes	yes
ind. action	yes	yes	yes	yes	no	yes
locat. invers.	yes	yes	yes	yes	yes	yes
measure	yes	yes	yes	yes	yes	yes
*adj. perf.	no	no	no	no	no	no
*cogn. object	no	no	no	no	no	no
zero nom.	yes	yes	no	yes	yes	yes

Table 2: Portuguese *slide/roll/run* and *roll/run* verbs with their alternations

tions may depend on which translation is chosen, potentially giving us different clusters, but it is uncertain to what extent this is a factor, and it also requires further investigation. In this experiment, we have tried to choose the Portuguese verb that is most closely related to the description of the English verb in the Levin class.

We expect these cross-linguistic features to be useful for capturing translation generalizations between languages as discussed in the literature (Palmer and Rosenzweig, 1996), (Copestake and Sanfilippo, 1993), (Dorr, 1997). In pursuing this goal, we are currently implementing features for motion verbs in the English Tree-Adjoining Grammar, TAG (Bleam et al., 1998). TAGs have also been applied to Portuguese in previous work, resulting in a small Portuguese grammar (Kipper, 1994). We intend to extend this grammar, building a more robust TAG grammar for Portuguese, that will allow us to build an English/Portuguese transfer lexicon using these features.

References

Tonia Bleam, Martha Palmer, and Vijay Shanker. 1998. Motion verbs and semantic features in tag. In *TAG+-98*, Philadelphia, PA. Submitted.

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

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

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

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

Douglas Jones, Robert Berwick, Franklin Cho, Zeeshan Khan, Karen Kohl, Naouky Nomura, Anand Radhakrishnan Ulrich Sauerland, and Bryan Ulicny. 1994. Verb classes and alternations in bangla, german, english, and korean. Technical report, Massachussets Institute of Technology.

Karin Kipper. 1994. Uma investigacao de utilizacao do formalismo das gramaticas de adjuncao de arvores para a lingua portuguesa. Master's Thesis, CPGCC, UFRGS.

B. Levin. 1993. *English Verb Classes and Alternations.*

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

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

Naoyuki Nomura, Douglas A. Jones, and Robert C. Berwick. 1994. An architecture for a universal lexicon: A case study on shared syntactic information in japanese, hindi, ben gali, greek, and english. In *Proceedings of COLING*, pages 243–249, Santa Cruz, California.

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

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

Patrick Saint-Dizier. 1996. Semantic verb classes based on 'alternations' and on wordnet-like semantic criteria: A powerful convergence. In *Proceedings the Workshop on Predicative Forms in Natural Language and Lexical Knowledge Bases*, Toulouse, France.