A Large-scale Classification of English Verbs

Karin Kipper
University of Colorado at Boulder
kipper@verbs.colorado.edu

Anna Korhonen
University of Cambridge
Computer Laboratory
alk23@cam.ac.uk

Neville Ryant
Computer and Information Science Department
University of Pennsylvania
nryant@unagi.upenn.edu

Martha Palmer
Department of Linguistics
University of Colorado at Boulder
martha.palmer@verbs.colorado.edu

Abstract. Lexical classifications have proved useful in supporting various natural language processing (NLP) tasks. The largest verb classification for English is Levin's (1993) work which defined groupings of verbs based on syntactic and semantic properties. VerbNet (Kipper et al., 2000; Kipper-Schuler, 2005) – the largest computational verb lexicon currently available for English – provides detailed syntactic-semantic descriptions of Levin classes. While the classes included are extensive enough for some NLP use, they are not comprehensive. Korhonen and Briscoe (2004) have proposed a significant extension of Levin's classification which incorporates 57 novel classes for verbs not covered (comprehensively) by Levin. Korhonen and Ryant (2005) have recently supplemented this with another extension including 53 additional classes. This article describes the integration of these two extensions into VerbNet. The result is a comprehensive Levin-style classification for English verbs providing over 90% token coverage of the Proposition Bank data (Palmer et al., 2005) and thus can be highly useful for practical applications.

Keywords: lexical classification, lexical resources, computational linguistics

1. Introduction

Lexical classes, defined in terms of shared meaning components and similar (morpho-)syntactic behavior of words, have attracted considerable interest in NLP (Pinker, 1989; Jackendoff, 1990; Levin, 1993). These classes are useful for their ability to capture generalizations about a range of (cross-)linguistic properties. For example, verbs which share the meaning component of 'manner of motion' (such as *travel*, *run*, *walk*), behave similarly also in terms of subcategorization (*I traveled/ran/walked*, *I traveled/ran/walked to London*, *I traveled/ran/walked five miles*) and usually have zero-related nominals (*a run*, *a walk*). Although the correspondence between the syntax and semantics of words is not perfect and these classes do not provide means for full semantic inferencing, their predictive power is nevertheless considerable.

NLP systems can benefit from lexical classes in a number of ways. Such classes define the mapping from surface realization of arguments to predicate-argument structure, and are therefore an important component of any system which needs the latter. As the classes can capture higher level abstractions (e.g. syntactic or semantic features) they can be used as a principled means to abstract away from individual words when required. They are also helpful in many operational contexts where lexical information must be acquired from small application-specific corpora. Their predictive power can help compensate for lack of sufficient data fully exemplifying the behavior of relevant words. Lexical classes have proved helpful in supporting a number of (multilingual) tasks, such as computational lexicography, language generation, machine translation, word sense disambiguation, semantic role labeling, and subcategorization acquisition (Dorr, 1997; Prescher et al., 2000; Korhonen, 2002). While this work has met with success, it has been small in scale. Large-scale exploitation of the classes has not been possible because no comprehensive classification is available.

The largest and most widely deployed classification in English is Levin's (1993) classification of verbs. VerbNet (VN) (Kipper et al., 2000; Kipper-Schuler, 2005) [1] – the most extensive on-line verb lexicon currently available for English – provides detailed syntactic-semantic descriptions of Levin classes organized into a refined taxonomy. While the original version of VN has proved useful for a variety of natural language tasks (e.g. semantic role labeling, robust semantic parsing, word sense disambiguation,) it has mainly dealt with Levin-style verbs (i.e. verbs taking noun (NP) and prepositional phrase complements (PP)) and thus has suffered from limited coverage. Some experiments have been reported which indicate that it should be possible, in the future, to automatically supplement VN with novel classes and member verbs from corpus data (Brew and Schulte im Walde, 2002; Korhonen et al., 2003; Kingsbury, 2004). While an automatic approach would avoid the expensive overhead of manual classification and enable application-specific tuning, the very development of the technology capable of large-scale classification requires access to a target gold standard classification more extensive than that available currently.

Korhonen and Briscoe (2004) (K&B) have proposed a substantial extension to Levin's original classification which incorporates 57 novel classes for verb types not covered (comprehensively) by Levin. Korhonen and Ryant (2005) (K&R) have recently supplemented this with another extension including 53 additional classes. While these novel classes are potentially very

[1] http://verbs.colorado.edu/verb-index/index.php

useful for the research community, their practical use is limited by the fact that no detailed syntactic-semantic descriptions are provided with the classes, and no attempt has been made to organize the classes into a taxonomy or to integrate them into Levin's taxonomy.

Our article addresses these problems: it describes the integration of these two sets of novel classes into VerbNet (Kipper et al., 2006a; Kipper et al., 2006b). Due to many differences between the three classifications their integration was a major linguistic task which had to be conducted largely manually to obtain any reliable result. The outcome is a freely available on-line resource which constitutes the most comprehensive and versatile Levin-style verb classification for English. After the two extensions VN has now also increased our coverage of PropBank tokens (Palmer et al., 2005) from 78.45% to 90.86%, making feasible the creation of a substantial training corpus annotated with VN thematic role labels and class membership assignments, to be released in 2007. This will finally enable large-scale experimentation on the utility of the classes for improving the performance of syntactic parsers and semantic role labelers on new domains.

We introduce Levin's classification in Section 2, VerbNet in Section 3 and the novel classes of K&B and K&R in section 4. Section 5 describes the integration of the new classes into VN, and section 6 describes how this integration affected VN and its coverage. Finally, section 7 discusses on-going and future work.

2. Levin's Classification

Levin's classification (Levin, 1993) provides a summary of the variety of theoretical research done on lexical-semantic verb classification over the past decades. In this classification, verbs which display the same or a similar set of diathesis alternations in the realization of their argument structure are assumed to share certain meaning components and are organized into a semantically coherent class. Although alternations are chosen as the primary means for identifying verb classes, additional properties related to subcategorization, morphology and extended meanings of verbs are taken into account as well.

For instance, the Levin class of "*Break* Verbs" (class 45.1), which refers to actions that bring about a change in the material integrity of some entity, is characterized by its participation (1-3) or non-participation (4-6) in the following alternations and other constructions (7-8):

1. **Causative/inchoative alternation**:
   Tony broke the window ↔ *The window broke*
2. **Middle alternation**:
   Tony broke the window ↔ *The window broke easily*

3. **Instrument subject alternation**:
Tony broke the window with the hammer ↔ *The hammer broke the window*
4. ***With/against alternation**:
Tony broke the cup against the wall ↔ **Tony broke the wall with the cup*
5. ***Conative alternation**:
Tony broke the window ↔ **Tony broke at the window*
6. ***Body-Part possessor ascension alternation**:
**Tony broke herself on the arm* ↔ *Tony broke her arm*
7. **Unintentional interpretation available (some verbs)**:
Reflexive object: **Tony broke himself*
Body-part object: *Tony broke his finger*
8. **Resultative phrase**:
Tony broke the piggy bank open, *Tony broke the glass to pieces*

Levin's taxonomy provides a classification of 3,024 verbs (4,186 senses) into 48 broad and 192 fine-grained classes according to their participation in 79 alternations involving NP and PP complements. Verbs taking ADJP, ADVP, ADL, particle, predicative, control and sentential complements are largely excluded, except where they show interesting behavior with respect to NP and PP complementation.

3. Description of VerbNet

VerbNet (VN) is a hierarchical domain-independent, broad-coverage verb lexicon with mappings to several widely-used verb resources, including WordNet (Miller, 1990; Fellbaum, 1998), Xtag (XTAG Research Group, 2001), and FrameNet (Baker et al., 1998). It includes syntactic and semantic information for classes of English verbs derived from Levin's classification and it is considerably more detailed than that included in the original classification.

Each verb class in VN is completely described by a set of members, thematic roles for the predicate-argument structure of these members, selectional restrictions on the arguments, and frames consisting of a syntactic description and semantic predicates with a temporal function, in a manner similar to the event decomposition of Moens and Steedman (1988). The original Levin classes have been refined and new subclasses added to achieve syntactic and semantic coherence among members. The resulting class taxonomy incorporates different degrees of granularity. This is an important quality given that the desired level of granularity varies from one NLP application to another.

3.1. Syntactic Frames

Each VN class contains a set of syntactic descriptions, or syntactic frames, depicting the possible surface realizations of the argument structure for constructions such as transitive, intransitive, prepositional phrases, resultatives, and a large set of diathesis alternations listed by Levin as part of each verb class. Each syntactic frame consists of thematic roles (such as *Agent*, *Theme*, *Location*), the verb, and other lexical items which may be required for a particular construction or alternation.

Semantic restrictions (such as *animate*, *human*, *organization*) are used to suggest preferences as to the types of thematic roles allowed in the classes. Each syntactic frame may also be constrained in terms of which prepositions are allowed.

Additionally, further restrictions may be imposed on thematic roles to indicate the syntactic nature of the constituent likely to be associated with the thematic role. Levin classes are characterized primarily by NP and PP complements. Some classes also refer to sentential complementation, although this extends only to the distinction between finite and nonfinite clauses, as in the various subclasses of *Verbs of Communication*. In VN the frames for class *Tell-37.2* shown in Examples (1) and (2) are illustrative of how the distinction between finite and nonfinite complement clauses is implemented.

(1) Sentential Complement (finite)
"Susan told Helen that the room was too hot."
Agent V Recipient Topic$[+sentential - infinitival]$

(2) Sentential Complement (nonfinite)
"Susan told Helen to avoid the crowd."
Agent V Recipient Topic[*+infinitival -wh_inf*]

3.2. Semantic Predicates

Each VN frame also contains explicit semantic information, expressed as a conjunction of boolean semantic predicates such as 'motion,' 'contact,' or 'cause.' Each of these predicates is associated with an event variable E that allows predicates to specify when in the event the predicate is true ($start(E)$ for preparatory stage, $during(E)$ for the culmination stage, and $end(E)$ for the consequent stage). Aspect in VN is captured by this event variable argument present in the predicates. For example, verbs that denote activities or processes, such as motion verbs, have predicates referring to the *during(E)* stage of the event. Relations between verbs (or classes) such as antonymy and entailment present in WordNet and relations between verbs (and verb classes) such as the ones found in FrameNet can be predicted by semantic predicates. For example, classes with change of location of the object, *Pocket* and *Remove*, have the same predicates *cause* and *location* used differently (negated in different places).

3.3. Status of VerbNet

Before integrating the novel classes, VN 1.0 had descriptions for 4,100 verb senses (over 3,000 lemmas) distributed in 191 first-level classes, and 74 new subclasses. These descriptions used 21 thematic roles, 36 selectional restriction preferences, 314 syntactic frames and 64 semantic predicates. The lexicon also relies on a shallow hierarchy of prepositions with 57 entries. The coverage of VN 1.0 has been evaluated through a mapping to almost 50,000 instances from Proposition Bank's corpus instances (Kingsbury and Palmer, 2002). VN syntactic frames account for over 78% of the exact matches found to the frames in PropBank. The information in the lexicon has proved useful for various NLP tasks such as word sense disambiguation and semantic role labeling (see Section 6). In VN 1.0 Levin's taxonomy has gained considerably in depth, but not in breadth. Verbs ADJP, ADVP, particle, predicative, control and sentential complements were still largely excluded. Many of these verb types are highly frequent in language and thus important for applications. As the new classes being proposed cover these verb types, it made sense to invest effort on incorporating them into VN.

4. Description of the new classes

4.1. The Classes of Korhonen and Briscoe (2004)

The resource of Korhonen and Briscoe (2004) includes a substantial extension to Levin's classification with 57 novel classes for verbs as well as 106 new diathesis alternations. The classes were created using the following semi-automatic approach[2]:

Step 1: A set of diathesis alternations were constructed for verbs not covered extensively by Levin. This was done by considering possible alternations between pairs of subcategorization frames (SCFs) in the comprehensive classification of Briscoe (2000) which incorporates 163 SCFs (a superset of those listed in the ANLT (Boguraev et al., 1987) and COMLEX Syntax dictionaries (Grishman et al., 1994)), focusing in particular on those SCFs not covered by Levin. The SCFs define mappings from surface arguments to predicate-argument structure for bounded dependency constructions, but abstract over specific particles and prepositions. 106 new alternations were identified manually, using criteria similar to Levin's.

[2] See Korhonen and Briscoe (2004) for the details of this approach and http://www.cl.cam.ac.uk/users/alk23/classes/ for the latest version of the classification.

Table I. ORDER VERBS

SCF 57:	*John ordered him to be nice*
SCF 104:	*John ordered that he should be nice*
SCF 106:	*John ordered that he be nice*
Alternating SCFs: 57 ↔ 104, 104 ↔ 106	

Step 2: 102 candidate lexical-semantic classes were selected for the verbs from linguistic resources of a suitable style and granularity: (Rudanko, 1996; Rudanko, 2000), (Sager, 1981), (Levin, 1993) and the LCS database (Dorr, 2001).

Step 3: Each candidate class was evaluated by examining sets of SCFs taken by its member verbs in syntax dictionaries (e.g. COMLEX) and whether these SCFs could be related in terms of diathesis alternations (from the 106 novel ones or Levin's original ones). Where one or several alternations were found which captured the sense in question, a new verb class was created.

Identifying relevant alternations helped to identify additional SCFs, which often led to the discovery of additional alternations. For those candidate classes which had an insufficient number of member verbs, new members were searched for in WordNet. These were frequently found among the synonyms, troponyms, hypernyms, coordinate terms and/or antonyms of the extant member verbs. The SCFs and alternations discovered during the identification process were used to create the syntactic-semantic description of each novel class. For example, a new class was created for verbs such as *order* and *require*, which share the approximate meaning of "direct somebody to do something". This class was assigned the description shown in Table I (where the SCFs are indicated by number codes from Briscoe's (2000) classification):

The work resulted in accepting, rejecting, combining and refining the 102 candidate classes and - as a by-product - identifying 5 new classes not included in any of the resources used. In the end, 57 new verb classes were formed, each associated with 2-45 member verbs. Table II shows a small sample of these classes along with example verbs.

4.2. THE CLASSES OF KORHONEN AND RYANT (2005)

While working on VerbNet and the integration of K&B classes, Korhonen and Ryant (2005) (K&R) uncovered 53 additional verb classes which deal with a wide range of different complements. Most of these classes cover prepositional complements and, as with the K&B classes, many of the classes add new frames with sentential complements. K&R classes also introduce a

Table II. Examples of K&B's Verb Classes

Class	**Example Verbs**
URGE	*ask, persuade*
FORCE	*manipulate, pressure*
WISH	*hope, expect*
ALLOW	*allow, permit*
FORBID	*prohibit, ban*
HELP	*aid, assist*
DEDICATE	*devote, commit*
LECTURE	*comment, remark*

large number of verb particles. Table III presents a small sample of these classes along with member verbs.

K&R classes were identified using the same methodology as in 3.1 (Step 3), associated with 2-37 member verbs and assigned similar syntactic descriptions as K&B classes. Table III presents a small sample of these classes along with member verbs.

Table III. Examples of K&R's Verb Classes

Class	**Example Verbs**
INTERROGATE	*interrogate, question*
ESTABLISH	*bring_about, open_up*
ADJUST	*adjust, adapt*
SUBJUGATE	*shut_up, subdue*
BEG	*request, supplicate*
COMPREHEND	*grasp, comprehend*

5. Incorporating the New Classes into VerbNet

Although the new classes of K&B and K&R are similar in style to the Levin classes included in VN, their integration to VN proved a major task. The first step was to assign the classes VN-style detailed syntactic-semantic descriptions. This was not straightforward because the K&B and K&R classes lacked explicit semantic descriptions and had syntactic descriptions not directly compatible with VN's descriptions. Also some of the descriptions available in VN had to be enriched for the new classes for this task. The second step was to incorporate the classes into VN. This was complicated by the fact that K&B and K&R are inconsistent in terms of granularity: some classes are broad while others are fine-grained. Also the comparison of the new classes to Levin's original classes had to be done on a class-by-class basis: some classes are entirely new, some are subclasses of existing classes, while others require reorganization of original Levin classes. These steps, which had to be conducted largely manually in order to obtain a reliable result, are described in the following sections.

5.1. Syntactic-Semantic Descriptions of Classes

Assigning syntactic-semantic descriptions to the new classes involved work on both VN and on the two new classifications. The set of SCFs in K&B and K&R is broad in coverage and relies, in many cases, on finer-grained treatment of sentential complementation than present in VN 1.0. Therefore, new VN syntactic descriptions had to be created and existing ones enriched with a more detailed treatment of sentential complementation. On the other hand, prepositional SCFs in K&B and K&R do not provide VN with explicit lists of allowed prepositions as required, so these had to be added to the classes. In addition, no syntactic description of the surface realization of the frames was included in K&B and K&R and had to be created. In some cases, the creation of new syntactic descriptions demanded a larger inventory of thematic roles than the existing one, therefore new roles were added. In addition, many new semantic predicates needed to be created to VN to convey the proper semantics of the integrated K&B and K&R new classes.

5.1.1. *Syntactic Descriptions*

Only 44 of VN's syntactic frames had a counterpart in the SCF classification assumed by K&B and K&R (Briscoe, 2000). This discrepancy is the by-product of differences in the design of the two resources. In his classification, Briscoe abstracts over prepositions and particles whereas VN differentiates between otherwise identical frames based on the precise types of prepositions that a given class of verbs subcategorizes for. Additionally, VN may distinguish two syntactic frames depending on thematic roles (e.g. there are two variants of the Material/Product Alternation Transitive frame differing on whether the object is the *Material* or *Product*).

Regarding sentential complements the opposite occurs, with VN conflating SCFs that Briscoe's classification considers distinct. In integrating the proposed classes into VN it was necessary to greatly enrich the set of possible syntactic restrictions VN allows on clauses. The original hierarchy contained only the valences ± *sentential*, ± *infinitival*, and ± *wh_inf*. The new set of possible syntactic restrictions consists of 57 such features accounting for object control, subject control, and different types of complementation (see Appendix 7 for a partial list of these features.)

Examples (3), (4), (5), and (6) show the VN realizations and the set of constraints for the proposed FORCE class (from K&B) which includes two frames with object control complements.

(3) Basic Transitive
“I forced him.”
Agent V Patient

(4) P-P-ING-OC (into-PP)
“I forced him Prep(into) coming.”
Agent V Patient into Proposition[+oc_ing]

(5) NP-PP (into-PP)
I forced John into the chairmanship.”
Agent V Patient into Proposition[-sentential]

(6) NP-TO-INF-OC
“I forced him to come.”
Agent V Patient Proposition[+oc_to_inf]

K&R classes also required the use of new SCFs not appearing in either VN or in any of the classes of the first candidate set given by K&B. These classes include USE, BASE, and SEEM with examples of new SCFs shown in Table IV.

Table IV. Examples of SCFs for K&R classes

Class	**SCF**	**Example**
BASE	NP-P-POSSING	*They based their plan on his seizing the base.*
BASE	NP-P-WH-S	*They based their claim on whether he happened to mention the danger.*
BASE	NP-P-NP-ING	*They based their objections on him failing to mention the dangers.*

5.1.2. *Thematic Roles*

In integrating the new classes, it was found that none of the 21 original VN thematic roles seemed to appropriately convey the semantics of the arguments for some classes. As an example, the members of the proposed URGE class (K&B) describe events in which one entity exerts psychological pressure on another to perform some action (*John urged Maria to go home*). While the urger (*John*) is assigned the role *Agent* as the volitional agent of the action and the urged entity (*Maria*) is assigned *Patient* as the affected participant, it is unclear what thematic role best suits the urged action (of going home). A new *Proposition* role was included which seemed to more appropriately describe the semantics of the "urge" action. Similar situations arose in the integration of 8 other classes. In the end, two new thematic roles were added to VN, *Content* and *Proposition*.

5.1.3. *Semantic descriptions*

Integrating the new classes also required enriching VN's set of semantic predicates. Whenever possible, existing VN predicates were reused. However, as many of the incoming classes represent concepts entirely novel to VN, it was necessary to introduce 30 new predicates to adequately provide descriptions of the semantics of these incoming classes. Examples of such predicates include *approve*, *spend*, *command*, and *attempt*.

5.2. Integrating the K&B Classes into VerbNet

After assigning the class descriptions, each of K&B's classes was thoroughly investigated to determine the feasibility of it being added to VN. Of the classes proposed, two were rejected as being either insufficiently semantically homogeneous or too small to be added to the lexicon, with the remaining 55 selected for incorporation. The classes fell into three different categories regarding Levin's classification: 1) classes that could be subclasses of existing Levin classes; 2) classes that would require a reorganization of Levin classes[3]; 3) entirely new classes.

5.2.1. *Entirely Novel Classes*

A total of 42 classes could be added to the lexicon as novel classes or subclasses without any restructuring. Some of these overlapped to an extent with existing VN classes semantically but syntactic behavior of the members was sufficiently distinctive to allow them to be added as new classes without restructuring of VN. 35 novel classes were actually added as new classes while 7 others were added as new subclasses (e.g. an additional novel subclass, *Continue-55.3*, was discovered in the process of subdividing *Begin-55.1*). The 35 new classes all share the quality of not overlapping to any appreciable extent with a pre-existing VN class from the standpoint of semantics. For instance, K&B's classes of FORCE, TRY, FORBID, and SUCCEED express entirely new concepts as compared to VN 1.0.

5.2.2. *Novel Sub-Classes*

Some of the proposed classes, such as CONVERT, SHIFT, INQUIRE, and CONFESS were considered sufficiently similar in meaning to current classes

[3] Levin focused mainly on NP and PP complements, but many verbs classify more naturally in terms of sentential complementation

and were added as new subclasses to existing VN classes. For example, both the proposed classes CONVERT and SHIFT are similar syntactically to the VN class *Turn-26.6*. However, whereas the members of *Turn-26.6* exclusively involve total physical transformations, the members of the proposed class CONVERT invariably exclude physical transformation, instead having a meaning that involves non-physical changes such as changes in the viewpoint of the Theme (*I converted the man to Judaism.*). Similarly, the verbs of SHIFT might be characterized as the class of verbs only taking the intransitive frames from CONVERT. Consequently, as both SHIFT and CONVERT are semantically similar, yet still distinct, from the existing VN class *Turn-26.6*, they were added as subclasses to 26.6, yielding the new classification *Turn-26.6.1*, *Convert-26.6.2*, and *Shift-26.6.3*.

5.2.3. *Classes Where Restructuring Was Necessary*

13 of the proposed classes overlapped significantly in some way with existing VN classes (either too close semantically or syntactically) and required restructuring of VN.

Classes such as WANT, PAY, and SEE obviously overlapped with existing VN classes *Want-32.1*, *Give-13.1*, and *See-30.1* in terms of meaning. Nor could the proposed classes be distinguished from the existing classes by recourse to syntactic behavior. Adding such classes required restructuring of VN to produce classes whose verb membership was the union of the overlapping proposed and existing classes and whose SCFs, similarly, were the union of those for each of the overlapping classes.

Broadly, the process of integrating the classes can be divided into two categories: 1) merging proposed classes with the related VN class; 2) adding the proposed class as a novel class but making modifications to existing VN classes.

Cases involving merger of a proposed class and an existing class: In considering these classes for addition to VN, it was observed that semantically their members patterned after a pre-existing class almost exactly. In the cases where the frames from the new classes were a superset of the frames recorded in VN, then the existing VN class was restructured by adding the new members and by enriching its syntactic description with the novel frames.

For example, both K&B's proposed WANT class and the VN class *Want-32.1* relate to the act of an experiencer desiring something. VN class *Want-32.1* differs from the proposed WANT class in its membership and in that it considers only alternations in NP and PP complements whereas the proposed class WANT also considered alternations in sentential complements, particularly control cases.

Added as new class but requiring restructuring of classes: K&B's work is of particular importance when considered in the context of classes of *Verbs With Predicative Complements*, whose members are frequent in language.

These verbs classify more naturally in terms of sentential rather than NP or PP complementation. The proposed class CONSIDER overlaps with four of VN's classes (*Appoint-29.1*, *Characterize-29.5*, *Declare-29.4*, and *Conjecture-29.6*), none of which were originally semantically homogeneous (see Figure 1.) The process of adding CONSIDER as another class of verbs with predicative complement gave us the opportunity to revise these four problematic classes making them more semantically homogeneous by using the more detailed coverage of complementation presented in K&B.

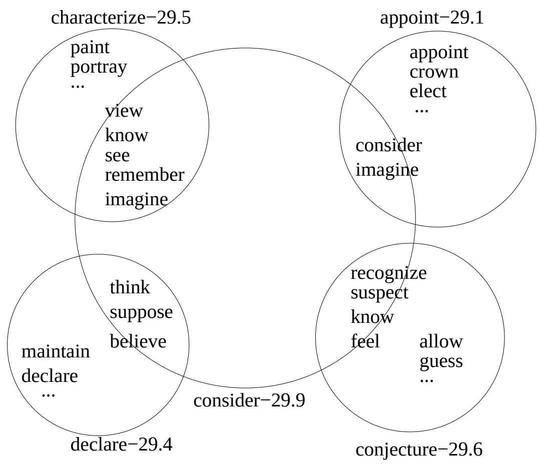

Figure 1. Original classes of Predicative Complement and the new *Consider-29.9* class

5.3. Integrating the K&R Classes into VerbNet

Integrating the second set of candidate classes proceeded much as the integration of the first set. Of the 53 suggested classes, 7 were omitted as they did not fully meet the requirements of Levin style syntactic-semantic classes, 11 were decided to overlap to a reasonable extent with a pre-existing class, and 36 were added as new classes (1 candidate class was divided into 2 new classes).

5.3.1. *Novel Classes and Subclasses*

In total, 35 classes from K&R were regarded as sufficiently novel for addition to VN without restructuring of an existing VN class. In addition, one class was divided into 2 new classes, PROMISE and ENSURE. As with K&B, 10 classes overlapped semantically, but not syntactically with existing VN classes, and hence were added as new subclasses. Examples of such classes include the proposed classes INTERROGATE and BEG, which were added as subclasses of the classes concerning Communication. The remaining 26 candidate classes were added as new classes. Examples include the classes RE-

Table V. Summary of the Lexicon's Extension

	VN 1.0	Extended VN
First-level classes	191	274
Thematic roles	21	23
Semantic Predicates	64	94
Select. Restr. (semantic)	36	36
Syntactic Restr. (on sent. compl.)	3	57
Lemmas	3445	3769
Verb senses	4656	5257

QUIRE, DOMINATE, SUBJUGATE, and HIRE, all of which express novel concepts.

5.3.2. *Additions to Existing Classes*

11 of the candidate classes overlapped significantly both syntactically and semantically with an existing class. Examples include CLARIFY (overlaps the EXPLAIN class of the first candidate set), DELEGATING POWER (overlaps ALLOW of first candidate set), BEING IN CHARGE OF (overlaps second candidate set DOMINATE). Unlike with K&B classes, very little restructuring was needed for these cases. In each of the 11 cases, the proposed class contained a subset of the SCFs in the class it overlapped with or contained one or two additional SCFs which were compatible with the pre-existing class.

6. The Extended Verbnet

A summary of how this integration affected VN and the result of the extended lexicon is shown in Table V. The figures show that our work enriched and expanded VN considerably. The number of first-level classes grew significantly (from 191 to 274). There was also a significant increase in the number of verb senses and lemmas, along with the set of semantic predicates and the syntactic restrictions on sentential complements.

We also examined the qualitative contributions of K&B and K&R to VN. The most salient difference among the two candidate sets is in the categories of activities they include. Many of the 42 classes of the K&B set tended to cluster among 3 broad categories:

1. **Classes describing the interaction of two animate entities**: There are 14 classes which describe interactions or relationships among entities in

some social context (see the following examples). The interaction can be either cooperative or non-cooperative and the two entities may or may not be thought to exist in some power relationship:

a) FORCE - *John forced Bill to go home.*

b) CONSPIRE - *John conspired with Bill to overthrow the government.*

c) BATTLE - *John battled with Bill over the insult.*

2. **Classes describing the degree of engagement of an entity with an activity**: There are 11 classes that involve an agent and an activity in which the agent is involved, but differ in how the Agent approaches the activity, e.g.

   a) TRY - *John tries to keep the house clean.*

   b) NEGLECT - *John neglected to wash the car.*

   c) FOCUS - *John focused on getting the car clean.*

3. **Classes describing the relation of an entity and some abstract idea**: There are 6 classes that describe relations between and abstract entities, such as whether the idea is a novel contribution of the entity or the entity's attitude toward the idea, e.g.

   a) DISCOVER - *John discovered that he can hold his breath for two minutes.*

   b) WISH - *John wishes to go home.*

The classes of the K&R set seem to address a much broader range of concepts (note that they also cover a wider range of complementation pattern types than the K&B classes). There, is, again, a group of 10 classes that could be considered broadly as describing social interactions among animate entities (i.e., DOMINATE, SUBJUGATE, HIRE). The remaining of the classes tend to form small clusters of 2-4 classes, or are among the 10 completely idiosyncratic classes.

1. **Small clusters**: For example, both ESTABLISH and PATENT classes describe activities of bringing into existence, but, unlike the existing Create-26.4 verbs, these new classes relate to the creation of abstractions such as organizations or ideas.

   a) ESTABLISH - *John tries to keep the house clean.*

   b) PATENT - *I patented my discovery with a gleeful smile.*

2. **Idiosyncratic classes**: Examples of these include classes such as USE, SEEM, and MULTIPLY.

   a) USE - *I utilized the new methodology in my research.*

   b) SEEM - *John seems a fool.*

   c) MULTIPLY - *The children divided each sum by the number of items in a simple exercise of statistics.*

With the integration of the new classes, which portray very diverse phenomena, the extended VN is now able to represent a much larger segment of the English language.

7. Conclusion and Future Work

Integrating the two recent extensions to Levin classes into VerbNet was an important step in order to address a major limitation of Levin's verb classification, namely the fact that verbs taking ADJP, ADVP, predicative, control and sentential complements were not included or addressed in depth in that work. This limitation excludes many verbs that are highly frequent in language.

An obvious question from the NLP point of view is the practical usefulness of the extended VN. When evaluating the usefulness of the current VN (extended with both K&B and K&R), the key issue is coverage, given the insufficient coverage has been the main limitation of the use of verb classes in practical NLP so far. In order to address this question, we investigated the coverage of the current VN over PropBank (Palmer et al., 2005) - the annotation of the Penn Treebank II with dependency structures. The list of verbs in VN before the class extensions included 3,445 lemmas which matched 78.45% of the verb tokens in the annotated PropBank data (88,584 occurrences). The new version of VN extended with verbs and classes included in K&B and K&R contains 3,769 lemmas. This greatly increased the coverage of VN to now match 90.86% of the PropBank verb occurrences (102,600 occurrences).

Korhonen and Briscoe (2004) showed that the K&B classes now incorporated in VN can be used to significantly aid subcategorization acquisition and that the extended classification has good coverage over WordNet. We can expect to see similar improved results on many NLP applications in the near future, given the wide use of VN in the research community. Currently, the use of verb classes in VN 1.0 is being attested in a variety of applications such as automatic verb acquisition (Swift, 2005), semantic role labeling (Swier and Stevenson, 2004), robust semantic parsing (Shi and Mihalcea, 2005), word sense disambiguation (Dang, 2004), building conceptual graphs (Hensman and Dunnion, 2004), and creating a unified lexical resource for knowledge extraction (Croch and King, 2005), among others.

In the future, we hope to extend VN's coverage further. We plan to search for additional novel classes and members using automatic methods, e.g. clustering. This is now realistic given the more comprehensive target and gold standard classification provided by VN. In addition, we plan to include in VN statistical information concerning the relative likelihood of different classes, SCFs and alternations for verbs in corpus data, using, e.g. the automatic methods proposed by McCarthy (2001) and Korhonen (2002). Such information can be highly useful for statistical NLP systems utilizing lexical classes.

Acknowledgments

This work was supported by National Science Foundation Grants NSF-9800658, VerbNet, NSF-9910603, ISLE, NSF-0415923, WSD, the DTO-AQUAINT NBCHC040036 grant under the University of Illinois subcontract to University of Pennsylvania 2003-07911-01 and DARPA grant N66001-00-1-8915 at the University of Pennsylvania. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, DARPA, or the DTO.

Appendix

VerbNet Screens

Appendix

List of syntactic features

+ac_bare_inf	*He helped bake the cake.*
+ac_ing	*She discussed writing novels.*
+ac_to_inf	*He helped to save the child.*
+acc_ing	*I kept them laughing.*
+adv_loc	*He put it there.*
+bare_inf	*He made her sing.*
+be_sc_ing	*She stopped smoking.*
-defi nite	*There raged a fire.*
+for_comp	*I need for her to be happy.*
+genitive	*The clown's antics amused the children.*
+gerund	*They limited smoking a pipe to the lounge.*
+how_extract	*He asked how she did it.*
+indicative	*For him to report the theft indicates that he wasn't guilty.*
+np_ing	*I discovered about him drinking.*
+np_omit_ing	*His hair needs combing.*
+np_p_ing	*They considered him as being stupid.*
+np_ppart	*He revealed the children found.*
+np_to_inf	*She relies on him to help.*
+np_tobe	*They allow us to be smokers.*
+oc_bare_inf	*He helped her bake the cake.*
+oc_ing	*I caught him stealing.*
+oc_to_inf	*I advised Mary to go.*
+plural	*The grocery carts thudded together.*
+pos_ing	*I saw their laughing and joking.*
+poss	*Nora pushed her way through the crowd.*
+poss_ing	*I loved him writing novels.*
+ppart	*He wanted the children found.*
+quotation	*Ellen warned Helen, 'Avoid that hole in the sidewalk.'*
+refl	*Marlene dressed herself.*
+rs_to_inf	*He seemed to come.*
+sc_ing	*He combed the woods looking for her.*
+sc_to_inf	*John promised Mary to resign.*
+sentential	*She gets through to him that he came.*
-sentential	*I worked as an apprentice cook.*
+small_clause	*I found him to smoke.*
-tensed_that	*They suggested to him that he go.*
+that_comp	*It annoys them that they left.*
+to_be	*I wished him to be nice.*
+to_inf	*It remains to find a cure.*
+to_inf_rs	*He continued to pack.*
+vc_to_inf	*They badgered him to go.*
+wh_comp	*They asked him whether he was going.*
+wh_inf	*He explained how to do it.*
+what_extract	*They made a great fuss about what they should do.*
+what_inf	*They made a great fuss about what to do.*
+wheth_comp	*They made a great fuss about whether they should participate.*
+wheth_inf	*They made a great fuss about whether to go.*

References

Baker, C. F., C. J. Fillmore, and J. B. Lowe: 1998, 'The Berkeley FrameNet project'. In: *Proceedings of the 17th International Conference on Computational Linguistics (COLING/ACL-98)*. Montreal, pp. 86–90.

Boguraev, B., T. Briscoe, J. Carroll, D. Carter, and C. Grover: 1987, 'The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English'. In: *Proceedings of the* 25^{th} *annual meeting of ACL*. Stanford, CA, pp. 193–200.

Brew, C. and S. Schulte im Walde: 2002, 'Spectral Clustering for German Verbs'. In: *Conference on Empirical Methods in Natural Language Processing*. Philadelphia, USA.

Briscoe, T.: 2000, *Dictionary and System Subcategorisation Code Mappings*. http://www.cl.cam.ac.uk/users/alk23/subcat/ subcat.html, University of Cambridge Computer Laboratory: Unpublished manuscript.

Croch, D. and T. H. King: 2005, 'Unifying Lexical Resources'. In: *In Proceedings of Interdisciplinary Workshop on the Identifi cation and Representation of Verb Features and Verb Classes*. Saarbruecken, Germany.

Dang, H. T.: 2004, 'Investigations into the Role of Lexical Semantics in Word Sense Disambiguation'. Ph.D. thesis, CIS, University of Pennsylvania.

Dorr, B. J.: 1997, 'Large-scale dictionary construction for foreign language tutoring and interlingual machine translation'. *Machine Translation* **12**(4), 271–325.

Dorr, B. J.: 2001, 'LCS Verb Database'. In: *Online Software Database of Lexical Conceptual Structures and Documentation*. University of Maryland.

Fellbaum, C. (ed.): 1998, *WordNet: An Eletronic Lexical Database*, Language, Speech and Communications. Cambridge, Massachusetts: MIT Press.

Grishman, R., C. Macleod, and A. Meyers: 1994, 'COMLEX syntax: building a computational lexicon'. In: *Proceedings of the International Conference on Computational Linguistics*. Kyoto, Japan.

Hensman, S. and J. Dunnion: 2004, 'Automatically building conceptual graphs using VerbNet and WordNet'. In: *In Proceedings of the 3rd International Symposium on Information and Communication Technologies (ISICT)*. Las Vegas, NV, pp. 115–120.

Jackendoff, R.: 1990, *Semantic Structures*. MIT Press, Cambridge, Massachusetts.

Kingsbury, P.: 2004, 'Verb clusters from PropBank annotation'. Technical report, University of Pennsylvania, Philadelphia, PA.

Kingsbury, P. and M. Palmer: 2002, 'From Treebank to PropBank'. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas, Canary Islands, Spain.

Kipper, K., H. T. Dang, and M. Palmer: 2000, 'Class-Based Construction of a Verb Lexicon'. In: *AAAI/IAAI*. pp. 691–696.

Kipper, K., A. Korhonen, N. Ryant, and M. Palmer: 2006a, 'Extending VerbNet with Novel Verb Classes'. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genova, Italy.

Kipper, K., A. Korhonen, N. Ryant, and M. Palmer: 2006b, 'A Large-Scale Extension of VerbNet with Novel Verb Classes'. In: *Proceedings of Euralex*. Turin, Italy.

Kipper-Schuler, K.: 2005, 'VerbNet: A broad-coverage, comprehensive verb lexicon'. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA.

Korhonen, A.: 2002, 'Semantically Motivated Subcategorization Acquisition'. In: *ACL Workshop on Unsupervised Lexical Acquisition*. Philadelphia.

Korhonen, A. and T. Briscoe: 2004, ' Extended Lexical-Semantic Classification of English Verbs '. In: *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*. Boston, MA.

Korhonen, A., Y. Krymolowski, and Z. Marx: 2003, 'Clustering Polysemic Subcategorization Frame Distributions Semantically'. In: *Proceedings of the 41st Annual Meeting of ACL*. Sapporo, Japan, pp. 64–71.

Korhonen, A. and N. Ryant: 2005, *Novel Lexical-Semantic Verb Classes*. http://www.cl.cam.ac.uk/users/alk23/classes2/: Unpublished manuscript.

Levin, B.: 1993, *English Verb Classes and Alternation, A Preliminary Investigation*. The University of Chicago Press.

McCarthy, D.: 2001, 'Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences'. Ph.D. thesis, University of Sussex.

Miller, G. A.: 1990, 'WordNet: An on-line lexical database'. *International Journal of Lexicography* **3**(4), 235–312.

Moens, M. and M. Steedman: 1988, 'Temporal Ontology and Temporal Reference'. *Computational Linguistics* **14**, 15–38.

Palmer, M., D. Gildea, and P. Kingsbury: 2005, 'The Proposition Bank: A Corpus Annotated with Semantic Roles'. *Computational Linguistics* **31**(1), 71–106.

Pinker, S.: 1989, *Learnability and Cognition: The acquisition of argument structure*. Cambridge: MIT Press.

Prescher, D., S. Riezler, and M. Rooth: 2000, 'Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution'. In: *18th International Conference on Computational Linguistics*. Saarbrücken, Germany, pp. 649–655.

Rudanko, J.: 1996, *Prepositions and Complement Clauses*. State University of New York Press, Albany.

Rudanko, J.: 2000, *Corpora and Complementation*. University Press of America.

Sager, N.: 1981, *Natural Language Information Processing: A Computer Grammar of English and Its Applications*. Addison-Wesley Publising Company, MA.

Shi, L. and R. Mihalcea: 2005, 'Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing'. In: *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, Mexico.

Swier, R. and S. Stevenson: 2004, 'Unsupervised Semantic Role Labelling'. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain, pp. 95–102.

Swift, M.: 2005, 'Towards automatic verb acquisition from VerbNet for spoken dialog processing'. In: *In Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. Saarbruecken, Germany.

XTAG Research Group: 2001, 'A Lexicalized Tree Adjoining Grammar for English'. Technical Report IRCS-01-03, IRCS, University of Pennsylvania.