

Establishing the Upper Bound and Inter-judge Agreement of a Verb Classification Task

Paola Merlo* and Suzanne Stevenson†

* University of Geneva
Department of Linguistics
2 rue de Candolle
1211 Genève 4
Switzerland
merlo@lettres.unige.ch

† Department of Computer Science,
and Center for Cognitive Science
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854-8019 USA
suzanne@cs.rutgers.edu

Abstract

Detailed knowledge about verbs is critical in many NLP and IR tasks, yet manual determination of such knowledge for large numbers of verbs is difficult, time-consuming and resource intensive. Recent responses to this problem have attempted to classify verbs automatically, as a first step to automatically build lexical resources. In order to estimate the upper bound of a verb classification task, which appears to be difficult and subject to variability among experts, we investigated the performance of human experts in controlled classification experiments. We report here the results of two experiments—using a forced-choice task and a non-forced choice task—which measure human expert accuracy (compared to a gold standard) in classifying verbs into three pre-defined classes, as well as inter-expert agreement. To preview, we find that the highest expert accuracy is 86.5% agreement with the gold standard, and that inter-expert agreement is not very high (K between .53 and .66). The two experiments show comparable results.

1. Introduction

Automatic lexical acquisition is a fundamental problem in natural language processing (Boguraev and Pustejovsky, 1996). Detailed knowledge about verbs in particular is critical in many NLP and IR tasks, yet manual determination of such knowledge for large numbers of verbs is a difficult, time-consuming and resource intensive task (Dang et al., 1998; Fellbaum, 1998; Levin, 1993; Miller et al., 1990). Furthermore, while syntactic properties of verbs such as subcategorization frames may be gleaned from a machine readable dictionary (Dorr, 1997), or from examples of usage in a corpus (Brent, 1993; Briscoe and Carroll, 1997; Lapata, 1999; Manning, 1993; McCarthy and Korhonen, 1998), the extraction of semantic properties of verbs poses a more challenging problem. On the assumption that syntactic properties of verb usage, and their frequency distributions, reflect underlying semantic properties of the verb, recent research has developed statistical corpus-based methods for learning selectional restrictions (Resnik, 1996; Riloff and Schmelzenbach, 1998), verbal aspect (Klavans and Chodorow, 1992; Siegel, 1999), and lexical semantic classification (Aone and McKee, 1996; Lapata and Brew, 1999; Schulte im Walde, 1998; Stevenson et al., 1999; Stevenson and Merlo, 2000).

Classification in particular has recently attracted attention, as it imposes a hierarchical organization that enables an NLP system to efficiently maintain and exploit generalizations over lexical items (Palmer, to appear). One type of classification of verbs that has inspired recent work is the

one in (Levin, 1993), which explores the hypothesis that the semantics of a verb determines the possible syntactic expressions of its arguments. Based on this assumption, Levin has developed a very influential framework, where verbs are classified into semantic classes that are revealed empirically by diathesis alternations—alternations in the syntactic expressions of arguments.

This kind of approach is very fruitful for NLP for two reasons. First, it organises the verbal lexicon around generalizations about argument structure—i.e., the thematic roles assigned to the arguments of a verb. These generalisations play a major role in many language engineering tasks. For example, knowledge about argument structure is crucial for dealing with dependency relations in parsing and generation (e.g., (Srinivas and Joshi, 1999; Stede, 1998)); for handling thematic divergences in machine translation (Dorr, 1997); for document profiling in information retrieval (Klavans and Kan, 1998); and for template filling in information extraction (Riloff and Schmelzenbach, 1998). Secondly, it supports corpus-based approaches. Corpus-based approaches to lexical semantic classification have drawn on Levin’s hypothesis (that the semantics of a verb determines the possible syntactic expressions of its arguments) by applying the converse reasoning—that is, assuming that similar subcategorization alternations correspond to similar underlying semantics. This method is especially promising for automatic lexical acquisition, since it is the syntactic properties of arguments that are more easily extractable from a corpus (Dorr and Jones, 1996; Lapata and

Brew, 1999; Schulte im Walde, 1998; Stevenson and Merlo, 2000).

These approaches have thus far led to some promising preliminary successes. For example, in using frequencies of syntactic features to automatically classifying verbs which share subcategorization frames but differ in argument structure, (Stevenson and Merlo, 2000) achieve 69.5% accuracy in a task whose baseline performance is 34%, and (Lapata and Brew, 1999) achieve 83.9% accuracy compared to a 61.3% baseline, both attaining reductions in error rate of over 50%. However, while accuracies of 70–84% seem like respectable initial results, the problem arises that we simply do not know what is a reasonable expectation for performance in this kind of task.

Our observation is that (human) classification based on alternations and argument structures of verbs engenders a lively theoretical debate on class membership of verbs, which requires complex linguistic information and expertise. This leads us to believe that the task is intrinsically difficult, and also likely to show differences in classification between experts. There is reason then to hypothesize that the expert-defined upper bound of classification algorithms will be lower than 100% accuracy, and show variability. Our goal then is to determine a realistic approximation to this upper bound, in order to enable more informed evaluation of verb classification algorithms.

We report here the results of two experiments—using a forced-choice task and a non-forced choice task—which measure expert accuracy (compared to a gold standard) in classifying verbs into three pre-defined classes, as well as inter-expert agreement. The forced-choice version corresponds most closely to the typical computational classification task, while the non-forced choice version provides a measure on what is assumed to be a more natural task for a human expert. To preview our results, we find that the highest expert accuracy is 86.5% agreement with the gold standard, and that expert agreement is not very high (a kappa value between .53 and .66). Moreover, the two experiments show comparable results.

This study provides several results of general interest:

1. It provides the experts' accuracy on a controlled classification task—providing the upper bound of performance for an automatic verb classification algorithm on a representative classification task.
2. It provides the degree of agreement between experts in both experiments—indicating how stable the expert performance is.
3. It provides results that are comparable across the two experiments—indicating that the controlled experiment is representative of a natural classification task.

2. The Classification Task

In choosing a representative type of classification task for our experiments on expert performance, we focused on the verb classification task that we have been addressing through automatic corpus-based learning methods (Stevenson and Merlo, 1999; Stevenson and Merlo, 2000). In our

computational experiments, we have been using frequencies of syntactic distributions from a very large corpus to train an automatic classifier to discriminate three lexical semantic classes from (Levin, 1993). This choice of task for our experiments with human experts clearly helps us in the practical problem of evaluating our own automatic classifier. However, the task is also representative of much of the recent work on verb classification, which is largely inspired by Levin's work, as discussed above.

Thus, we adopted a definition of verb classes based on their argument structure, and we used Levin's classification as the benchmark. Our experimental focus is also noteworthy in that the differences in the classes lies in their *argument structure*, not in their possible subcategorizations. That is, the three verb classes represented in the stimuli each support both the transitive and intransitive subcategorization frame. This allows us to manipulate only the argument structure variable, while holding subcategorization constant across the classes.

The stimuli were thus chosen from the following three classes, as exemplified in these sentences:

Unergative: The horse raced past the barn.
The jockey raced the horse past the barn.

Unaccusative: The butter melted in the pan.
The cook melted the butter in the pan.

Object-drop: The boy washed the hall.
The boy washed.

Each class is distinguished by the content of the thematic roles assigned by the verb. For object-drop verbs, the subject is an Agent and the optional object is a Theme, yielding the thematic assignments (Agent, Theme) and (Agent) for the transitive and intransitive alternants respectively. Unergatives and unaccusatives differ from object-drop verbs in participating in the causative alternation, and also differ from each other in their core thematic argument. In an intransitive unergative, the subject is an Agent, and in an intransitive unaccusative, the subject is a Theme. In the causative transitive form of each, this core semantic argument is expressed as the direct object, with the addition of a Causal Agent (the causer of the action) as subject in both cases.

The fact that the classes from which we draw the experimental stimuli are distinguished by argument structure and not subcategorization is important for two reasons. First, the task represents a key problem in verb classification, as alluded to above, since the ability to automatically learn thematic distinctions among verbs is a necessary component of automatic lexicon acquisition. Thus, it is important to determine a realistic upper bound on performance for this kind of task. Second, this property entails a fine-grained discrimination task that provides a challenging testbed for automatic classification, and (we hypothesize) for human expert classification as well. The verbs cannot be distinguished simply by the subcategorization frames they allow,

	E1		E2		E3	
	%Agr	K	%Agr	K	%Agr	K
E2	75%	.59				
E3	70%	.53	77%	.66		
LEVIN	71%	.56	86.5%	.80	83%	.74

Table 1: Percent Agreement (%Agr) and Pair-wise Agreement (K , Calculated by the Kappa Statistic) of Three Experts (E1, E2, E3), Compared to Each Other and to a Gold Standard (Levin).

enabling us to explore the level of performance that can be expected when discriminating verbs based on argument structure alone.

3. Experiments

The format, the task and the stimuli of both experiments were chosen to produce a controlled experimental situation. We decided to measure the upper bound for a very narrowly characterised classification task first, and to measure a slightly more relaxed version of it in comparison. Moreover, we needed to be able to compare responses. Therefore, we performed a closed-form questionnaire study, where the number and types of the target classes are defined in advance, for which we prepared a forced-choice and a non-forced-choice variant. The forced-choice study provides data for a maximally restricted experimental situation, which corresponds most closely to the automatic verb classification task. However, we are also interested in slightly more natural results—provided by the non-forced-choice task—where the experts can assign the verbs to an “Others” category.

3.1. Forced-Choice Experiment

We asked three experts in lexical semantics (all native speakers of English) to complete a forced-choice electronic questionnaire study. Materials consisted of individually randomized lists of 59 verbs selected using Levin’s index (Levin, 1993). The verbs were to be classified into three target classes—unergative, unaccusative, and object-drop—which were described in the instructions. The definitions of the classes were as follows. Unergative: A verb that assigns an agent theta role to the subject in the intransitive. If it is able to occur transitively, it can have a causative meaning. Unaccusative: A verb that assigns a patient/theme theta role to the subject in the intransitive. When it occurs transitively, it has a causative meaning. Object-Drop: A verb that assigns an agent role to the subject and patient/theme role to the object, which is optional. When it occurs transitively, it does not have a causative meaning. There were 20 unergative verbs, 19 unaccusative verbs, and 20 object-drop verbs. The list of verbs is given in the appendix; complete materials and instructions are available from the authors.

We calculated the pairwise agreement of the experts’ classifications, with the gold standard (Levin) and with each other. The results are indicated in Table 3.1. The percentage of verbs to which they all gave the same classification (60%) is smaller than any of the pairwise agreements,

indicating that the experts do not all agree on the same subset of verbs. Assessing the percentage of verbs on which the experts agree gives us an intuitive measure. However, this measure does not take into account how much the experts agree *over* the expected agreement by chance. This value is provided by the Kappa statistic, K , which we calculated following (Klauer, 1987, pages 55-57), to measure the experts’ degree of agreement over chance, with the gold standard and with each other (using the z distribution to determine significance; $p < 0.001$ for all reported results). These results are also shown in Table 3.1.

Expected chance agreement varies with the number and the relative proportions of categories used by the experts. This means that two given pairs of experts might reach the same percent agreement on a given task, but not have the same expected chance agreement, if they assigned verbs to classes in different proportions. The Kappa statistic ranges from 0, for no agreement above chance, to 1, for perfect agreement. The interpretation of the scale of agreement depends on the domain. (Carletta, 1996) cites the convention from the domain of content analysis indicating that $.67 < K < .8$ indicates marginal agreement, while $K \geq .8$ is an indication of good agreement. However, we can observe that only one of our agreement figures almost reaches what would be considered “good” under this interpretation. Given the very high level of expertise of our human experts, we suspect then that this is too stringent a scale for our task, which is qualitatively quite different from content analysis.

Evaluating the experts’ performance summarized in Table 3.1., we can remark two main things, which confirm our expectations. First, the percent agreement with the gold standard is not close to 100% even by trained experts. The best accuracy of a human expert (in comparison with the gold standard) is 86.5%, and the average accuracy of the three experts is 80%. Second, with respect to comparison of the experts among themselves, the rate of agreement is never very high, and the variability in agreement is considerable, ranging from .53 to .66. We can conclude then that the task we have set is very difficult. Based on these results, we note that 86.5% is a more realistic upper bound (than 100%) for evaluating machine performance in this three-way classification task.

Examining the pattern of average agreement between the experts and Levin, we find agreement on 17.7 (of 20) unergatives, 16.7 (of 19) unaccusatives, and 11.3 (of 20) object-drops. This clearly indicates that the object-drop class is the most difficult for the human experts to define. This class is the most heterogeneous in our verb list, consisting of verbs from several subclasses of the “unspecified object alternation” class in (Levin, 1993). We conclude that the verb classification task is likely easier for very homogeneous classes, and more difficult for more broadly defined classes, even when the exemplars share the critical syntactic behaviours.

The results from this first experiment are useful as they provide an upper bound for our task in a very controlled experimental situation, and they confirm the original expectations. Nonetheless, one possible shortcoming of the above experiment is that the forced-choice task, while maximally comparable to our computational experiments, may not be

a natural one for human experts. To explore this issue, we performed a non-forced choice version of the experiment.

3.2. Non-Forced Choice Experiment

We asked two additional experts in lexical semantics to complete the non-forced-choice electronic questionnaire study. (One expert was a native speaker of English and one bi-lingual; neither participated in the first experiment.) In addition to the three verb classes of interest, an answer of “Others” was allowed, for verbs that the experts could not classify within the three target classes. Materials consisted of individually randomized lists of 119 target and filler verbs taken from the electronic version of Levin’s index, available through Chicago University Press. The targets were again the same 59 verbs used in the forced-choice experiment. To avoid unwanted priming of target items, the 60 fillers were carefully selected from the set of verbs that do not share any class in Levin’s index with any of the senses of the 59 target verbs.

In this task, if we take only the target items into account, the experts agreed 74.6% of the time ($K=0.64$) with each other, and 86% ($K=0.80$) and 69% ($K=0.57$) with the gold standard. (If we take all the verbs into consideration, they agreed in 67% of the cases ($K=0.56$) with each other, and 68% ($K=0.55$) and 60.5% ($K=0.46$) with the gold standard, respectively.) These results show that the forced-choice and non-forced-choice task are comparable in accuracy of classification and inter-judge agreement on the target classes, giving us confidence that the forced-choice results provide a reasonably stable upper bound for computational experiments.

4. General Discussion

Going back to the original issues raised in the introduction, we can then draw the following conclusions. First, accuracy by experts in the 3-way forced choice task is between 71% and 86.5%, confirming the initial intuition that the task is difficult. Since it is not likely for the classification task to be solved at more than expert accuracy by automatic methods, we can conclude that 86.5% is a more realistic upper limit on the performance that we can expect in this task from an automatic classifier.

Second, the degree of agreement among experts in both tasks is relatively low. Carletta (1996) suggests that a kappa value of 0.8 or greater is an indication of good agreement. In neither of our tasks do highly trained experts reach that degree of agreement. The K values we obtain (between .53 and .66) are in a way rather striking because in fact they are quite low.

One might think that if classification is so difficult, then it might not be useful. We remark however that an individual expert consistently over- or under-estimates specific classes (compared to the gold standard). The fact that the experts make consistent mistakes suggests that they interpreted the defining properties given in the instructions, such as agent or theme, differently from each other. If this is the case, the disagreement can probably be partly solved by discussion and negotiation, as is done for instance in tagging and bracketing a corpus. If establishing a new classification, we can therefore recommend to follow corpus anno-

tation practice, and adopt a two stage classification process, where experts can discuss results after the first stage, and reclassify verbs after discussion.

Furthermore, it might be objected that the classes we have chosen make the task too hard. We should recall that the way we have chosen the verbs kept their subcategorization alternations constant between the classes, while changing only their argument structure. This set-up likely constitutes a more difficult discrimination task than when both argument structures and subcategorization frames can vary. From existing classification work, we know that many verbs can be classified by only looking at their subcategorization frames (Lapata and Brew, 1999). However, in some cases, subcategorization information might not be sufficiently informative for the purposes of classification. Our verbs are a case in point, since the three classes share subcategorization frames but differ in their pattern of thematic assignments. Thus, it seemed preferable to us to keep the subcategorization frame constant rather than varying both argument structures and subcategorization frames at the same time, as it provided a controlled experimental situation.

Finally, we observe that the measure of agreement across the two tasks is similar for the target items: in the first task, $K=.53$, $.59$, and $.66$ among experts, and $K=.56$, $.74$, and $.80$ with the gold standard; in the second task, $K=.64$ among experts and $K=.57$ and $.80$ with the gold standard. The average accuracy in the two experiments is also very close: 80% in the first, and 77% in the second. We conclude that, although not ideal, a forced-choice task can be representative of a natural task, and is informative about the difficulty of more open-ended classifications.

5. Conclusions

In conclusion, we have demonstrated an expert-based upper bound of 86.5%, far below the default maximum accuracy of 100%, in a lexical semantic classification task applied to verbs that share subcategorization frames but differ in argument structure (i.e., in the thematic roles they assign). Our results indicate that in a verb classification task of this kind, which depends on complex linguistic information and relationships, it is important to experimentally determine a realistic upper bound of accuracy for the task from human expert performance, in order to enable informed evaluation of automatic methods.

Appendix

The unergatives are manner of motion verbs: *floated*, *galloped*, *glided*, *hiked*, *hopped*, *hurried*, *jogged*, *jumped*, *leaped*, *marched*, *paraded*, *raced*, *rushed*, *scooted*, *scurried*, *skipped*, *tiptoed*, *trotted*, *vaulted*, *wandered*.

The unaccusatives are verbs of change of state: *boiled*, *changed*, *cleared*, *collapsed*, *cooled*, *cracked*, *dissolved*, *divided*, *exploded*, *flooded*, *folded*, *fractured*, *hardened*, *melted*, *opened*, *simmered*, *solidified*, *stabilized*, *widened*.

The object-drop verbs are unspecified object alternation verbs: *borrowed*, *called*, *carved*, *cleaned*, *danced*, *inherited*, *kicked*, *knitted*, *organised*, *packed*, *painted*, *played*,

reaped, rented, sketched, studied, swallowed, typed, washed, yelled.

6. References

- Aone, Chinatsu and Douglas McKee, 1996. Acquiring predicate-argument mapping information in multilingual texts. In Branimir Boguraev and James Pustejovsky (eds.), *Corpus Processing for Lexical Acquisition*. MIT Press, pages 191–202.
- Boguraev, Branimir and James Pustejovsky, 1996. Issues in text-based lexicon acquisition. In Branimir Boguraev and James Pustejovsky (eds.), *Corpus Processing for Lexical Acquisition*. MIT Press, pages 3–20.
- Brent, Michael, 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.
- Briscoe, Ted and John Carroll, 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Applied Natural Language Processing Conference*.
- Carletta, Jean, 1996. Assessing agreement on classification tasks: the Kappa statistics. *Computational Linguistics*, 22(2):249–254.
- Dang, Hoa Trang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig, 1998. Investigating regular sense extensions based on intersective Levin classes. In *Proc. of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*. Montreal, Canada: Université de Montreal.
- Dorr, Bonnie, 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12:1–55.
- Dorr, Bonnie and Doug Jones, 1996. Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proc. of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark.
- Fellbaum, Christiane, 1998. *Wordnet: an Electronic Lexical Database*. MIT Press.
- Klauser, Klaus J., 1987. *Kriteriumsorientierte Tests*. Göttingen, Germany: Verlag für Psychologie.
- Klavans, Judith and Min-Yen Kan, 1998. Role of verbs in document analysis. In *Proc. of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*. Montreal, Canada: Université de Montreal.
- Klavans, Judith L. and Martin Chodorow, 1992. Degrees of stativity: The lexical representation of verb aspect. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*.
- Lapata, Maria, 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*. College Park, MD.
- Lapata, Maria and Chris Brew, 1999. Using subcategorization to resolve verb class ambiguity. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language*. College Park, MD.
- Levin, Beth, 1993. *English Verb Classes and Alternations*. Chicago, IL: University of Chicago Press.
- Manning, Christopher D., 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Ohio State University, Association for Computational Linguistics.
- McCarthy, Diana and Anna Korhonen, 1998. Detecting verbal participation in diathesis alternations. In *Proc. of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*. Montreal, Canada: Université de Montreal.
- Miller, George, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, 1990. Five papers on Wordnet. Technical report, Cognitive Science Lab, Princeton University.
- Palmer, Martha, to appear. Consistent criteria for sense distinctions. *Computing for the Humanities*.
- Resnik, Philip, 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61(1–2):127–160.
- Riloff, Ellen and Mark Schmelzenbach, 1998. An empirical approach to conceptual case frame acquisition. In *Proceedings of the Sixth Workshop on Very Large Corpora*.
- Schulte im Walde, Sabine, 1998. Automatic semantic classification of verbs according to their alternation behaviour. Technical Report AIMS Report 4(3), Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
- Siegel, Eric, 1999. Corpus-based linguistic indicators for aspectual classification. In *Proceedings of ACL'99*. College Park, MD: University of Maryland.
- Srinivas, Bangalore and Aravind K. Joshi, 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.
- Stede, Manfred, 1998. A generative perspective on verb alternations. *Computational Linguistics*, 24(3):401–430.
- Stevenson, Suzanne and Paola Merlo, 1999. Verb classification using distributions of grammatical features. In *Proc. of the 9th Conference of the European Chapter for Computational Linguistics (EACL'99)*. Bergen, Norway.
- Stevenson, Suzanne and Paola Merlo, 2000. Automatic lexical acquisition based on statistical distributions. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbruecken, Germany.
- Stevenson, Suzanne, Paola Merlo, Natalia Kariaeva, and Kamin Whitehouse, 1999. Supervised learning of lexical semantic verb classes using frequency distributions. In *Proceedings of SigLex99: Standardizing Lexical Resources (SigLex'99)*. College Park, Maryland.