Using Semantic Preferences to Identify Verbal Participation in Role Switching Alternations.

Diana McCarthy
Cognitive & Computing Sciences,
University of Sussex
Brighton BN1 9QH, UK
dianam@cogs.susx.ac.uk

Abstract

We propose a method for identifying diathesis alternations where a particular argument type is seen in slots which have different grammatical roles in the alternating forms. The method uses selectional preferences acquired as probability distributions over WordNet. Preferences for the target slots are compared using a measure of distributional similarity. The method is evaluated on the causative and conative alternations, but is generally applicable and does not require a priori knowledge specific to the alternation.

1 Introduction

Diathesis alternations are alternate ways in which the arguments of a verb are expressed syntactically. The syntactic changes are sometimes accompanied by slight changes in the meaning of the verb. An example of the causative alternation is given in (1) below. In this alternation, the object of the transitive variant can also appear as the subject of the intransitive variant. In the conative alternation, the transitive form alternates with a prepositional phrase construction involving either *at* or *on*. An example of the conative alternation is given in (2).

1. The boy broke the window $\leftrightarrow$ The window broke.
2. The boy pulled at the rope $\leftrightarrow$ The boy pulled the rope.

We refer to alternations where a particular semantic role appears in different grammatical roles in alternate realisations as "role switching alternations" (RSAs). It is these alternations that our method applies to.

Recently, there has been interest in corpus-based methods to identify alternations (McCarthy and Korhonen, 1998; Lapata, 1999), and associated verb classifications (Stevenson and Merlo, 1999). These have either relied on a priori knowledge specified for the alternations in advance, or are not suitable for a wide range of alternations. The fully automatic method outlined here is applied to the causative and conative alternations, but is applicable to other RSAs.

2 Motivation

Diathesis alternations have been proposed for a number of NLP tasks. Several researchers have suggested using them for improving lexical acquisition. Korhonen (1997) uses them in subcategorization frame (SCF) acquisition to improve the performance of a statistical filter which determines whether a SCF observed for a particular verb is genuine or not. They have also been suggested for the recovery of predicate argument structure, necessary for SCF acquisition (Briscoe and Carroll, 1997; Boguraev and Briscoe, 1987). And Ribas (1995) showed that selectional preferences acquired using alternations performed better on a word sense disambiguation task compared to preferences acquired without alternations. He used alternations to indicate where the argument head data from different slots can be combined since it occupies the same semantic relationship with the predicate.

Different diathesis alternations give different emphasis and nuances of meaning to the same basic content. These subtle changes of meaning are important in natural language generation (Stede, 1998).

Alternations provide a means of reducing redundancy in the lexicon since the alternating SCFs need not be enumerated for each individual verb if a marker is used to specify which verbs the alternation applies to. Alternations also provide a means of generalizing patterns of behaviour over groups of verbs, typically the group members are semantically related. Levin (1993) provides a classification of over 3000 verbs according to their participation in alternations involving NP and PP constituents. Levin's classification is not intended to be exhaustive. Automatic identification of alternations would be a useful tool for extending the classification with new participants. Levin's taxonomy might also be used alongside observed behaviour, to predict unseen behaviour.

Levin's classification has been extended by other NLP researchers (Dorr and Jones, 1996; Dang et al.,

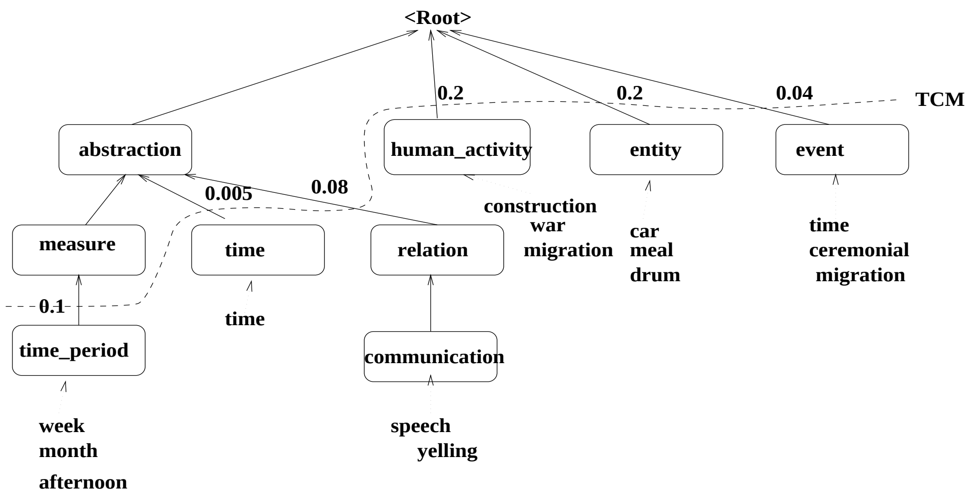

Figure 1: TCM for the object slot of the transitive frame of *start*.

1998). Dang et al. (1998) modify it by adding new classes which remove the overlap between classes from the original scheme. Dorr and Jones (1996) extend the classification by using grammatical information in LDOCE alongside semantic information in WordNet. What is missing is a way of classifying verbs when the relevant information is not available in a manmade resource. Using corpora by-passes reliance on the availability and adequacy of MRDs. Additionally, the frequency information in corpora is helpful for estimating alternation productivity (Lapata, 1999). Estimations of productivity have been suggested for controlling the application of alternations (Briscoe and Copestake, 1996). We propose a method to acquire knowledge of alternation participation directly from corpora, with frequency information available as a by-product.

3 Method

We use both syntactic and semantic information for identifying participants in RSAs. Firstly, syntactic processing is used to find candidates taking the alternating SCFs. Secondly, selectional preference models are acquired for the argument heads associated with a specific slot in a specific SCF of a verb.

We use the SCF acquisition system of Briscoe and Carroll (1997), with a probabilistic LR parser (Inui et al., 1997) for syntactic processing. The corpus data is POS tagged and lemmatised before the LR parser is applied. Subcategorization patterns are extracted from the parses, these include both the syntactic categories and the argument heads of the constituents. These subcategorization patterns are then classified according to a set of 161 SCF classes. The SCF entries for each verb are then subjected to a statistical filter which removes SCFs that have occurred with a frequency less than would be expected by chance. The resulting SCF lexicon lists each verb with the SCFs it takes. Each SCF entry includes a frequency count and lists the argument heads at all slots.

Selectional preferences are automatically acquired for the slots involved in the role switching. We refer to these as the target slots. For the causative alternation, the slots are the direct object slot of the transitive SCF and the subject slot of the intransitive. For the conative, the slots are the direct object of the transitive and the PP of the `np v pp` SCF.

Selectional preferences are acquired using the method devised by Li and Abe (1995). The preferences for a slot are represented as a tree cut model (TCM). This is a set of disjoint classes that partition the leaves of the WordNet noun hypernym hierarchy. A conditional probability is attached to each of the classes in the set. To ensure the TCM covers all the word senses in WordNet, we modify Li and Abe's original scheme by creating hyponym leaf classes below all WordNet's hypernym (internal) classes. Each leaf holds the word senses previously held at the internal class. The nominal argument heads from a target slot are collected and used to populate the WordNet hierarchy with frequency information. The head lemmas are matched to the classes which contain them as synonyms. Where a lemma appears as a synonym in more than one class, its frequency count is divided between all classes for which it has direct membership. The frequency counts from hyponym classes are added to the count for each hypernym class. A root node, created above all the WordNet roots, contains the total frequency count for all the argument head lemmas found within WordNet. The minimum description length principle (MDL) (Rissanen, 1978) is used to find the best TCM by considering the cost (in bits) of describing both the model and the argument head data encoded in the model. The cost (or description length) for a TCM is calculated according to equation 1. The number of parameters of the model is given by k, this is the number of classes in the TCM minus one. S is the sample size of the argument head data. The cost of describing each argument head (n) is calculated using the log of the probability estimate for the classes on the TCM that n belongs to (c_n).

$$\textit{description length} = \frac{k}{2} \times \log |S| - \sum_{n \in S} \log p(c_n) \quad (1)$$

A small portion of the TCM for the object slot of *start* in the transitive frame is displayed in figure 1. WordNet classes are displayed in boxes with a label which best reflects the sense of the class. The probability estimates are shown for the classes along the TCM. Examples of the argument head data are displayed below the WordNet classes with dotted lines indicating membership at a hyponym class beneath these classes.

We assume that verbs which participate will show a higher degree of similarity between the preferences at the target slots compared with non-participating verbs. To compare the preferences we compare the probability distributions across WordNet using a measure of distributional similarity. Since the probability distributions may be at different levels of WordNet, we map the TCMs at the target slots to a common tree cut, a "base cut". We experiment with two different types of base cut. The first is simply a base cut at the eleven root classes of WordNet. We refer to this as the "root base cut" (RBC). The second is termed the "union base cut" (UBC). This is obtained by taking all classes from the union of the two TCMs which are not subsumed by another class in this union. Duplicates are removed. Probabilities are assigned to the classes of a base cut using the estimates on the original TCM. The probability estimate for a hypernym class is obtained by combining the probability estimates for all its hyponyms on the original cut. Figure 2 exemplifies this process for two TCMs (TCM1 and TCM2) in an imaginary hierarchy. The UBC is at the classes B, C and D.

To quantify the similarity between the probability distributions for the target slots we use the α-skew divergence (αSD) proposed by Lee (1999). [1] This measure, defined in equation 2, is a smoothed version of the Kulback-Liebler divergence. $p1(x)$ and $p2(x)$ are the two probability distributions which are being compared. The α constant is a value between 0 and 1 which smooths $p1(x)$ with $p2(x)$ so that αSD is always defined. We use the same value (0.99) for α as Lee. If α is set to 1 then this measure is equivalent to the Kulback-Liebler divergence.

$$\alpha\mathrm{sd}(p1(x), p2(x)) = D(p2(x) || ((\alpha \times p1(x)) + ((1-\alpha) \times p2(x)))) \quad (2)$$

4 Experimental Evaluation

We experiment with a SCF lexicon produced from 19.3 million words of parsed text from the BNC (Leech, 1992). We used the causative and conative alternations, since these have enough candidates in our lexicon for experimentation. Evaluation is performed on verbs already filtered by the syntactic processing. The SCF acquisition system has been evaluated elsewhere (Briscoe and Carroll, 1997).

We selected candidate verbs which occurred with 10 or more nominal argument heads at the target slots. The argument heads were restricted to those which can be classified in the WordNet hypernym hierarchy. Candidates were selected by hand so as to obtain an even split between candidates which did participate in the alternation (positive candidates) and those which did not (negative candidates). Four human judges were used to determine the "gold standard". The judges were asked to specify a *yes* or *no* decision on participation for each verb. They were also permitted a *don't know* verdict. The kappa statistic (Siegel and Castellan, 1988) was calculated to ensure that there was significant agreement between judges for the initial set of candidates. From these, verbs were selected which had 75% or more agreement, i.e. three or more judges giving the same *yes* or *no* decision for the verb.

For the causative alternation we were left with 46 positives and 53 negatives. For the conative alternation we had 6 of each. In both cases, we used the Mann Whitney U test to see if there was a significant relationship between the similarity measure and participation. We then used a threshold on the similarity scores as the decision point for participation to determine a level of accuracy. We experimented with both the mean and median of the scores as a threshold. Seven of the negative causative candidates were randomly chosen and removed to ensure an even split between positive and negative candidates for determining accuracy using the mean and median as thresholds.

The following subsection describes the results of the experiments using the method described in section 3 above. Subsection 4.2 describes an experiment on the same data to determine participation using a similarity measure based on the intersection of the lemmas at the target slots.

[1] We also experimented with euclidian distance, the L1 norm, and cosine measures. The differences in performance of these measures were not statistically significant.

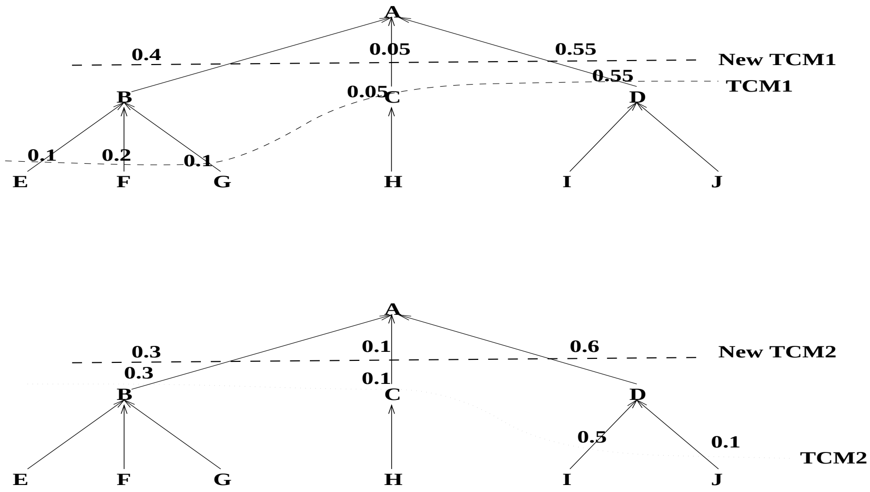

Figure 2: New TCMs at the union base cut

4.1 Using Syntax and Selectional Preferences

The results for the causative alternation are displayed in table 1 for both the RBC and the UBC. The relationship between participation and αSD is highly significant in both cases, with values of p well below 0.01. Accuracy for the mean and median thresholds are displayed in the fourth and fifth columns. Both thresholds outperform the random baseline of 50%. The results for the UBC are slightly improved, compared to those for the RBC, however the improvement is not significant.

The numbers of false negative (FN) and false positive (FP) errors for the mean and median thresholds are displayed in table 2, along with the threshold and accuracy. The outcomes for each individual verb for the experiment using the RBC and the mean threshold are as follows:

- True negatives:
  add admit answer believe borrow cost declare demand expect feel imagine know notice pay perform practise proclaim read remember sing survive understand win write
- True positives:
  accelerate bang bend boil break burn change close cook cool crack decrease drop dry end expand fly improve increase match melt open ring rip rock roll shatter shut slam smash snap spill split spread start stop stretch swing tilt turn wake
- False negatives:
  flood land march repeat terminate
- False positives:
  ask attack catch choose climb drink eat help kick knit miss outline pack paint plan prescribe pull remain steal suck warn wash

The results for the UBC experiment are very similar.

If the median is used, the number of FPs and FNs are evenly balanced. This is because the median threshold is, by definition, taken midway between the test items arranged in order of their similarity scores. There are an even number of items on either side of the decision point, and an even number of positive and negative candidates in our test sample. Thus, the errors on either side of the decision point are equal in number.

For both base cuts, there are a larger number of false positives than false negatives when the mean is used. The mean produces a higher accuracy than the median, but gives an increase in false positives. Many false positives arise where the preferences at both target slots are near neighbours in WordNet. For example, this occurred for *eat* and *drink*. There verbs have a high probability mass (around 0.7) under the **entity** class in both target slots, since both *people* and types of *food* occur under this class. In cases like these, the probability distributions at the RBC, and frequently the UBC, are not sufficiently distinctive.

The polysemy of the verbs may provide another explanation for the large quantity of false positives. The SCFs and data of different senses should not ideally be combined, at least not for coarse grained sense distinctions. We tested the false positive and true negative candidates to see if there was a relationship between the polysemy of a verb and its misclassification. The number of senses (according to WordNet) was used to indicate the polysemy of a verb. The Mann Whitney U test was performed on

	Mann Whitney z	significance (p)	mean	median
RBC	-4.03	0.0003	71	63
UBC	-4.3	0.00003	73	70

Table 1: Causative results

base cut	threshold type	threshold	accuracy %	num FPs	num FNs
UBC	mean	0.38	73	21	4
UBC	median	0.20	70	14	14
RBC	mean	0.32	71	22	5
RBC	median	0.15	63	17	17

Table 2: Error analysis for the causative experiments

the verbs found to be true negative and false positive using the RBC. A significant relationship was not found between participation and misclassification. Both groups had an average of 5 senses per verb. This is not to say that distinguishing verb senses would not improve performance, provided that there was sufficient data. However, verb polysemy does not appear to be a major source of error, from our preliminary analysis. In many cases, such as *read* which was classified both by the judges, and the system as a negative candidate, the predominant sense of the verb provides the majority of the data. Alternate senses, for example, *the book reads well*, often do not contribute enough data so as to give rise to a large proportion of errors. Finding an appropriate inventory of senses would be difficult, since we would not wish to separate related senses which occur as alternate variants of one another. The inventory would therefore require knowledge of the phenomena that we are endeavouring to acquire automatically.

To show that our method will work for other RSAs, we use the conative. Our sample size is rather small since we are limited by the number of positive candidates in the corpus having sufficient frequency for both SCFs. The sparse data problem is acute when we look at alternations with specific prepositions. A sample of 12 verbs (6 positive and 6 negative) remained after the selection process outlined above. For this small sample we obtained a significant result ($p = 0.02$) with a mean accuracy of 67% and a median accuracy of 83%. On this occasion, the median performed better than the mean. More data is required to see if this difference is significant.

4.2 Using Syntax and Lemmas

This experiment was conducted using the same data as that used in the previous subsection. In this experiment, we used a similarity score on the argument heads directly, instead of generalizing the argument heads to WordNet classes. The venn diagram in figure 3 shows a subset of the lemmas at the transitive and intransitive SCFs for the verb *break*.

The lemma based similarity measure is termed lemma overlap (LO) and is given in equation 3, where A and B represent the target slots. LO is the size of the intersection of the multisets of argument heads at the target slots, divided by the size of the smaller of the two multisets. The intersection of two multisets includes duplicate items only as many times as the item is in both sets. For example, if one slot contained the argument heads {*person, person, person, child, man, spokeswoman*}, and the other slot contained {*person, person, child, chair, collection*}, then the intersection would be {*person, person, child*}, and LO would be $\frac{3}{5}$. This measure ranges between zero (no overlap) and 1 (where one set is a proper subset of that at the other slot).

$$\textsc{lo}(\mathrm{A},\ \mathrm{B}) = \frac{|multiset\ intersection(A\ B)|}{|smallest\ set(A, B)|} \qquad (3)$$

Using the Mann Whitney U test on the LO scores, we obtained a z score of 2.00. This is significant to the 95% level, a lower level than that for the class-based experiments. The results using the mean and median of the LO scores are shown in table 3. Performance is lower than that for the class-based experiments. The outcome for the individual verbs using the mean as a threshold was:-

- True negatives:
  add admit answer borrow choose climb cost declare demand drink eat feel imagine notice outline pack paint perform plan practise prescribe proclaim read remain sing steal suck survive understand wash win write
- True positives:
  bend boil burn change close cool dry end fly improve increase match melt open ring roll shut slam smash start stop tilt wake
- False negatives:
  accelerate bang break cook crack decrease drop expand flood land march repeat rip rock shatter

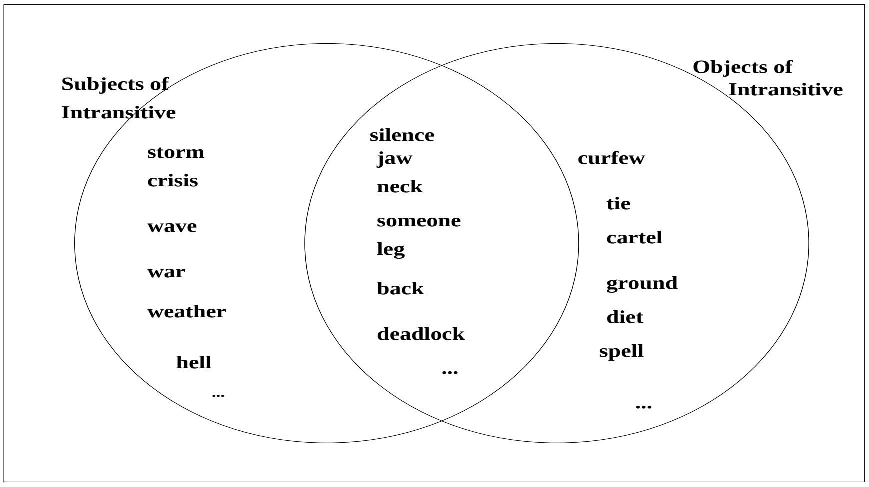

Figure 3: Lemmas at the causative target slots of *break*

snap spill split spread stretch swing terminate turn

- False positives:
ask attack believe catch expect help kick knit know miss pay pull remember warn

Interestingly, the errors for the LO measure tend to be false negatives, rather than false positives. The LO measure is much more conservative than the approach using the TCMs. In this case the median threshold produces better results.

For the conative alternation, the lemma based method does not show a significant relationship between participation and the LO scores. Moreover, there is no difference between the sums of the ranks of the two groups for the Mann Whitney U test. The mean produces an accuracy of 58% whilst the median produces an accuracy of 50%.

5 Related Work

There has been some recent interest in observing alternations in corpora (McCarthy and Korhonen, 1998; Lapata, 1999) and predicting related verb classifications (Stevenson and Merlo, 1999). Earlier work by Resnik (1993) demonstrated a link between selectional preference strength and participation in alternations where the direct object is omitted. Resnik used syntactic information from the bracketing within the Penn Treebank corpus. Research into the identification of other diathesis alternations has been advanced by the availability of automatic syntactic processing. Most work using corpus evidence for verb classification has relied on a priori knowledge in the form of linguistic cues specific to the phenomena being observed (Lapata, 1999; Stevenson and Merlo, 1999). Our approach, whilst being applicable only to RSAs, does not require human input specific to the alternation at hand.

Lapata (1999) identifies participation in the dative and benefactive alternations. Lapata's strategy is to identify participants using a shallow parser and various linguistic and semantic cues, which are specified manually for these two alternations. PP attachments are resolved using Hindle and Rooth's (1993) lexical association score. Compound nouns, which could be mistaken for the double object construction, were filtered using the log-likelihood ratio test. The semantic cues were obtained by manual analysis. The relative frequency of a SCF for a verb, compared to the total frequency of the verb, was used for filtering out erroneous SCFs.

Lapata does not report recall and precision figures against a gold standard. The emphasis is on the phenomena actually evident in the corpus data. Many of the verbs listed in Levin as taking the alternation were not observed with this alternation in the corpus data. This amounted to 44% of the verbs for the benefactive, and 52% for the dative. These figures only take into account the verbs for which at least one of the SCFs were observed. 54% of the verbs listed for the dative and benefactive by Levin were not acquired with either of the target SCFs. Conversely, many verbs not listed in Levin were identified as taking the benefactive or dative alternation using Lapata's criteria. Manual analysis of these verbs revealed 18 false positives out of 52 candidates.

Stevenson and Merlo (1999) use syntactic and lexical cues for classifying 60 verbs in three verb classes: unergative, unaccusative and verbs with an optional direct object. These three classes were chosen be-

threshold type	threshold	accuracy %	num FPs	num FNs
mean	0.26	60	14	23
median	0.23	63	17	17

Table 3: Accuracy and error analysis for lemma based experiments

cause a few well defined features, specified a priori, can distinguish the three groups. Twenty verbs from Levin's classification were used in each class. They were selected by virtue of having sufficient frequency in a combined corpus (from the Brown and the WSJ) of 65 million words. The verbs were also chosen for having one predominant intended sense in the corpus. Stevenson and Merlo used four linguistically motivated features to distinguish these groups. Counts from the corpus data for each of the four features were normalised to give a score on a scale of 1 to 100. One feature was the causative non-causative distinction. For this feature, a measure similar to our LO measure was used. The four features were identified in the corpus using automatic POS tagging and parsing of the data. The data for half of the verbs in each class was subject to manual scrutiny, after initial automatic processing. The rest of the data was produced fully automatically. The verbs were classified automatically using the four features. The accuracy of automatic classification was 52% using all four features, compared to a baseline of 33%. The best result was obtained using a combination of three features. This gave an accuracy of 66%.

McCarthy and Korhonen (1998) proposed a method for identifying RSAs using MDL. This method relied on an estimation of the cost of using TCMs to encode the argument head data at a target slot. The sum of the costs for the two target slots was compared to the cost of a TCM for encoding the union of the argument head data over the two slots. Results are reported for the causative alternation with 15 verbs. This method depends on there being similar quantities of data at the alternating slots, otherwise the data at the more frequent slot overwhelms the data at the less frequent slot. However, many alternations involve SCFs with substantially different relative frequencies, especially when one SCF is specific to a particular preposition. We carried out some experiments using the MDL method and our TCMs. For the causative, we used a sample of 110 verbs and obtained 63% accuracy. For the conative, a sample of 16 verbs was used and this time accuracy was only 56%. Notably, only one negative decision was made because of the disparate frame frequencies, which reduces the cost of combining the argument head data.

6 Conclusion

We have discovered a significant relationship between the similarity of selectional preferences at the target slots, and participation in the causative and conative alternations. A threshold, such as the mean or median can be used to obtain a level of accuracy well above the baseline. A lemma based similarity score does not always indicate a significant relationship and generally produces a lower accuracy.

There are patterns of diathesis behaviour among verb groups (Levin, 1993). Accuracy may be improved by considering several alternations collectively, rather than in isolation. Complementary techniques to identify alternations, for example (Resnik, 1993), might be combined with ours.

Although we have reported results on only two RSAs, our method is applicable to other such alternations. Furthermore, such application requires no human endeavour, apart from that required for evaluation. However, a considerably larger corpus would be required to overcome the sparse data problem for other RSA alternations.

7 Acknowledgements

Some funding for this work was provided by UK EPSRC project GR/L53175 'PSET: Practical Simplification of English Text'. We also acknowledge Gerald Gazdar for his helpful comments on this paper.

References

Bran Boguraev and Ted Briscoe. 1987. Large lexicons for natural language processing: Utilising the grammar coding system of LDOCE. *Computational Linguistics*, 13(3-4):203–218.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Fifth Applied Natural Language Processing Conference*, pages 356–363.

Ted Briscoe and Ann Copestake. 1996. Controlling the application of lexical rules. In E Viegas, editor, *SIGLEX Workshop on Lexical Semantics - ACL 96 Workshop.*

Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosensweig. 1998. Investigating regular sense extensions based on intersective Levin classes. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 293–299.

Bonnie J. Dorr and Doug Jones. 1996. Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, pages 322–327.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Kentaro Inui, Virach Sornlertlamvanich, Hozumi Tanaka, and Takenobu Tokunaga. 1997. A new formalization of probabilistic glr parsing. In *5th ACL/SIGPARSE International Workshop on Parsing Technologies*, pages 123–134, Cambridge, MA.

Anna Korhonen. 1997. *Acquiring Subcategorisation from Textual Corpora.* Master's thesis, University of Cambridge.

Maria Lapata. 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*, pages 397–404.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Geoffrey Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Beth Levin. 1993. *English Verb Classes and Alternations: a Preliminary Investigation.* University of Chicago Press, Chicago and London.

Hang Li and Naoki Abe. 1995. Generalizing case frames using a thesaurus and the MDL principle. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 239–248, Bulgaria.

Diana McCarthy and Anna Korhonen. 1998. Detecting verbal participation in diathesis alternations. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguists.*, volume 2, pages 1493–1495.

Philip Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania.

Francesc Ribas. 1995. *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy.* Ph.D. thesis, University of Catalonia.

Jorma. Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14:465–471.

Sidney Siegel and N. John Castellan, editors. 1988. *Non-Parametric Statistics for the Behavioural Sciences.* McGraw-Hill, New York.

Manfred Stede. 1998. A generative perspective on verb alternations. *Computational Linguistics*, 24(3):401–430.

Suzanne Stevenson and Paola Merlo. 1999. Automatic verb classification using distributions of grammatical features. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–52.