Identifying Participation of Individual Verbs or VerbNet Classes in the Causative Alternation

Esther Seyffarth
Heinrich Heine University Düsseldorf
Düsseldorf, Germany
esther.seyffarth@hhu.de

Abstract

Verbs that participate in diathesis alternations have different semantics in their different syntactic environments, which need to be distinguished in order to process these verbs and their contexts correctly. We design and implement 8 approaches to the automatic identification of the causative alternation in English (3 based on VerbNet classes, 5 based on individual verbs). For verbs in this alternation, the semantic roles that contribute to the meaning of the verb can be associated with different syntactic slots. Our most successful approaches use distributional vectors and achieve an F1 score of up to 79% on a balanced test set. We also apply our approaches to the distinction between the causative alternation and the unexpressed object alternation. Our best system for this is based on syntactic information, with an F1 score of 75% on a balanced test set.

1 Introduction

English verbs impose syntactic and semantic restrictions on their arguments, but some verbs are more flexible than others. A number of verbs in English have different syntactic frames (subcategorization frames, SCFs) that are associated with different semantics. This behavior of a subset of the verbs in a language is known as *diathesis alternations*, or verb alternations (Levin, 1993).

Verbs that participate in one or more alternations are potentially problematic in the context of natural language processing tasks. In order to be able to process an instance of an alternating verb in a text, it is necessary to distinguish between the different possible uses, so that the correct meaning can be assigned to the given instance.

The causative alternation is one of the regular verb alternations in English. Verbs in this alternation can be used intransitively, with an inchoative meaning, or transitively, leading to a causative meaning. The choice of a transitive or intransitive syntactic frame has an impact on the semantic roles that are part of the meaning of the verb. Consider sentences (1) and (2).

(1) The number of students has decreased.

(2) We have decreased the number of students.

Both sentences make a statement about the number of students having changed. In (1), the only semantic role that is explicitly encoded is the THEME (the thing that has decreased). In addition to this role, sentence (2) also specifies an AGENT that has causative control over the event.

The automatic identification of verbs that participate in the causative alternation is not trivial, because the semantic roles involved in the events described by the different uses of these verbs are encoded in the syntactic frames in different ways. In the sentences above, the THEME is located in the syntactic subject position in (1), but in the syntactic direct object position in (2).

It is insufficient to use the presence or absence of syntactic arguments, such as the direct object, as indicators for or against a verb's participation in the alternation, since verbs can occur with different sets of arguments for other reasons. Many verbs in English have optional direct objects; in the terminology of Levin (1993), they participate in the *unexpressed object alternation*. Distinguishing between different alternations is essential for tasks that rely on correct semantic analyses of a verb and its arguments. Examples for applications where this is important are question answering, information extraction, or summarization.

This paper describes, compares and evaluates a total of 8 approaches to the automatic identification of verbs in the causative alternation. Of particular interest to us is the comparison of different evaluation conditions and different test sets, which give

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 146-155.
New York City, New York, January 3-6, 2019

a wide range of accuracy scores depending on the setup.

Our results show that some of our setups are more robust than others against different evaluation conditions. The different evaluation conditions are useful because they can expose problems of individual setups, such as a tendency to assign false positive labels.

We also evaluate the performance of our systems on the distinction between verbs in the causative alternation and verbs in the unexpressed-object alternation. Although these alternations resemble each other in the SCFs they allow, our results show that some systems perform well on both classification tasks.

While English alternating verbs can be identified by looking them up in one of the resources that exist for such purposes, we find it desirable to create dynamic systems for the identification of alternating verbs, for two main reasons. First, the phenomenon may be productive to a certain degree, which means that resources can become outdated; and second, there are other languages with similar phenomena for which no (large) resources like this exist. A system that does well on English alternation identification can be helpful in building this type of resource for other languages.

2 Related Work

While diathesis alternations have been a topic of linguistic discussion for some time, the idea of identifying these alternations automatically has mostly been discussed after the publication of Levin (1993). Her lists of verb alternations and of verb classes made it possible to design systems that cluster verbs automatically and to evaluate the outputs of those systems against Levin's data.

The creation of resources like WordNet (Miller, 1998) and VerbNet (Kipper et al., 2000) made this even easier. For instance, approaches like the one by McCarthy (2000, 2001) use the WordNet hierarchy to predict whether a verb participates in the causative or conative alternation. Using a subcategorization lexicon derived from the BNC, she calculates the similarity of the role fillers for each position in the verb's syntactic frame to identify cases where different slots have a systematic overlap in their semantic preferences, which is seen as an indicator for the alternation. McCarthy (2000) hand-picks a set of 46 positive and 53 negative verbs that are being classified. Human annotators decide whether each verb participates in the alternation. The author describes several different setups; the highest accuracy on her test set for the causative alternation is 73%.

Merlo and Stevenson (2001) distinguish three types of optionally intransitive verbs using a number of features. Their distinction is between unergative, unaccusative, and object-drop verbs. Their features include semantic features like animacy or causativity, as well as syntactic features like passive voice or presence of a VBN tag. The combination of all features leads to a classification accuracy of 69.8% on a test set of 20 unergative verbs, 19 unaccusative verbs, and 20 object-drop verbs. From their experiments with human annotators, they derive an "expert-based upper bound" accuracy around 86.5% for the task.

Sun et al. (2013) present an unsupervised approach to the semantic classification of verbs that uses approximations of diathesis alternations as features. While they evaluate their system on verb class induction tasks, their method for approximating diathesis alternations is also of interest in isolation from the clustering results. The approximation takes into account the subcategorization frames that are observed for each verb in a corpus and the likelihood of individual verbs and individual SCFs. As each diathesis alternation gives rise to several SCFs, the joint probability can be used to predict whether the SCFs of a verb are observed by chance or due to the verb's participation in an alternation.

An area of research that is related to, but distinct from the task we discuss here is the clustering of verbs into Levin classes, VerbNet classes, or other semantic classes. Examples for this are presented in Lapata and Brew (1999); Schulte Im Walde (2000); Joanis (2002); Joanis and Stevenson (2003); Stevenson and Joanis (2003); Schulte Im Walde (2006); Joanis et al. (2008); Sun et al. (2008); Korhonen (2009). While verb classes are associated with diathesis alternations, there is no one-to-one relation between a verb class and an alternation. Thus, strategies that are useful for verb class prediction cannot always be used in the same way for the prediction of verb alternations. In this paper, we develop systems for the identification of diathesis alternations because we are specifically interested in properties of verbs at the syntax-semantics interface.

One of the contributions of our work that has not been discussed in depth in the literature is the

comparison of the performance of different classification methods on the lists by Levin (1993) and on VerbNet classes, and on different subsets of verbs from those lists. We show that the scores of each method sometimes vary by a large margin, which makes it difficult to compare the performance of previous approaches when the evaluation technique is not described in detail.

3 Methods

We develop three class-based systems and five verb-based systems for the classification of English verbs regarding their participation in the causative alternation. Class-based systems classify verbs that belong to the same VerbNet class together. Verb-based systems classify each verb lemma individually.

3.1 Data

Syntactic patterns and frequency counts are derived from a dependency-parsed version of the BNC corpus (Burnard, 2007)[1]. The approaches that classify individual verbs refer to the dependency annotations in the corpus and in some cases to distributional word vectors, but do not make use of any external, manually-created resources.

For the class-based approaches, we collect information about individual verbs from verb class definitions in VerbNet 3.3 (Kipper et al., 2000). The information we gather from the resource is limited to determining which verbs form a class together; nothing else is being read from VerbNet.

We evaluate our systems on two different test sets[2], each including examples of alternating verbs as well as examples of non-alternating verbs. One of the test sets is the list given by Levin (1993, pp. 27–32), where Levin lists 365 examples of verbs that participate in the alternation and 238 examples of verbs that do not participate. The other test set is extracted from VerbNet; the set of alternating verbs consists of all members of all VerbNet classes that are marked as having causative and inchoative uses. As negative examples for this test set, we use verbs that are marked in VerbNet as having transitive and intransitive uses (i.e., they are syntactically flexible), but not as having causative and inchoative uses. This leads to 469 positive examples and 454 negative examples.

[1] 112,298,424 tokens in 6,026,307 sentences. Parsed with the Stanford CoreNLP Library (Manning et al., 2014).

[2] The full test sets are available from the author upon request.

Our strategy for collecting verbs outside the causative alternation creates a negative set that is not too large and contains only verbs that exhibit behavior similar to alternating verbs. Effectively, we select verbs that participate in the unexpressed-object alternation (see also Section 6).

One difficulty in evaluating our classification systems is that verbs are not distinguished by verb sense. For instance, Levin lists *advance* as an alternating verb in its (CAUSED-)CHANGE-OF-STATE sense, but also as a non-alternating verb in its FUTURE-HAVING sense. We plan to examine these issues and their implications for the task more closely in future work. For now, in order to achieve a clear separation of alternating and non-alternating verbs, we decide to drop verbs with multiple class membership from the test sets for our experiments, leading to a set of 358 positive and 236 negative examples[3].

While the positive examples from both sources share a specific syntactic and semantic behavior, the negative examples are less homogeneous, and it is not necessarily useful to treat the negative examples as a closed set. Theoretically, all verbs that are not listed as positive examples should be usable as negative examples, provided the list of positive examples is complete and covers all verbs in the language that participate in the alternation.

However, using all unlisted verbs as negative examples would make the classification much more difficult to evaluate, due to the overwhelming majority of negative items. Instead, we use the sets described above and propose three test conditions to base our evaluations on. The test conditions are described in Section 4.

3.2 Class-Based Methods

We define three class-based approaches to the task. The information these approaches rely on is extracted from VerbNet 3.3. The measures used to calculate the likelihood of each verb to alternate are inspired by Bonial et al. (2011). There, the authors define scores that are used as indicators for how likely the members of each VerbNet class are to occur in the CAUSED-MOTION construction. In a similar way, we define scores that indicate the likelihood of participation in the causative alternation for the members of each VerbNet class.

[3] Levin also lists some verbs more than once, such as *bleed*. We group words with the same surface forms together without performing word sense disambiguation.

The motivation for the approaches in this section is that the causative alternation is closely related to the question whether a verb can be used transitively and intransitively. Although transitive and intransitive uses can also be an indicator of optional complements or of the unexpressed object alternation, we implement three setups based on the transitivity behavior of verbs to determine how indicative the use of transitive and intransitive SCFs is for the causative alternation.

3.2.1 VNType

Our first classifier, VNTYPE, assigns the alternating label to all members of the VerbNet class if all members occur in both SCFs at least once in the corpus, following the rule in Equation 1.

$$\mathsf{a}_1(\mathsf{v} \in \mathsf{C}) = 1 \text{ iff } \forall \mathsf{v}' \in \mathsf{C} : \mathsf{c}(\mathsf{v}'_{\mathsf{trans}}) > 0 \wedge \mathsf{c}(\mathsf{v}'_{\mathsf{intrans}}) > 0 \quad (1)$$

The alternatability score 1 (verb participates in the alternation) is given to the verb v iff all verbs in the same VerbNet class C occur at least once in a transitive SCF and at least once in an intransitive SCF.

In all other cases, including the case that one or more verbs in the class are not found in the corpus at all, the classifier assigns the non-alternating label to all members of the VerbNet class. This approach is type-based: Frequencies of transitive and intransitive uses are not taken into account.

3.2.2 VNRank

Our second approach, VNRANK, derives the alternatability score from the percentage of verb types in a VerbNet class that were observed at least once transitively and at least once intransitively, following Equation 2.

$$\mathsf{a}_2(\mathsf{v} \in \mathsf{C}) = \frac{|\{\mathsf{v}' \in \mathsf{C} : \mathsf{c}(\mathsf{v}'_{\mathsf{trans}}) > 0 \wedge \mathsf{c}(\mathsf{v}'_{\mathsf{intrans}}) > 0\}|}{|\mathsf{C}|} \quad (2)$$

a_2 assigns each verb v an alternatability score between 0 and 1 that reflects the percentage of verbs in the same VerbNet class C that occur at least once in a transitive SCF and at least once in an intransitive SCF.

If a class C contains one or more verbs v' that are not observed in the corpus, the likelihood for that class and its member v to be labeled as alternating becomes lower.

3.2.3 VNToken

Our third approach, VNTOKEN, is a token-based variation of VNRANK, assigning a score following Equation 3.

$$\mathsf{a}_3(\mathsf{v} \in \mathsf{C}) = \frac{\sum_{\mathsf{v}' \in \mathsf{C}} \min(\mathsf{c}(\mathsf{v}'_{\mathsf{trans}}), \mathsf{c}(\mathsf{v}'_{\mathsf{intrans}}))}{\sum_{\mathsf{v}' \in \mathsf{C}} \mathsf{c}(\mathsf{v}')} \quad (3)$$

a_3 assigns each verb v an alternatability score that is based on the frequency of transitive or intransitive uses (whichever is lower) of verbs in the same VerbNet class C.

Since we use the lower of the two numbers as indicator for how strong the verb's ability to alternate is, the highest possible score that verbs in a VerbNet class C can achieve according to a_3 is 0.5, meaning that the overall number of transitive occurrences of members of C in the corpus was exactly as high as the overall number of intransitive occurrences of verbs in C. The lowest score is 0, which is the case when all members of C occur exclusively transitively or exclusively intransitively in the corpus.

Verbs that do not occur in the corpus do not impact the score calculated by a_3, since they do not add anything to either the numerator or the denominator in the equation.

3.3 Verb-Based Methods

The following methods classify or rank verbs individually, without taking their VerbNet classes into account.

3.3.1 SCFFlag

Our first verb-based approach, SCFFLAG, is a variation of VNTYPE that classifies verbs individually, not based on their VerbNet classes. Equation 4 shows how the score is calculated.

$$\mathsf{a}_4(\mathsf{v}) = 1 \text{ iff } \mathsf{c}(\mathsf{v}_{\mathsf{trans}}) > 0 \ \wedge \mathsf{c}(\mathsf{v}_{\mathsf{intrans}}) > 0 \quad (4)$$

The alternatability score 1 (verb participates in the alternation) is given to a verb v iff v is observed at least once with a transitive SCF and at least once with an intransitive SCF.

Since this approach is purely syntactic, it is prone to misclassifying verbs that have optional direct objects as verbs that participate in the causative alternation. We include it here in order to compare other systems against it. Unattested verbs are labeled as non-alternating, as they do not fulfill the criteria defined here for alternating verbs.

3.3.2 SCFRatio

Our second verb-based approach, SCFRATIO, is a variation on SCFFLAG that takes the frequency of each syntactic frame into account, as shown in Equation 5.

$$a_5(v) = \frac{c(v_{trans})}{c(v_{intrans})} \quad (5)$$

The result of a_5 is a continuous value that reflects the relation between the number of transitive and intransitive occurrences of v. In the special case that the denominator is zero, a value of 0 is assigned.

While SCFFLAG only checks whether both transitive and intransitive frames are possible, this approach looks at their relative frequency, as observed in the corpus. If a verb is mostly used transitively and is only intransitive in one instance, it does not necessarily participate in the causative alternation; the one intransitive instance might be due to an error or a coerced usage of the verb. If it occurs transitively and intransitively with similar frequencies, it is more likely for that verb to participate in the causative alternation.

Verbs that are not found in the corpus are treated like syntactically inflexible verbs, that is, they are unlikely to alternate. Like SCFFLAG, we expect this setup to be prone to misclassifying verbs with optional objects as participating in the causative alternation.

The method is similar to that of Sun et al. (2013), but unlike that system, it does not take the overall frequency of each SCF over all verbs in the corpus into account. In SCFFLAG and SCFRATIO, the SCFs are simplified and distinguished by (in)transitivity; the presence or absence of other dependents and the order of the items in the SCF are disregarded. The main reason for this simplification is that it minimizes sparsity problems by generalizing over different SCFs that share the property of being (in)transitive.

The following approaches explore the idea that verbs in the alternation will impose the same selectional preferences on their direct objects as on their intransitive subjects, as illustrated by examples (1) and (2). Therefore, the set of possible direct objects should be more similar to the set of possible intransitive subjects if a verb alternates than it is for verbs that do not alternate. We implement two vector-based approaches to test this.

3.3.3 CentroidDistance

Our third verb-based approach, CENTROIDDISTANCE, makes use of distributional information about the arguments observed for each verb in the BNC corpus. Our source for distributional information is a set of pre-trained word vectors derived from a 100-billion word portion of the Google News dataset[4]. The vector set covers 3 million words and phrases and represents each item with a 300-dimensional vector created with word2vec (Mikolov et al., 2013). The formula that calculates the alternatability score is given in Equation 6.

$$a_6(v) = \cos(\overrightarrow{\text{objects}}, \overrightarrow{\text{intr-subjects}}) \quad (6)$$

a_6 derives a verb's alternatability score from the difference between the centroid of the verb's direct objects in the corpus, and the centroid of its intransitive subjects. The difference is represented by the cosine similarity between the centroids. The method is an unsupervised approximation of the WordNet-based approach by McCarthy (2001).

Verbs that are unattested in the corpus or occur with only one SCF are treated as unlikely to alternate. Some verb argument slot fillers, particularly proper names, are not covered by the Google News vectors. These arguments are disregarded.

3.3.4 CentroidSubjVsObj

Our fourth verb-based approach, CENTROIDSUBJVSOBJ, extends the CENTROIDDISTANCE approach by another criterion. Verbs can impose similar selectional preferences on different argument slots, e.g. when they require animate subjects as well as animate direct objects. CENTROIDDISTANCE may not be successful in classifying such verbs. This approach takes that fact into account by comparing the centroid distance calculated in a_6 above to the distance between the centroids for transitive and intransitive subjects of the given verb, as shown in Equation 7.

$$a_7(v) = a_6(v) - \cos(\overrightarrow{\text{tr-subjects}}, \overrightarrow{\text{intr-subjects}}) \quad (7)$$

The assumption is that verbs whose subjects tend to be more similar to each other than their intransitive subjects are to their objects are likely to participate in the object-drop alternation, not the causative alternation. Unattested verbs and arguments that cannot be associated with a vector are handled as in the CENTROIDDISTANCE approach.

[4] Available from https://code.google.com/archive/p/word2vec/.

3.3.5 RNN-LM

Our fifth verb-based approach, RNN-LM, determines the alternatability score of a verb based on automatically-generated acceptability scores for the verb's uses in transitive and intransitive environments. We extract ordered, lemma-based argument sequences from the corpus and use a script to turn transitive sentences into intransitive sentences by deleting the direct object, and to turn intransitive sentences into transitive sentences by adding dummy arguments in the form of personal pronouns. Thus, the sentence "Pat invites Kim" becomes the (transitive) argument sequence "invite-Pat-Kim" and is turned into the (intransitive) sequence "invite-Pat". The sentence "Sandy slept" becomes "sleep-Sandy", and is turned into the sequence "sleep-Sandy-her".

Having extended the input data with these alternated sentences, the system now compares the average acceptability of all transitive and intransitive uses of each verb, as shown in Equation 8.

$$a_8(v) = \frac{1}{|\mathsf{avg\_acc}(v_{\mathsf{trans}}) - \mathsf{avg\_acc}(v_{\mathsf{intrans}})|} \quad (8)$$

Acceptability is approximated by a modified version of the Recurrent Neural Network from Lau et al. (2015). That work is concerned with the prediction of acceptability judgments that are normalized for factors like sentence length or word frequency. This makes the measure more useful for our task than the mere probability of an input sequence (because changing the transitivity in a sentence always changes the length of the sentence, and thus, its probability). For details on the algorithm, see the original paper (Lau et al., 2015). By using argument sequences instead of sentences, we reduce noise. Unattested verbs are treated as unlikely to alternate.

We expect that the artificially-alternated sentences will receive lower acceptability scores for non-alternating verbs, while verbs that participate in the alternation are more acceptable in their generated alternate uses. Thus, the difference in acceptability scores should be lower for alternating verbs.

4 Results

We report the scores of our systems evaluated in three different conditions, to show the impact of the size and contents of the test set on the score.

The first test condition, ALL, evaluates the accuracy of each setup on the full verb sets. It uses all positive and negative examples from Levin, and all positive examples and syntactically-flexible negative examples from VerbNet.

The second test condition, FREQ, evaluates the accuracy of each setup on the 300 most frequent verbs from the full Levin and VerbNet sets. The selection is based on the number of occurrences of each verb in the BNC corpus. For the Levin set, the frequency cut-off leads to a slight majority of alternating verbs in the test set (177 alternating, 123 non-alternating). For the VerbNet set, it leads to a slight majority of non-alternating verbs (138 alternating, 162 non-alternating).

The third test condition, BALANCED, evaluates the accuracy of each setup on the 150 most frequent verbs from each class for each source, so that we force a balance of 150 positive and 150 negative examples per source, where each verb occurs with high frequency in the corpus.

Since the ALL condition has the largest majority of alternating verbs, the setups that perform better in this condition are the ones that are more likely to assign the alternating label. We find that the BALANCED condition gives the best indicator of the accuracy of a setup.

Table 1 shows the F1 scores of our setups. For comparison, we also include the results of a random baseline. Bold font indicates the highest scores achieved in each condition.

For the systems that employ a ranking to classify the verbs, we report the scores achieved by splitting the ranked lists at a pre-defined index. The results reported here correspond to the split index that leads to the same number of alternating and non-alternating verbs as in the test sets.

Initially, we had split the ranked lists at the index leading to the optimal score. This introduced a strong bias in favor of the ranking approaches. In another setup, we used the mean ranking score to separate the classes, which sometimes also achieved better results than the fixed-index split. We decided to report the score as described above because we find that it best reflects the ability of each approach to predict the data in the test sets.

5 Discussion

For those of our systems that use corpus data for the classification, the frequency of individual verbs has an impact on the likelihood of the system to

	Levin			VerbNet		
	ALL	FREQ	BALANCED	ALL	FREQ	BALANCED
RANDOM BASELINE	0.51	0.54	0.52	0.53	0.47	0.56
VNTYPE	0.20	0.31	0.32	0.10	0.18	0.17
VNRANK	0.67	0.63	0.52	0.60	0.42	0.52
VNTOKEN	0.61	0.59	0.50	**0.83**	0.68	0.71
SCFFLAG	**0.71**	**0.74**	**0.67**	0.59	0.63	0.67
SCFRATIO	**0.71**	0.72	0.65	0.68	0.57	0.60
CENTROIDDISTANCE	0.62	0.60	0.62	0.64	0.78	**0.79**
CENTROIDSUBJVSOBJ	0.63	0.63	0.57	0.64	**0.79**	**0.79**
RNN-LM	0.66	0.69	0.59	0.66	0.78	**0.79**

Table 1: F1 scores of all setups

assign the correct label. Verbs that are listed in the Levin or VerbNet sets as participating in the alternation cannot be classified correctly with these approaches if they are unattested in the corpus.

All setups except VNTYPE outperform the random baseline. While this approach had a precision of roughly 50% in each of the setups, the recall was extremely low, leading to the low F1 scores shown in the table. VNTYPE is generally prone to mislabeling alternating verbs as non-alternating. The reason for the prevalence of such errors is the lack of tolerance in assigning the alternating label. Recall that a verb is only labeled as alternating by this setup if it is a member of a VerbNet class in which *all* members are observed in transitive and intransitive SCFs in the corpus. Data sparsity may lead to individual VerbNet class members not being attested at all, or being attested only in one syntactic configuration, even when the other one is possible. Thus, the criterion that *all* members of the class need to be observed in both types of SCFs is apparently too strict for successful classification.

Relaxing this condition leads to the approach VNRANK, where a high percentage of syntactically flexible members in a VerbNet class leads to all members of that class being classified as alternating verbs, but it is not necessary for all verbs in the class to be observed with both types of SCFs. Table 1 shows that relaxing the conditions for syntactic flexibility leads to high performance gains. Compared to VNTYPE, this approach assigns a lower number of false negatives.

VNTOKEN performs slightly worse than VNRANK on the Levin test sets, but is among the best setups on the VerbNet test sets. A closer look at its false positives and false negatives shows that they consist mainly of verbs that are unattested or infrequent in the BNC.

The approaches SCFFLAG and SCFRATIO perform surprisingly well on the task, particularly on the Levin test sets. These approaches derive a verb's alternatability score from a comparison of the numbers of transitive and intransitive occurrences of the verb in the corpus. The fact that SCFFLAG is one of the best-scoring approaches on the Levin test sets, while its accuracy is less impressive for the VerbNet test sets, indicates one of the main differences in the two test sets. The approach simply checks whether the verb was used at least once in each SCF type. That it performs well on the Levin set mainly tells us that the syntactic flexibility of the negative examples used in that test set is lower than that of the positive examples.

SCFFLAG and SCFRATIO are also discussed below, in Section 6.

On the VerbNet test set, the systems that rely on word vectors achieved good scores, together with the RNN-LM setup. In some other evaluation setups, the RNN method slightly outperformed the vector-based ones.

The RNN-LM approach generally preferred transitive verb uses over intransitive ones. This was even the case when lexically-intransitive verbs were involved. For instance, "John-sleep" received a lower acceptability score than "John-sleep-it", even though the verb *to sleep* rarely takes a direct object in English. This shows that the way the RNN-LM scores were calculated is not necessarily a good approximation of acceptability.

Thus, even though the RNN setup was among the best-scoring ones on the balanced VerbNet test set, we find the vector-based approaches preferable. For the Levin test sets, the SCFFLAG setup achieves the highest scores throughout the test conditions.

A common source of errors for many of our approaches were mistakes in the parse trees in the corpus. They influenced our results in two ways. First, some tokens were erroneously annotated as arguments of a verb. These cases added noise to the modeling of the verb's behavior and argument preferences. They may also be responsible for the preference of the RNN-LM approach for transitive sentences; possibly, the preference is simply a result of the dependency parser having the same preference. Second, some verbs in our test sets were systematically not annotated as being verbs, including verbs that have noun homographs, such as *circle* or *drip*, and verbs that have adjective homographs, such as *yellow* or *awake*. Since our approaches rely on the dependency trees observed in the corpus in order to model the behavior of each verb, the absence of correctly-parsed sentences for these verbs had a negative impact on our overall performance. These issues may be resolved by adding a word-sense disambiguation step to the process. We will explore this in future work.

Our vector-based systems perform much better on VerbNet than on the Levin test set. This is probably because the VerbNet classes are organized semantically, which means that even when one verb is not attested frequently enough to draw conclusions about it based on its arguments, the availability of vectors for other verbs in the same VerbNet class improves the score by a large margin.

We expected our class-based approaches that rely on VerbNet class groupings (VNTYPE, VNRANK, VNTOKEN) to generally perform better when evaluated against VerbNet than when evaluated against the Levin test set. This is the case for VNTOKEN, but the other approaches in the set achieve better scores on the Levin set. Particularly surprising is the better performance on the Levin set of the VNTYPE method. However, since this approach achieved scores well below the random baseline, we do not explore it further.

Our strategy for assembling the negative test set for the VerbNet setup by collecting verbs from classes that are annotated as having both transitive and intransitive uses, but not as having both causative and inchoative uses, has an influence on the results reported in Table 1. In particular, the approaches that use the presence or absence of transitive and intransitive SCFs (SCFFLAG, SCFRATIO, VNTYPE, VNRANK, VNTOKEN) have a high risk of false positives on this test set. This is not ideal, since the ability to distinguish between the causative alternation and the unexpressed-object alternation is one of the motivations for the alternation identification task. If the identification of one of these alternations relies on features that both classes of verbs share, that distinction may not be feasible.

The following section shows how our approaches perform when they do not assign alternating and non-alternating labels, but instead separate verbs in the causative alternation from verbs in the unexpressed-object alternation.

6 Distinguishing Causative and Unexpressed-Object Alternation

The way we collected the VerbNet test set already filters the verbs in such a way that optionally-transitive verbs that do not participate in the causative alternation make up the negative set. This criterion has the result that we effectively separate causatively-alternating verbs from verbs in the unexpressed-object alternation.

Table 2 shows the performance of our systems on the task of distinguishing the causative alternation from the unexpressed-object alternation on verbs from Levin. The verbs that participate in the unexpressed-object alternation were taken from Levin (1993, pp. 33–40). The full set contained 343 verbs, which we reduced for the FREQ and BALANCED conditions as in the previous task.

The performance of our systems on this task differs only slightly from the scores reported above. Against our expectations, the vector-based approaches CENTROIDDISTANCE and CENTROIDSUBJVSOBJ perform slightly worse when applied to the distinction between the two alternations. These approaches are designed to take each verb's semantic preferences for its argument slots into account. Since CENTROIDSUBJVSOBJ performs better than CENTROIDDISTANCE, the poor performance of the latter is probably due to overlapping semantic preferences for the different slots.

To assess the merit of our approaches, we compare them to the work of Merlo and Stevenson (2001). They achieve an accuracy of 69% on the three-way distinction of unaccusative, unergative, and unexpressed-object verbs (our distinction combines the first two of these). Our verb-based approach SCFRATIO outperforms that of Merlo and Stevenson by 6%. Our setup is simpler than that of Merlo and Stevenson, requiring only a dependency-

	all	freq	balanced
RANDOM BASELINE	0.48	0.50	0.49
VNTYPE	0.19	0.30	0.30
VNRANK	**0.79**	0.67	0.68
VNTOKEN	0.61	0.51	0.51
SCFFLAG	0.64	0.66	0.67
SCFRATIO	0.68	**0.73**	**0.75**
CENTROIDDISTANCE	0.53	0.55	0.55
CENTROIDSUBJVSOBJ	0.59	0.61	0.61
RNN-LM	0.58	0.63	0.63

Table 2: F1 scores of all setups when applied to the causative-versus-unexpressed-object task, on verbs from Levin (1993)

parsed corpus from which to derive SCF statistics, whereas their system makes use of features like animacy or causativity, which pose problems when a verb is attested infrequently.

Our best-performing setup SCFRATIO assigned the causative-alternation label for verbs that occur in transitive or intransitive SCFs with very dissimilar frequencies (that is, where one of the SCFs was dominant in the corpus), while verbs whose frequencies with transitive and intransitive SCFs were closer to each other were more likely to participate in the unexpressed-object alternation.

7 Conclusion

Our results are difficult to compare to those reported in previous work. Early work, such as McCarthy (2000); Schulte Im Walde (2000); McCarthy (2001); Merlo and Stevenson (2001), tends to evaluate on small test sets. Some of our approaches exceed the performance of the systems for the identification of the causative alternation presented in these publications. Note that our test conditions achieve varying scores even with minimal changes in the test sets; previous work in this area often did not specify the exact conditions of the evaluation.

More recent work in this area has focused less on the identification of diathesis alternations, and more on the induction of Levin-like verb classes, with a stronger focus on semantics. In contrast to this, our work focuses on the classification of verbs based on properties at the syntax-semantics interface. The good results of the SCFRATIO approach on the distinction between verbs in the causative alternation and verbs in the unexpressed-object alternation shows that this system learns something about the verbs that are being classified.

The experiments in this paper were performed on English verbs because gold data for the task was readily available for the two alternations in English. However, since many of our approaches do not rely on manually-compiled resources, they can be applied to other languages with little effort, as long as dependency parsers and corpora to use as a source for distributional information are available. However, the quality of available dependency parsers for the language will always influence the performance of our methods.

The methods presented here may be applied to any role-switching alternation (McCarthy, 2001) in any language for which the necessary preprocessing tools are available. Comparing the performance of our approaches to the performance they achieve on other languages will be informative from a typological perspective. We plan to conduct similar experiments on at least one language from another language family in the future.

Due to the interest in transferring our findings to similar phenomena in other languages, we find that verb-based methods are preferable for this task. As Table 1 shows, they perform reasonably well in comparison with our other systems, and the cases where class-based methods perform better (VNTOKEN on the VerbNet test sets) are likely the result of overfitting on the structure of VerbNet.

Acknowledgments

The work presented in this paper was financed by the Deutsche Forschungsgemeinschaft (DFG) within the CRC 991 “The Structure of Representations in Language, Cognition, and Science”. The author wishes to thank Laura Kallmeyer, Kilian Evang, Jakub Waszczuk, and three anonymous reviewers for their valuable feedback and helpful comments.

References

Claire Bonial, Susan Windisch Brown, Jena D Hwang, Christopher Parisien, Martha Palmer, and Suzanne Stevenson. 2011. Incorporating Coercive Constructions into a Verb Lexicon. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 72–80. Association for Computational Linguistics.

Lou Burnard. 2007. Reference Guide for the British National Corpus (XML Edition) .

Eric Joanis. 2002. Automatic Verb Classification Using a General Feature Space. *Master's thesis, Department of Computer Science, University of Toronto*.

Eric Joanis and Suzanne Stevenson. 2003. A General Feature Space for Automatic Verb Classification. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 691–696.

Anna Korhonen. 2009. Automatic Lexical Classification - Balancing between Machine Learning and Linguistics. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*.

Maria Lapata and Chris Brew. 1999. Using Subcategorization to Resolve Verb Class Ambiguity. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 266–274.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised Prediction of Acceptability Judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628. Association for Computational Linguistics.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago press.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. Association for Computational Linguistics.

Diana McCarthy. 2000. Using Semantic Preferences to Identify Verbal Participation in Role Switching Alternations. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 256–263. Association for Computational Linguistics.

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

Paola Merlo and Suzanne Stevenson. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–408.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119. Curran Associates, Inc.

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Sabine Schulte Im Walde. 2000. Clustering Verbs Semantically According to their Alternation Behaviour. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, pages 747–753. Association for Computational Linguistics.

Sabine Schulte Im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

Suzanne Stevenson and Eric Joanis. 2003. Semi-supervised Verb Class Discovery Using Noisy Features. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 71–78. Association for Computational Linguistics.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008. Verb Class Discovery from Rich Syntactic Data. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 16–27. Springer.

Lin Sun, Diana McCarthy, and Anna Korhonen. 2013. Diathesis alternation approximation for verb clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 736–741. Association for Computational Linguistics.