Corpus-based Identification of Verbs Participating in Verb Alternations Using Classification and Manual Annotation

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

Laura Kallmeyer
Heinrich Heine University Düsseldorf
Düsseldorf, Germany
kallmeyer@phil.hhu.de

Abstract

English verb alternations allow participating verbs to appear in a set of syntactically different constructions whose associated semantic frames are systematically related. We use ENCOW and VerbNet data to train classifiers to predict the instrument subject alternation and the causative-inchoative alternation, relying on count-based and vector-based features as well as perplexity-based language model features, which are intended to reflect each alternation's felicity by simulating it. Beyond the prediction task, we use the classifier results as a source for a manual annotation step in order to identify new, unseen instances of each alternation. This is possible because existing alternation datasets contain positive, but no negative instances and are not comprehensive. Over several sequences of classification-annotation steps, we iteratively extend our sets of alternating verbs. Our hybrid approach to the identification of new alternating verbs reduces the required annotation effort by only presenting annotators with the highest-scoring candidates from the previous classification. Due to the success of semi-supervised and unsupervised features, our approach can easily be transferred to further alternations.

1 Introduction

Verb alternations are a phenomenon at the syntax-semantics interface in which several syntactic patterns are available for certain verbs, assigning different semantic roles to the elements in individual slots of the constructions. English verbs that participate in the instrument subject alternation (ISA, Levin (1993)) can take an optional instrument argument, encoded as a *with*-PP or *using*-phrase, which can also be realized in subject position (see (1)). Replacing *with* with *using* is also a standard test for instrumenthood, i.e., a means to exclude other cases of *with*-PPs. Not all verbs with an (optional) instrument participate in ISA (see (2)). This is usually explained by the specific nature of the instrument in the context of the verb (Van Hooste, 2018). The English causative-inchoative alternation (CIA) involves two main constructions: Verbs can appear intransitively, as in (3-a), or transitively, as in (3-b). The syntactically intransitive construction is associated with an inchoative sense, while the transitive construction has a causative sense.

(1) a. She broke the glass **with/using a hammer**. b. **The hammer** broke the glass.

(2) a. She ate the soup **with/using a spoon**. b. ***The spoon** ate the soup.

(3) a. **The door** opens. b. I open **the door**.

Verb alternations pose a challenge to tasks like semantic role labeling or frame induction, and knowledge about which verbs participate in which alternations is highly useful for such tasks: Certain syntactic patterns need to be interpreted differently, depending on whether or not the verb participates in the alternation. For instance, the syntactic subject in a sentence with the syntactic structure of (3-a) would

Proceedings of the 28th International Conference on Computational Linguistics, pages 4044–4055
Barcelona, Spain (Online), December 8-13, 2020

Corpus	Sentence structure	Sentences	Example
WITH-SC	nsubj V (dobj) <with ...>	761,326 (150,360)	*She begins the account of natural law with Aristotle.*
USING-SC	nsubj V (dobj) <using ...>	21,332 (2,015)	*We pay the bills using our computers.*
SUBJ-SC	nsubj V (dobj)	11,173,394 (2,981,835)	*The rest of my experience at college followed that pattern.*
INTRANS-SC	nsubj V	438,002 (323,860)	*A secret exit appears.*
TRANS-SC	nsubj V dobj	1,262,165 (787,043)	*I opened the kitchen door.*

Table 1: Our five subcorpora of ENCOW. Sentences with the required elements in a different order are also used. Bracketed numbers refer to the reduced subcorpora (max. sent. length 10). *with*-phrases must be connected to the root verb of the sentence with a `prep` dependency relation, *using*-phrases with an `xcomp` relation. Additional filters are applied to INTRANS-SC and TRANS-SC to discard edge cases. The ISA subcorpora contained 4,813 verb types and the CIA subcorpora contained 4,162 verb types.

predominantly align with an agent-like semantic role, but by recognizing that the sentence instantiates a construction belonging to the causative-inchoative alternation, we can determine that the correct semantic role here is THEME. For tasks like frame induction, where semantic frames are learned systematically from unlabeled corpus data, alternations mean that the arguments that appear in specific syntactic slots of the verb are not necessarily instances of the same semantic role.

Levin (1993) and VerbNet (Kipper et al., 2000) list alternating verbs, but are not comprehensive. The identification of verb alternations is difficult due to the many-to-many relationship between syntactic forms and semantic frames. For instance, the appearance of a verb in transitive and intransitive syntactic environments may signal participation in CIA, but it could also be explained by a number of other reasons, e.g. the object of the verb being a non-obligatory argument (Levin, 1993).

Since verb alternations are associated with different constructions, the automatic identification of verb alternations is usually studied for particular alternations. CIA is an extensively researched alternation and has been identified using WordNet (McCarthy, 2000; McCarthy, 2001; Tsang and Stevenson, 2004), distributional data (Baroni and Lenci, 2009), feature engineering (Merlo and Stevenson, 2001; Joanis and Stevenson, 2003; Stevenson and Joanis, 2003; Seyffarth, 2019), and neural networks (Kann et al., 2019). To our knowledge, there has not been any comparable work on ISA identification so far.

The contribution of this paper is twofold: First, we develop classifiers for identifying ISA and CIA verbs, and second, we find new ISA/CIA verbs by using these classifiers in a subsequent iteration of manual annotation and reclassification. Our classifiers perform slightly better on the CIA identification task than on the ISA task; the annotation results yield promising new candidates for both alternations.

Our classifiers use count-based, perplexity-based and vector-based features obtained from the sentence types in (1) and (3). The perplexity features are based on a language model and intended to approximate acceptability judgements of sentence variations (e.g., in the case of ISA replacing *with* with *using* or moving a *with*-NP to the subject position; in the case of CIA discarding a transitive sentence's subject and promoting the object to the subject position). This has become possible with the new, transformer-based generation of language models (Devlin et al., 2019; Dai et al., 2019).

2 Method

2.1 Data

For our ISA/CIA identification approach, we collect sentences from ENCOW (Schäfer and Bildhauer, 2012; Schäfer, 2015)[1] that instantiate the syntactic patterns exemplified in (1) and (3). Table 1 describes

[1] English web text from 2012/14, available in a sentence-shuffled version, containing roughly 9.6 billion tokens, dependency-parsed with MaltParser (Nivre et al., 2006).

the collected subcorpora.[2] For SUBJ-SC, we only use sentences whose root verbs are observed at least once in WITH-SC or USING-SC. We also create a reduced version of each corpus containing only sentences with up to 10 tokens. The aim of these reduced corpora is to determine to what extent misparsed sentences or overly long constituents can impact the success of the annotation step. We run that part of our experiments once on reduced corpora and once on the full corpora for each alternation.

We use data from VerbNet 3.3 (VN, Kipper et al. (2000)) as ground truth to train our classifiers. 14 VerbNet (sub)classes are labeled as participating in ISA: *hit-18.1*, *hit-18.1-1*, *poke-19*, *destroy-44*, *carve-21.2-1*, *carve-21.2-2*, *cooking-45.3*, *bend-45.2*, *cut-21.1*, *cut-21.1-1*, *break-45.1*, *murder-42.1-1*, *other_cos-45.4*, *other_cos-45.4-1*. Out of the 558 verbs[3] contained in these classes, we only use the 408 verbs that occur at least once in WITH-SC or USING-SC, and at least once in SUBJ-SC. VN does not mark CIA verbs with a dedicated alternation tag, so we select as alternating set all 9 (sub)classes that contain at least one frame labeled as "Causative" and one labeled as "Inchoative": *cooking-45.3*, *bend-45.2*, *suffocate-40.7*, *break-45.1*, *knead-26.5*, *other_cos-45.4*, *other_cos-45.4-1*, *turn-26.6.1*, *turn-26.6.1-1*. Out of the 465 verbs[4] contained in these classes, we only use the 331 verbs that occur at least once in INTRANS-SC and at least once in TRANS-SC.

We discard unattested positive instances for each alternation to avoid teaching the classifier that verbs can participate in the alternation without being observed in all relevant syntactic patterns. While that is the case for the verbs we discard from VerbNet when creating our initial training set, it should not be generalized to other verbs that are only attested in one of the syntactic patterns. Thus, appearing in all relevant patterns at least once is a necessary condition for us to view a verb as an alternation candidate.

Our negative sets are sourced from the corpora such that each ISA/CIA verb is complemented by a verb with roughly the same frequency (across relevant subcorpora for the given alternation). Our annotation and retraining steps will test whether they are in fact negative instances; to train and evaluate our initial classifiers, we treat them as negatives. Since all our negative instances are (superficially) distributed like the positive instances, our alternation prediction task is especially challenging (see also Sec. 3.1).

2.2 Classification

Features We treat participation in the alternations as a lexical property, and use our alternation-specific subcorpora to derive values for a set of linguistically-informed features. Our features are designed to quantify to which extent the observed syntactic patterns for each verb are indeed indicative of the relevant alternation. Table 2 describes our features for ISA classification, Table 3 our features for CIA classification.

Count-based (cnt) features are based on the frequency of specific syntactic patterns in our subcorpora. **Vector-based (vec) features** are intended to approximate the selectional preferences for the argument slots involved in the alternation.[5] In ISA, we expect the possible elements in a verb's *with/using* phrase (cf. (1-a)) to be similar to the possible elements in that verb's subject position in sentences without a *with/using* phrase (cf. (1-b)). In CIA, the expected similarity is between the subjects of intransitive sentences as in (3-a), and the objects of transitive sentences as in (3-b). **Perplexity-based (pplx) features** are based on the idea that only verbs that participate in a given alternation may appear in all constructions associated with that alternation. Wrt. ISA, for each sentence in WITH-SC and USING-SC, we generate an alternate sentence by moving the potential instrument to the subject position. Concerning CIA, for each sentence in TRANS-SC, we generate an alternate sentence in which the object filler is moved to the subject position; furthermore, for sentences in INTRANS-SC, we also generate alternate sentences, moving subjects to the object position and inserting pronouns as new subjects. Where possible, we select pronouns that were observed at least once as subjects of the current verb. Basic pre-defined agreement rules ensure grammaticality for alternated sentences. For alternating verbs, transforming sentences that

[2] The sentence IDs for our corpora, as well as our verb sets, are available online at `https://github.com/eseyffarth/alternation-identification-coling2020`.

[3] The classes have a total of 588 members, but this includes duplicates. We use the 558 surface forms contained in the set.

[4] The classes have a total of 479 members including duplicates, 465 surface forms.

[5] We use pre-trained, 300d word2vec vectors (Mikolov et al., 2013) from the Google News dataset, available from `code.google.com/archive/p/word2vec/`.

instantiate one syntactic pattern to alternate sentences instantiating the other syntactic pattern should result in acceptable sentences, as in (1). For non-alternating verbs, the resulting sentences should be less acceptable, see (2). Note that constructions are form-meaning pairs, but our subcorpus filters and sentence-alternating script operate purely on syntactic patterns. We use a transformer-based language model (Dai et al., 2019), trained on the One Billion Word benchmark (Chelba et al., 2013), to calculate perplexity scores for the original sentences and their reordered counterparts, as an approximation of acceptability. We expect the average perplexity scores of generated alternate sentences with alternating verbs to be more similar to the original sentences' perplexity, and less so for non-alternating verbs.

While studies like Lau et al. (2015) and Warstadt et al. (2019) have shown that acceptability judgments based on language models do not perform on par with human annotators, and that human acceptability judgments also do not produce perfect inter-annotator agreement, we nevertheless use perplexity scores as an approximation of acceptability; in doing so, we are less interested in the acceptability of individual sentences, and more in overall trends across all attestations of a given verb. Our results show that the perplexity scores yield useful features for the task.

Cnt features (L^v_w, L^v_u = (head) lemmas in *with*-PPs in S^v_w resp. in *using*-phrases in S^v_u, L^v_{sub} = head lemmas occurring as subjects with v in SUBJ-SC, $T^v_{...}$ same for tokens)

1. $\frac{|(L^v_{sub} \cap L^v_w)|}{|L^v_w|}$ 2. $\frac{|(L^v_{sub} \cap L^v_u)|}{|L^v_u|}$ 3. $\frac{|(L^v_{sub} \cap L^v_w)|}{|L^v_{sub}|}$ 4. $\frac{|(L^v_{sub} \cap L^v_u)|}{|L^v_{sub}|}$

5. $\frac{|(T^v_{sub} \cap T^v_w)|}{|T^v_w|}$ 6. $\frac{|(T^v_{sub} \cap T^v_{ug})|}{|T^v_u|}$ 7. $\frac{|(T^v_{sub} \cap T^v_w)|}{|T^v_{sub}|}$ 8. $\frac{|(T^v_{sub} \cap T^v_u)|}{|T^v_{sub}|}$

Vec features (for sets $X; Y$: $dist(X, Y) = 1 - cos(\vec{c}(X), \vec{c}(Y))$, where $\vec{c}(X) =$ centroid of all vectors of $x \in X$; L^v_{sub-w} and L^v_{sub-u} the sets of all subj. lemmas in S^v_w and S^v_u resp.)

1. $dist(L^v_w, L^v_{sub})$ 2. $dist(L^v_u, L^v_{sub})$ 3. $dist(L_{sub-w}, L^v_{sub})$ 4. $dist(L_{sub-u}, L^v_{sub})$

Pplx features ($pplx(s)$ = perplexity of s, $s_{w \to u}$ = *using*-alternate of $s \in S_w$, $s_{w \to s}$ = subject -alternate of $s \in S_w$, $s_{u \to w}$ = *with*-alternate of $s \in S_u$, $s_{u \to s}$ = subject-alternate of $s \in S_u$)

1. average of $\frac{pplx(s)}{pplx(s_{w \to u})}$ for all $s \in S^v_w$, 2. average of $\frac{pplx(s)}{pplx(s_{w \to s})}$ for all $s \in S^v_w$,

3. average of $\frac{pplx(s)}{pplx(s_{u \to w})}$ for all $s \in S^v_u$, 4. average of $\frac{pplx(s)}{pplx(s_{u \to s})}$ for all $s \in S^v_u$

Table 2: Description of features for each verb v used for ISA classification, S^v_w, S^v_u containing the sentences with v from WITH-SC and USING-SC resp.

We use a Support Vector Machine (SVM), implemented with `scikit-learn` (Pedregosa et al., 2011), to classify the small number of instances involved here.[6] The SVM hyperparameters are determined in advance with 10-fold cross-validation, based on feature values derived from the full versions of our subcorpora. Because the alternations behave differently in the corpora, we use different sets of hyperparameters: C=3000 and gamma=0.001 for ISA, C=100 and gamma=1 for CIA.

We compare the classifiers to a simple baseline that labels a verb as participating in the alternation if its less frequent syntactic pattern appears at least 10% as often as its more frequent pattern. For ISA, we group the *with* and *using* patterns together, and compare them to the subject pattern. The 10% threshold is meant to exclude one-offs, misparses, and idiosyncratic usages: Only verbs that are observed with sufficiently comparable frequencies count as participating in the alternation.

2.3 Annotation

The second contribution of this paper is finding new verbs participating in the two alternations, in addition to the ones labeled as such in VerbNet. Since we source the negative sets for our classifiers from the

[6] A pilot study comparing the performance of SVM, k-Nearest Neighbor, and Naive Bayes classifiers resulted in the selection of SVM as the most reliable algorithm.

Cnt features (L^v_{dobj} = head lemmas in object position in S^v_t, $L^v_{subj\_i}$ (and $L^v_{subj\_t}$ resp.) = head lemmas in subject position in S^v_i (resp. S^v_t), $L^v_{nsubj} = L^v_{subj\_i} \cup L^v_{subj\_t}$; $T^v_{...}$ same for tokens) 1. $\frac{\lvert(L^v_{dobj} \cap L^v_{nsubj})\rvert}{\lvert L^v_{dobj}\rvert}$ 2. $\frac{\lvert(L^v_{dobj} \cap L^v_{nsubj})\rvert}{\lvert L^v_{nsubj}\rvert}$ 3. $\frac{\lvert(T^v_{dobj} \cap T^v_{nsubj})\rvert}{\lvert T^v_{dobj}\rvert}$ 4. $\frac{\lvert(T^v_{dobj} \cap T^v_{nsubj})\rvert}{\lvert T^v_{nsubj}\rvert}$
Vec features 1. $dist(L^v_{subj\_t}, L^v_{subj\_i})$ 2. $dist(L^v_{obj}, L^v_{subj\_i})$ 3. $dist(L^v_{obj}, L^v_{nsubj})$
Pplx features ($s_{i \to t}$ = transitive alternate of $s \in S_i$, $s_{t \to i}$ = intransitive alternate of $s \in S_t$) 1. average of $\frac{pplx(s)}{pplx(s_{t \to i})}$ for all $s \in S^v_t$, 2. average of $\frac{pplx(s)}{pplx(s_{i \to t})}$ for all $s \in S^v_i$, 3. average of $pplx(s)$ for all $s \in S^v_t$, 4. average of $pplx(s)$ for all $s \in S^v_i$, 5. (average of $pplx(s)$ for all $s \in S^v_t$) $-$ (average of $pplx(s)$ for all $s \in S^v_i$)

Table 3: Description of features for each verb v used for CIA classification, S^v_i, S^v_t containing the sentences with v from INTRANS-SC and TRANS-SC resp.

corpus, there is a possibility that new alternating verbs appear in the negative sets. We hypothesize that false positives (FPs), ie., verbs that were members of the negative set for the current alternation, but have been labeled by the classifier as belonging to the positive set, are a useful source of new alternating verbs.

To test that hypothesis, each classification round is followed by a manual annotation step. In this phase, we ask annotators to assign one of four labels per construction to the 40 highest-scoring FPs. Each verb's distance from the separating hyperplane learned by the classifier during training is treated as that verb's score. Verbs identified as alternating by this annotation are then added to the gold data and the classifier is retrained, which in turn leads to a new round of annotation.

The threshold of 40 verbs selected for annotation per iteration produces a reasonable amount of manual annotation work. We only annotate the top 40 candidates because we expect these verbs to be most similar to known alternating verbs, without (initially) being classified as positive instances. Each verb is presented to each annotator at most once, even if it appears in the top 40 list again in later iterations.[7]

For ISA, the list of sentences from WITH-SC (as a whole) receives one of the labels 1a–4a from Table 4, the one from USING-SC is labeled with 1b–4b, and the one from SUBJ-SC with 1c–4c. For instance, in sentence (1-a), annotators select one of the four labels to decide whether or not the *hammer* in the *with/using* phrase is an appropriate instrument to bring about the event described by the verb *break*.

For CIA, the labels concern the degree of causative impact of the subject on the object it appears with for the list of sentences from TRANS-SC (labels 1a–4a in Table 4), and the degree of inchoativity that the subject undergoes for sentences from INTRANS-SC (labels 1b–4b). For instance, in sentence (3-a), annotators select one of the four labels to decide to what extent the *door* in the subject phrase of the sentence is inchoatively impacted by the verb *open*; that is, whether or not the event is accompanied by a change of state that applies to the argument.

After the annotation step is finished, the labels are processed to identify verbs that are good candidates for the alternation. Verbs that receive labels 1 or 2 for all relevant constructions are treated as alternation candidates, or *actual positives*, and are moved from the negative set to the positive set. Verbs that receive the label 4 in all relevant constructions are treated as *known negatives*. They incorrectly received a positive label during the classification step because they exhibit properties associated with the current alternation, but annotators strongly interpret them as not participating in the alternation. We regard them as atypical examples of the negative class, because they were in the false positive set in the previous classification step. Due to their atypicality, they are discarded from the negative set. Whenever the negative set is reduced, we insert new candidates from the corpus to ensure the classes stay balanced.

[7] In a previous design stage, we always annotated 40 previously-unannotated verbs per iteration, instead of selecting only the unannotated verbs from the top 40. This resulted in extremely long annotation rounds that yielded roughly the same number of new alternation candidates. We therefore decided to restrict annotation to the top 40 verbs in our final setup.

ISA		
1. Sentences have a strong	instrument candidate in the	a. *with* phrase. (WITH-SC)
2. Sentences have a good		b. *using* phrase. (USING-SC)
3. Sentences sometimes have an		c. subject slot. (SUBJ-SC)
4. Sentences don't have a good		

CIA	
1. The majority of subjects are/have	
2. Some subjects are/have	a. inchoatively impacted by the verb. (INTRANS-SC)
3. Subjects are rarely/rarely have	b. a causative impact on their objects. (TRANS-SC)
4. Subjects are not/do not have	

Table 4: Labels 1a–4a, 1b–4b, and 1c–4c for ISA and 1a–4a and 1b–4b for CIA available to annotators during the annotation step (any row from the first column is combined (with, if applicable, the middle column and) with any row in the right column, yielding one of the labels). Structurally misparsed or unattested constructions for a verb receive the label 4.

Balancing the classes in each round serves two main purposes. First, it prevents the majority class preference that is observed in standard, unweighted SVM setups with unbalanced classes (preliminary experiments that attempted to classify the full set of observed verb types in the corpora led to weak results). Second, our goal is to use the outcome of each annotation round to find new alternation candidates, so we are interested in continuously extending the set of instances to be classified.

Finally, a new classifier is trained on the new verb sets and applied to the training set to identify promising candidates for the next annotation round. In this phase, the training set and test set for the classifiers are identical because the goal is not to generalize to new data, but to identify new positives among the instances labeled as negative. Throughout the classification-annotation rounds, the positive set grows as more alternation candidates are identified. The iterations stop when a round yields no new *actual positives*. Note that due to the way we set up our iterations, it is actually desirable for a classifier to not achieve perfect accuracy: The next annotation step only takes place if new candidates are found.

Since each annotator's results from an annotation round directly influence the contents of the training set for the next round, and thus also the pool of verbs that will be annotated in the next round, it is possible for verbs to be seen by only one annotator. We mitigate this by introducing a final annotation round after all regular annotations are finished. In the final annotation round, each annotator is presented with all verbs that were only seen by other annotators. This ensures that each verb eventually receives the same number of annotation labels.

3 Results and Discussion

3.1 Classification results

We evaluate each classifier before and after annotation, having included the new alternation candidates in the classifiers' training sets. We always evaluate on balanced sets, which we achieve by selecting one similarly-frequent verb from the corpora for each verb in the known positive class for each alternation. Since it is always possible for instances assumed to be negative to actually participate in the alternation, we view the classifier scores as a lower bound: Some instances that we treat as false positives may in fact be actual positives that we have not identified yet, and assigning the correct label to them would raise the measured classifier accuracy. This also means that a high recall is more important than high precision.

Table 5 presents the scores for the full corpora before and after annotation, determined with 10-fold cross-validation. The vec features were particularly useful for identifying ISA, while the pplx features gave the best results for CIA identification. The likely explanation for the relatively poor performance of pplx features for ISA is that changing one of the observed sentences to the alternate syntactic pattern may often result in acceptable sentences, even when the sentences do not involve any instruments, which is possible due to the polysemy of *with* (more on this in Sec. 3.2.1).

The initial CIA classifier more accurately identifies known alternating verbs than the initial ISA clas-

ISA classifier scores (full corpora)												
Feat. set	**before annotation (class size: 408)**				**after annot., incl. strong (class size: 437)**				**after annot., incl. strong + weak (class size: 458)**			
	A	R	P	F1	A	R	P	F1	A	R	P	F1
c(nt)	53.02	24.27	57.32	33.54	54.23	29.06	58.60	38.51	55.02	31.23	59.13	40.69
v(ec)	56.92	67.71	55.75	**60.99**	58.13	68.68	56.62	**61.89**	58.63	69.00	57.07	**62.23**
p(plx)	48.55	35.80	51.20	35.88	47.48	20.71	50.08	24.05	51.41	30.32	52.44	37.84
c+v	56.41	59.90	55.67	57.47	57.19	56.28	57.09	56.30	60.26	62.25	59.45	60.42
c+p	54.23	31.30	57.21	40.15	54.80	35.48	57.81	43.73	57.31	39.52	60.23	47.49
p+v	55.58	64.77	54.53	58.99	56.63	65.91	55.34	59.84	58.84	65.95	57.51	61.09
all	53.75	56.02	52.99	54.04	57.64	55.61	56.86	55.95	60.58	59.41	60.26	59.55
Basel.	49.88	28.88	49.79	36.56	49.54	28.38	49.21	35.99	49.56	28.17	49.24	35.83
CIA classifier scores (full corpora)												
Feat. set	**before annotation (class size: 331)**				**after annot., incl. strong (class size: 337)**				**after annot., incl. strong + weak (class size: 349)**			
	A	R	P	F1	A	R	P	F1	A	R	P	F1
c(nt)	60.42	43.78	66.18	51.89	61.42	45.67	67.38	53.79	62.46	46.70	68.91	54.98
v(ec)	60.09	44.39	64.38	51.78	59.61	44.14	63.80	51.43	62.02	46.70	67.73	54.78
p(plx)	64.95	74.93	62.62	**68.10**	65.30	74.22	63.25	**68.11**	64.76	75.09	62.54	**68.10**
c+v	62.07	48.62	66.62	55.50	62.00	48.96	65.70	55.64	64.61	50.68	70.13	58.39
c+p	65.27	65.58	65.91	65.36	65.29	65.90	65.62	65.40	65.77	67.08	66.10	66.27
p+v	64.49	62.85	65.93	63.81	65.43	64.40	66.58	65.00	65.05	61.89	66.83	63.72
all	67.21	65.56	69.17	66.46	66.17	63.80	67.80	64.99	67.06	64.16	68.93	65.86
Basel.	55.59	88.22	53.38	66.51	56.23	88.43	53.79	66.89	57.31	88.54	54.50	67.47

Table 5: Scores (accuracy, recall, precision, F1) for the ISA and CIA classifiers 1. before and 2. after annotation, adding either only the strong or the strong and the weak candidates, using each feature subset. Scores on full corpora are reported, so only the 29 strong and 21 weak candidates (ISA) resp. 6 strong and 12 weak candidates (CIA) from full corpora are added.

sifier, with an F1 score of up to 68.10 when using pplx features. While the score of the simple baseline for CIA actually outperforms some of the feature sets (F1=66.51), the same baseline performs poorly on ISA (F1=36.56). Together with the better overall CIA scores, this suggests that ISA identification is a more challenging task.

The performance of the ISA classifier improves as new candidates are added to the gold data (F1=60.99 to F1=62.23). Note that the final evaluation also includes more negative instances, as we always balance the classes. The CIA classifier, which already performs well initially, improves slightly on some feature sets when adding more instances, but the best-performing pplx feature set is stable at F1=68.10. This is not surprising, as the final verb set differs from the initial one to a smaller degree than in the ISA task.

The vec features (most indicative for ISA) are derived in an unsupervised manner from the corpora, and the pplx features (most indicative for CIA) are derived in a semi-supervised manner. The pplx features involve the initial manual development of a set of reordering rules that specify how each syntactic pattern involved in the alternation can be transformed to the alternation's other syntactic pattern(s). These features allow us to determine individual verbs' participation in a given alternation with relatively little manual effort, which means our setup is well-suited to be transferred to other alternations as well.

To our knowledge, this is the first work to approach the task of CIA identification based on **(a)** balanced verb classes, **(b)** such that positive as well as negative instances are observed in the relevant syntactic patterns in a large corpus, **(c)** without relying on lexical information from an external resource.

Concerning **(a)**, the positive sets used by Baroni and Lenci (2009) and one setting in Seyffarth (2019) are larger than their negative sets. Concerning **(b)**, Kann et al. (2019) work with manually-constructed sentence pairs, while Baroni and Lenci (2009) and another setting in Seyffarth (2019) use the negative

		ISA	CIA
full corpora	strong/weak candid.	29/21	6/12
	non-candidates	117	64
	Cohen's κ	0.65	0.44

		ISA	CIA
reduced corpora	strong/weak cand.	8/11	7/10
	non-candidates	94	53
	Cohen's κ	0.55	0.5

Table 6: Overview of the new strong, weak and non-candidates from the ISA and CIA annotations.

examples from Levin (1993) as negative set, the majority of which cannot appear in both CIA constructions. Concerning **(c)**, older approaches such as McCarthy (2000), McCarthy (2001), Tsang and Stevenson (2004) use WordNet to compare argument fillers.

For the ISA identification task, we are not aware of any comparable work.

3.2 Annotation results

We ran the experiments for each alternation with two annotators, performing the annotations once on the full corpora and once on the reduced corpora containing only sentences up to length 10. The agreement between annotators (Cohen's κ) across these 4 experiments was between 0.44 and 0.65 (see Table 6).

We count a verb as a new alternation candidate for an annotator if each of the relevant constructions received a good label (either 1 or 2). For ISA, we group the *with* and *using* constructions together, so it is sufficient for a verb to receive a good label in either one of these in addition to receiving a good label for the subject construction. Verbs that were candidates for both annotators are **strong alternation candidates**; verbs that were candidates for only one annotator are **weak alternation candidates**.

Our results partially support the initial hypothesis that using only short sentences (up to 10 tokens) leads to better annotation results. For CIA, a similar number of strong/weak candidates are found in the full and reduced corpora (full: 6/12, reduced: 7/10), with the reduced corpora yielding a slightly higher percentage of strong or weak candidates out of all annotated verbs (full: 22.0%, reduced: 24.3%). For ISA, the full corpora yield a larger set of new candidates (full: 29/21, reduced: 8/11), which also corresponds to a larger percentage of either strong or weak candidates out of all annotated verbs (full: 29.9%, reduced: 16.8%). Annotation increased the number of known positive instances for ISA by 9.1% (resp. 16.9% incl. weak candidates), and for CIA by 3.9% (resp. 6.6% incl. weak candidates).

3.2.1 Discussion of new ISA candidates

(4) lists all the verbs that were found to be new candidates for ISA across both full and reduced corpora: 37 strong and 32 weak candidates, with a relatively low annotation effort (appr. 6 hours per annotator).

(4) a. **Strong candidates:** *amuse, unite, insulate, jam, incapacitate, overwhelm, deafen, conceal, endear, alleviate, infect, distort, erase, upset, propel, unlock, eliminate, tickle, transform, publicise, shape, alert, suppress, cultivate, instill, bombard, excavate, smother, buffer, refute, substantiate, bombard, shield, disguise, sustain, model, dress*
b. **Weak candidates:** *redraw, alienate, tie, stir, ferment, tarnish, clamp, clinch, suspend, sustain, endanger, anger, relax, revisit, circle, anchor, color, rouse, kick-start, mop, ventilate, print, facilitate, terminate, return, wrap, permit, educate, insult, pollute, agitate, pin*

Two verb classes were particularly difficult to annotate: experiencer verbs and verbs that belong to the Locatum Subject Alternation (LSA). For experiencer verbs (*amuse*, *endear*, *upset*), their participation in ISA depends on the extent of intentionality associated with the event: If the agent in (5-a) uses her stories to achieve an intended goal of amusing someone, then *her stories* can be considered an instrument. (5-b), however, leaves the degree of intentionality open and only describes a change of state. Annotators assigned labels to sentences without context, and thus had to use sentence content and world knowledge to decide how likely each of these cases were to involve intentionality on the part of the agent.

(5) a. She amused me with **her stories**. b. **Her stories** amused me.

(6) a. He decorates the garden with **trees**. b. **Trees** decorate the garden.

Verbs that belong to the Locatum Subject Alternation (Levin, 1993) (*bombard*, *decorate*, *smother*) allow a locatum – an entity whose location is described by the verb – to appear either in a *with*-phrase or in the subject position (see (6)). Levin and Rappaport (1988) argue that locatum *with*-phrases may be interpreted as instrument phrases, but they may also be combined with an additional *with*-phrase containing an instrument, in which case the locatum element is not an instrument. Furthermore, the subject pattern may also receive a stative interpretation, which would also exclude ISA. In our data, LSA verbs usually appeared in locatum contexts as well as in instrument contexts.

Some new ISA candidates do not take a lexically obligatory instrument. For instance, the event described by *infect* can be achieved with or without the use of an instrument, intentionally or unintentionally. It was labeled as belonging to ISA even though its instrument argument is optional.

The new ISA candidates do not exhibit a strict selectional preference on the semantic type of instruments. While prototypical instruments (like *hammer* in (1) or *spoon* in (2)) are typically concrete objects, this was not always the case for instrument arguments of our candidates. For instance, one observed instrument for the verb *shape* is *our ideas*, which describes an abstract set of entities. Annotators viewed this as a felicitous instrument for that verb: In the situation described by the sentence, the *ideas* were used by an explicit or implicit agent to bring about the event described by the verb *shape*. Events were also occasionally used as instruments (e.g. *kick to the shin* as instrument for *incapacitate*).

The diversity of possible semantic types of instruments is one of the challenges of identifying new ISA verbs: Instruments can be either canonical to a verb (e.g. opening a door with a key), objects that receive an instrument function by way of coercion (e.g. opening it with a crowbar), or events that precede or accompany the event described by the verb and thus facilitate it (e.g. opening it with a kick).

Most non-candidates found in our experiments appeared with *with*-phrases that signified something other than instrument usage, e.g. accompaniment. *with* is a highly polysemous preposition; the Preposition Project (Litkowski and Hargraves, 2006) lists 10 main senses.[8] Such verbs superficially seem to instantiate the constructions associated with ISA and are accordingly labeled as such by the classifier, but do not carry the corresponding semantics, and are thus bad candidates for the alternation.

Issues of attachment ambiguity, where the parser has to decide whether a *with*-phrase is attached under the verb or under an object or another element in the sentence, were mostly circumvented by our original corpus filter that only selected sentences with *with*-phrases directly under the verb.

3.2.2 Discussion of new CIA candidates

The verbs in (7) were found to be new candidates for CIA across both full and reduced corpora: 13 strong and 22 weak candidates, with a relatively low annotation effort (appr. 3 hours per annotator).

(7) a. **Strong candidates:** *convulse, merge, resume, commence, dress, stall, deploy, panic, prick, spin, ruffle, spill, start*
b. **Weak candidates:** *award, disconnect, circulate, stick, delight, curl, spell, spread, depress, queue, relocate, label, transfer, engage, pass, still, graduate, bubble, pitch, bounce, spring, stream*

The CIA annotation resulted in a smaller set of new alternation candidates than the ISA annotation. This is partially due to its smaller initial training set (408 for ISA, 331 for CIA); furthermore, the CIA classifier already performed reasonably well on the initial training data – most iterations did not yield as many as 40 false positives, which also explains why annotation took less time for CIA.

The new CIA candidates *resume*, *commence*, *start* belong to the VerbNet class *begin-55-1*. While that class has no explicit causative frame, Levin (1993) does state that some members of the *Begin* class of verbs participate in the causative-inchoative alternation. Our setup identified these alternating verbs independently and without annotators consulting Levin or VerbNet, which shows that our approach is successful at identifying verbs that are also discussed as good alternation candidates in the literature.

[8] The list of senses is available at `https://web.archive.org/web/20191203143759/http://www.clres.com/cgi-bin/onlineTPP/find_prep.cgi?lemma=with`.

The new candidates *spread*, *circulate*, *relocate*, *transfer*, *pass*, *bounce*, *spring* relate to physical state changes. Such verbs seem to be particularly well-suited for CIA, since this type of state change can be induced either by an agent (causative use) or by the theme itself (inchoative use).

Most non-candidates were not verbs with non-obligatory arguments, as we expected, but rather verbs for which many corpus sentences were misparsed. This was often the case for verbs with noun homographs, such as *ruin*, *limit*, *paint*, *sequence*, *guess*, *treat*, *axe*, *scream*.

3.3 Limitations and Future Work

An open issue with our approach to both alternations is the impact of ambiguity and polysemy on the reliability of our results. The observed usages in different syntactic patterns for each verb do not necessarily involve the same sense: In the current setup, we do not distinguish word senses, and thus cannot verify whether the observed instances actually relate to the same lexical item. A word sense disambiguation (WSD) step may help; however, WSD is usually applied on the document level or based on immediate sentence context (Raganato et al., 2017), neither of which is available in ENCOW.

Our use of corpus data allows us to identify new alternation candidates with some success, but our annotation scheme is not suited for the explicit identification of non-alternating verbs. The absence of usages of a given verb in a particular syntactic pattern in our corpus does not entail the infelicity of that pattern for the given verb. This is a well-known issue in corpus linguistics (Kübler and Zinsmeister, 2014).

Furthermore, the work described here relies on two annotators. While the results are promising, this proof-of-concept requires an extension involving more annotators, which will yield more reliable results. Another possible improvement is the development of a more detailed annotation scheme (including, for instance, a confidence score for the annotator and/or an estimate of the degree of polysemy for the verb in the presented corpus examples). However, this would also increase the required annotation effort.

4 Conclusion

The contribution of this work is two-fold. First, we present feature-based classifiers for the instrument subject alternation and the causative-inchoative alternation. Second, we show that an iterative, hybrid approach to identifying alternations involving a classification step and a manual annotation step is a promising way to find more alternating verbs.

Our classifiers outperform the baseline for both alternations (by a small margin for CIA and a larger margin for ISA). Our features are good predictors of the alternations, even though our data makes the tasks particularly challenging because both positive and negative instances are observed in all relevant syntactic patterns in our corpora. The best performance on the alternation identification task is achieved with vector-based features for ISA, and perplexity-based features for CIA. The features are derived from corpora in an unsupervised (vec) or semi-supervised (pplx) manner, which means that our approach can be transferred to similar phenomena. The pplx features in particular seem to be a reliable approximation of the felicity of alternate sentences constructed from original corpus data.

Our gold data from VerbNet only contains positive instances, but no reliable negative instances. Our manual annotation step allows us to verify whether hypothesized negative examples are truly negative instances, or instead good candidates for the alternation that should be added to the positive set. The combination of our classification step and annotation step minimizes the necessary annotation effort while leading to a number of new alternation candidates.

Acknowledgements

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” and the individual DFG project “Unsupervised Frame Induction (FInd)”. We wish to thank the anonymous reviewers for their constructive feedback and helpful comments.

References

Marco Baroni and Alessandro Lenci. 2009. One distributional memory, many semantic spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 1–8, Athens, Greece, March. Association for Computational Linguistics.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. 2013. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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*, Budapest, Hungary, April. Association for Computational Linguistics.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

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.

Sandra Kübler and Heike Zinsmeister. 2014. *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Publishing.

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, Beijing, China, July. Association for Computational Linguistics.

Beth Levin and Malka Rappaport. 1988. Nonevent -er nominals: A probe into argument structure. *Linguistics*, 26:1067–1084, 01.

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

Kenneth C. Litkowski and Orin Hargraves. 2006. Coverage and Inheritance in The Preposition Project. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*.

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 C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 3111–3119. Curran Associates, Inc.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain, April. Association for Computational Linguistics.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, and Andreas Witt, editors, *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster. UCREL, IDS.

Esther Seyffarth. 2019. Identifying participation of individual verbs or VerbNet classes in the causative alternation. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 146–155.

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.

Vivian Tsang and Suzanne Stevenson. 2004. Using selectional profile distance to detect verb alternations. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*, pages 30–37, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Koen Van Hooste. 2018. *Instruments and Related Concepts at the Syntax-Semantics Interface*. Düsseldorf University Press GmbH.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, March.