Aspectuality Across Genre: A Distributional Semantics Approach

Thomas Kober
Rasa
t.kober@rasa.com

Malihe Alikhani
University of Pittsburgh
malihe@pitt.edu

Matthew Stone
Rutgers University
mdstone@rutgers.edu

Mark Steedman
University of Edinburgh
steedman@inf.ed.ac.uk

Abstract

The interpretation of the lexical aspect of verbs in English plays a crucial role for recognizing textual entailment and learning discourse-level inferences. We show that two elementary dimensions of aspectual class, *states* vs. *events*, and *telic* vs. *atelic* events, can be modelled effectively with distributional semantics. We find that a verb's local context is most indicative of its aspectual class, and demonstrate that closed class words tend to be stronger discriminating contexts than content words. Our approach outperforms previous work on three datasets. Lastly, we contribute a dataset of human–human conversations annotated with lexical aspect and present experiments that show the correlation of telicity with genre and discourse goals.

1 Introduction

One of the fascinating aspects of studying aspectual class of verbs in English is its relation with non-verbal categories. Thus, although in origin a property of the verb, the aspectual class interacts in a tight-knit fashion with other words in a sentence. Previous research has discussed the importance of predicting the aspectual classes of verbs for predicting coherence relations in text and imagery (Alikhani and Stone, 2019), predicting links in entailment graphs (Hosseini et al., 2019) and interpreting sign languages (Wilbur, 2003). In addition, knowledge about the aspectual class of a verb phrase, and its influence on the temporal extent and entailments that it licenses, has been leveraged in the past for a number of natural language understanding tasks such as temporal relation extraction (Costa and Branco, 2012), event ordering (Chambers et al., 2014; Modi and Titov, 2014), and statistical machine translation (Loáiciga and Grisot, 2016).

The *Aktionsart* (Vendler, 1957) of a verb determines the temporal extent of the predication as well as whether it causes a change of state for the entities involved (Filip, 2012). As *Aktionsart* typically refers to the lexical aspect of a verb in isolation, we adopt the terminology of Verkuyl (2005), and refer to the compositionally formed *Aktionsart* of a verb phrase as *predicational aspect*.

One of the most important distinctions of the predicational aspect of a verb is between **states**, such as *to know* or *to love*, and **events**, such as *visit* or *swim*. This distinction is important for identifying the entailments that a given verb phrase licenses, as *stative* predications do not, by definition, entail any change of state. This property has important consequences for a number of natural language understanding tasks such as question answering. For example, if it is known that *John has arrived in Vienna*, a system leveraging aspectual information will be able to infer that the completion of the event of *arriving in Vienna*, indicated by the perfect VP *having arrived in*, has caused a change of state which entails *being in*. Therefore, when asked *Where is John?*, the system will be able to produce the correct answer: *Vienna*. On the other hand, a predominantly stative verb such as *to know*, as in *Eve knows a lot about quantum mechanics*, does not cause a change of state for either *Eve* or *quantum mechanics*.

Telic predicates do not license consequent state inferences from their progressive VP forms to corresponding non-progressive forms.[1] Thus, telic/atelic classifications are supported by contrastive pairs like the following:

---

[1]This is also known as the *Imperfective Paradox* (Dowty, 1979).

(1) Mary *was drawing* a circle $\nrightarrow$ Mary *drew* a circle (telic)

(2) Mary *was pushing* a cart $\rightarrow$ Mary pushed a cart (atelic)

In this paper we propose to approach the problem of classifying predicational aspect with distributional semantics. Our hypothesis is that the meaning distinctions of a verb that relate to its aspectual class should be reflected in its distribution when composed with its context. We therefore intersect word vectors with their context in order to determine a VP's predicational aspect, and show that we achieve a new state-of-the-art on two datasets. We further evaluate our approach on two new genres: image captions and situated human–human conversations, thereby extending the validity of our findings across a variety of genres.

2   Related Work

An early approach to classifying the lexical aspectual class of a verb in context was proposed by Passonneau (1988), who applied a decompositional analysis of the verb to determine the aspectual class for verb occurrences in a restricted domain. The first general-purpose study was conducted by Siegel and McKeown (2000), who built up on earlier work by Klavans and Chodorov (1992), and collected *linguistic indicators* for lexical aspect from a large corpus. These include the presence of *in-* or *for*-adverbials, the tense of the verb or its frequency. Siegel and McKeown (2000) subsequently applied different supervised machine learning algorithms to classify the extracted feature vectors into either *states* or *events*, or *telic* or *atelic* events. Siegel and McKeown (2000) show that their method substantially improves over a majority-class baseline. The first approach to include features derived from a distributional semantic model has been proposed by Friedrich and Palmer (2014). In addition to the linguistic indicator features of Siegel and McKeown (2000), Friedrich and Palmer (2014) extract representative stative, dynamic or mixed verbs from the *lexical conceptual structure (LCS)* database (Dorr and Olsen, 1997) and subsequently use distributional representations to derive similarity scores for the mined verbs.

Another extension to the work of Friedrich and Palmer (2014) has been proposed by Heuschkel (2016), who refines the distributional similarity features by first contextualising a target verb with its subject or object, and only then computing the distributional similarities to the set of representative verbs from the LCS database as in Friedrich and Palmer (2014). All else being equal, Heuschkel (2016) shows that contextualising the distributional representations improves performance on the **Asp-ambig** dataset of Friedrich and Palmer (2014).

In contrast to this line of research we do not make explicit use of any hand-engineered linguistic indicator features but show that these can be picked up in an unsupervised way by composing distributional semantic word representations. The linguistic indicators are furthermore frequently collected on the verb *type* level instead of on the *token* level. Similar to Falk and Martin (2016), we are concerned with classifying *verb readings*; however, we do not use engineered features as Falk and Martin (2016) do, but directly leverage local contextual information in the form of distributional representations. Our approach is also not reliant on the availability of a parallel corpus as in Friedrich and Gateva (2017). The major difference between our approach of using distributional word representations and previous approaches is that we are using the word representations *directly* for classification, rather than *indirectly* by computing similarity scores and using these as features. This furthermore liberates us from the requirement of having a representative seed set of verbs per class to compute the distributional similarities from.

3   Dataset

We introduce a new dataset, **DIASPORA**,[2] of human-human conversations annotated with predicational aspect, representing the first dialogue dataset annotated with aspectual information. As reviewed in Section 2, what partly makes computational research on predicational aspect difficult is the lack of diverse annotated corpora. With the release of **DIASPORA** we provide a situated conversational perspective on

---

[2]**DI**alogue **ASP**ectuality from **ORA**l conversations. The dataset is publicly available from `https://go.rutgers.edu/cb1e5c1`.

predicational aspect research, and thereby extend the existing evaluation repertoire by a very important genre.

Annotation effort. As our starting point, we sampled 2000 utterances from the Walking Around corpus (Brennan et al., 2013) uniformly at random. The Walking Around corpus is a dataset of human–human phone conversations where one party needs to find certain landmarks on a university campus and receives directions via phone from the other party. Table 1 lists part of a conversation from the Walking Around corpus, where the speakers identify a landmark that the second speaker needs to reach.

...	
Speaker 1	*You're looking for the ship scul- sculpture*
Speaker 2	*Okay*
Speaker 1	*it should be (..) yeah*
Speaker 2	*I think I see it but I'm not close enough yet*
Speaker 1	*It should be right in front of the Heavy Engineering Building*
Speaker 2	*It's a ship?*
Speaker 1	*Yeah it looks like one (..) like it's got (..) how do you describe i- it looks really weird*
...	

Table 1: Part of an example dialogue from the Walking Around corpus (Brennan et al., 2013).

We chose the Walking Around corpus because its conversations are situated and in real-time, and because it contains a good distribution of *stative*, *telic*, and *atelic* verb phrases. After sampling the initial set of 2000 utterances, we filtered multi-sentence utterances and utterances that did not contain a verb. We furthermore removed any filled pauses (indicated by "*(..)*" in Table 1) that have been transcribed and marked in the dataset. Following Alikhani and Stone (2019), we annotated the *first* VP for predicational aspect in all utterances. For example, the last utterance of Speaker 1 in Table 1 contains multiple verbs, and we have annotated the phrase *Yeah it looks like one*—the first VP in the utterance.

The study has been approved by Rutgers's IRB. Expert annotators annotated the whole dataset and were paid an hourly rate of 15 USD. They were final year linguistics undergraduate students and were provided with an annotation protocol for their task.[3] To assess the inter-annotator agreement, we determine Cohen's κ value. We randomly selected 200 sentences and assigned each to two annotators, obtaining a Cohen's κ of 0.81, which indicates almost perfect agreement (Viera and Garrett, 2005).

Overall statistics. The final **DIASPORA** dataset contains 927 annotated utterances, consisting of 400 utterances labelled as expressing *stative* predicational aspect (43%), 279 labelled as *telic* (30%), and 248 labelled as *atelic* (27%). The overall average utterance length is 15.58. Table 2 lists length statistics

Label	Mean	Median	Min	Max
State	16.34	13	4	94
Telic	14.65	11	3	80
Atelic	15.38	12	2	74

Table 2: Utterance length statistics per label in **DIASPORA**.

for **DIASPORA** per individual label. The means and medians are relatively similar across all classes, suggesting that there is no bias in terms of utterance lengths for any individual class.

The **DIASPORA** dataset contains 98 unique verb forms, spanning 69 lemmas, with the top 10 most frequent verb lemmas making up ≈78% of all verbs in the corpus.

The characteristic of few verbs making up a large proportion of the overall data has already been observed for captions (Alikhani and Stone, 2019). This is an expected property in **DIASPORA** and is due to the single-domain nature of the Walking Around corpus. Figure 1 shows the frequency distribution of the top 10 most frequent verbs and their associated label distribution. The large proportion of various forms of *be* is due to many utterances of either speaker referring to the current location of the subject looking for the landmark (e.g. utterances like *I am currently at . . .*). The label distribution of the 10 most frequent verbs shows that there are some highly skewed verbs, such as *get*, *see* or *know*, which have a

---

[3]The annotation protocol is published with the dataset at `https://go.rutgers.edu/cb6le5c1`.
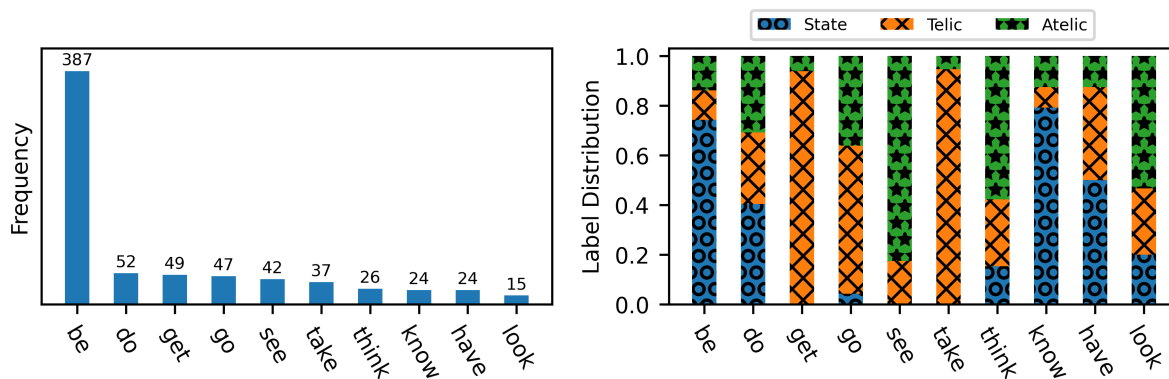
Figure 1: Frequency distribution of the 10 most frequent verbs (left) and associated label distribution for the 10 most frequent verbs (right).

clear majority class, whereas *do* or *look* exhibit a much more balanced, and therefore ambiguous label distribution.

4 Experiments

The utility of distributional semantic word representations has been shown in a large body of works in recent years (Weeds et al. (2014), Nguyen et al. (2017), Socher et al. (2013), Bowman et al. (2015); *passim*). In order to compose a verb with its context we apply pointwise addition as a simple distributional composition function. Pointwise addition in neural word embeddings approximates the intersection of their contexts[4] (Tian et al., 2017), and has been shown to be an efficient function for contextualising a word in a phrase (Arora et al., 2016; Kober et al., 2017).

4.1 Distributional Models for Predicational Aspect

Following previous work on modelling the aspectual class of a verb (Siegel and McKeown, 2000; Friedrich and Palmer, 2014), we treat the problem as a supervised classification task, $y = f(x)$, where y represents the aspectual class of a verb, f represents a classification algorithm, and x an input vector representation of a verb in context. For all of our experiments, f is a logistic regression classifier,[5] with default hyperparameter settings. In all our experiments, the input vector x is based on 300-dimensional pre-trained skip-gram word2vec (Mikolov et al., 2013) vectors.[6] We lowercase all words, but do not apply any other form of morphological preprocessing, which means that we retain different representations for different inflected forms of a verb—i.e. *look*, *looks*, *looking*, and *looked* are represented by 4 distinct vectors.

4.1.1 Classifying Aspect with Distributional Semantics

For this approach we obtain a word2vec representation x for a given verb v and feed x into a logistic regression classifier in order to predict the aspectual class of v. This approach represents a rather naïve baseline that assumes that the aspectual class of a verb is a purely lexical phenomenon on the type level and can be determined independent of any context.

4.1.2 Incorporating Context with Distributional Composition

Let x be a word2vec representation for a given verb v, and C be the set of context words extracted for v, with $c \in C$ denoting the vector representation for an extracted context word of v. The composed

---

[4]Note that for sparse count-based distributional representations, addition would correspond to a union of contexts and multiplication to an intersection (Kober, 2018).

[5]We use scikit-learn (Pedregosa et al., 2011)—which, much like ourselves, is relying on numpy (Harris et al., 2020).

[6]The pre-trained vectors are availble from `https://code.google.com/archive/p/word2vec/`, and we used gensim (Řehůřek and Sojka, 2010) for processing the word2vec representations.

representation of *v*, denoted by x', can then be expressed as a simple sum:

$$x' = x + \sum_{c \in C} c \tag{1}$$

Subsequently, x' is passed through a logistic regression classifier in order to predict the aspectual class of *v*. This model aims to capture the compositional nature of predicational aspect by integrating local contextual information into the model.

Types of Context

We investigate two different kinds of context: simple linear context windows of varying length and first-order dependency contexts. For example for the sentence in Figure 2, a linear context window of size 1 would extract *Jane* and *to* for the target verb *decided*, whereas a dependency-based context would extract *Jane* and *leave*. We used the Stanford NLP pipeline (Manning et al., 2014) with default settings for parsing the sentences in our datasets.
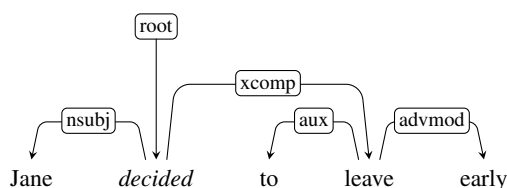
Figure 2: With a linear context window of size 1, *Jane* and *to* would be extracted as contexts for the verb *decided*. With a dependency-based context, *Jane* and *leave* would be extracted.

For linear context windows we use sizes $\{1, 2, 3, 5, 10\}$, and for first-order dependency-based contexts we experiment with using only the head[7] of the verb, only its children, or the full first-order context.

Incorporating the Full Sentence

We furthermore test a model that incorporates the whole sentential context into a vector representation. The approach simply uses all words from a given sentence and composes their corresponding word2vec representations as in Equation 1 above to create an embedding for the whole sentence. Embedding a sentence by adding word vectors has been shown to be an effective method for other NLP tasks such as sentiment analysis (Iyyer et al., 2015) and recognising textual entailment (Wieting et al., 2016).

The underlying rationale behind this approach is that the aspectual class of a verb is a function of the sentence as a whole, rather than dependent on local context alone (Moens, 1987; Moens and Steedman, 1988; Dowty, 1991).

5 Experiments

We perform experiments that assess the suitability of distributional representations for distinguishing *states* from *events* (§ 5.1), and *telic* from *atelic* events (§ 5.2). Only a completed and *telic* event licenses a new consequent *state*. Therefore, modelling predicational aspect is important for deeper text understanding, for example for modelling cause and effect, and especially for inferring consequent states.

5.1 Experiment 1 — States vs. Events

For the distinction between *states* and *events* we perform experiments on 5 datasets in total. We use the **Asp-ambig** dataset by Friedrich and Palmer (2014), the **SitEnt** dataset by Friedrich et al. (2016), our own sub-sampled version of the **SitEnt** dataset, the **Captions** dataset by Alikhani and Stone (2019), and our own **DIASPORA** dataset, proposed in this work.

The **Asp-ambig** dataset is sampled from the Brown corpus (Francis and Kucera, 1979) and is based on 20 frequently occurring verbs whose predicational aspect changes depending on context. For each verb, Friedrich and Palmer (2014) collected 138 sentences, resulting in 2760 examples in total. The dataset

---

[7]If the target verb is the root, then no context is used for that verb.

contains the annotations of whether the verb in context expresses a *state*, *event*, or whether it could be *both*.[8] Following Friedrich and Palmer (2014), we report accuracy using leave-one-out cross-validation.[9]

We furthermore evaluate our approach on the **SitEnt** dataset (Friedrich et al., 2016). The **SitEnt** dataset contains 40k sentences from the MASC corpus (Ide et al., 2008) and English Wikipedia, split into separate training and test sets. We evaluate our approach on the test set, using the original split of Friedrich et al. (2016). The **SitEnt** dataset contains annotations for verbs in context as either expressing a *state* or an *event*, but not both. Following Friedrich et al. (2016), we report class-based F1-scores.

During model development we noticed an idiosyncrasy in the **SitEnt** dataset, where only 900 verb types out of 4.5k occurred with both class labels, and only 267 of them had a balanced (i.e. ambiguous) class distribution. We considered this as likely problematic as a classifier might just pick up this artefact. We therefore created a downsampled dataset — **SitEnt-ambig** — that only contains verb types with a balanced class distribution, randomly sub-sampling the majority class for verb types with an imbalanced class distribution[10]. The resulting dataset consists of 6547 examples in the training set and 1402 examples in the test set.[11] As for the original **SitEnt** dataset, we report class-based F1-scores for **SitEnt-ambig**.

In order to cover a wider variety of genres, we also evaluate our approach on the **Captions** dataset of Alikhani and Stone (2019). The dataset is based on a number of image captions corpora and contains annotations for verbs being used as *states*, *telic events* and *atelic events*. For this experiment we merge the *telic* and *atelic* class, resulting in a 2-class problem with 2687 instances with a class distribution of 22:78 (*state*:*event*). The dataset does not contain pre-defined training/evaluation splits. We therefore evaluate using 10-fold cross-validation and report class-based F1-scores.

Finally, we evaluate on our proposed **DIASPORA** dataset, again merging the *telic* and *atelic* classes for this experiment, resulting in a class distribution of 43:57 (*state*:*event*). We again report class-based F1-scores over 10-fold cross-validation.

5.1.1 Results

Table 3 below shows the results on all datasets. We compare classifying the representation of a verb without any context, the verb with local context, and the full sentence, with a majority-class baseline and previous results in the literature. The results of using a local context are based on the best performing context window around the verb, an overview of the effect of the size of the context window is shown in Figure 3. A result table comparing the best linear context window window with the best performing dependency context window is presented in Table 8 in Appendix C.

Dataset	Verb only	Local Context	Full Sentence	Maj. Class	FP14	H16	F16
Asp-ambig Accuracy	65.9	**74.2**	60.0	65.9	72.0	72.8	-
SitEnt F1 (State)	**84.0**	81.3	26.4	0.0	-	-	80.6
SitEnt F1 (Event)	**86.6**	84.5	71.9	68.9	-	-	78.6
SitEnt-ambig F1 (State)	44.0	**62.6**	0.0	0.0	-	-	-
SitEnt-ambig F1 (Event)	62.4	66.2	68.3	**68.4**	-	-	.-
Captions F1 (State)	0.1 ($\pm$ 0.07)	**58.8 ($\pm$ 0.05)**	23.4 ($\pm$ 0.06)	0.0	-	-	-
Captions F1 (Event)	87.3 ($\pm$ 0.01)	**89.7 ($\pm$ 0.02)**	86.7 ($\pm$ 0.02)	87.6	-	-	-
DIASPORA F1 (State)	76.4 ($\pm$ 0.07)	**86.5 ($\pm$ 0.04)**	80.1 ($\pm$ 0.06)	0.0	-	-	-
DIASPORA F1 (Event)	83.5 ($\pm$ 0.03)	**89.8 ($\pm$ 0.03)**	84.8 ($\pm$ 0.03)	72.5	-	-	-

Table 3: Results on classifying *states* vs. *events*. **FP14** refers to Friedrich and Palmer (2014), **H16** to Heuschkel (2016), and **F16** to Friedrich et al. (2016).

In general, a local context window exhibits the strongest performance, even achieving a new state-of-the-art on the **Asp-ambig** dataset, despite the simplicity of our setup. The strong results of the verb-only model on the **SitEnt** dataset, that substantially outperforms the sequence model of Friedrich et al. (2016) and the local context model, confirms our suspicion that the classifier learnt the fact that most verbs in

---

[8]The instances labelled as both have either been labelled by the annotators as both because both readings are possible, or have been labelled as both when the annotators disagreed in their judgement (Friedrich and Palmer, 2014).

[9]A performance overview per verb type is included in Appendix A

[10]We allowed the imbalance to be at most 60:40.

[11]The sub-sampled dataset is available from https://go.rutgers.edu/cb6le5c1

the dataset occur unambiguously with their target label. This is furthermore reflected in the results on the **SitEnt-ambig** dataset, where using only the verb leads to considerably worse performance than when taking a local context window around the verb into account. While the results on **SitEnt-ambig** are generally low, this reflects the increased difficulty of the task as well as the simplicity of our setup, and we expect to improve on these results with higher capacity models in future work.

5.1.2 Analysis

In Figure 3 we show class-based F1-score performance trajectories for varying sizes of the linear context window and the dependency context across all datasets. We observe that performance typically peaks at a narrow context window of taking 1-3 surrounding words into account, with performance dropping steeply when increasing the context window.[12] Our results also exhibit that linear window contexts are typically better predictors for predicational aspect than dependency contexts.
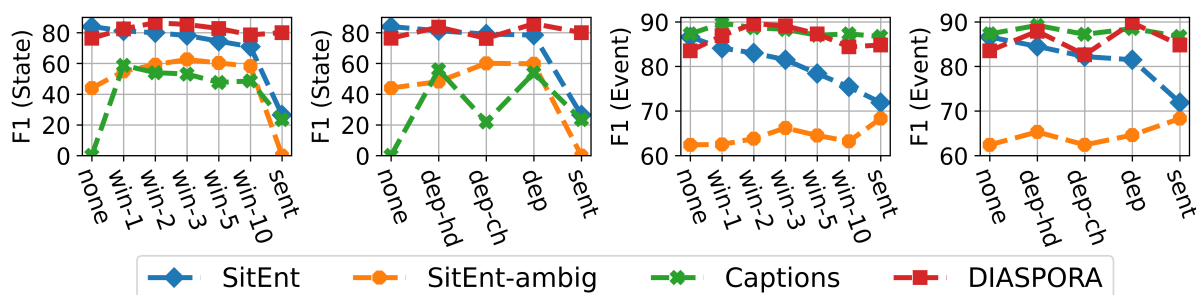

Figure 3: Class-based F1-score performance trajectories for varying sizes of the context window across all datasets.

This is an interesting result as dependency contexts are more likely to yield content words, such as nouns, adjectives or other verbs as context,[13] as opposed to linear context windows yielding more closed class words. We investigate this effect further by comparing the general overall performance of closed class words with content words. Figure 4 provides empirical evidence that closed class words are strong predictors of predicational aspect. The figure shows accuracies for PoS tags belonging to a closed-class group, in comparison to ones belonging to open class content words. We calculated PoS-based accuracy by counting how often a word with a given PoS tag contributed to a correct classification as opposed to an incorrect one. For example if the PoS tag `IN`[14] occurs 8 times as part of correctly classified context windows and 2 times as part of incorrectly classified ones, we estimate its accuracy as 0.8. We count the participation of a PoS tag for a correct or incorrect classification decision as evidence that the given word is a reliable predictor for a given class. We expect that words with high predictive capacity will more often occur in correctly classified context windows than in incorrectly classified ones.

Figure 4 highlights that closed class words are typically more reliable predictors for predicational aspect than content words. This is a very interesting result, given our model solely operates on the basis of composed word vectors, thus indicating that distributional representations for closed class words encode a substantial amount of information that can potentially be leveraged for fine-grained directional inferences. In order to assess the generalisation capability of distributional representations we performed a zero-shot experiment on the **Asp-ambig** dataset where we held out all annotated data for a specific verb for evaluation, and trained the model on the remaining data. Table 10 in Appendix D provides evidence that distributional representations capture predicational aspect of unseen verbs to a surprising extent.

Table 4 shows example sentences for two ambiguous verbs from our datasets. In the first and third sentence the preposition *at* and the particle *up*, respectively, cause the predicate to express an event. Without a preposition, verbs such as *look* can express stative aspect as in the second sentence. The

---

[12]We note that the F1-score for event on the **SitEnt-ambig** datasets appears to peak by taking the whole sentence as input, however this is an effect of its performance degrading to that of the majority class baseline, as can also be seen on its near-0 performance for predicting states.

[13]As one of our reviewers pointed out, this likely is due to our use of Universal Dependencies (Osborne and Gerdes, 2019). We show the distribution of extracted context words for the linear and dependency context windows per PoS tag in Appendix B.

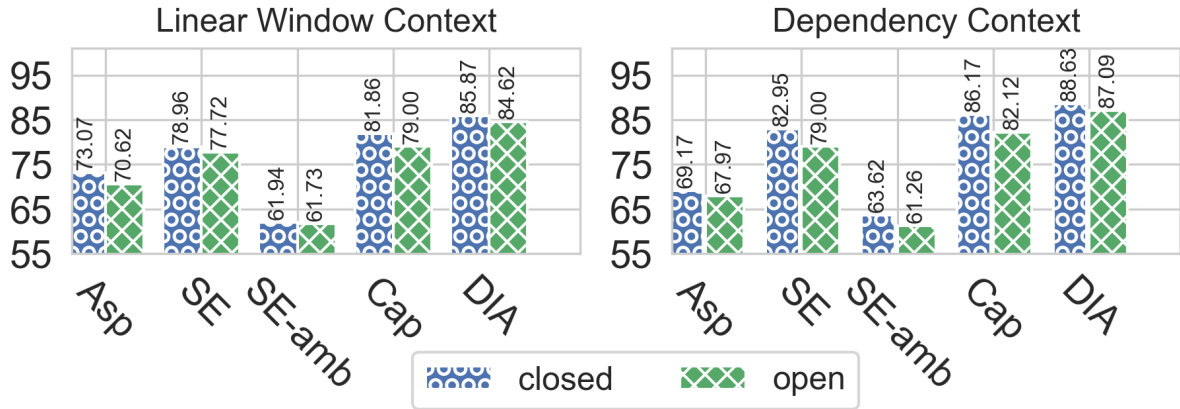[14]`IN` refers to a preposition or subordinating conjunction.

Figure 4: Averaged accuracy scores for closed class context words in comparison to content word contexts.

Aspect	Verb	Example Sentences
Event	look	(1) Shane *looked at* his watch.
State	look	(2) She sure *looks* good.
Event	stand	(3) Jefferson shook his head and *stood up*.
State	stand	(4) A metal table and four chairs *stood in* the center.

Table 4: Example Sentences from the state vs. event datasets.

last sentence is an interesting case where the verb *stand* occurs in the context of a preposition, yet the combination remains stative, as the sentence describes the arrangement of inanimate objects.

5.2 Experiment 2 — Telic vs. Atelic Events

For classifying *telic* and *atelic* events we are using the *Telicity* dataset of Friedrich and Gateva (2017), the **Captions** dataset of Alikhani and Stone (2019), as well as our own proposed **DIASPORA** dataset.

The **Telicity** dataset contains 1863 sentences extracted from the MASC corpus, where a verb in context is labelled as either *telic* or *atelic*. The dataset is imbalanced with 82% of verb occurrences being labelled as *telic*. We follow the experimental protocol of Friedrich and Gateva (2017) and report accuracy and class-based F1-scores, using document-based cross-validation. During our experimental work we again noticed that only 70 out of approximately 570 distinct verbs in the dataset occur with both labels. However, applying the same strategy as for the **SitEnt-ambig** dataset would have resulted in too little data.[15] Therefore, given this characteristic, we again expect the classifier using the verb without any context to achieve artificially high performance.

For the **Captions** dataset, we omit the examples labelled as *stative*, leaving us with 2092 captions in total, of which 800 are annotated as *telic* (38%), and 1292 as *atelic* (62%). We perform 10-fold cross-validation and report accuracy and class-based F1-scores.

Finally, for our **DIASPORA** dataset, we also omit the utterances annotated as expressing *stative* predicational aspect, leaving us 527 examples in total, 279 instances labelled as *telic* (53%), and 248 instances labelled as *atelic* (47%), thus representing the most balanced dataset among the three.

5.2.1 Results

Table 5 shows the results for all three datasets, comparing a model that only has access to the distributional representation of the target verb itself, with models that have access to a local context window and the full sentence, as well as to previous results in the literature. A result table comparing the best linear context window window with the best performing dependency context window is presented in Table 9 in Appendix C.

---

[15]Using only sentences with verbs that occur with both labels regardless of their class distribution resulted in only 295 examples in total.

Dataset	Verb only	Local Context	Full Sentence	Maj. Class	FG17	FG17+IC
Telicity Accuracy	**87.2**	86.4	85.8	82.0	86.7	82.3
Telicity F1 (Telic)	92.1	91.6	91.1	90.1	**92.2**	88.6
Telicity F1 (Atelic)	**62.6**	60.3	60.6	0.0	53.7	61.4
Captions Accuracy	78.8 (± 0.03)	**79.1 (± 0.02)**	78.8 (± 0.02)	61.8	-	-
Captions F1 (Telic)	70.7 (± 0.05)	**72.3 (± 0.03)**	71.2 (± 0.04)	55.3	-	-
Captions F1 (Atelic)	**83.3 (± 0.03)**	83.0 (± 0.02)	83.1 (± 0.02)	0.0	-	-
Dialogue Accuracy	64.3 (± 0.04)	**69.3 (± 0.05)**	65.1 (± 0.05)	52.9	-	-
DIASPORA F1 (Telic)	64.5 (± 0.06)	**70.3 (± 0.06)**	66.5 (± 0.05)	69.2	-	-
DIASPORA F1 (Atelic)	63.3 (± 0.05)	**67.8 (± 0.05)**	62.9 (± 0.07)	0.0	-	-

Table 5: Results on classifying telic vs. atelic events. **FG17** refers to the best performing model of Friedrich and Gateva (2017), and **FG17+IC** refers to the model of Friedrich and Gateva (2017) with access to additional data.

Our purely distributional models achieve competitive results, with the expected strong performance for the verb-only model, that is even beating the current state-of-the-art in terms of accuracy and F1-score for the *atelic* class. For the **Captions** and **DIASPORA** datasets we observe similar trends as for the *state* vs. *event* datasets above, with the models that operate over a local context window typically achieving the strongest performance. Notably, the verb-only models are able to perform competitively with local context windows across all datasets. While telicity itself is not part of the morphology of English verbs, telic events frequently correlate with the past tense, such that the distributional representation for the inflected verb already encodes a substantial amount of information.

5.2.2 Analysis

Figure 5 shows a class-based F1-score performance trajectory across all datasets and varying context window sizes. Unlike for distinguishing states from events in Figure 3 above, predicting telicity appears to be less dependent on a small local context window surrounding the target verb. This is reflected in Figure 5 which does not contain such clear performance peaks, but is more uniform across different sizes of context windows.
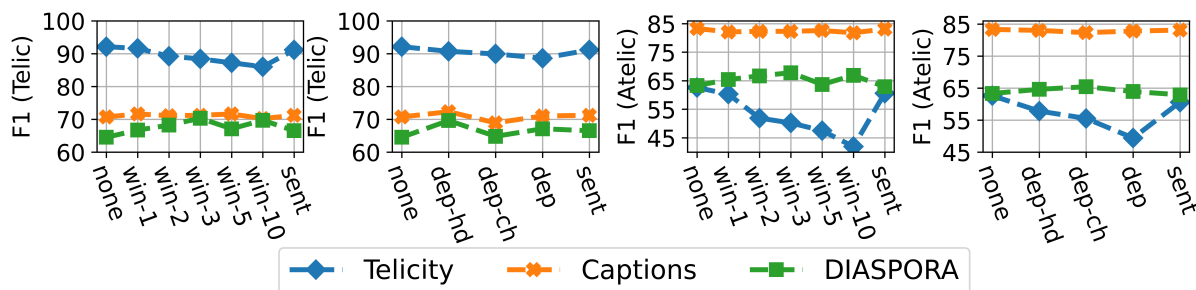

Figure 5: Class-based F1-score performance trajectories for varying sizes of the context window across all datasets.

We furthermore show the averaged PoS-based accuracy plot in Figure 6. For predicting telicity, closed class words are less reliable predictors in comparison to content words than for modelling states and events above.

This result becomes more transparent when analysing actual sentences from our dataset.

Aspect	Verb	Example Sentences
Telic	leave	(1) Okay, I *have left* the building.
Atelic	walk	(2) Okay, I'*m still walking* towar- oh is it blue?
Telic	turn	(3) Fans *turned* on the players and manager.
Atelic	paddle	(4) Four kayakers *paddle* through the water.

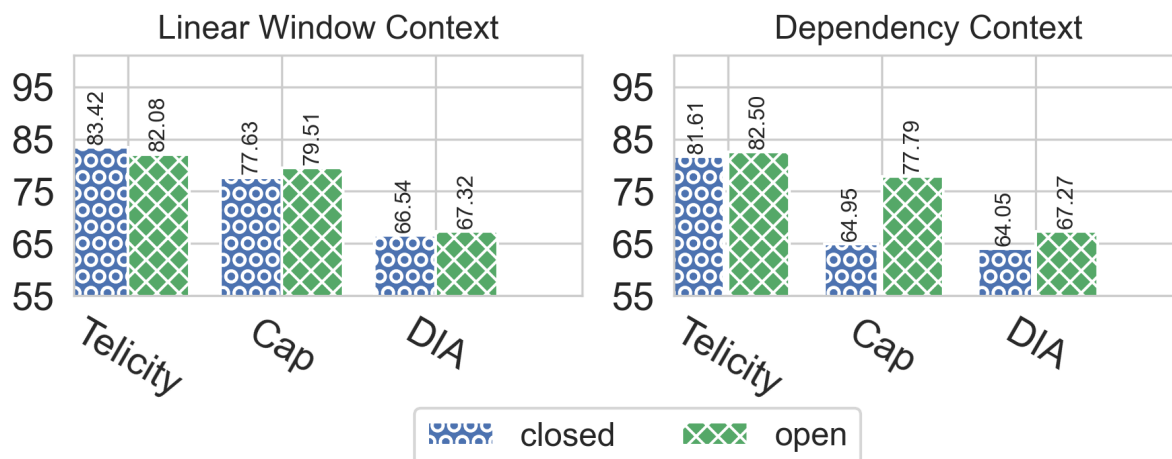Table 6: Example Sentences from the telic vs. atelic datasets.

Figure 6: Averaged accuracy scores for closed class context words in comparison to content word contexts.

Table 6 shows some example sentences from the datasets annotated for telicity. The sentences show that telicity in English is frequently associated with tense, with present tenses indicating atelic eventualities and past tense indicating a completed event. This suggests that frequently the verb by itself might be sufficient for inferring telicity as in sentences (3) and (4). In many other cases, the verb interacts with its auxiliary in a tensed construction as in sentences (1) and (2).

6 Conclusion

In this work, we have proposed the first dataset of human-human dialogues annotated with the aspectual class of verbs. We have proposed a compositional distributional approach for modelling the aspectual class of English verbs in context. Our results indicate that distributional models are able to learn concise representations for closed class words such as particles and prepositions, and that classifiers using composed distributional representations achieve a new state-of-the-art on three recently proposed datasets. We have furthermore contributed a qualitative analysis, providing empirical evidence for the long standing insight of semanticists that the presence of prepositions or particles in a verb phrase, tend to be very reliable indicators of the verb's aspectual class (Vendler (1957), Dowty (1979), Moens and Steedman (1988), *passim*). Our model setup was intentionally kept simple as we were primarily concerned with the question whether predicational aspect can be captured with a distributional semantics approach *in principle*. We note that using more sophisticated models might yield even stronger results, although in preliminary tests, we did not observe any meaningful performance difference when replacing our bag-of-embeddings approach with ELMo (Peters et al., 2018) or BERT (Devlin et al., 2019).

While this work was done on English, we aim to use our methodology in a multilingual setup in future work as distributional approaches scale well with growing amounts of data and across languages.

Aspect, alongside tense, is a crucial indicator of the temporal extent of a verb as well as the entailments it licenses. In future work we plan to integrate aspectual information for improving the unsupervised construction of entailment graphs (Berant et al., 2010; Hosseini et al., 2018), as well as temporal reasoning, which has been shown recently to be difficult for distributional semantic models (Kober et al., 2019).

Aspectual information can be utilised for directional entailment detection by inferring that the event of *buying something* entails the state of *owning that thing*, but not the other way round. Determining the *telicity* of an event also enables fine-grained inferences about whether an event caused a change of state. For example, while the telic context of *writing a sonnet in fifteen minutes* entails a change to a state where a finished sonnet exists, the atelic context of *writing a sonnet for fifteen minutes* does not.

Acknowledgements

References

Malihe Alikhani and Matthew Stone. 2019. "caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. *CoRR*, abs/1601.03764.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2010. Global learning of focused entailment graphs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1220–1229, Uppsala, Sweden, July. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Susan Brennan, Katharina Schuhmann, and Karla Batres. 2013. Entrainment on the move and in the lab: The walking around corpus. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Francisco Costa and António Branco. 2012. Aspectual type and temporal relation classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 266–275, Avignon, France, April. 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.

Bonnie J. Dorr and Mari Broman Olsen. 1997. Deriving verbal and compositonal lexical aspect for nlp applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 151–158, Madrid, Spain, July. Association for Computational Linguistics.

David Dowty. 1979. *Word Meaning and Montague Grammar*. Dordrecht, Holland.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67:547–619.

Ingrid Falk and Fabienne Martin. 2016. Automatic identification of aspectual classes across verbal readings. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 12–22, Berlin, Germany, August. Association for Computational Linguistics.

Hana Filip, 2012. *The Oxford Handbook of Tense and Aspect*, chapter Lexical Aspect. Oxford University Press.

W. N. Francis and H. Kucera. 1979. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Annemarie Friedrich and Damyana Gateva. 2017. Classification of telicity using cross-linguistic annotation projection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2559–2565, Copenhagen, Denmark, September. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 517–523, Baltimore, Maryland, June. Association for Computational Linguistics.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768, Berlin, Germany, August. Association for Computational Linguistics.

Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585:357–362.

Liesa Heuschkel. 2016. Automatic classification of lexical aspectual class using distributional and rule-based methods. Master's thesis, Saarland University.

Mohammad Javad Hosseini, Nathanael Chambers, Siva Reddy, Xavier R. Holt, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2018. Learning typed entailment graphs with global soft constraints. *Transactions of the Association for Computational Linguistics*, 6:703–717.

Mohammad Javad Hosseini, Shay B Cohen, Mark Johnson, and Mark Steedman. 2019. Duality of link prediction and entailment graph induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4736–4746.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. Masc: the manually annotated sub-corpus of american english. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. 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 1681–1691, Beijing, China, July. Association for Computational Linguistics.

Judith Klavans and Martin Chodorov. 1992. Degrees of stativity: The lexical representation of verb aspect. In *In Proceedings of the 14th International Conference on Compoutational Linguistics*, pages 1126–1131, Nantes, France. Association for Computational Linguistics.

Thomas Kober, Julie Weeds, John Wilkie, Jeremy Reffin, and David Weir. 2017. One representation per word - does it make sense for composition? In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 79–90, Valencia, Spain, April.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th Conference on Computational Semantics (IWCS 2019)*, Gothenburg, Sweden.

Thomas Kober. 2018. *Inferring Unobserved Co-occurrence Events in Anchored Packed Trees*. Ph.D. thesis, University of Sussex.

Sharid Loáiciga and Cristina Grisot. 2016. Predicting and using a pragmatic component of lexical aspect of simple past verbal tenses for improving english-to-french machine translation. *Linguistic Issues in Language Technology*, 13(3):1–34.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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 26*, pages 3111–3119. Curran Associates, Inc.

Ashutosh Modi and Ivan Titov. 2014. Inducing neural models of script knowledge. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 49–57, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

Marc Moens. 1987. *Tense, Aspect and Temporal Reference*. Ph.D. thesis, University of Edinburgh.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, Copenhagen, Denmark, September. Association for Computational Linguistics.

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of universal dependencies (ud). *Glossa: A Journal of General Linguistics*, 1(4):516–547.

Rebecca Passonneau. 1988. A computational model of the semantics of tense and aspect. *Computational Linguistics*, 14(2):44–60.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, November.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Eric Siegel and Kathleen McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–628, December.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Ran Tian, Naoaki Okazaki, and Kentaro Inui. 2017. The mechanism of additive composition. *Machine Learning*, 106(7):1083–1130.

Zeno Vendler. 1957. Verbs and times. *Linguistics in Philosophy*, pages 97–121.

Henk Verkuyl. 2005. How (in-)sensitive is tense to aspectual information? In Bert Hollerbrandse, Angeliek van Hout, and Co Vet, editors, *Crosslinguistic Views on Tense, Aspect and Modality*, pages 145–169. Rodopi.

Anthony J. Viera and Joanne M. Garrett. 2005. Understanding interobserver agreement: The kappa statistic. *Family medicine*, 37:360–3, 06.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *Proceedings of the International Conference on Learning Representations*.

Ronnie Wilbur. 2003. Representations of telicity in asl. In *Proceedings from the annual meeting of the Chicago Linguistic Society*, volume 39, pages 354–368. Chicago Linguistic Society.

A Supplemental Material — Performance per Verb Type in Asp-ambig

Table 7 below lists the results per verb type for all 20 ambiguous verbs in the **Asp-ambig** dataset, comparing the majority class baseline, the models of Friedrich and Palmer (2014) and Heuschkel (2016) to our window-1 and dependency (full) approaches.

Verb	Majority class	Friedrich and Palmer (2014)	Heuschkel (2016)	window-1	dependency (full)
feel	**96.1**	93.8	95.5	94.2	93.5
say	**94.9**	93.5	94.2	**94.9**	**94.9**
make	91.9	91.2	**92.0**	90.6	90.6
come	**88.0**	87.2	**88.0**	85.5	86.2
take	85.4	85.4	85.5	85.5	**87.0**
meet	83.9	87.7	85.9	87.7	**90.6**
stand	80.0	83.1	81.8	**87.7**	82.6
find	74.5	68.8	**76.6**	75.4	69.6
accept	**70.9**	65.7	67.4	68.8	61.6
hold	56.0	49.3	57.8	**62.3**	55.8
carry	55.9	58.1	60.3	58.0	**63.8**
look	55.8	74.6	65.9	**79.7**	61.6
show	54.9	**68.4**	65.9	67.4	65.2
appear	52.2	61.0	56.6	**70.3**	62.3
follow	51.6	65.6	61.8	61.6	**69.6**
consider	50.7	70.3	67.4	**77.5**	59.4
cover	50.4	54.5	55.0	**57.2**	50.0
fill	47.8	62.7	**69.4**	60.1	48.6
bear	47.4	67.4	**77.2**	68.1	63.8
allow	37.8	**51.9**	51.1	50.7	44.9
Average	66.3	72.0	72.8	**74.2**	70.1

Table 7: Per verb Accuracies on the **Asp-ambig** dataset (Friedrich and Palmer, 2014).

For strongly imbalanced classes as in the case of *feel*, which almost always functions as a *state*, the majority baseline is very difficult to beat. Interestingly, the window-1 and dependency (full) approaches frequently exhibit complementary performance. For example, while for *stand* or *look* a window-based context works substantially better, for *follow* or *carry* a dependency-based context is preferable. One explanation for this behaviour is that for *stand* or *look* prepositions are frequently the most salient indicator of aspectual class as shown in Section 5. On the other hand, for *follow* or *carry* a content word, such as the subject or direct object, is frequently more salient.

B   Supplemental Material — PoS Tag Distribution of Extracted Contexts

Figure 7 shows the PoS tag distribution of extracted contexts of the linear context window in comparison to dependency contexts. Dependency contexts, based on Universal Dependencies, overwhelmingly extract content words, whereas the linear context window predominantly tends to extract more closed class words.
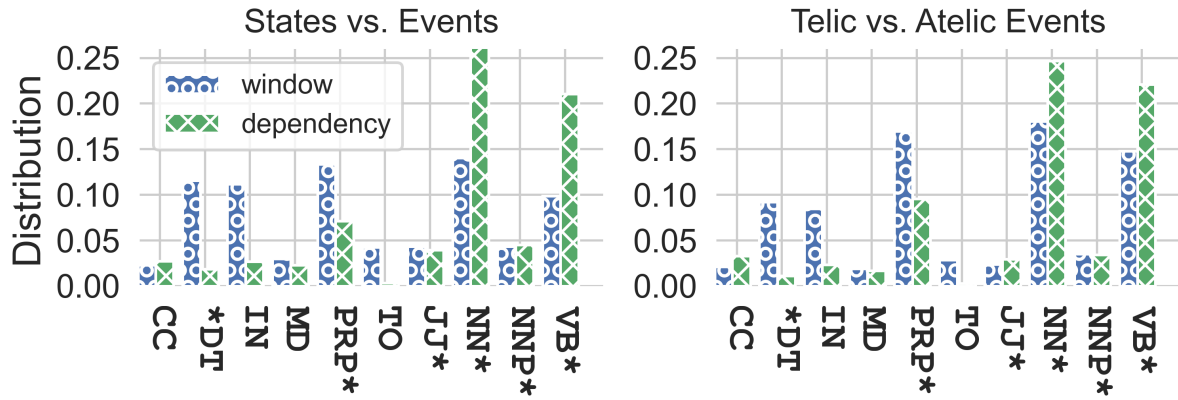

Figure 7: PoS tag distribution of the extracted contexts of the linear context window vs. dependency contexts.

C   Supplemental Material — Window Contexts vs. Dependency Contexts

Tables 8 & 9 present the performance of the best performing linear context window in comparison to the best performing dependency context window, as well as the verb-only and full-sentence and majority class baselines.

Dataset	Best Linear	Best Dependency	Verb only	Full Sentence	Majority Class
Asp-ambig Accuracy	**74.2**	70.1	65.9	60.0	65.9
SitEnt F1 (State)	81.2	81.3	**84.0**	26.4	0.0
SitEnt F1 (Event)	84.1	84.5	**86.6**	71.9	68.9
SitEnt-ambig F1 (State)	**62.6**	60.1	44.0	0.0	0.0
SitEnt-ambig F1 (Event)	66.2	65.3	62.4	68.3	**68.4**
Captions F1 (State)	**58.8**	55.9	0.1	23.4	0.0
Captions F1 (Event)	**89.7**	89.2	87.3	86.7	87.6
DIASPORA F1 (State)	**86.5**	85.9	76.4	80.1	0.0
DIASPORA F1 (Event)	89.6	**89.8**	83.5	84.8	72.5

Table 8: Comparison between linear contexts windows and dependency contexts on classifying *states* vs. *events*.

Overall linear context window perform slightly better on average than dependency context windows — and as highlighted in Section 5 — this can be explained by linear context window extracting more closed class context words (see Figure 7 in Appendix B), which tend to be stronger disambiguation signals than content words.

Dataset	Best Linear	Best Dependency	Verb only	Full Sentence	Majority Class
Telicity F1 (Telic)	91.6	90.7	**92.1**	91.1	90.1
Telicity F1 (Atelic)	60.3	57.8	**62.6**	60.6	0.0
Captions F1 (Telic)	71.6	**72.3**	70.7	71.2	55.3
Captions F1 (Atelic)	82.6	83.0	**83.3**	83.1	0.0
DIASPORA F1 (Telic)	**70.3**	69.6	64.5	66.5	69.2
DIASPORA F1 (Atelic)	**67.8**	65.4	63.3	62.9	0.0

Table 9: Comparison between linear contexts windows and dependency contexts on classifying *telic* vs. *atelic* events.

D Supplemental Material — Zero Shot Generalization

For assessing the generalisation capabilities of our methodology we are performing a zero-shot setup on the **Asp-ambig** dataset. Instead of running an evaluation for each verb individually as originally proposed by Friedrich and Palmer (2014), we are evaluting the model on the data for one particular verb, say *look*, and train the model on all available data, **except** the data for the heldout verb *look*. This way, we investigate whether distributional representations truly capture the underlying semantics of predicational aspect.

We use the same simple setup as in Section 5, with a linear regression classifier that operates on the basis of averaged word2vec embeddings. We used a linear context window of size 1 for this experiments as this was the best performing setup for the **Asp-ambig** dataset in the evaluation in Section 5. Table 10

Verb	Majority class	Zero Shot
feel	**96.1**	43.0
say	**94.9**	55.8
make	**91.9**	86.2
come	**88.0**	78.1
take	**85.4**	82.6
meet	**83.9**	76.8
stand	**80.0**	22.5
find	**74.5**	69.6
accept	**70.9**	68.1
hold	**56.0**	26.3
carry	**55.9**	54.0
look	55.8	**77.5**
show	**54.9**	52.2
appear	**52.2**	51.4
follow	**51.6**	39.9
consider	**50.7**	**50.7**
cover	**50.4**	41.6
fill	47.8	**48.2**
bear	**47.4**	47.1
allow	37.8	**39.4**
Average	**66.3**	55.5

Table 10: Per verb Accuracies on the **Asp-ambig** dataset (Friedrich and Palmer, 2014).

shows the results of the zero-shot experiment in comparison to the majority class baseline. While for the majority of verbs, our model underperforms the majority class baseline — which is difficult to beat especially for the very skewed verbs such as *feel* or *say*, our approach beats the baseline for 3 verbs and achieves comparable performance for more than half of the verbs, while not having encountered any annotated data for the target verb during training at all.

Given the simplicity of our setup, we regard that as strong evidence that a model based on distributional semantics does indeed capture a substantial amount of predicational aspect in its representations.