Automatic prediction of aspectual class of verbs in context

Annemarie Friedrich and Alexis Palmer
Department of Computational Linguistics
Saarland University, Saarbrücken, Germany
`{afried,apalmer}@coli.uni-saarland.de`

Abstract

This paper describes a new approach to predicting the aspectual class of verbs in context, i.e., whether a verb is used in a stative or dynamic sense. We identify two challenging cases of this problem: when the verb is unseen in training data, and when the verb is ambiguous for aspectual class. A semi-supervised approach using linguistically-motivated features and a novel set of distributional features based on representative verb types allows us to predict classes accurately, even for unseen verbs. Many frequent verbs can be either stative or dynamic in different contexts, which has not been modeled by previous work; we use contextual features to resolve this ambiguity. In addition, we introduce two new datasets of clauses marked for aspectual class. *(updated March 2015)*

1 Introduction

In this work, we focus on the automatic prediction of whether a verb in context is used in a *stative* or in a *dynamic* sense, the most fundamental distinction in all taxonomies of *aspectual class*. The aspectual class of a discourse's finite verbs is an important factor in conveying and interpreting temporal structure (Moens and Steedman, 1988; Dorr, 1992; Klavans and Chodorow, 1992); others are tense, grammatical aspect, mood and whether the utterance represents an event as completed. More accurate temporal information processing is expected to be beneficial for a variety of natural language processing tasks (Costa and Branco, 2012; UzZaman et al., 2013).

While most verbs have one predominant interpretation, others are more flexible for aspectual class and can occur as either stative (1) or dynamic (2) depending on the context. There are also cases that allow for both readings, such as (3).

*(1) The liquid **fills** the container. (stative)*

*(2) The pool slowly **filled** with water. (dynamic)*

*(3) Your soul was made to be **filled** with God Himself. (both)* `(Brown corpus, religion)`

Cases like (3) do not imply that there is a third class, but rather that two interpretations are available for the sentence, of which usually one will be chosen by a reader.

Following Siegel and McKeown (2000), we aim to automatically classify clauses for *fundamental aspectual class*, a function of the main verb and a select group of complements, which may differ per verb (Siegel and McKeown, 2000; Siegel, 1998b). This corresponds to the aspectual class of the clause's main verb when ignoring any aspectual markers or transformations. For example, English sentences with perfect tense are usually considered to introduce states to the discourse (Smith, 1991; Katz, 2003), but we are interested in the aspectual class before this transformation takes place. The clause *John has kissed Mary* introduces a state, but the fundamental aspectual class of the 'tenseless' clause *John kiss Mary* is dynamic.

In contrast to Siegel and McKeown (2000), we do not conduct the task of predicting aspectual class solely at the type level, as such an approach ignores the minority class of ambiguous verbs. Instead we predict the aspectual class of verbs in the context of their arguments and modifiers. We show that this method works better than using only type-based features, especially for verbs with ambiguous aspectual class. In addition, we show that type-based features, including novel distributional features based on representative verbs, accurately predict predominant aspectual class for unseen verb types. Our work differs from prior work in that we treat the problem as a three-way classification task, predicting DYNAMIC, STATIVE or BOTH as the aspectual class of a verb in context.

2 Related work

Aspectual class is well treated in the linguistic literature (Vendler, 1957; Dowty, 1979; Smith, 1991, for example). Our notion of the stative/dynamic distinction corresponds to Bach's (1986) distinction between states and non-states; to states versus occurrences (events and processes) according to Mourelatos (1978); and to Vendler's (1957) distinction between states and the other three classes (activities, achievements, accomplishments).

Early studies on the computational modeling of aspectual class (Nakhimovsky, 1988; Passonneau, 1988; Brent, 1991; Klavans and Chodorow, 1992) laid foundations for a cluster of papers published over a decade ago (Siegel and McKeown, 2000; Siegel, 1998b; Siegel, 1998a). Since then, it has mostly been treated as a subtask within temporal reasoning, such as in efforts related to TimeBank (Pustejovsky et al., 2003) and the TempEval challenges (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013), where top-performing systems (Jung and Stent, 2013; Bethard, 2013; Chambers, 2013) use corpus-based features, WordNet synsets, parse paths and features from typed dependencies to classify events as a joint task with determining the event's span. Costa and Branco (2012) explore the usefulness of a wider range of explicitly aspectual features for temporal relation classification.

Siegel and McKeown (2000) present the most extensive study of predicting aspectual class, which is the main inspiration for this work. While all of their linguistically motivated features (see section 4.1) are type-based, they train on and evaluate over labeled verbs in context. Their data set taken from medical discharge summaries comprises 1500 clauses containing main verbs other than *be* and *have* which are marked for aspectual class. Their model fails to outperform a baseline of memorizing the most frequent class of a verb type, and they present an experiment testing on unseen verb types only for the related task of classifying completedness of events. We replicate their method using publicly available software, create a similar but larger corpus,[1] and show that it is indeed possible to predict the aspectual class of unseen verbs. Siegel (1998a) investigates a classification method for the verb *have* in context; in-

| | COMPLETE | | W/O *have/be/none* | |
genre	clauses	κ	clauses	κ
jokes	3462	0.85	2660	0.77
letters	1848	0.71	1444	0.62
news	2565	0.79	2075	0.69
all	7875	0.80	6161	0.70

Table 1: **Asp-MASC**: Cohen's observed unweighted κ.

	DYNAMIC	STATIVE	BOTH
DYNAMIC	4464	164	9
STATIVE	434	1056	29
BOTH	5	0	0

Table 2: **Asp-MASC**: confusion matrix for annotators, without *have/be/none* clauses, $\kappa = 0.70$.

spired by this work, our present work goes one step further and uses a larger set of *instance-based* contextual features to perform experiments on a set of 20 verbs. To the best of our knowledge, there is no previous work comprehensively addressing aspectual classification of verbs in context.[2]

3 Data

3.1 Verb type seed sets

Using the **LCS Database** (Dorr, 2001), we identify sets of verb types whose senses are only stative (188 verbs, e.g. *belong, cost, possess*), only dynamic (3760 verbs, e.g. *alter, knock, resign*), or mixed (215 verbs, e.g. *fill, stand, take*), following a procedure described by Dorr and Olsen (1997).

3.2 Asp-MASC

The Asp-MASC corpus[3] consists of 7875 clauses from the letters, news and jokes sections of MASC (Ide et al., 2010), each labeled by two annotators for the aspectual class of the main verb. Texts were segmented into clauses using SPADE (Soricut and Marcu, 2003) with some heuristic post-processing. We parse the corpus using the Stanford dependency parser (De Marneffe et al., 2006) and extract the main verb of each segment. We use 6161 clauses for the classification task, omitting clauses with *have* or *be* as the main verb and those where no main verb could be identified due to parsing errors (*none*). Table 1 shows

---

[1] Direct comparison on their data is not possible; feature values for the verbs studied are available, but full texts and the English Slot Grammar parser (McCord, 1990) are not.

[2] Zarcone and Lenci (2008) classify 3129 occurrences of 28 Italian verbs according to Vendler's classes *state, process, accomplishment* and *achievement*. Not citing this work in the original version of the paper was an unfortunate oversight on our side. (March 2015)

[3] `www.coli.uni-saarland.de/~afried`.

	DYNAMIC	STATIVE	**both**
DYNAMIC	1444	201	54
STATIVE	168	697	20
BOTH	44	31	8

Table 3: **Asp-Ambig**: confusion matrix for two annotators. Cohen's κ is 0.6.

inter-annotator agreement; Table 2 shows the confusion matrix for the two annotators. Our two annotators exhibit different preferences on the 598 cases where they disagree between DYNAMIC and STATIVE. Such differences in annotation preferences are not uncommon (Beigman Klebanov et al., 2008). We observe higher agreement in the jokes and news subcorpora than for letters; texts in the letters subcorpora are largely argumentative and thus have a different rhetorical style than the more straightforward narratives and reports found in jokes. Overall, we find substantial agreement.

The data for our experiments uses the label DYNAMIC or STATIVE whenever annotators agree, and BOTH whenever they disagree or when at least one annotator marked the clause as BOTH, assuming that both readings are possible in such cases. Because we don't want to model the authors' personal view of the theory, we refrain from applying an adjudication step and model the data as is.

3.3 Asp-Ambig: (Brown)

In order to facilitate a first study on ambiguous verbs, we select 20 frequent verbs from the list of 'mixed' verbs (see section 3.1) and for each mark 138 sentences. Sentences are extracted randomly from the Brown corpus, such that the distribution of stative/dynamic usages is expected to be natural. We present entire sentences to the annotators who mark the aspectual class of the verb in question as highlighted in the sentence. The data is processed in the same way as Asp-MASC, discarding instances with parsing problems. This results in 2667 instances. κ is 0.6, the confusion matrix is shown in Table 3. Details are listed in Table 10.

4 Model and Features

For the three-way classification task of predicting the aspectual class of verbs in context (STATIVE, DYNAMIC, BOTH), we assume a supervised learning setting and explore features mined from a large background corpus, distributional features, and instance-based features. If not indicated otherwise, experiments use a Random Forest classifier

FEATURE	EXAMPLE	FEATURE	EXAMPLE
frequency	-	continuous adverb	*continually endlessly*
present	*says*	evaluation adverb	*better horribly*
past	*said*		
future	*will say*		
perfect	*had won*	manner adverb	*furiously patiently*
progressive	*is winning*		
negated	*not/never*	temporal adverb	*again finally*
particle	*up/in/...*		
no subject	-	in-PP	*in an hour*
		for-PP	*for an hour*

Table 4: **LingInd** feature set and examples for lexical items associated with each indicator.

FEATURE	VALUES
part-of-speech tag of the verb	VB, VBG, VBN, ...
tense	present, past, future
progressive	true/false
perfect	true/false
voice	active/passive
grammatical dependents	WordNet lexname/POS

Table 5: Instance-based (**Inst**) features

(Breiman, 2001) trained with the implementation and standard parameter settings from Weka (Hall et al., 2009).

4.1 Linguistic indicator features (LingInd)

This set of corpus-based features is a reimplementation of the linguistic indicators from (Siegel and McKeown, 2000), who show that (some of) these features correlate with either stative or dynamic verb types.[4] We parse the AFE and XIE sections of Gigaword (Graff and Cieri, 2003) with the Stanford dependency parser. For each verb type, we obtain a normalized count showing how often it occurs with each of the indicators in Table 4, resulting in one value per feature per verb. For example, for the verb *fill*, the value of the feature `temporal-adverb` is 0.0085, meaning that 0.85% of the occurrences of *fill* in the corpus are modified by one of the temporal adverbs on the list compiled by Siegel (1998b). Tense, progressive, perfect and voice are extracted using a set of rules following Loaiciga et al. (2014).

4.2 Distributional Features (Dist)

We aim to leverage existing, possibly noisy sets of representative stative, dynamic or mixed verb types extracted from LCS (see section 3.1), making up for unseen verbs and noise by averaging over distributional similarities. Using an exist-

---

[4]We refer the reader to that paper for detailed analysis, description, and motivation.

FEATURES	ACCURACY (%)
Baseline (Lemma)	83.6
LingInd	83.8
Inst	70.8
Inst+Lemma	83.7
Dist	83.4
LingInd+Inst+Dist+Lemma	**84.1**

Table 6: **Experiment 1**: SEEN verbs, following Siegel, using **Asp-MASC**. Baseline: memorizes most frequent class per verb type in training folds.

	FEATURES	ACCURACY (%)
1	Baseline	72.5
2	Dist	78.3*
3	LingInd	80.4*
4	LingInd+Dist	**81.9*†**

Table 7: **Experiment 2**: UNSEEN verb types, Logistic regression, **Asp-MASC**. Baseline labels everything with the most frequent class in the training set (DYNAMIC). *Significantly[5] different from line 1. †Significantly[5] different from line 2.

ing large distributional model (Thater et al., 2011) estimated over the set of Gigaword documents marked as stories, for each verb type, we build a syntactically informed vector representing the contexts in which the verb occurs. We compute three numeric feature values per verb type, which correspond to the average cosine similarities with the verb types in each of the three seed sets.

4.3 Instance-based features (Inst)

Table 5 shows our set of instance-based syntactic and semantic features. In contrast to the above described type-based features, these features do not rely on a background corpus, but are extracted from the clause being classified. Voice, tense and grammatical aspect are extracted from dependency parses as described above. For feature encoding grammatical dependents, we focus on a subset of grammatical relations: the feature value is either the WordNet lexical filename (e.g. *noun.person*) of the given relation's argument or its POS tag, if the former is not available. We simply use the most frequent sense for the dependent's lemma. We also include features that indicate, if there are any, the particle of the verb and its prepositional dependents. For the sentence *A little girl had just finished her first week of school*, the instance-based feature values would include `tense`:*past*, `subj`:*noun.person*, `dobj`:*noun.time* or `particle`:*none*.

DATA	FEATURES	ACC. (%)
one-label verbs (1966 inst.)	Baseline	**92.8**
	LingInd	92.8
	Dist	92.6
	Inst+Lemma	91.4*
	LingInd+Inst+Lemma	92.4
multi-label verbs (4195 inst.)	Baseline	78.9
	LingInd	79.0
	Dist	79.0
	Inst	67.4*
	Inst+Lemma	79.9
	LingInd+Inst+Lemma	**80.9***
	LingInd+Inst+Lemma+Dist	80.2*

Table 8: **Experiment 3**: 'ONE-LABEL VS. MULTI-LABEL' verbs, **Asp-MASC**. Baseline as in Table 6. *Indicates that result is significantly[5] different from the respective baseline.

5 Experiments

The experiments presented in this section aim to evaluate the effectiveness of the feature sets described in the previous section, focusing on the challenging cases of verb types unseen in the training data and highly ambiguous verbs. The feature **Lemma** indicates that the verb's lemma is used as an additional feature.

Experiment 1: SEEN verbs

The setting of our first experiment follows Siegel and McKeown (2000). Table 6 reports results for 10-fold cross-validation, with occurrences of all verbs distributed evenly over the folds. No feature combination significantly[5] outperforms the baseline of simply memorizing the most frequent class of a verb type in the respective training folds.

Experiment 2: UNSEEN verbs

This experiment shows a successful case of semi-supervised learning: while type-based feature values can be estimated from large corpora in an unsupervised way, some labeled training data is necessary to learn their best combination. This experiment specifically examines performance on verbs not seen in labeled training data. We use 10-fold cross validation but ensure that all occurrences of a verb type appear in the same fold: verb types in each test fold have *not* been seen in the respective training data, ruling out the Lemma feature. A Logistic regression classifier works better here (using only numeric features), and we present re-

---

[5]According to McNemar's test with Yates' correction for continuity, $p < 0.01$.

CLASS	ACC.(%)	P	R	F
Baseline (Lemma)				
micro-avg.	78.9	0.75	0.79	0.76
LingInd+Inst+Lemma				
DYNAMIC		0.84	0.95	0.89
STATIVE		0.76	0.69	0.72
BOTH		0.51	0.24	0.33
micro-avg.	**80.9***	0.78	0.81	**0.79**

Table 9: **Experiment 3**: 'MULTI-LABEL', precision, recall and F-measure, detailed class statistics for the best-performing system from Table 8.

VERB	# OF INST.	MAJORITY CLASS[6]		Inst +Lemma	Inst +Lemma +LingInd +Dist
feel	128	**96.1**	STAT	93.0	93.8
say	138	**94.9**	DYN	93.5	93.5
make	136	**91.9**	DYN	91.9	91.2
come	133	**88.0**	DYN	87.2	87.2
take	137	**85.4**	DYN	85.4	85.4
meet	130	83.9	DYN	86.2	**87.7**
stand	130	80.0	STAT	79.2	**83.1**
find	137	**74.5**	DYN	69.3	68.8
accept	134	**70.9**	DYN	64.9	65.7
hold	134	**56.0**	BOTH	43.3	49.3
carry	136	55.9	DYN	55.9	**58.1**
look	138	55.8	DYN	72.5	**74.6**
show	133	54.9	DYN	**69.2**	68.4
appear	136	52.2	STAT	**64.7**	61.0
follow	122	51.6	BOTH	**69.7**	65.6
consider	138	50.7	DYN	61.6	**70.3**
cover	123	50.4	STAT	46.3	**54.5**
fill	134	47.8	DYN	**66.4**	62.7
bear	135	47.4	DYN	**70.4**	67.4
allow	135	37.8	DYN	48.9	**51.9**
micro-avg.	2667	66.3		**71.0***	**72.0***

Table 10: **Experiment 4**: INSTANCE-BASED. **Accuracy** (in %) on **Asp-Ambig**. *Differs significantly[5] from the majority class baseline.

sults in Table 7. Both the LingInd and Dist features generalize across verb types, and their combination works best.

Experiment 3: ONE-LABEL VS. MULTI-LABEL verbs

For this experiment, we compute results separately for one-label verbs (those for which all instances in Asp-MASC have the same label) and for multi-label verbs (instances have differing labels in Asp-MASC). Otherwise, the experimental setup is as in experiment 1. Results appear in Table 8. For one-label verbs, which we can expect are likely to have a strong predominant aspectual class, the linguistic indicator features work well, though they again fail to beat the baseline. For multi-label verbs, the feature combination Lemma+LingInd+Inst leads to significant[5] improvement of 2% gain in accuracy over the baseline. Table 9 reports detailed class statistics for this feature combination and reveals a gain in F-measure of 3 points over the baseline. To sum up, Inst features are essential for classifying multi-label verbs, and the LingInd features provide some useful prior. These results motivate the need for an instance-based approach.

Experiment 4: INSTANCE-BASED classification

For verbs with ambiguous aspectual class, type-based classification is not sufficient, as this approach selects a dominant sense for any given verb and then always assigns that. Therefore we propose handling ambiguous verbs separately. As Asp-MASC contains only few instances of each of the ambiguous verbs, we turn to the Asp-Ambig dataset (see section 3.3). We perform a Leave-One-Out (LOO) cross validation evaluation, with results reported in Table 10.[6] Using the Inst fea-

---

[6] The third column also shows the outcome of using either only the Lemma, only LingInd or only Dist in LOO; all have almost the same outcome as using the majority class,

tures alone (not shown in Table 10) results in a micro-average accuracy of only 58.1%: these features are only useful when combined with the feature Lemma. For classifying verbs whose most frequent class occurs less than 56% of the time, Lemma+Inst features are essential. Whether or not performance is improved by adding LingInd/Dist features, with their bias towards one aspectual class, depends on the verb type. It is an open research question which verb types should be treated in which way.

6 Discussion and conclusions

We have described a new, context-aware approach to automatically predicting aspectual class including a new set of distributional features, at the same time introducing two new data sets of clauses labeled for aspectual class. Our experiments show that in any setting where labeled training data is available, improvement over the most frequent class baseline can only be reached by integrating instance-based features, though type-based features (LingInd, Dist) may provide useful priors for some verbs and successfully predict predominant aspectual class for unseen verb types. In or-

---

numbers differ only after the decimal point.

der to arrive at a globally well-performing system, we envision a multi-stage approach, treating verbs differently according to whether training data is available and whether or not the verb's aspectual class distribution is highly skewed.

Acknowledgments

References

Emmon Bach. 1986. The algebra of events. *Linguistics and philosophy*, 9(1):5–16.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 2–7. Association for Computational Linguistics.

Steven Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 10–14.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Michael R. Brent. 1991. Automatic semantic classification of verbs from their syntactic contexts: an implemented classifier for stativity. In *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*, pages 222–226. Association for Computational Linguistics.

Nathanael Chambers. 2013. Navytime: Event and time ordering from raw text. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 73–77.

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. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Bonnie J. Dorr and Mari Broman Olsen. 1997. Deriving verbal and compositional lexical aspect for NLP applications. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 151–158. Association for Computational Linguistics.

Bonnie J. Dorr. 1992. A two-level knowledge representation for machine translation: lexical semantics and tense/aspect. In *Lexical Semantics and Knowledge Representation*, pages 269–287. Springer.

Bonnie J. Dorr. 2001. LCS verb database. Online software database of Lexical Conceptual Structures, University of Maryland, College Park, MD.

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

David Graff and Christopher Cieri. 2003. English gigaword.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: a community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73. Association for Computational Linguistics.

Hyuckchul Jung and Amanda Stent. 2013. ATT1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 20–24.

Graham Katz. 2003. On the stativity of the english perfect. *Perfect explorations*, pages 205–234.

Judith L. Klavans and Martin Chodorow. 1992. Degrees of stativity: the lexical representation of verb aspect. In *Proceedings of the 14th conference on Computational Linguistics*, pages 1126–1131. Association for Computational Linguistics.

Sharid Loaiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. In *Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*.

Michael C. McCord. 1990. *Slot Grammar*. Springer.

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

Alexander P.D. Mourelatos. 1978. Events, processes, and states. *Linguistics and philosophy*, 2(3):415–434.

Alexander Nakhimovsky. 1988. Aspect, aspectual class, and the temporal structure of narrative. *Computational Linguistics*, 14(2):29–43.

Rebecca Passonneau. 1988. A computational model of the semantics of tense and aspect. *Computational Linguistics*, Spring 1988.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

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

Eric V. Siegel. 1998a. Disambiguating verbs with the WordNet category of the direct object. In *Proceedings of Workshop on Usage of WordNet in Natural Language Processing Systems*, Universite de Montreal.

Eric V. Siegel. 1998b. *Linguistic Indicators for Language Understanding: Using machine learning methods to combine corpus-based indicators for aspectual classification of clauses*. Ph.D. thesis, Columbia University.

Carlota S. Smith. 1991. *The Parameter of Aspect*. Kluwer, Dordrecht.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 149–156. Association for Computational Linguistics.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *IJCNLP*, pages 1134–1143.

Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second joint conference on lexical and computational semantics (* SEM)*, volume 2, pages 1–9.

Zeno Vendler, 1957. *Linguistics in Philosophy*, chapter Verbs and Times, pages 97–121. Cornell University Press, Ithaca, New York.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. Association for Computational Linguistics.

Alessandra Zarcone and Alessandro Lenci. 2008. Computational models for event type classification in context. In *LREC*.