Event and Temporal Relation Extraction from Croatian Newspaper Texts

Mladen Marović, Jan Šnajder, Goran Glavaš

University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
mladen.marovic@gmail.com, {jan.snajder, goran.glavas}@fer.hr

Abstract

Event extraction and temporal relation extraction are the subjects of extensive research, which has been additionally motivated by focused evaluation exercises such as TempEval. In this paper we present the work on supervised event and temporal relation extraction from Croatian newspaper texts. Taking into account the limited availability of linguistic tools for Croatian, we focus our research around simple lexical features. We manually annotated a newspaper corpus or events and temporal relations in Croatian according to the TimeML and TimeBank guidelines. Experimental evaluation yielded promising results: F1 scores of up to 77% for event identification, 48% for event classification, and 51% for temporal relation classification.

Luščenje dogodkov in časovnih relacij iz hrvaških časopisnih besedil

Luščenje dogodkov in časovnih relacij je zelo živahno raziskovalno področje, ki se je še posebej razmahnilo s pojavom skupinskih evalvacijskih pobud, kot je na primer TempEval. V pričujočem prispevku predstavljamo sistem za nadzorovano luščenje dogodkov in časovnih relacij iz hrvaških časopisnih besedil. Glede na to, da je dostopnost jezikovnih orodij za hrvaščino omejena, se v raziskavi osredotočamo zgolj na enostavne leksikalne lastnosti. Pri ročnem označevanju dogodkov in časovnih relacij v korpusu časopisnih besedil smo uporabljali smernice TimeML in TimeBank. Eksperimentalno vrednotenje rezultatov je zelo spodbudno, saj za prepoznavanje dogodkov F1 znaša 77%, za klasifikacijo dogodkov 48% in za klasifikacijo časovnih relacij 51%.

1. Introduction

Event extraction and temporal relation extraction are non-trivial information extraction (IE) tasks that often play an important role in various practical natural language processing (NLP) applications, such as question answering (Saurí et al., 2005) and document summarization (Lee et al., 2003). Event and temporal relation extraction tasks have attracted a lot of attention, in particular within the TempEval evaluation exercises. Event extraction task refers to the classification of words into events and non-events (event identification), or, more commonly, to identifying the events and determining their types (event classification). Temporal relation extraction refers to classification of temporal relations between extracted pairs of events.

In this paper we present the work on supervised event and temporal relation extraction from Croatian newspaper texts. As part of this work, we manually annotated a newspaper corpus following the guidelines from TimeML (Pustejovsky et al., 2003a) and TimeBank (Pustejovsky et al., 2003b). We then evaluated the event and temporal relation extraction performance of several classifiers. Because of the limited availability of linguistic tools for Croatian language, we use mostly lexical features. We achieved promising results, showing that using only simple features can yield satisfactory results for event and temporal relation extraction tasks for Croatian language. To the best of our knowledge, the work described in this paper is the first work on event and temporal relation extraction for the Croatian language, and Slavic languages in general.

The rest of the paper is structured as follows. The next section gives an overview of related research. In Section 3 we describe the corpus annotation. Supervised machine learning for event and temporal relation extraction is described in Section 4. Section 5 describes the experimental setting and discusses the results. Section 6 concludes the paper and suggests future work.

2. Related Work

2.1. Event extraction

Events in sentences were first studied in linguistic literature (Vendler, 1957; Verkuyl, 2005; Pustejovsky, 1991). Different event properties were defined based on event structure, duration, and telicity. Vendler (1957) proposed the distinction between four types of events (states, activities, accomplishments, and achievements), while Pustejovsky (1991) proposed a structural event hierarchy. Work focusing on practical IE and NLP tasks combines statistical and machine learning methods to automatically determine various properties of verbs and events occurring in text. Siegel and McKeown (2000) used machine learning methods based on features such as verb tense, presence of a negation word or absence of a subject to determine the aspectual properties of verbs. Classification of verbs into states and events resulted in an accuracy of 93.9%, while the classification of events into aspectual categories resulted in an accuracy of 74.0%.

To aid further research on event and temporal relation extraction, Pustejovsky et al. (2003a) developed TimeML, a rich specification language for event and temporal expressions in natural language text. In TimeML, events are defined as "situations that *happen* or *occur*" and can either be punctual or last for a period of time. The events are divided into eight classes: OCCURRENCE, PERCEPTION, REPORTING, ASPECTUAL, STATE, I_STATE, I_ACTION, and MODAL. Based on TimeML, Pustejovsky et al. (2003b) created the TimeBank corpus consisting of 300 documents (60k words) manually annotated for events. Since the introduction of TimeML and TimeBank, a number of event extraction methods have been proposed (Saurí et al., 2005; Boguraev and Ando, 2005; Bethard and Martin, 2006). Boguraev and Ando (2005) additionally performed a quantitative analysis of the TimeBank corpus and criticized its small size and the unbalanced distribution of event classes, which they tried to compensate for with a word profiling technique.

2.2. Temporal relation extraction

The problem of defining events in the context of time was addressed early in linguistic literature (C. Bruce, 1972; Reichenbach, 1980). Allen (1983) introduced interval temporal algebra, which was widely accepted and further improved on by Galton (1990). In interval algebra, temporal relations between two events are defined as relations between the corresponding beginning points and end points.

The study of temporal relations is closely related to event extraction research. In TimeML, Pustejovsky et al. (2003a) introduced eight labels based on Allen's interval algebra for labeling relations between events, as well as relations between events and temporal expressions: *before*, *immediately before*, *includes*, *holds*, *simultaneous*, *identity*, *begins*, and *ends*. These labels were also used to annotate the TimeBank corpus, although with low inter-annotator agreement. Mani et al. (2006) expanded the TimeBank relations using a temporal closure algorithm (Verhagen, 2005), and used machine learning methods to classify the temporal relations. Lapata and Lascarides (2004) chose a somewhat different approach: they selected sentences containing temporal connectives, such as *during*, *when*, *while*, etc., and used a simple probabilistic model to insert the appropriate connectives in place of the removed ones. In their subsequent work, Lapata and Lascarides (2006) determined the mapping between the connectives and the temporal relations defined in TimeML and used this mapping for relation classification.

To encourage further research in temporal text processing, the TempEval (Verhagen et al., 2007) and TempEval-2 (Verhagen et al., 2010) evaluation exercises were organized in 2007 and 2010, respectively. In order to achieve a satisfactory inter-annotator agreement, in TempEval-2 the following reduced subset of relations was used: BEFORE, AFTER, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER, and VAGUE. Two TempEval-2 tasks are relevant to the work described in this paper: (1) determining the temporal relation between two main events in consecutive sentences, and (2) determining the temporal relation between two events where one event syntactically dominates the other. Diverse systems participated in the competition, achieving F1 scores of up to 58% for the first task and 66% for the second task.

3. Corpus Annotation

The data we used for training and testing of models consists of 230 newspaper articles from the daily Croatian newspaper *Vjesnik* from years 1999–2009. We chose the articles based on their length, type, and topic. The average article length was around 500 tokens (including words and punctuation marks). The chosen topics were daily news from Croatia and the world, sports, politics, and culture. We did not consider articles such as columns, reviews, and other types of opinionated text. The resulting corpus consists of 118,900 tokens (102,830 words), with 26,095 word form types and 10,963 lemma types.

3.1. Events and event classes

In our research an event is a "cover term for situations that *happen* or *occur*," as defined by Pustejovsky et al. (2003a). Because our research focuses on newspaper texts and articles that report about specific instances of events, we introduced three modifications to the TimeML guidelines. First, we consider only *realis* events, i.e., events that are asserted to have already happened, are happening, or will happen (Polanyi and Zaenen, 2006). Events modified by a modal verb are not realis, so we disregard such events. Secondly, we do not consider generic events – events that describe classes of events instead of specific events and are not anchored in time. For instance, the word *playing* in the sentence *Playing with knives is dangerous* is a generic event. Contrary to that, words representing sets of events of the same type that are anchored in time are considered events. For example, the word *matches* in the sentence *The semifinal matches of the Euro 2012* is an event because the matches are all unique events, anchored in time, specific, and well-defined in the context of the Euro 2012 semifinals. Thirdly, in the newspaper domain, we do not consider states as relevant events. For example, the word *knows* in the sentence *He knows a secret* is not an event. However, we do consider changes of states as events. For example, the word *discovered* in the sentence *He discovered a secret* is an event because it refers to a state change, i.e., a transition from the state of not knowing to the state of knowing. Events are considered on a single word basis; each word is classified separately as either event or non-event. We use the following seven event classes:

1. REPORTING – events of narrative character (*reći – say, objaviti – declare*);
2. ASPECTUAL – events that describe the aspect (*početi – begin, završiti – finish*);
3. PERCEPTION – events that describe physical perception (*vidjeti – see, čuti – hear*);
4. I_ACTION – events that introduce another event (*istražiti umorstvo – investigate a murder, pokušati pobjeći – attempt to escape*);
5. OCCURRENCE – single events that happen or occur (*umorstvo – murder, pobjeći – escape*);
6. HALF_GENERIC – words denoting sets of unique events defined in a context (*The semifinal matches of the Euro 2012*);
7. STATE_CHANGE – events that describe the transition from one state to another (*otkriti – discover, prihvatiti – accept*).

Table 1: Event annotation summary

Event Class	Frequency	IAA
OCCURRENCE	6867	0.6537
REPORTING	1303	0.8207
I_ACTION	1124	0.3341
HALF_GENERIC	642	0.2080
STATE_CHANGE	348	0.2349
ASPECTUAL	301	0.4272
PERCEPTION	58	0.3383
Total	10,643	

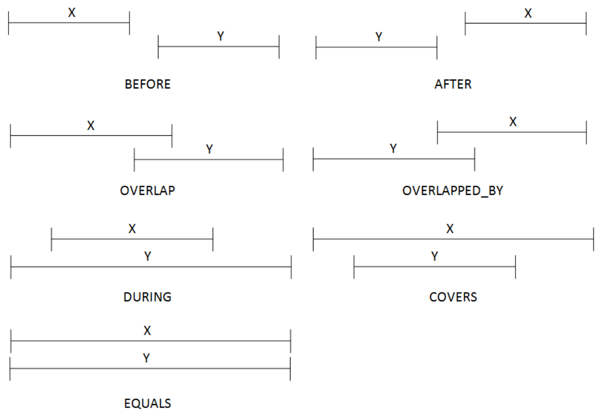

Figure 1: Types of temporal relations

The first five classes are similar to those defined by Pustejovsky et al. (2003a), whereas HALF_GENERIC and STATE_CHANGE are the newly introduced classes, as described previously.

Some events can belong to more than one class. For example, the event *saw* in the example *He saw the car crash* can be labeled as PERCEPTION, as well as I_ACTION (because of the event *crash*). To avoid inconsistencies, we introduced priority levels for event classes as follows (listed from highest to lowest priority):

1. PERCEPTION, ASPECTUAL, REPORTING;
2. I_ACTION;
3. STATE_CHANGE;
4. OCCURRENCE, HALF_GENERIC.

Classes of the same priority level are mutually exclusive. We established these priority levels based on the frequency of event classes in the TimeBank corpus (Boguraev and Ando, 2005). The priority levels ensure that classes with lower frequencies, such as ASPECTUAL and PERCEPTION, will not be neglected by choosing a more frequent class, such as I_ACTION. The most general event classes – OCCURRENCE and HALF_GENERIC – are assigned the lowest priority level.

Five annotators annotated the corpus. They were given detailed guidelines and performed preliminary annotation in two calibration rounds to improve the inter-annotator agreement (IAA). In each round they annotated a set of ten articles. Following the annotation, they met, discussed borderline cases, and resolved the disagreements. After the second round, the IAA on the event extraction task (computed as the average of the pairwise F1 scores) was 0.7951. The remaining 210 articles were then distributed evenly among the annotators and each article was independently annotated by a single annotator. The summary of event annotation is given in Table 1.

3.2. Temporal relations

Our definition of types of temporal relations is based on Allen's interval algebra (Allen, 1983). Because some of the interval relations seem too specific to be determinable from text, we conjoined some relations with the appropriate more general relations: we conjoined MEET with BEFORE, MEETS_INVERSE with AFTER, STARTS and FINISHES with DURING, and STARTS_INVERSE and FINISHES_INVERSE with COVERS. The resulting relation types are shown in Fig. 1. We used the label NON_DETERMINABLE for relations whose type could not be determined based on the information provided in the text.

Table 2: Temporal relation annotation summary

Relation Type	Frequency	IAA
BEFORE	4860	0.7660
AFTER	3500	0.8676
EQUALS	1880	0.4968
COVERS	1597	0.5847
DURING	1341	0.5775
NON_DETERMINABLE	763	0.1813
OVERLAP	46	0.0000
OVERLAPPED_BY	24	0.0833
Total	14,011	

After labeling the events, four annotators, who participated in event annotation, annotated the temporal relations between all pairs of events within the same sentence. After two calibration rounds, they achieved the IAA of 0.5855, measured in terms of Cohen's κ (Cohen, 1960) – a moderate agreement according to Landis and Koch (1977). The temporal relation annotation summary is given in Table 2. Relation types OVERLAP and OVERLAPPED_BY occurred rarely, which can be traced down to two causes: (1) a generally small number of such relations in newspaper articles, and (2) the specific nature of such relations that makes them easily confusable with other relations. This was confirmed by the annotators, who reported that they had usually labeled the potential OVERLAP and OVERLAPPED_BY relations as other relations, such as BEFORE and AFTER, because the context did not explicitly indicate that the events were overlapping.

4. Classifiers and Features

For the event extraction task, we experimented with two approaches: binary classification of words into events and non-events (event identification) and multiclass classification of words into one of eight event classes (an additional class was introduced for non-event words). The temporal relation extraction task was framed as multiclass classification of all pairs of events occurring in the same sentence.

For both extraction tasks, we experimented with the following classification algorithms: naive Bayes (NB), k-nearest neighbors (k-NN), and support vector machines (SVM) with a linear kernel. Based on preliminary experiments, $k = 3$ yielded the best results for k-NN. We used *RapidMiner*[1] implementations of naive Bayes and k-NN, and *LibLinear*[2] implementation of the SVM. The baseline for event extraction is a simple classifier that labels each word with its most frequent label in the training data. For relation extraction, a majority class classifier is used as a baseline (with BEFORE being the majority class).

Because the linguistic tools for Croatian are of limited availability, we used mostly lexical features to build the classification models: word, lemma, stem, POS tag, case, number, modality, auxiliary verbs, verb form, verb valence class, negation, and the surrounding words. For lemmatization and (ambiguous) POS tagging we used the semi-automatically acquired morphological lexicon by Šnajder et al. (2008). For verb valence classes we used the Croatian verb valence lexicon *Crovallex* developed by Preradović et al. (2009). Verb form may be either indicative, imperative, conditional, infinitive, or participle. The features we used for temporal relation classification were similar to those used for event extraction. For each event in the pair, we used the following features: word, lemma, stem, POS tag, modality, auxiliary words, *Crovallex* class, and event class. Moreover, we used a binary feature vector indicating which words occurred between the two event words.

5. Experimental Evaluation

We performed two experiments to evaluate event extraction: binary classification (event identification) and multiclass classification (event classification). We evaluated temporal relation classification by determining types of relations between all pairs of events within the same sentence. Performance estimates are obtained using ten-fold cross-validation; the reported results are macro-averaged F1 scores averaged over ten folds.

5.1. Results

Table 3 shows the results for event extraction. All classifiers outperformed the baseline, with SVM performing best in most cases. Event identification achieved the F1 score of up to $77.40 \pm 0.80\%$. As expected, event classification proved to be a more difficult task than event identification, yielding the much lower overall F1 score of $48.04 \pm 3.21\%$. The classifiers performed better on the OCCURRENCE, REPORTING and ASPECTUAL event classes. This is in accordance with the per class inter-annotator agreement (cf. Table 1), indicating that inter-annotator agreement correlates with classifier performance.

The results for temporal relation extraction are given in Table 4. All classifiers outperformed the majority class baseline. SVM performed best, outperforming the naive Bayes and k-NN classifiers by a margin of 12% and 19%, respectively, and yielding the F1 score of $51.16 \pm 2.94\%$. The performance for classes BEFORE, AFTER, DURING, and COVERS is higher than for other classes, which again is in accordance with the per class inter-annotator agreement (cf. Table 1).

5.2. Discussion

The comparison of our event extraction results with the results of others is difficult due to the differences in the annotation schemes. For example, Saurí et al. (2005) achieved the F1 score of 80% for event identification, which is slightly better than the 77% we achieved, but their approach was based on word chunking, whereas we consider events as single words. Similarly, Saurí et al. (2005) report the F1 score of 86% for event classification, which is much better than our 48%, but they used an entirely different set of event classes (states, general occurrences, reporting, intensional, and perception). Moreover, as noticed by Bethard and Martin (2006), their method included a check to determine whether a word occurs as an event in the TimeBank corpus, which resulted in unfair performance estimates. Bethard and Martin (2006) report F1 scores of up to 76% for event extraction and 58% for event classification. Verhagen et al. (2010) achieved somewhat higher F1 scores for Spanish and English: 88% and 80% for event extraction, and 66% and 79% for event classification.

Temporal relation extraction results are also not directly comparable to other reported results because of the differences in the relation types and the pairs of events considered. Verhagen et al. (2010) report F1 scores of 58% and 66% for the tasks relevant to temporal relation extraction. Only the second task can be related to our research because it considers event pairs within the same sentence. However, only specific event pairs are considered, therefore these results are expected to be better than ours.

6. Conclusion and Future Work

Event and temporal relation extraction are widely researched IE tasks, which can be used in various NLP application. In this paper, we studied event and temporal extraction from Croatian newspaper texts. To that end we manually annotated a corpus and used it to evaluate the event and temporal relation extraction performance of several different classifiers. We achieved the F1 scores of 77% for event identification, 48% for event classification, and 51% for temporal relation classification, significantly outperforming the baseline on all three tasks. A direct comparison to the results for English is difficult, nonetheless we consider the results to be satisfactory given that we were using only simple lexical features. We believe that our results are also indicative for other Slavic languages.

There are several directions for further research. First, considering the relatively low inter-annotator agreement,

[1] http://rapid-i.com/content/view/181/190/

[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Table 3: Event extraction performance (% F1)

	Baseline	NB	k-NN	SVM
Event identification	5.82 ± 0.69	68.85 ± 0.63	71.07 ± 1.31	$\mathbf{77.40 \pm 0.80}$
Event classification	0.84 ± 0.20	33.56 ± 1.27	43.63 ± 2.93	$\mathbf{48.04 \pm 3.21}$
OCCURRENCE	3.10 ± 0.76	55.40 ± 2.04	53.33 ± 2.14	$\mathbf{62.34 \pm 1.23}$
REPORTING	0.69 ± 0.87	75.28 ± 1.09	76.44 ± 3.32	$\mathbf{79.91 \pm 2.36}$
ASPECTUAL	0.39 ± 1.24	12.25 ± 2.31	58.42 ± 7.37	$\mathbf{59.21 \pm 5.48}$
PERCEPTION	0.00 ± 0.00	24.66 ± 8.65	50.80 ± 15.64	$\mathbf{56.32 \pm 18.18}$
I_ACTION	0.99 ± 0.90	$\mathbf{28.63 \pm 1.88}$	24.21 ± 4.26	24.28 ± 2.40
STATE_CHANGE	0.72 ± 0.93	$\mathbf{25.11 \pm 3.62}$	23.18 ± 8.04	23.17 ± 6.49
HALF_GENERIC	0.00 ± 0.00	13.59 ± 2.11	18.99 ± 5.70	$\mathbf{31.04 \pm 6.15}$

Table 4: Temporal relation extraction performance (% F1)

	Baseline	NB	k-NN	SVM
Temporal relation classification	6.44 ± 0.01	38.77 ± 1.87	32.17 ± 1.86	$\mathbf{51.16 \pm 2.94}$
BEFORE	51.43 ± 0.05	63.63 ± 1.74	59.61 ± 3.47	$\mathbf{73.12 \pm 0.85}$
AFTER	–	59.35 ± 2.18	56.16 ± 3.83	$\mathbf{71.08 \pm 1.46}$
OVERLAP	–	11.88 ± 11.70	0.00 ± 0.00	$\mathbf{32.07 \pm 19.44}$
OVERLAPPED_BY	–	2.64 ± 2.37	0.00 ± 0.00	$\mathbf{20.67 \pm 23.82}$
DURING	–	55.59 ± 3.14	46.16 ± 4.26	$\mathbf{60.41 \pm 2.89}$
COVERS	–	36.51 ± 2.52	24.49 ± 3.72	$\mathbf{50.83 \pm 3.49}$
EQUALS	–	36.91 ± 3.54	33.87 ± 2.30	$\mathbf{46.01 \pm 3.43}$
NON_DETERMINABLE	–	43.63 ± 3.71	37.05 ± 7.22	$\mathbf{55.11 \pm 8.20}$

further work should be focused on a more detailed analysis of the annotations and the improvement in annotation guidelines. Secondly, the emergence of new linguistic tools for Croatian language presents an opportunity for using more complex features for classification. Finally, we intend to work on methods for relating the events to normalized temporal expressions (TIMEXes) extracted from text, which should aid in the classification of temporal relations between events.

7. Acknowledgments

The authors would like to thank Damir Cvetovac, Latica Čačković, Marijana Marović, and Kristijan Pavlović for manually annotating the corpus. This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under Grant 036-1300646-1986.

8. References

J.F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

S. Bethard and J.H. Martin. 2006. Identification of event mentions and their semantic class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 146–154.

B. Boguraev and R.K. Ando. 2005. Timebank-driven TimeML analysis. *Annotating, Extracting and Reasoning about Time and Events*, (05151).

B. C. Bruce. 1972. A model for temporal references and its application in a question answering program. *Artificial intelligence*, 3:1–25.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

A. Galton. 1990. A critical examination of Allen's theory of action and time. *Artificial Intelligence*, 42(2-3):159–188.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

M. Lapata and A. Lascarides. 2004. Inferring sentence-internal temporal relations. In *Proceedings of HLT-NAACL*, pages 153–160.

M. Lapata and A. Lascarides. 2006. Learning sentence-internal temporal relations. *Journal of Artificial Intelligence Research*, 27(1):85–117.

C.S. Lee, Y.J. Chen, and Z.W. Jian. 2003. Ontology-based fuzzy event extraction agent for Chinese e-news summarization. *Expert Systems with Applications*, 25(3):431–447.

I. Mani, M. Verhagen, B. Wellner, C.M. Lee, and J. Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 753–760.

L. Polanyi and A. Zaenen. 2006. Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, pages 1–10.

N.M. Preradović, D. Boras, and S. Kisićek. 2009. CROVALLEX: Croatian verb valence lexicon. In *Information Technology Interfaces, 2009. ITI'09. Proceedings of the ITI 2009 31st International Conference on*, pages 533–538. IEEE.

J. Pustejovsky, J. Castano, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev. 2003a. TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering*, 2003:28–34.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. 2003b. The TimeBank corpus. In *Corpus Linguistics*, volume 2003, page 40.

J. Pustejovsky. 1991. The syntax of event structure. *Cognition*, 41(1):47–81.

H. Reichenbach. 1980. *Elements of symbolic logic*. Dover Publications.

R. Saurí, R. Knippen, M. Verhagen, and J. Pustejovsky. 2005. Evita: a robust event recognizer for QA systems. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 700–707.

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

Z. Vendler. 1957. Verbs and times. *The philosophical review*, 66(2):143–160.

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80.

M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62.

M. Verhagen. 2005. Temporal closure in an annotation environment. *Language Resources and Evaluation*, 39(2):211–241.

H. Verkuyl. 2005. Aspectual composition: Surveying the ingredients. *Perspectives on aspect*, pages 19–39.

J. Šnajder, B. Dalbelo Bašić, and M. Tadić. 2008. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing and Management*, 44(5):1720–1731.