

Can We Create a Tool for General Domain Event Analysis?

Siim Orasmaa

Institute of Computer Science,
University of Tartu
siim.orasmaa@ut.ee

Heiki-Jaan Kaalep

Institute of Computer Science,
University of Tartu
heiki-jaan.kaalep@ut.ee

Abstract

This study outlines a question about the possibility of creation of a tool for general domain event analysis. We provide reasons for assuming that a TimeML-based event modelling could be a suitable basis for general domain event modelling. We revise and summarise Estonian efforts on TimeML analysis, both at automatic analysis and human analysis, and provide an overview of the current challenges/limitations of applying a TimeML model in an extensive corpus annotation. We conclude with a discussion on reducing complexity of the (TimeML-based) event model.

1 Introduction

It has been hypothesised in language comprehension research that human understanding of natural language involves a mental representation of events (situations) described in texts (Zwaan and Radvansky, 1998). As many texts can be interpreted as stories/narratives that are decomposable into events, the hypothesis gains further support from research in communication (Fisher, 1984) and in computer science (Winston, 2011), which emphasises the importance of the capability of understanding stories/narratives in natural language understanding. Following this, a creation of an automatic tool that analyses texts for events and their characteristics (e.g. participants and circumstances of events) can be seen as a prerequisite for applications involving text understanding, such as automatic question answering and summarisation. Furthermore, considering the vast amount of information created in online news media on daily basis, one can argue for a clear need of such tool, as it would help to provide a human intuitive overview (e.g. focusing on questions *who did what, when and where?*) on what is reported in online media (Vossen et al., 2014).

Since the Message Understanding Conferences (MUC) and the initiation of information extrac-

tion (IE) research, numerous works have attacked the problem from a domain-specific side, focusing on automatic analysis of specific “events of interest”. Following Cunningham (2005), this is due to automatic analysis of complex information (such as events) requires restricting focus to a specific domain (on specific events) to maintain an acceptable performance level. However, a thread of research, initiated by TimeML—a framework for time-oriented event analysis—(Pustejovsky et al., 2003a), suggests a possibility that event analysis (the annotation of events in texts) could be considered as an extensive automatic language analysis task approachable in a general domain manner, “not restricted to a specific domain” (Saurí et al., 2005). The TimeML-driven fine-grained (word- and phrase-level) event analysis has gained increasing research interest ever since, with the analysis being conducted for different languages (Bittar, 2010; Xue and Zhou, 2010; Caselli et al., 2011; Yaghoobzadeh et al., 2012), tested in several text domains (Pustejovsky et al., 2003b; Bethard et al., 2012; Galescu and Blaylock, 2012) and sub-domains (Bittar, 2010), and extended beyond time-oriented analysis and towards generic event analysis (Bejan and Harabagiu, 2008; Moens et al., 2011; Cybulska and Vossen, 2013; Fokkens et al., 2013). However, the question whether this thread of research should lead to a creation of *a tool for general-domain automatic event analysis*—a tool allowing similar extensive automatic analysis as grammatical level analysis tools (part-of-speech tagging, morphological analysis and syntactic parsing) allow—has not been outlined.

The current work outlines this question, revises and summarises the Estonian efforts on TimeML-based text annotation, both on automatic annotation (Orasmaa, 2012) and human annotation (Orasmaa, 2014a; Orasmaa, 2014b), and interprets the results in the context of creation of a tool for general domain event analysis (Orasmaa, 2016). As the human performance (inter-annotator agreement) on text analysis can be seen as an upper limit for what automatic analysis can

achieve, this provides an overview of current challenges/limitations of applying a TimeML model in an extensive corpus annotation. Observing these limitations, we also discuss a simplified model that could be explored in the future: a model that approximates event annotations to syntactic predicates, and focuses straightforwardly on the annotation of (temporal) relations, without the decomposition of the task.

This paper has the following structure. The next section gives a very general outline to the problem of event analysis, and also the motivation to pursue the problem from the perspective of time-oriented analysis. Section 3 introduces the TimeML model, and gives reasons why it could be considered as a suitable basis for general domain event model. Section 4 gives details on the basic assumptions in TimeML markup, and also revises the Estonian experience in contrast to these assumptions. Subsections of Section 4 focus on event mention, temporal relation and temporal expression annotation. Finally, Section 5 provides a discussion on reducing the complexity of (TimeML-based) event model, and a conclusion that attempts to put the time-oriented event modelling to a broader perspective.

2 The Problem of Event Analysis

Although not often emphasised, “the definition of an “event” is ill-defined” in Natural Language Processing (Bracewell, 2015), and the research progress on event analysis has been hindered by “linguistic and ontological complexity” of events (Nothman, 2013). The struggle with the definition of “event” can also be encountered in other fields, notably in philosophy, where “there is significant disagreement concerning the precise nature” of events (Casati and Varzi, 2014). In philosophy, important characteristics of events could be outlined, perhaps, only when contrasting events against “entities from other metaphysical categories”, such as objects, facts, properties, and times (Casati and Varzi, 2014).

Despite the lack of common theoretical understanding on the concept of event, ever-growing volumes of digital and digitised natural language texts provide a motivation to pursue the research on event analysis. As our understanding of natural language texts can be seen as residing in understanding the “eventive” meanings encoded in texts (Zwaan and Radvansky, 1998), successes in au-

tomatic event analysis promise to open up more human-intuitive ways of automatically organising and summarising large volumes of texts, e.g. providing an overview about events described in on-line news media (Vossen et al., 2014).

While choosing a strong theoretical basis for a tool for automatic analysis of events is rather difficult, one could note that there seems to be an agreement among philosophers that events are generally related to time (“events /- - / have relatively vague spatial boundaries and crisp temporal boundaries”) (Casati and Varzi, 2014). Verbs—a linguistic category most commonly associated with events—often convey markers of temporal meaning at the grammatical level, e.g. Estonian verb tenses provide a general distinction between past and present. Furthermore, some influential theoretical works have generalised from lexical and grammatical properties of verbs to models of time: Reichenbach argued that tenses of verbs can be abstracted to the level of temporal relations (Reichenbach, 1947), and Vendler proposed that verbs can be classified by their temporal properties (Vendler, 1957). This does suggest that it is reasonable to start out approaching general domain event analysis focusing on modelling temporal characteristics of events in natural language, and this is also the approach used in the TimeML framework (Pustejovsky et al., 2003a).

3 TimeML as a Base Model for General-domain Event Analysis

TimeML (and also its revised version: ISO-TimeML (Pustejovsky et al., 2010)) proposes a fine-grained (word- and phrase-level) approach to event analysis: firstly, event-denoting words, such as verbs (e.g. *meet*), nouns (e.g. *meeting*) and adjectives (e.g. *(be) successful*), and temporal expressions (such as *on 1st of February* or *from Monday morning*) are annotated in text, and then, temporal relations holding between events, and also between events and temporal expressions are marked. For example, a TimeML annotation would formalise that the sentence “*After the meeting, they had a lunch at a local gourmet restaurant*” expresses temporal precedence: the event of *meeting* happened before the event of *lunch*.

One can argue that TimeML’s approach is a particularly suitable basis for a general-domain event analysis for the following reasons:

- TimeML’s event is ”simply something that can be related to another event or temporal expression”, and, given this very generic definition, a TimeML-compliant event representation could be used for ”different genres, styles, domains, and applications” (Pustejovsky et al., 2010);
- In TimeML, only a word that best represents the event is annotated in text (Xue and Zhou, 2010), without the full mark up / analysis of event’s argument structure (except time-related arguments: temporal expressions). Following Cunningham (2005), there is a trade-off between an event model’s complexity and its general applicability: an accurate automatic analysis of an event’s complex argument structure requires focusing on a specific domain; however, TimeML’s light-weight commitment to modelling argument structure does suggest a possibility that an accurate analysis could be extended beyond specific domains;
- TimeML follows a principle that in case of complex syntactic structures, only the head of a construction is annotated as an event mention (Saurí et al., 2009). As Robaldo et al. (2011) argue, this makes it particularly feasible to build TimeML annotations upon (dependency) syntactic structures. In case of a successful grounding of event annotations on syntactic structures, one could inherit the general domain analysis capabilities from a syntactic analysis;
- The extensions and derivations of TimeML event model indicate its potential as a generic event model. For instance, TimeML-based event models have been enriched with additional relations holding between events, such as subevent and causal relations (Bejan and Harabagiu, 2008) and spatial relations (Pustejovsky et al., 2011). A TimeML-derived model has been extended with other generic arguments, referring to participants and locations of events, resulting in a four component event model (expressing semantics: *who* did *what*, *when* and *where*?) (Fokkens et al., 2013; Cybulska and Vossen, 2013).

Considering the aforementioned reasons, we as-

sumed in this work that a TimeML model is a suitable basis for developing a general domain event analysis tool.

4 Estonian Experience

In the next subsections, we will discuss the Estonian experience on adapting the TimeML annotation framework. Data and experimental results we use as a basis are from Estonian TimeML-annotated corpus (Orasmaa, 2014b; Orasmaa, 2014a).¹ The corpus has the following characteristics important to our study:

- The corpus is fully annotated by three independent annotators (2 annotators per text), thus it can be used for retrospective inter-annotator agreement studies. Human agreements on analysis indicate the possible upper limits that automatic analysis could achieve;
- The corpus builds upon manually corrected morphological and dependency syntactic annotations of Estonian Dependency Treebank (Muischnek et al., 2014), thus it can be used for studying how well event annotations can be grounded on (gold standard) grammatical annotations;
- The corpus is compiled from news domain texts and covers different sub-genres of news, including local and foreign news, sports, and economy news. Given the heterogeneity of news texts, we assume the corpus is varied enough for using it as a testbed for a general domain event modelling;

In the current work, the inter-annotator agreement experiments on the corpus are revised, and the results are interpreted in the context of creation of a tool for general domain event analysis.

In addition, we also discuss Estonian experience on automatic temporal expression tagging: we contrast the Estonian results (Orasmaa, 2012) with the state-of-the-art results in English, and open up a discussion on the theoretical scope of TimeML’s concept of temporal expression.

4.1 The Annotation of Event Mentions

Assumptions. TimeML assumes that before one can capture semantics of events in text, e.g. the temporal ordering of events and the placement

¹The corpus is available at: <https://github.com/soras/EstTimeMLCorpus> (Last accessed: 2017-01-13)

on a timeline, one needs to establish a consistent **event mention** annotation, upon which semantic relation annotation can be built. At the linguistic level, the range of potential event-denoting units is assumed to be wide, covering “tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases” (Pustejovsky et al., 2003a).

When examining more closely, however, one could note that TimeML’s modelling of events is leaning towards the verb category. Firstly, the guidelines (Saurí et al., 2009) instruct to mark up surface-grammatical attributes for characterising the event, and most of these attributes describe verb-related (or verb phrase related) properties (e.g. *tense*, *aspect*², *polarity*, or *modality*). For instance, the attribute *modality* indicates whether the event mention is in the scope of a modal auxiliary, such as *may*, *must*, *should*. Secondly, if we make a rough generalisation from English TimeML annotation guidelines (Saurí et al., 2006; Saurí et al., 2009), with an admitted loss of some specific details, it appears that: 1) most of the annotation of non-verb event mentions focuses on nouns, adjectives and pre-positions; 2) out of the three parts-of-speech, only noun annotations cover a wide range of syntactic positions, as event mention annotations on adjectives and prepositions are limited to predicative complement positions.

Considering this rough outline of the TimeML event model, it is interesting to ask, how well does one extend the annotation of event mentions beyond the category of verbs, which could be considered as a “prototypical” category for event mentions. The Estonian TimeML-annotated corpus allows us to examine this question more closely.

Estonian experience. The Estonian TimeML annotation project aimed for a relatively extensive event mention annotation, attempting to maximise the coverage on syntactic contexts interpretable as “eventive”. The corpus was created on top of a gold standard grammatical annotations, and it contains (independent) annotations of three different human annotators. Thus, the corpus allows to take out grammatically constrained subsets of event mention annotations, and to study the inter-annotator agreements on these subsets.

Table 1 shows how the inter-annotator agree-

²Note that not all languages have the grammatical aspect as a property of the verb, and this is also the case with Estonian.

ment and the coverage on event mention annotations changes when the annotations are extended beyond prototypically “eventive” syntactic contexts. The highest agreement, F1-score 0.982, was obtained in covering syntactic predicates with event mention annotations. The syntactic predicate consists of the root node of the syntactic tree (mostly a finite verb), and, in some cases, also its dependents: an auxiliary verb (in case of negation) or a finite verb (e.g. in case of modal verb constructions, where an infinite verb dominates the modal finite verb). The agreement remained relatively high (F1-score 0.943) if all verbs, regardless of their syntactic function, were allowed to be annotated as event mentions. However, including part-of-speech categories other than verbs in the event model caused decrease in agreements, and the largest decrease (F1-score falling to 0.832) was noted if nouns were included as event mentions. The high-agreement model (verbs as event mentions) covered only ~65% of all event mentions annotated, and obtaining a high coverage (more than 90% of all event annotations) required the inclusion of the problematic noun category in the model.

4.2 Enriching Event Annotations: Providing Temporal Relation Annotations

Assumptions. Temporal semantics of events in text can be conveyed both by explicit and implicit means. Main **explicit** temporality indicators are verb tense, temporal relationship adverbials (e.g. *before*, *after* or *until*), and explicit time-referring expressions (e.g. *on Monday at 3 p.m.*). The interpretation of **implicit** temporal information usually requires world knowledge (e.g. knowledge about typical ordering of events), and/or applying temporal inference (inferring new relations based on existing ones).

It is stated that the ultimate goal of TimeML annotation is to capture/encode all temporal relations in text, “regardless of whether the relation is explicitly signaled or not” (Verhagen et al., 2009). The TempEval-1 and TempEval-2 evaluation campaigns (Verhagen et al., 2009; Verhagen et al., 2010) have approached this goal by dividing the task into smaller subtasks, and by providing systematic (relatively extensive in the coverage) annotations for these subtasks. Notably in

³In cases of counting EVENT coverage, each token with a unique position in text was counted once, regardless of how many different annotators had annotated it.

| EVENT subset description | EVENT coverage ³ | IAA on EVENT extent |
|-----------------------------|-----------------------------|---------------------|
| syntactic predicates | 57.16% | 0.982 |
| verbs | 65.18% | 0.943 |
| verbs and adjectives | 70.18% | 0.916 |
| verbs and nouns | 93.69% | 0.832 |
| verbs, adjectives and nouns | 98.64% | 0.815 |
| all syntactic contexts | 100.0% | 0.809 |

Table 1: How the annotation coverage and inter-annotator agreement (F1-score) changed when extending EVENT annotations beyond (syntactic predicates and) verbs. Gold standard grammatical annotations were used as a guide in selecting subsets of EVENT annotations provided by three independent human annotators, and inter-annotator agreements and coverages (of all EVENT annotations provided by the annotators) were measured on these subsets. This is a revised version of the experiment firstly reported by Orasmaa (2014b).

TempEval-2, the relation annotations were guided by syntactic relations, e.g. one of the subtasks required the identification of temporal relations between two events in all contexts where one event mention syntactically governed another.

Estonian experience. Following the TempEval-2 (Verhagen et al., 2010) example, the Estonian TimeML annotation project split the temporal relation annotation into syntactically guided subtasks, and attempted to provide a relatively extensive/systematic annotation in these subtasks. However, the resulting inter-annotator agreements showed that approaching the task in this way is very difficult: on deciding the type of temporal relation, the observed agreement was 0.474, and the chance-corrected agreement (Cohen’s kappa) was even lower: 0.355.

Still, the systematic coverage of the temporal annotations and the availability of gold standard syntactic annotations enabled us to investigate whether there existed grammatically constrained subsets of annotations exhibiting higher than average agreements. It was hypothesised that the human agreements were affected by explicit temporal cues: verb tenses encoded in morphology and temporal expressions syntactically governed by verb event mentions⁴. Table 2 shows how the quality of temporal relation annotation, measured in terms of the proportion of VAGUE relations used by annotators and the inter-annotator agreement, was affected by the presence of these

⁴Important explicit cues would also be temporal relationship adverbials, such as *before* or *until*, however, these temporal signals were not annotated in the Estonian corpus.

explicit temporal cues.

| EVENT subset description | Proportion of VAGUE relations | Avg ACC | Avg κ |
|--------------------------------|-------------------------------|--------------|--------------|
| EVENTs in simple past tense | 3.5% | 0.574 | 0.333 |
| EVENTs in present tense | 28.5% | 0.43 | 0.271 |
| EVENTs governing TIMEX | 4.04% | 0.607 | 0.476 |
| EVENTs not governing any TIMEX | 21.1% | 0.447 | 0.291 |

Table 2: How presence of explicit temporal cues affected the quality of manual temporal relation annotation. The quality was measured in terms of the proportion of VAGUE relations used by annotators, and the average inter-annotator agreement (accuracy and Cohen’s kappa) on specifying temporal relation type. This is a revised version of the experiment firstly reported by Orasmaa (2014a).

The results showed that the presence of temporal expressions contributed most to the inter-annotator agreements: the observed agreement rose to 0.607 (kappa to 0.476), and the usage of VAGUE relations dropped to 4.04% (from 21.1%). The morphologically encoded verb tense, however, provided to be an ambiguous indicator of temporal semantics: simple past contributed to

making temporal relations more clearer for annotators, while the present tense contributed to increased temporal vagueness. This can be explained by the Estonian simple past serving mostly a single function—expressing what happened in the past—, while the present tense is conventionally used to express temporal semantics of present, future, recurrence, and genericity.

4.3 Annotation of Temporal Expressions

Assumptions. Temporal expressions are usually seen as an important part of event’s structure, providing answers to questions such as when did the event happen (e.g. *on 2nd of February* or *on Monday morning*), how long did the event last (e.g. *six hours*), or how often did the event happened (e.g. *three times a week*)?

The research on temporal expression (TIMEX) annotation has a long tradition, starting along side with named entity recognition in the MUC competitions (Nadeau and Sekine, 2007), where the focus was mainly on mark-up of temporal expression phrases, and leading to the annotation schemes TIMEX2 (Ferro et al., 2005) and TimeML’s TIMEX3 (Pustejovsky et al., 2003a), where, in addition to the mark-up, also expressions’ semantics are represented in a uniform format. The representation of semantics (normalisation) in TIMEX2 and TIMEX3 builds upon a calendric time representation from the ISO 8601:1997 standard. It allows to encode meanings of common date and time expressions (such as *on 20th of May*, *last Wednesday*, or *12 minutes after midday*), as well as meanings of calendric expressions with fuzzy temporal boundaries (e.g. *in the summer of 2014*, or *at the end of May*), and generic references to past, present or future (e.g. *recently* or *now*). The TimeML scheme assumes a relatively clear separation between temporal expressions and event mentions, with the encoding of semantics of temporal expressions being considered as a straightforward task, while the encoding of semantics of event expressions being considered a complex task of involving mark-up of events, temporal expressions, and temporal relations connecting them.

From the practical point of view, the TimeML TIMEX3 scheme has proven to be relatively successful if one considers performance levels of automatic approaches. A recent evaluation of automatic temporal expression tagging in news do-

main, TempEval-3 evaluation exercise (UzZaman et al., 2013), reports 90.32% as the highest F1-score on detecting temporal expressions in English (82.71% as the highest F1-score for detection with *strict* phrase boundaries), and 77.61% as the highest F1-score on the task involving both detection and normalisation of expressions.

Estonian experience. A large-scale evaluation of an Estonian TimeML-based automatic temporal expression tagger was reported by Orasmaa (2012). We took the results on the news portion of that evaluation (a corpus in size of approximately 49,000 tokens and 1,300 temporal expressions), and recalculated precisions and recalls as TempEval-3 compatible F1-scores. The resulting scores are in the Table 3.

| Subcorpus | F1 | F1 (strict) | normalisation (F1) |
|------------------------------|-------|-------------|--------------------|
| Local news | 89.38 | 84.19 | 80.98 |
| Foreign news | 91.83 | 88.44 | 85.68 |
| Opinions | 87.77 | 80.19 | 75.13 |
| Sport | 94.48 | 89.29 | 81.44 |
| Economics | 86.16 | 79.92 | 77.99 |
| Culture | 86.86 | 81.36 | 76.61 |
| <i>Total (macro-average)</i> | 89.41 | 83.90 | 79.64 |

Table 3: The state-of-the-art performance of Estonian automatic temporal expression tagging on different subgenres of news. The scores are based on precisions and recalls reported by Orasmaa (2012), recalculated as TempEval-3 (UzZaman et al., 2013) compatible F1-scores.

The results indicate that the performance levels on automatic temporal expression tagging in English (UzZaman et al., 2013) and Estonian compare rather well. Although the evaluation settings are not fully comparable, the initial comparison confirms the potential of the TimeML’s TIMEX3 scheme in enabling high accuracy general domain automatic temporal expression tagging across different languages. From the theoretical point of view, however, we note that there is a room for a discussion on how well the information-extraction-oriented approach of TimeML scheme covers the language phenomenon.

The Grammar of Estonian (Erelt et al., 1993) describes a linguistic category similar to TimeML’s temporal expressions: *temporal adverbials*. Temporal adverbials also express occurrence times, durations and recurrences. While

Maršić (2012) states that temporal expressions form “the largest subclass” of temporal adverbials, we note that in addition to the large overlap, the two categories also have notable differences. Temporal adverbials in The Grammar of Estonian are syntactically restricted to sentence constituents that modify the meaning of the main verb or the sentence. Temporal expressions, on the other hand, are not restricted to the syntactic role of an adverbial, e.g. they can also modify the meaning of a single constituent in the sentence, such as the expression *today* in the phrase *today’s meeting*. Semantically, the class of temporal adverbials in The Grammar of Estonian is open: it also includes time expressions with no explicit calendric information (such as *in a stressful era*) and event-denoting time expressions (such as *since the congress*). This contrasts to TimeML’s information extraction perspective that restricts the focus mainly on temporal expressions conveying calendric information.

5 Discussion

TimeML proposes a compositional approach to event analysis: first event mentions should be identified in text, and then, temporal semantics of the events should be encoded via markup of temporal relations.

It can be argued that temporal annotation in TimeML is inherently a very complex task, even for humans (Maršić, 2012), and that a high consistency in the process may not come from a single effort, but rather from an iterative annotation development process. An iteration in this process involves modelling the phenomenon, annotating texts manually according to the model, performing machine learning experiments on the annotations, and finally revising both the model and the machine learning algorithms before starting a new iteration (Pustejovsky and Moszkowicz, 2012; Pustejovsky and Stubbs, 2012). However, the aforementioned strategy may still not be sufficient to tackle the problem, as one could humbly remind that problems related to natural language understanding “have not been studied in linguistics nor anywhere else in the systematic way that is required to develop reliable annotation schemas” (Zaenen, 2006).

Reversing the compositional approach of TimeML, we can argue that a perceivable presence of explicit temporal information is actually

one important indicator of “eventiveness”: that one can interpret text units as “event mentions” with a high degree of certainty only in contexts that allow to place events reliably on a time-line or temporally order with respect to each other. However, the Estonian experience on manual annotation indicates these contexts are not pervasive in news texts, like the grammatically analysable contexts are. Rather, the evidence shows that higher than average consistency can be obtained only in certain syntactic contexts characterised by explicit temporal cues, such as temporal expressions and past-indicating verb tenses. This calls for a discussion for an alternative modeling of events, with the aim of reducing the complexity of the model.

Studies of narratology propose that the semantics of events have a lot to do with events’ relations to other events. One could even go as far as to argue that events “become meaningful” only “in series”, and “it is pointless to consider whether or not an isolated fact is an event” (Bal, 1997). This suggests that the perspective that considers a single event as an atomic unit for analysis could be revised, and events could be analysed in series from the beginning. A minimal unit to be annotated/detected would then be a pair of events connected by a relation, e.g. by a temporal or a causal relation. Note that while the ultimate aim of TimeML is capturing temporal relations, because of the decomposition of the task, someone employing the framework could easily get stuck with the problems of event mention annotation (e.g. how to reliably ground the concept of event at the grammatical level), and may be hindered from reaching temporal relation annotation.

A simpler annotation model could focus directly on annotation of relations between text units, without the decomposition of annotations into events and relations. Before the creation of TimeML, a similar idea was proposed by Katz and Arosio (2001), who did not use event annotation and simply marked temporal relations on verbs in their annotation project. The Estonian annotation experience also showed a high inter-annotator agreement on verbs as event mentions, and the highest agreement on syntactic predicates (main verbs). This suggests that syntactic predicates could be a reasonable (although, admittedly, very rough) approximation for event mentions, and the simple model involving mark-up of relations on syntactic predicates could be the first one to be de-

veloped and tested out in a general domain analysis, before developing more complex models, e.g. adding nouns as event mentions.

Lefeuve-Halftermeyer et al. (2016) make a similar proposal to "characterize eventualities not at the text level, but on the syntactic structures of a treebank", i.e. to mark nodes in a syntactic tree as event mentions. The benefit would be that the syntactic structure would already approximate the event structure, and (to an extent) would provide an access to event's arguments without the need for an explicit markup of event-argument relations. However, the authors do not discuss reducing the complexity of the event model, which, in our view, would also be worth experimenting with. Focusing straightforwardly on the annotation of relations could enable more simple designs both for human annotation and machine learning experiments, which, in turn, could foster more experimentation and, hopefully, improvements on the current results.

In the markup of temporal relations, the Estonian experience showed increased agreements and also less vagueness in the contexts of temporal expressions. As the results of automatic temporal expression tagging in Estonian (reported in Table 3) were also rather encouraging, indicating that satisfactory practical performance levels (95% and above) may not be very far from the reach, one could argue for focusing future temporal relation annotation efforts on contexts with temporal expressions, taking advantage of their high accuracy pre-annotation.

However, contrasting TimeML-compatible temporal expressions with temporal adverbials distinguished in Estonian grammatical tradition revealed that the TIMEX (TIMEX2, TIMEX3) annotation standards have been, to a large extent, "optimised" for capturing "calendric" temporal expressions, i.e. expressions whose semantics can be modeled in the calendar system. A syntax-based view suggests that TimeML's temporal expressions do not cover non-calendric temporal references and also event mentions appearing in the syntactic positions of temporal adverbials. Instead, event mentions in TimeML are considered as markables clearly separable from temporal expressions.

If we are to step back, and attempt to put the problem in a broader philosophical context, we may note that historically, (calendric) temporal

expressions also originate from event mentions. They refer to "major cyclic events of the human natural environment on earth", such as "the alternation of light and dark, changes in the shape of the moon, and changes in the path of the sun across the sky (accompanied by marked climatic differences)" (Haspelmath, 1997). One could say that (driven by the need for expressing time) the natural language has developed rather systematic and relatively unambiguous ways for expressing "calendric events".

This may also offer an explanation why the task of generic event analysis is so difficult to establish—compared to the task of analysing "calendric events" / temporal expressions. Temporal expression tagging builds on the part of human language usage that is already systematic, as it is based on a well-defined conventional system of time-keeping. Yet, it is still an open question whether there is a similar convention of expressing "events in general" in natural language, upon which a systematic general-domain event analyser can be built. While tending towards answering this question, we believe that it is also worthwhile to revise the existing event models for their complexity, and to test out simpler models building straightforwardly on the syntactic structure, and centring them on the explicit temporal cues available in texts.

Acknowledgments

This work was supported by Estonian Ministry of Education and Research (grant IUT 20-56 "Computational models for Estonian").

References

- Mieke Bal. 1997. *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press. <https://archive.org/details/BalNarratologyIntroductionToTheTheoryOfNarrative> (Date accessed: 2017-01-10).
- Cosmin Adrian Bejan and Sanda M Harabagiu. 2008. A Linguistic Resource for Discovering Event Structures and Resolving Event Coreference. In *LREC*.
- Steven Bethard, Oleksandr Kolomiyets, and Marie-Francine Moens. 2012. Annotating Story Timelines as Temporal Dependency Structures. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

- André Bittar. 2010. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard*. Ph.D. thesis, Université Paris Diderot, Paris, France.
- David B Bracewell. 2015. Long nights, rainy days, and misspent youth: Automatically extracting and categorizing occasions associated with consumer products. *SocialNLP 2015 @ NAACL*, pages 29–38.
- Roberto Casati and Achille Varzi. 2014. Events. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2014 edition. <http://plato.stanford.edu/archives/fall2014/entries/events/> (Date accessed: 2017-01-20).
- Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Linguistic Annotation Workshop*, pages 143–151. The Association for Computer Linguistics.
- Hamish Cunningham. 2005. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics*, 5:665–677.
- Agata Cybulska and Piek Vossen. 2013. Semantic Relations between Events and their Time, Locations and Participants for Event Coreference Resolution. In *RANLP*, pages 156–163.
- Tiiu Ereht, Ülle Viks, Mati Ereht, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, and Silvi Vare. 1993. *Eesti keele grammatika. 2., Süntaks (Grammar of Estonian: The syntax)*. Tallinn: Eesti TA Keele ja Kirjanduse Instituut.
- Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. 2005. TIDES 2005 standard for the annotation of temporal expressions. <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-timex2-guidelines-v0.1.pdf> (Date accessed: 2017-01-15).
- Walter R Fisher. 1984. Narration as a human communication paradigm: The case of public moral argument. *Communications Monographs*, 51(1):1–22.
- Antske Fokkens, Marieke Van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. 2013. GAF: A grounded annotation framework for events. In *NAACL HLT*, volume 2013, pages 11–20. Cite-seer.
- Lucian Galescu and Nate Blaylock. 2012. A corpus of clinical narratives annotated with temporal information. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 715–720. ACM.
- Martin Haspelmath. 1997. *From space to time: Temporal adverbials in the world’s languages*. Lincom Europa.
- Graham Katz and Fabrizio Arosio. 2001. The annotation of temporal information in natural language sentences. In *Proceedings of the Workshop on Temporal and Spatial Information Processing*, volume 13, pages 15–22. Association for Computational Linguistics.
- Anaïs Lefeuvre-Halftermeyer, Jean-Yves Antoine, Alain Couillault, Emmanuel Schang, Lotfi Abouda, Agata Savary, Denis Maurel, Iris Eshkol-Taravella, and Delphine Battistelli. 2016. Covering various Needs in Temporal Annotation: a Proposal of Extension of ISO-TimeML that Preserves Upward Compatibility. In *LREC 2016*.
- G. Marşic. 2012. Syntactically Motivated Task Definition for Temporal Relation Identification. *Special Issue of the TAL (Traitement Automatique des Langues) Journal on Processing of Temporal and Spatial Information in Language - Traitement automatique des informations temporelles et spatiales en langage naturel*, vol. 53, no. 2:23–55.
- Marie-Francine Moens, Oleksandr Kolomiyets, Emanuele Pianta, Sara Tonelli, and Steven Bethard. 2011. D3. 1: State-of-the-art and design of novel annotation languages and technologies: Updated version. Technical report, TERENCE project-ICT FP7 Programme-ICT-2010-25410. http://www.terenceproject.eu/c/document_library/get_file?p_l_id=16136&folderId=12950&name=DLFE-1910.pdf (Date accessed: 2017-01-15).
- Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, and Dage Särg. 2014. Estonian Dependency Treebank and its annotation scheme. In *Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 285–291.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.
- Joel Nothman. 2013. *Grounding event references in news*. Ph.D. thesis, The University of Sydney.
- Siim Orasmaa. 2012. Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides (*Automatic Recognition and Normalization of Temporal Expressions in Estonian Language Texts*). *Eesti Rakenduslingvistika Ühingu aastaraamat*, (8):153–169.
- Siim Orasmaa. 2014a. How Availability of Explicit Temporal Cues Affects Manual Temporal Relation Annotation. In *Human Language Technologies—The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014*, volume 268, pages 215–218. IOS Press.

- Siim Orasmaa. 2014b. Towards an Integration of Syntactic and Temporal Annotations in Estonian. In *LREC*, pages 1259–1266.
- Siim Orasmaa. 2016. *Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience*. Ph.D. thesis, University of Tartu, Estonia.
- James Pustejovsky and Jessica Moszkowicz. 2012. The Role of Model Testing in Standards Development: The Case of ISO-Space. In *LREC*, pages 3060–3063.
- James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O’Reilly Media, Inc.
- James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5)*.
- James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The TimeBank corpus. In *Corpus Linguistics*, volume 2003, pages 647–656.
- James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *LREC*.
- James Pustejovsky, Jessica L Moszkowicz, and Marc Verhagen. 2011. ISO-Space: The annotation of spatial information in language. In *Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 1–9.
- Hans Reichenbach. 1947. *Elements of symbolic logic*. Macmillan Co.
- Livio Robaldo, Tommaso Caselli, Irene Russo, and Matteo Grella. 2011. From Italian text to TimeML document via dependency parsing. In *Computational Linguistics and Intelligent Text Processing*, pages 177–187. Springer.
- Roser Saurí, Robert Knippen, Marc Verhagen, and James 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. Association for Computational Linguistics.
- Roser Saurí, Jessica Littman, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML annotation guidelines, version 1.2.1. http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf (Date accessed: 2017-01-20).
- Roser Saurí, Lotus Goldberg, Marc Verhagen, and James Pustejovsky. 2009. *Annotating Events in English. TimeML Annotation Guidelines*. <http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/EventGuidelines-050409.pdf> (Date accessed: 2017-01-15).
- 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. <http://derczynski.com/sheffield/papers/tempeval-3.pdf> (Date accessed: 2017-01-15).
- Zeno Vendler. 1957. Verbs and times. *The philosophical review*, pages 143–160.
- Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The TempEval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179.
- 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.
- Piek Vossen, German Rigau, Luciano Serafini, Pim Stouten, Francis Irving, and Willem Robert Van Hage. 2014. Newsreader: recording history from daily news streams. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31.
- Patrick Henry Winston. 2011. The Strong Story Hypothesis and the Directed Perception Hypothesis. In Pat Langley, editor, *Technical Report FS-11-01, Papers from the AAI Fall Symposium*, pages 345–352, Menlo Park, CA. AAI Press.
- Nianwen Xue and Yuping Zhou. 2010. Applying Syntactic, Semantic and Discourse Constraints in Chinese Temporal Annotation. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING ’10*, pages 1363–1372, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yadollah Yaghoobzadeh, Gholamreza Ghassem-Sani, Seyed Abolghasem Mirroshandel, and Mahbaneh Eshaghzadeh. 2012. ISO-TimeML Event Extraction in Persian Text. In *COLING*, pages 2931–2944.
- Annie Zaenen. 2006. Mark-up barking up the wrong tree. *Computational Linguistics*, 32(4):577–580.
- Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162.