DISSERTATIONES
MATHEMATICAE
UNIVERSITATIS
TARTUENSIS
108

SIIM ORASMAA

Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
108

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
108

SIIM ORASMAA

Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience

UNIVERSITY OF TARTU
Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (PhD) on November 16th, 2016 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor:

PhD.	Heiki-Jaan Kaalep University of Tartu Tartu, Estonia

Opponents:

Assoc. Prof. PhD.	Agata Savary Département des Réseaux et Télécommunication Université François Rabelais Tours Tours, France
PhD.	Uuno Vallner e-Governance Academy Tallinn, Estonia

The public defense will take place on January 3rd, 2017 at 16:15 in Liivi 2-405.

The publication of this dissertation was financed by Institute of Computer Science, University of Tartu.

ISSN 1024-4212
ISBN 978-9949-77-298-8 (print)
ISBN 978-9949-77-299-5 (pdf)

University of Tartu Press
www.tyk.ee

Contents

LIST OF ORIGINAL PUBLICATIONS

- Orasmaa, S. (2012). Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides *(Automatic Recognition and Normalization of Temporal Expressions in Estonian Language Texts)*. Eesti Rakenduslingvistika Ühingu aastaraamat, (8):153–169.

  DOI: http://dx.doi.org/10.5128/ERYa8.10

- Orasmaa, S. (2014). Towards an Integration of Syntactic and Temporal Annotations in Estonian. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1259–1266.

- Orasmaa, S. (2014). 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.

  DOI: http://dx.doi.org/10.3233/978-1-61499-442-8-215

- Orasmaa, S. (2015). Event coreference detection in Estonian news articles: preliminary experiments. Eesti Rakenduslingvistika Ühingu aastaraamat, 11:189–203.

  DOI: http://dx.doi.org/10.5128/ERYa11.12

ABSTRACT

Due to massive scale digitalisation processes and a switch from traditional means of written communication to digital written communication, vast amounts of human language texts are becoming machine-readable. Machine-readability holds a potential for easing human effort on searching and organising large text collections, allowing applications such as automatic text summarisation and question answering. However, current tools for automatic text analysis do not reach for text understanding required for making these applications generic. It is hypothesised that automatic analysis of events in texts leads us closer to the goal, as many texts can be interpreted as stories/narratives that are decomposable into events.

This thesis explores event analysis as broad-coverage and general domain automatic language analysis problem in Estonian, and provides an investigation starting from time-oriented event analysis and tending towards generic event analysis. We adapt TimeML framework to Estonian, and create an automatic temporal expression tagger and a news corpus manually annotated for temporal semantics (event mentions, temporal expressions, and temporal relations) for the language; we analyse consistency of human annotation of event mentions and temporal relations, and, finally, provide a preliminary study on event coreference resolution in Estonian news.

The current work also makes suggestions on how future research can improve Estonian event and temporal semantic annotation, and the language resources developed in this work will allow future experimentation with end-user applications (such as automatic answering of temporal questions) as well as provide a basis for developing automatic semantic analysis tools.

CHAPTER 1

INTRODUCTION

Due to massive scale digitalisation processes and a switch from traditional means of written communication to digital written communication, vast amounts of human language texts are becoming readable to both humans and computers. While human readers have the capability of understanding the texts, they are having increasing difficulties in coping with the sheer amount of text data available. Society would benefit greatly if we could extend computer's large scale text processing capabilities with the capability of understanding (at least to an extent) natural language, allowing us to search and organise vast text collections by the very meanings they convey.

It has been hypothesised in communication research (Fisher, 1984), and in computer science research (Winston, 2011), that the kernel of understanding natural language is in the capability of understanding stories/narratives, which are the most common form of human communication. As events are the basic building blocks of stories, automatic analysis of events in texts can be seen as a prerequisite for applications involving text understanding, such as automatic summarisation, question answering, and construction of event chronologies.

Since the introduction of the time-oriented event analysis framework TimeML (Pustejovsky et al., 2003a), fine-grained (word- and phrase-level) automatic event analysis has gained increasing research interest, with analysis being performed in different languages (Bittar, 2010; Xue and Zhou, 2010; Caselli et al., 2011; Marovic et al., 2012; Yaghoobzadeh et al., 2012), tested in several text domains (Pustejovsky et al., 2003b; Bethard et al., 2012; Galescu and Blaylock, 2012), and extended beyond the time-oriented analysis towards generic event analysis (Bejan and Harabagiu, 2008; Moens et al., 2011; Cybulska and Vossen, 2013; Fokkens et al., 2013). However, the question about whether this thread of research would give a basis for creating a **tool for broad-coverage and general-domain automatic event analysis**, which supports a range of different purposes/applications, in a similar way that grammatical level automatic analysis (part-of-speech tag-

ging, syntactic parsing) supports, has not been extensively researched, especially in the context of less resourced languages. We contribute by exploring this question in the context of automatic event analysis of Estonian – a language which has scarce support in terms of language resources for event analysis – providing an investigation starting from time-oriented event analysis and moving towards generic event analysis.

The contributions of this thesis are the following:

- In contrast to previous research on event analysis in Estonian (Müürisep et al., 2008; Õim et al., 2010), which mainly focused on frame-based approaches, we contributed to the investigation of **light-weight event models**, which are unrestricted to a specific set of frames and decompose the problem into analyses of separate event components (event mentions, temporal expressions, location expressions and participant mentions). These components can be combined relatively unrestrictedly to experiment with different event representations. We hypothesised that such an approach is a necessary basis for automatic broad-coverage and general-domain event analysis;

- We provided a detailed study on an event analysis subtask: **automatic temporal expression (TIMEX) tagging** in Estonian. We outlined the specifics of an Estonian TIMEX annotation format, developed a language-specific rule-based temporal expression tagger for Estonian, and provided a thorough evaluation of its general domain analysis capabilities;

- We created **a TimeML annotated corpus for Estonian**, which contains manually provided event mention, temporal expression and temporal relation annotations. This focus on event analysis from the perspective of temporal semantics is novel in the context of previous works on semantic analysis of Estonian. In a wider context, the corpus is distinctive because:

    1. It is thoroughly annotated by multiple annotators (2 annotators + 1 judge per text), so it provides a basis for large-scale inter-annotator agreement studies;
    2. It contains manually corrected morphological and syntactical annotations, thus providing a basis for investigations of bridging grammatical and semantic annotations;
    3. It contains event mention annotations in an attempt to maximise coverage to all syntactic contexts that can be interpreted as "eventive", thus serving as an example of a relatively exhaustive fine-grained event annotation;

- We conducted two large-scale automated **inter-annotator agreement studies** on the TimeML annotations in the corpus. In the first study, we investigated how much agreement was maintained when event mention annotations were extended beyond prototypical verbal event mentions; in the second study, we investigated how the availability of explicit temporal cues affects the results of temporal relation annotation (in terms of the vagueness encountered, and inter-annotator agreements). These investigations give an overview of the practical limitations of applying a TimeML model in an exhaustive corpus annotation;

- Finally, we made a preliminary exploration on **cross-document event coreference detection** in Estonian, outlining the setup, providing initial investigations, and proposing questions to be investigated in the future;

This monograph has the following structure.

In the following subsections of Chapter 1, we give a brief overview regarding what philosophers have considered an event, introduce automated event analysis as a fine-grained natural language processing task, and outline the desiderata for automated event analysis, and its potential applications.

In Chapter 2, we introduce the task of automatic tagging of temporal expressions, the standard TIMEX3 annotation format used for the task, and Estonian-specific divergences from the standard; we give an overview of a language-specific rule-based temporal tagger developed for Estonian, and its evaluation; we also discuss related works in Estonian and possible future developments in the context of recent research in English. This chapter is partly based on the publication Orasmaa (2012).

In Chapter 3, we provide a study of TimeML-based manual annotation of event mentions and temporal relations in Estonian. We first provide an overview of TimeML-based annotation formats, introduce the theoretical background of TimeML event models, and provide an overview of commonly employed event annotation principles. We then describe the creation of an Estonian TimeML annotated corpus, and provide two detailed discussion threads: the first on event mention annotation specifics, and inter-annotator agreement studies on the task; and the second on temporal relation annotation specifics and inter-annotator agreement studies on the task. Finally, we outline potential applications of the corpus. This chapter contains revisions of the work first published in Orasmaa (2014b) and Orasmaa (2014a).

In Chapter 4, we introduce the preliminary work on cross-document event coreference detection in Estonian. We begin with a discussion of different perspectives on news events, then introduce the problem of event coreference detection, and our preliminary experiments to find events related to a specific person from the daily news. In the discussion part, we consider possible future re-

search directions (investigating the influence of media coverage, and distinctions between summary events and content events) based on the experiments and the literature. The chapter is based on a revision of the work published in Orasmaa (2015).

The Chapters 2–4 are concluded in a two-fold manner: in the "Conclusions" part, we bring out the contributions and conclusions specific to our work, and in the "Philosophical Notes" part, we attempt to place the problem into a broader, philosophical context, as we believe the topic also requires such a viewpoint.

In Chapter 5, we summarise the general problems that have emerged in fine-grained event analysis based on our work and the literature, and outline possible future investigations of these problems in the Estonian language.

Chapter 6 provides an overall conclusion.

1.1 "Events" according to Philosophers

Before we consider event analysis as a practical natural language processing task, we ought to consider the broader context of the phenomenon: What can be considered events? What are the possible difficulties encountered when one tries to establish an abstract definition of events at a more concrete level? We believe that philosophy can shed some light on these questions, and in the following, we give a brief overview of events from philosophical perspectives.[1]

According to Dowden (2009), one can contrast the usual way people understand (and speak about) events, and how events are understood and defined in more formal accounts, specifically in physics. In ordinary discourse, an event is usually understood as "a happening lasting a finite duration during which some object changes its properties". For example, the event of warming milk causes the milk to change from un-warmed to warmed. So, in ordinary discourse, "events are not basic, but rather are defined in terms of something more basic – objects and their properties". This contrasts with physics, where events are considered as "basic, and objects are defined in terms of them". In physics, a basic event is called a point event, defined as "a property (value of a variable) at an instant of time and at a point in space". So, all "real world events" can be considered as decomposable into point events. Physicists do not require that "an event must involve a change in some property", as in ordinary discourse; however, they require that before one can register point events, one must choose a coordinate system – "reference frame" – according to which one can measure and judge about points in space and "the phenomena that take place there" (Dowden, 2009).

[1]The overview is based on The Stanford Encyclopedia of Philosophy (`http://plato.stanford.edu` , 2015-04-01), and The Internet Encyclopedia of Philosophy (`http://www.iep.utm.edu`, 2015-04-01)

The notion of point events in physics seems to be "metaphysically unacceptable to many philosophers, in part because it deviates so much from the way "event" is used in ordinary language" (Dowden, 2009). Events seem to have much more in common than the fact that they can be, theoretically, decomposed into point events, and philosophers "have inquired into what this common nature is" (Schneider, 2005). According to Schneider (2005), "the main aim of a theory of events in philosophy is to propose and defend an identity condition on events; that is, a condition under which two events are identical". According to Casati and Varzi (2014), "there is significant disagreement concerning the precise nature [of events]", and one can, perhaps, shed some light on the matter when comparing events "against entities belonging to other, philosophically more familiar, metaphysical categories", such as objects, facts, properties, and times.

According to Schneider (2005), philosopher Jaegwon Kim argues that events "are constituted by an object (or number of objects), a property or relation, and a time (or an interval of time)". Identity condition for events specifies that two events are identical iff their structural parts (object(s), property, time) are identical. Therefore, according to Kim, events are "property exemplifications": they emerge due to objects having specific properties at specific times. However, this offers a main source of criticism of Kim's approach: i.e. it remains unclear which properties are specific enough to result in an event, and which properties should be merged with other, more generic properties. For example, if we have a verb designating some generic event, e.g. "walking", and a modifier, e.g. "slowly", does applying the modifier on the verb result in a new generic event ("walking slowly"), or does it just indicate that the event "walking" exemplifies the property "being slow"? According to Kim's perspective, the "stabbing of Caesar" and "killing of Caesar" can be considered as different events (with "stabbing" and "killing" being distinct actions), although critics note that "it is a historical fact that the method of killing was a stabbing" (Schneider, 2005). This leads to a debate about the relationship between events and facts, which has not yet been settled by philosophers. Facts seem to be "characterized by features of abstractness and a-temporality" – i.e. facts are more universal and fixed: re-describing a fact can be problematic, as it might lose its truth value – events seem to be more actual than facts, and more firmly tied to a concrete spatiotemporal location of occurrence – thus events are more open to rephrasing/re-describing (Casati and Varzi, 2014).

According to Schneider (2005), philosopher Donald Davidson first proposed that the identity condition for events relies on causality relations: "events are identical iff they have exactly the same causes and effects". This condition, however, can be shown to lead to a problematic circularity. Suppose one wants to determine whether events e and e' are identical. This requires first checking whether their causes – d and d' – are identical. However, because causes are also events,

one can only determine their identity by checking the identity of their effects leading back to checking the identity of e and e'. Later, Davidson rejected the account and adopted a spatiotemporal identity condition: "events are identical iff they occur in the same space at the same time". Although, he noted that this point of view can also be criticised. For example, regarding the sentence "A metal ball becomes warmer during a certain minute, and during the same minute rotates through 35 degrees", should we interpret it as a description of a single event, despite the two event-describing verbs ("becomes [warmer]" and "rotates") in the sentence? Another problem is that two objects can also be considered identical iff they occupy the same spatiotemporal location, so one could ask: What distinguishes events from objects? Davidson tried to find support from natural language on this matter: "basic grammar and predicates" in natural language make it possible to convey that "an object remains the same object through changes", while an event is "a change in an object or objects" (Schneider, 2005). Most philosophers seem to agree that while both objects and events relate to space and time, some fundamental differences can be perceived. Ordinary objects seem to have "relatively crisp spatial boundaries and vague temporal boundaries", while "events, by contrast, would have relatively vague spatial boundaries and crisp temporal boundaries" (Casati and Varzi, 2014).

In the matter of relating events to time, one can even go as far as interpreting an event as a time with a description, "i.e., as a temporal instant or interval during which a certain statement holds", thus giving time the status of a "primitive ontological category". However, this approach seems to be problematic and unintuitive, because "events can be perceived but times cannot". A reversed interpretation would make an event a primitive ontological category, and would consider time instants/intervals as entities derived from events: either as relationships between events, or as specific systematically reoccurring events (Casati and Varzi, 2014).

Philosophers who accept the position that events are particulars – "individuals" that can be counted, compared, described and referred to (and natural language seems to provide support for accepting this position) – often make an attempt to divide events into different classes (Casati and Varzi, 2014).

A broad classification could separate "events" that "proceed in time" or have a "culmination" (such as described in the sentences "John ran for an hour", "John drew a circle", and "John won the race") from "states" for which "it makes no sense to ask how long they took or whether they culminated" (such as described in the sentence "John knows the town map by heart") (Casati and Varzi, 2014). However, arguments can also be made for accepting "states" as "events", especially when states have causal effects (e.g. "Not knowing the town, John got lost in the labyrinth of streets").

It is also an open question how important is the agency component of events:

e.g. whether one should distinguish events that involve the actions of agents from other events that occur regardless of the will of agents. Some philosophers have distinguished between "actions proper" (e.g. "John raises his arm") and "bodily movements" (e.g. "John's arm is rising"); and between "intentional actions" ("John is walking on a field") and "unintentional ones" ("John stumbles on a rock"), which are argued as being "necessary for explaining important facts of human behavior" (Casati and Varzi, 2014). Distinguishing between unintentional and intentional actions has a special importance in the domain of law, where an intentional action (by a person) is subject to legal regulations.

Intentionality can be, from a broader perspective, considered as a *mental process*, and it is unclear what is the exact relationship between mental processes and *"physical / physiological" events* (events that can be explained by general laws of nature and logic). It can be argued that the distinction between mental and physical events can only be encountered in natural language, in vocabulary, and in the "real world", only physical events occur. However, it still seems necessary to distinguish the two types of events, and as mentioned previously, in some domains, such as law, it is also a practical problem. In philosophy, this leads to the question about causal relationships between mental and physical events, which is still open to debate (Casati and Varzi, 2014).

Although events are usually regarded as "things that happen", in certain situations, one also needs to consider events that do/did not happen: "negative events". In natural language, expressing such events also seems relatively easy, and so negative events (e.g. failures / omissions / refrainments) can be counted, compared, and causal relationships can be indicated between negative and positive events. Negative causation (e.g. "Because John did not turn off the gas, an explosion occurred") is considered especially problematic, and it also holds practical importance in the domains of ethics/law and general moral responsibility. Some philosophers have also tried to make fine-grained distinctions between negative events, e.g. "several ways in which an agent may fail to do something: [trying and] not succeeding, refraining, omitting, and allowing" (some negative event). However, if one considers all negative events as "real", the line between real and unreal events becomes blurred, and one has to deal with questions such as how to "refrain from treating all omissions, including non-salient ones, as causes" (Casati and Varzi, 2014).

In conclusion, philosophical discussions on events provide us an overview of the *general scope* and *complexity* of the problem, as many sub-problems brought up in these discussions can be taken to the level of natural language sentences we intend to analyse automatically. However, these discussions are, perhaps, little aware of the empirical side of the problem: i.e. issues related to broad-coverage corpus-based analysis and annotation of events, and the work on natural language

processing that supports these tasks. It is these issues that are the focal point of this study.

1.2 Event Analysis as an Automated Language Analysis Task

Information extraction. Despite the lack of common theoretical understanding of what constitutes an event phenomenon (exemplified in philosophical disagreements), massive and ever-growing volumes of natural language data online have given rise to a wide practical need for automatic event analysis. As events can be viewed as the very basic units upon which our understanding of natural language texts resides (Zwaan and Radvansky, 1998), it is desirable to automatically organise large volumes of textual data based on contained "eventive" information, e.g. to get an overview about all the mentions of some specific event in the media, to present a chronological overview of events, or to provide customised summaries of events (e.g. summaries on events related to a specific person).

In general, automatic natural language understanding is a very complex problem. Thus, rather than directly aiming at text understanding, research has focused on simpler tasks, such as extracting snippets of meaningful information from texts (Cunningham, 2005). Much of the work on automatic event analysis has its roots in information extraction (IE) research, which has largely focused on extraction of events that are relevant to specific information needs.

During the Message Understanding Conferences (MUC) – a series of conferences that gave rise to the information extraction research – event analysis was considered as a template filling task, where a relatively small set of predefined templates was used to extract information about specific "events of interest". For example, when analysing news reports for *narcotics-smuggling* events, one could define an event template which contained slots for information about the source (location, country) and the destination (location, country) of the *smuggling* event, perpetrators (persons) involved in the event, and status of the event (e.g. whether the perpetrators were "arrested", "on trial", or "convicted") (Cunningham, 2005). These slots were then to be filled with the information gathered over the whole input text.

In Automatic Content Extraction (ACE)[2], the scope of event analysis is fixed to sentences: i.e. rather than gathering event details throughout the whole article (as in MUC), only one sentence is assumed to be the extent of the event's description (Linguistic Data Consortium et al., 2005). Targeted events are restricted to

[2]Automatic Content Extraction programme – a programme that was established in United States for developing automatic information extraction technologies, see also `https://www.ldc.upenn.edu/collaborations/pastprojects/ace` (2015-03-24)

eight broad types, such as "Life (subtypes: Be-Born, Marry, Divorce, Injure, Die)" and "Transaction (subtypes: Transfer-Ownership, Transfer-Money)" events, each of which have a set of predefined argument roles (Ahn, 2006). For example, the sentence "John and Kate were married in Spain" can be analysed as describing the event of type "Marry", which has argument roles for the Persons who married (*John* and *Kate*), for the Location where the marriage took place (*Spain*), and for the Time when the marriage happened (not available in the example sentence) (Linguistic Data Consortium et al., 2005).

A limitation of event models "in which an event is a complex structure, relating arguments that are themselves complex structures" (Ahn, 2006), is that these models have a fixed event inventory, which restricts their general applicability. For example, in the domain[3] of financial news, one could rarely find usage for the extraction of "Marry" events, as described above. According to Cunningham (2005), there is a performance trade-off between the complexity of information extraction model and its general applicability. If a model's complexity is increased (e.g. by focusing on "complex events that involve multiple participants", rather than just extracting names of people), its applicability must be restricted domain-wise (to a specific domain) in order to deliver acceptable performance. If a model's complexity is decreased (e.g. by focusing only on the extraction of "people's names", and not attempting to directly associate these with complex events), its general applicability can be widened (perhaps even to a relatively unrestricted domains) while maintaining acceptable performance.

There is a counter-argument to the above reasoning regarding trade-offs: one could claim that an extractor focusing on "Marry" events and obtaining high-accuracy in a variety of domains, including the financial news, is still relatively domain-independent, i.e. it can find the specific "Marry" events in a domain-independent manner. This argument calls for another distinction: between broad and narrow coverage automatic language analysis. An event extractor focusing on "Marry" events over several domains would have narrow coverage, as it would miss most of the other events discussed in these domains (e.g. in financial news it would miss all the business events, such as company merges, or acquisitions), although it could (arguably) deliver relatively domain-independent results on the focused events.

In this work, we are interested in event models tending towards both **broad coverage** and **domain-independence**. Before continuing the discussion on such models, we will describe the other areas of natural language processing, where broad coverage and domain-independence have (arguably) been achieved. We will do this by introducing state-of-the-art natural language processing in Estonian.

[3]In this work, we use the terms "domain" and "genre" interchangeably, considering them synonymous.

Broad coverage and general domain natural language analysis. Automatic analysis of natural language at a grammatical level[4] is often characterized as broad coverage and/or general domain. Tools, such as morphological analysers, part of speech taggers, and syntactic analysers / parsers, have become almost standard parts of language processing pipelines, and such pipelines serve as the basis for numerous language technology experiments and applications. One often assumes that these pipelines provide "general-domain / open-domain / domain-independent language analysis", "broad-coverage / wide-coverage language processing", or "robust processing of unrestricted text". For example, let us consider two automatic analysis steps when analysing Estonian: morphological and syntactic analysis.

Morphological analysis plays an important part in analysing Estonian texts, because the language has "a complicated morphology featuring rich inflection and marked and diverse morpheme variation, applying both to stems and formatives" (Uibo, 2002; Viks, 2000). Thus, the task of an Estonian morphological analyser is to analyse each word token and to determine: 1) how the word can be segmented into morphemes (separate the stem, prefixes, and suffixes); 2) what is the part of speech of the word; and 3) what other grammatical information is encoded in the morphemes (e.g. nominal cases; voice, tense, and mood of verbs).[5] In practice, the majority of words can be analysed using a dictionary and a set of morphological rules, and only approximately 3% of encountered words are unknown to the automatic analyser (Kaalep and Vaino, 2001). This makes the morphological analysis inherently a "broad-coverage" task, as the available linguistic knowledge can be encoded to cover the majority of language phenomena "appearing in the wild". The task can also be considered relatively "domain-independent": even if the scope of analysis is set to "contemporary written Estonian" (Kaalep and Vaino, 2001), a variety of text domains is covered, including news, fiction, and scientific texts (Kaalep et al., 2010).

Syntactic analysis takes analysis "above words", addressing sentence-internal grammatical relations between words: what are the syntactic roles of each word (e.g. subject, object, or predicate), how can words be grouped into phrases and clauses, and what is the underlying sentence structure (tree) created by these relations. This task is also challenging for Estonian, due to "the relatively free word order" and "wide extent and variety of grammatical homonymy" of words (Müürisep et al., 2003); however, recent advances in automatic analysis have shown that syntactic labels can be assigned to words with an accuracy of ap-

[4] While other interpretations are possible, we interpret the "grammatical level" as mainly concerning morphology and syntax.

[5] Ambiguities often arise with morphological analysis of Estonian, so complete morphological analysis is a two-step process: 1) determining all possible morphological categories for each word; 2) disambiguating morphologically ambiguous words based on contextual cues.

proximately 88%, covering the three domains: fiction, news, and scientific texts (Muischnek et al., 2014a). Thus, one can also consider syntactic analysis as a task tending towards "broad-coverage" and "domain-independence", as the majority of words in text can be (relatively accurately) analysed according to syntactic formalism, and accurate analysis is applicable over a variety of domains.

Thus, we can draw some conclusions about "broad-coverage" and "domain-independent" natural language analysis at a grammatical level.

"Broad-coverage" in this context refers to the coverage of word tokens: a large percentage of words in input text can be accurately automatically analysed (labelled) according to a given formalism. This also requires the formalism to be both theoretically well-defined, and tested in numerous empirical experiments and shown to be robust; language analysis formalisms at grammatical level tend to exhibit such properties[6].

The "broad-coverage" analysis does not always imply "domain-independence". Even for robust language analysis at a grammatical level, there are specialised language usage domains (e.g. biomedical, internet language, or speech transcripts), which can differ from the general written text domain and thus may require domain adaptation for analysis to be accurate. However, one often assumes that if a tool can accurately analyse texts from the "newspaper domain", it is relatively safe to call it "supporting general domain analysis", as the "newspaper domain" can be thought to represent a large amount of heterogeneity of written language.

Broad coverage and general domain event analysis. The previous research on event analysis in Estonian has mainly focused on the theoretical modelling of events, aiming for a model as close to "real" human language understanding as possible (Õim et al., 2010). Similar to information extraction approaches, the event inventory has been restricted to specific types of events, motion events; these events are represented as predefined argument-structures (sets of semantic roles), and (similar to the approach used in MUC) the details of the event are gathered over different sentences (Müürisep et al., 2008; Õim et al., 2010). As we have argued, such an approach is likely limited to only operating accurately in specific domains, and has a narrow coverage due to not addressing events other than motion events mentioned in the text.

In contrast to previous works in Estonian, this thesis explores approaches that go beyond the analysis of specific (pre-defined) events. A central question of this thesis is: could we make a tool that provides a broad-coverage and general-domain event analysis, and supports a range of different purposes/applications, in a similar way that grammatical level analysis supports?

[6]Although, quoting Edward Sapir "All grammars leak", thus one still cannot expect perfect coverage of a language phenomenon by its grammar.

As noted regarding the philosophical disagreements on the concept of event, a strong theoretical basis for developing such a tool is difficult to find. Yet, there seems to be agreement among philosophers that events are generally related to time ("events /- - -/ have relatively vague spatial boundaries and crisp temporal boundaries" (Casati and Varzi, 2014)). This agreement is also supported by linguistic analysis stemming from the grammatical level. Verbs – a category of words most directly associated with "eventiveness" in common interpretation – can be analysed for their temporal properties, e.g. Estonian verb tenses provide a general distinction between events that happened in the past or are happening in the present. Temporal adverbials (e.g. *tomorrow*, *24th of February*, or *on Monday*) can combine with verbs relatively unrestrictedly, indicating that time could be associated with any event.[7] This suggests that wide coverage general domain event analysis can be grounded in natural language via temporality cues, and we also took this assumption as our starting point.

While we discuss theoretical works that generalise from natural language cues to models of time (such as Vendler's classification of verbs by their temporal properties (Vendler, 1957) and Reichenbach's encoding of grammatical tenses (Reichenbach, 1947)), the essence of our work was an empirical corpus-based annotation study, driven by concerns regarding manual annotation consistency and automatic annotation replicability, rather than of a temporal theory.

We based our annotations on the TimeML (Pustejovsky et al., 2003a) framework, focusing on the annotation of **temporal expressions**, **event mentions**, and **temporal relations**. Compared to event models that focus on modelling "complex argument structures" (such as the models of MUC and ACE), TimeML proposes the most fine-grained approach yet to event analysis: analysis that focuses on verbs (e.g. *married*), nouns (e.g. *marriage*), and adjectives (e.g. *(be) pregnant*) as event mentions. While the ultimate goal of TimeML research is to detect temporal relations between events (e.g. to automatically discover that the sentence "*After their marriage, John and Kate moved to Florida*" expresses temporal precedence: the event "*marriage*" temporally precedes the event "*moved*"), it also provides a decomposition of the event analysis problem, allowing one to analyse event predicates separately from their arguments. This separation allows one to ground event mentions at grammatical structures, and similar to grammatical level analysis, to approach the problem in a broad coverage and general domain manner.

As we have seen from research on English, the TimeML's model covering event mentions (predicates) and temporal expressions (time arguments) can be extended with additional arguments referring to location and participants, arriving at the generic four component model (expressing semantics: *who* did *what*,

[7] Note that locational adverbials can also be potentially associated with any event verb, so this does provide an argument against "events being more related to time than to location".

where, and *when*) (Fokkens et al., 2013; Cybulska and Vossen, 2013, 2014b). An extension could also be made towards models with more sophisticated argument structures, e.g. if TimeML's event mentions were aligned with predicate-argument structures from PropBank (Palmer et al., 2005; Pustejovsky et al., 2005b) or VerbNet (Schuler, 2005). TimeML's model can also be extended with additional semantic relations that connect events, such as subevent and causal relations (Bejan and Harabagiu, 2008), and can be integrated with a fine-grained model of spatial semantics (Pustejovsky et al., 2011).

Though the ultimate aim of the research is "broad-coverage" analysis similar to grammatical level analysis (covering accurately a large percentage of words in the input text), our current work focuses on providing a basis for four component light-weight event models, modelling the generic "*who* did *what*, *when*, and *where*" semantics. The central idea is to first focus on developing "broad-coverage" analysis/annotation for separate event components – for event mentions, temporal expressions, location expressions, and participant mentions – that can then be used for exploring different light-weight event representations. Figure 1.1 provides an illustrative example of a sentence annotated both at grammatical and four event component levels, and lists all (semantically plausible) light-weight event representations that can be created by combining the event component annotations.[8] Appendix A provides definitions of the grammatical tags used in the example.

Let us clarify parts of Figure 1.1 and the theoretical scope of the light-weight models. First, we leave open the specific details of how event components should be realised. For example, the temporal expression component `TIMEX t1` in Figure 1.1 can be realised simply as a lemma of the expression ("täna"), but it can also be realised as a specific date string (such as "2014-04-12") corresponding to the date the expression refers to. We provide a TimeML-based realisation of temporal expression annotation in Chapter 2, and a TimeML-based realisation for event mention annotations in Chapter 3, but we do not rule out the possibility of using only subsets of these realisations, or using alternative realisations. Second, we also leave open the details of how event components should be combined into event models. The figure lists all "semantically plausible" combinations, illustrating an ideal situation where these combinations are known beforehand. In the empirical work, we experiment with two approaches: in Chapter 3, we use the two component TimeML model and make combinations based on dependency syntactic relations between the components (event mentions and temporal expressions), and in Chapter 4, we employ a two component model of co-occurring temporal and location expressions, but we place no syntactic restrictions on co-

[8] Including representations consisting of a single event component – we will discuss this in more detail in Chapter 2.

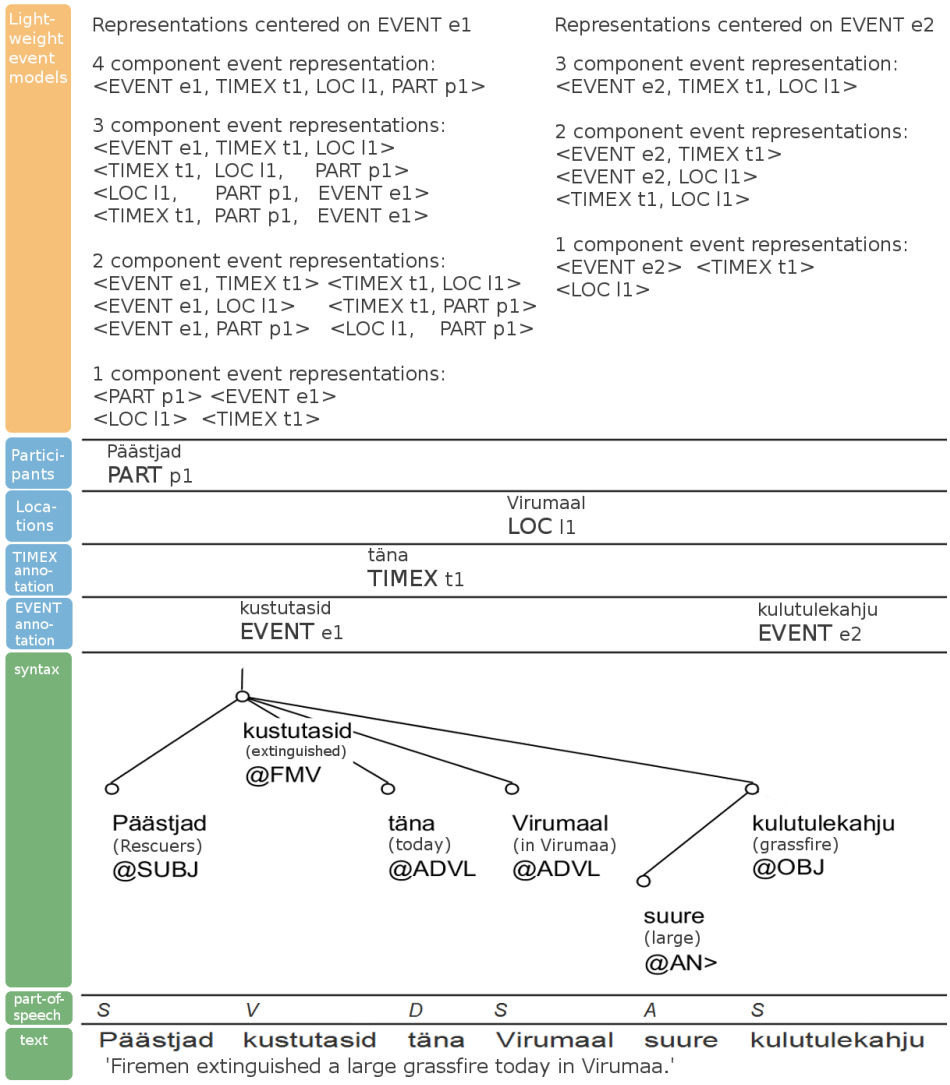

Figure 1.1: An example illustrating a sentence annotated with grammatical-level annotations (green layers) and event component annotations (blue layers), and listing light-weight event representations that can be constructed based on the event component annotations (the orange layer). Only semantically plausible event representations are listed.

occurring components.

Most of our work focuses on news texts, with the assumption that the news domain covers a variety of sub-domains of language usage, and thus can be representative of the "general domain". An exception is Chapter 2, where we focus on annotation of temporal expressions, and this allows us to extend our evaluation

to other domains of formal written language, such as legalese texts and parliamentary transcripts, and to divide news into sub-domains, such as foreign news and economic news.

Chapter 4 represents a divergence from the others, as we take the problem of fine-grained event analysis from single document level (as it is in Chapters 2 and 3) to cross-document level: i.e. from event detection to **cross-document event coreference detection** (i.e. finding event mentions referring to the same events across documents). While our empirical work in that chapter is limited and preliminary, it does provide another perspective on the problem: a perspective that suggests event analysis in news might be rather domain-specific, and must consider the structure of articles, and media coverage patterns of news events. A full empirical backup of these suggestions, however, remains out of the scope of this study.

1.3 Desiderata for Automatic Event Analysis

In this section, we briefly summarise our desiderata for automatic fine-grained event analysis, and also briefly point to known problems regarding them.

- *Support for general domain analysis.* The analysis is not limited to some specific domain (e.g. financial news or encyclopaedias), but can address the event phenomenon in a variety of written text domains;
- *Support for broad-coverage analysis.* The analysis should cover a majority of words in the input text, indicating each word's role in an event structure (e.g. whether it is an event mention, or an argument of event, such as time or location);
- *Compatibility with other linguistic resources/annotations*. The event analysis should be compatible both with *underlying annotations* (morphological and syntactic annotations), and with *overlying annotations* (e.g. frame structure annotations, such as VerbNet (Schuler, 2005) or FrameNet (Baker et al., 1998) annotations).

The aspect of general domain. As we discussed in the previous section, "general domain" could have different scopes, from "contemporary written language" to the news domain (a domain covering a variety of sub-genres). However, setting a large scope prompts the question whether it is even appropriate/meaningful to analyse every text for events, especially when the concern of text is not on *narrating a story*. For example, a geographical entry in an encyclopaedia, which only focuses on a spatial description of the location, could be ignored entirely from

the perspective of event analysis.[9] Therefore, one could suggest that the research should first focus on domains commonly associated with events and stories, such as news, and then, after progress is made on these domains, move on to other domains, where ”the eventive” interpretations are more arguable.

The aspect of broad-coverage. Having a broad-coverage similar to grammatical level annotation does require semantic annotation formalism, which is theoretically well-defined, commonly understood, and practically robust (has shown to be applicable over several domains). To our knowledge, such formalism has not yet arrived even for well-studied languages, such as English, so event analysis must start with simple models, such as the TimeML model and the four component event model described in the previous section, and must to evolve towards more complex models.

The aspect of compatibility. The compatibility between linguistic resources is difficult to obtain, because different resources tend to focus on different perspectives of the analysis, and thus are not inherently compatible. For example, while the English FrameNet (Baker et al., 1998), VerbNet (Schuler, 2005) and PropBank (Palmer et al., 2005) can all be considered as resources modelling predicate-argument structure (which could be approximated to ”event structure”), integrating these resources into a common semantic representation is considered a separate problem to be investigated (de la Calle et al., 2014). In a similar way, merging TimeML annotations with other semantic annotations, such as with PropBank annotations, does require a separate study (Pustejovsky et al., 2005b). In the case of limited linguistic resources, such as is the case with Estonian, one could set a goal of achieving compatibility with wide coverage linguistic resources at grammatical level, because most semantic level resources have yet to obtain wide coverage (see Liin et al. (2012) for a recent overview on Estonian language resources).

1.4 Target Applications

The ultimate goal of automatic event analysis would be to support many different natural language applications, and to be a standard part of natural language processing pipelines, as grammatical level analysis currently is. However, during the early stages of the research, it is more feasible to focus on a few key application areas. Here, we consider two.

1.4.1 Temporal ordering of events

Following the ideas behind TimeML (Pustejovsky et al., 2003a), the fine-grained event analysis discussed in this work aims at providing support for applications

[9] We would like to thank Haldur Õim for providing this example.

that require determining the temporal order of events: either a relative ordering of events with respect to other events, or absolute positioning of events on a global timeline.

Temporal ordering. "News stories are seldom if ever told in chronological order", events are either reported in reverse chronological order (latest events/developments are reported first) or, frequently, there seems to be no chronological order at all (Bell, 1999). Thus, extracting event mentions and providing a chronological (re-)ordering of these events can be considered as an important application of event analysis of news texts. Some application-oriented works have also focused on more restricted tasks, in which only explicit temporal information (dates, temporal expressions) is employed for building chronologies (Kessler et al., 2012), or for visualizing events on a timeline (Alonso et al., 2010a).

Overview of (possible) future events. News (and social media) texts frequently discuss people's future plans, goals, or predictions: i.e. references to possible future events. Therefore, some authors have considered extraction of future events as a separate task (Baeza-Yates, 2005; Jatowt and Au Yeung, 2011).

1.4.2 Extraction and aggregation of event information

While time is an important aspect of event analysis, one often needs to acquire more information about an event: about participants, location, and other circumstances of an event. Such information needs can be formulated via various information seeking, retrieval and re-organising tasks, including:

Automatic Question Answering. This is an area of research that aims to build systems that can automatically answer questions formulated in natural language. As questions (and especially factoid questions) usually address some aspect of an event (e.g. *when did it happen? where did it happened? who did it?*), event analysis should naturally provide support for building such systems. The TimeML annotation framework was also created in the context of developing question-answering systems, e.g. for answering questions such as "*When did Bill Gates became CEO of Microsoft?*"(Pustejovsky et al., 2003a).

Automatic Summarization. In order to reduce the load of information encountered when manually browsing a large document collection (e.g. a news archive, or encyclopaedia), documents need to be represented to the user in some compact, summarised form. The problem of how to reduce documents into such a form is addressed by automatic summarization, and there also an event-centric approach can be used: the aim is to produce a summary that only contains the mentions of salient events (Vanderwende et al., 2004).

Event-centric media monitoring. While many automatic media monitoring systems focus on analysing trends (present users with timelines and maps of trending keywords), event-centric media monitoring (centric to the questions *who did*

what, when and where?) would provide a more detailed overview of the development and media coverage of events (Vossen et al., 2014b).

CHAPTER 2

ARGUMENTS FOR EVENTS OR ARGUMENTS AS EVENTS: A CASE STUDY OF TEMPORAL TAGGING

2.1 Temporal Tagging and its Relationship to Event Analysis

Temporal expressions can be seen as playing an important role in event structure, providing answers to questions such as *when* did the event happen (e.g. *on the 25th of July 2013*, or *tomorrow*), *how long* did the event last (e.g. *three hours*), or *how often* did the event happen (e.g. *twice a week*)? If a temporal expression appears in a sentence, a compositional interpretation suggests that it adds details to the event described in the sentence, specifying the temporal location of the event. This motivates one treat the analysis of temporal expressions as a separate task – the task of *temporal tagging* – upon which more sophisticated event models can be built.

The task of temporal tagging aims to automatically extract temporal expressions from text and to normalise semantics of these expressions based on some annotation format.

If one considers events as "complex structures relating arguments" (Ahn, 2006), temporal tagging is only a subtask in event analysis, and one needs other stages of analysis, such as detection of event participant and location mentions, in order to create a complete event analysis process. However, the light-weight event models aimed at in this work (exemplified in Figure 2.1) do not set requirements for fixed event structures, and even suggest the possibility of a "minimalist" event model, where an event is only represented by its temporal location (denoted by the temporal expression).

As most of the temporal locations mentioned in a (news) discourse are as-

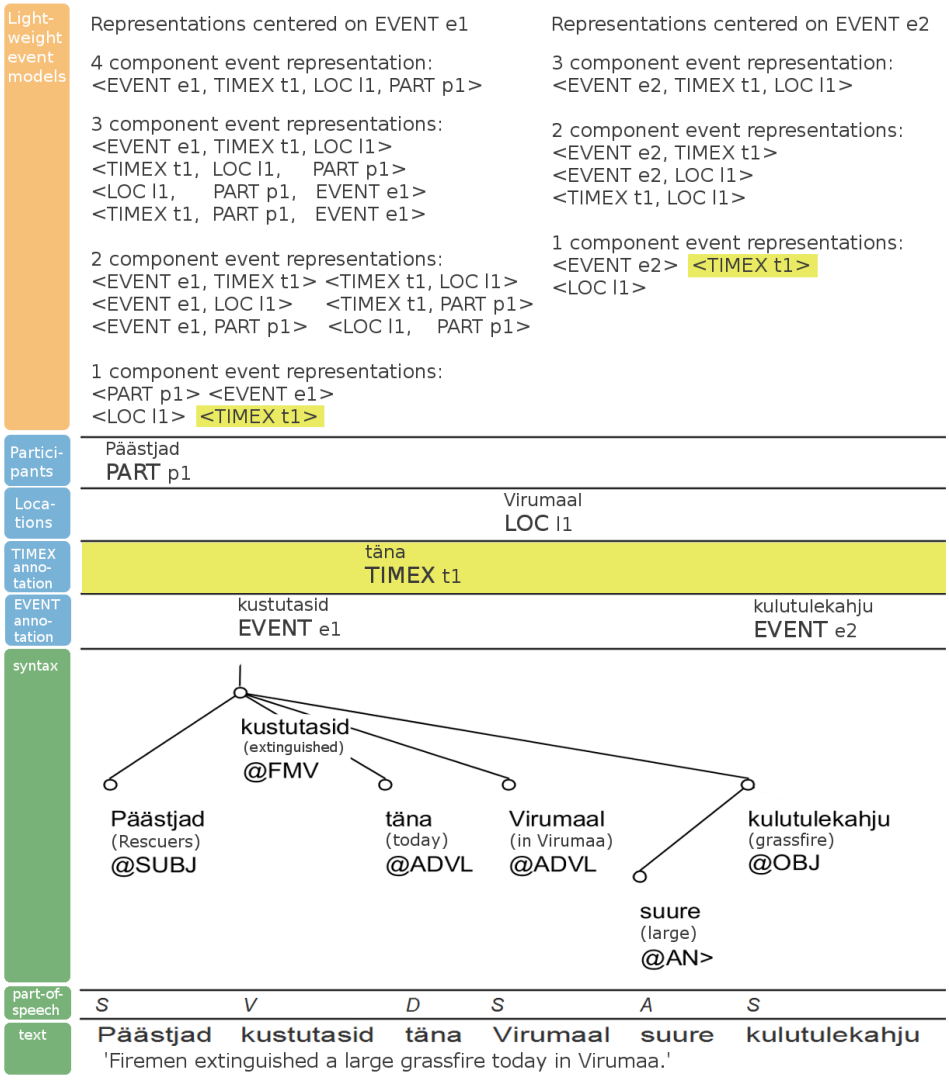

Figure 2.1: An example illustrating a sentence annotated with four layers of event component annotations, and (semantically plausible) light-weight event representations that can be constructed based on these annotations. Annotations and event representations directly supported by *temporal tagging* are marked with a yellow background

sociated with many events (e.g. in a stream of daily news, a large number of event descriptions use the expression *today* for referring to their temporal location), a "minimalist" event model is likely to have limited discriminative power in practice. The model could be most successful at discriminating historical events, which have a distant temporal location and are salient enough (within the dis-

course) that their temporal location is unambiguously associated with them. For example, if we consider a contemporary discourse of Estonian news, most of the temporal expressions referring to the date 1918-02-24 can be assumed to refer to the Estonian Declaration of Independence,[1] or to its subevents (such as printing and distributing the manifesto). An example from English news discourse could be the date 2001-09-11, which is relatively unambiguously associated with the *September 11 attacks*[2] (a series of events which are conventionally even named after their temporal location, i.e. *9/11*).

Figure 2.1 illustrates a sentence annotated for event components and brings out all (semantically plausible) light-weight event representations, highlighting (with the yellow background) the annotation layer and the ("minimalist") event representations directly supported by temporal tagging. There are two events mentioned in the example sentence (*kustutasid* 'extinguished' and *kulutulekahju* 'grassfire'), both of which can be associated with the date corresponding to the temporal expression *täna* 'today'. Thus, if we use a "minimalist" TIMEX based representation for these events, we will get two identical "events" (and thus an exemplification of the discrimination problem common to single component models).

The "minimalist" TIMEX based model can be extended to a two component event model, which likely has more discriminative power. Strötgen (2015) proposed a model of spatio-temporal events, where an event expression is formed if "a geographic and a temporal expression co-occur in a textual document or within a specific window in the document [e.g., in a sentence]". The author shows that this model offers a relatively strong baseline for event extraction (if all pairs of geographic and temporal expressions co-occurring in a sentence are interpreted as events, valid events can be extracted with an F1-score of 82.1%), and is useful for various search and exploratory tasks of document collections.

Alternatively to light-weight event models operating at the word/sentence level, temporal expressions can also be used for event modelling at a document level. Alonso et al. (2010b) proposed a framework for measuring similarity between documents, based on overlaps in the time periods mentioned in both documents. Assuming that each time period (temporal expression) uniquely stands for some event, one could also interpret this measure as reflecting the degree to which the two documents are discussing the same events.

These examples suggest that temporal tagging offers direct support to event analysis and in some scenarios, temporal expressions alone can stand out as event references.

[1] `https://en.wikipedia.org/wiki/Estonian_Declaration_of_Independence` (2016-01-09)

[2] `https://en.wikipedia.org/wiki/September_11_attacks` (2016-01-09)

2.2 The Annotation Format for Temporal Expressions

In this section, we give an overview of the annotation format used in this work. The first subsection gives a brief overview of the history of temporal expression annotation formats, the second subsection describes in detail the annotation format used for modelling Estonian temporal expressions, and the third subsection discusses the specifics of Estonian annotation, contrasting these to the standard (English) annotation.

2.2.1 History of the TIMEX annotation formats

The task of annotating temporal expressions in text can be traced back to the sixth Message Understanding Conference in 1995, where named entity recognition was introduced for the first time as a sub-task of Information Extraction. Along with other named entities, such as person, organization, and location names, the task also required detection of temporal expressions ("timexes") from text (Nadeau and Sekine, 2007). However, the scope of the task was limited, requiring only detection of expressions referring to TIMEs (time points having a temporal granularity finer than a day, such as *20 minutes after 9*, or *11:15 PM*), and DATEs (time points of a day or a coarser temporal granularity, e.g. *the 10th of October*, *1995*, or *the 20th century*) (Mazur, 2012).

The task definitions in MUC competitions did not address the issue of temporal expression normalization, that is, representing the semantics of the expressions in a uniform format. This issue was targeted by Mani and Wilson (2000a,b), who extended the MUC TIMEX format with a calendric time representation from the ISO 8601:1997 standard. Their annotation approach captured both expressions that can be normalized independently from the context (e.g. *the 20th March 1995*, or *20.03.1995*), and context-dependent expressions, which can be normalized "depending on the speaker and some reference time" (e.g. *last Friday*, or *the 20th of March*). Their work contributed to the development of the TIMEX2 annotation scheme (Ferro et al., 2005), which has become "an informal standard (of temporal annotation) in the research community" (Mazur, 2012).

The TIMEX2 annotation scheme widened greatly the range of temporal expressions that could be annotated/extracted, and introduced an interlingual representation for expressing the temporal semantics of the expressions. In addition to "points in time" (temporal expressions answering "when"-questions, such as *on the 20th of March*), duration expressions (expressions answering "how long"-questions, e.g. *three months*, or *a year*), and recurrence expressions (expressions answering "how often"-questions, e.g. *every Friday*) were also considered as target expressions. Semantic representation was extended to represent anchored expressions (such as durations anchored to speech time: *in the past two years*, or

three weeks from now), expressions with fuzzy temporal boundaries (such as *in the spring of 2004*, or *at the end of April*), and expressions with non-specific semantics (e.g. *a day in August*) (Wilson et al., 2001; Ferro et al., 2005).

The TimeML framework (Pustejovsky et al., 2003a) aims to extend temporal annotation beyond temporal expressions. The framework proposes that in addition to temporal expressions, event mentions in text should also be annotated as temporal objects, and that temporal relations between temporal objects should be made explicit in the annotation (e.g. relationships indicating the temporal order of events mentioned in the text). To model temporal expressions, TimeML uses a variation of TIMEX2 (called TIMEX3), which introduced several changes compared to the TIMEX2 scheme. In TimeML, complex anchored temporal expressions (e.g. event-anchored expressions, such as *the day after our meeting*, and timex-anchored expressions, such as *three weeks from now*[3]) are decomposed into smaller annotation segments. For example, the event-anchored expression *the day after our meeting* is decomposed into: 1) a TIMEX3 annotation covering the phrase *the day*; 2) an EVENT annotation covering the phrase *(our) meeting*; and 3) a temporal relation (TLINK) annotation connecting the two entities (the TIMEX3 and the EVENT) and indicating that the point in time (*the day*) is located after the event (*(our) meeting*).

In the following section, we introduce the TimeML TIMEX3 annotation format in more detail.

2.2.2 The TIMEX3-based Annotation Format

Our annotation format is based on the TIMEX3 tag in TimeML (Pustejovsky et al., 2003a). Following the TimeML specification, we distinguish four types of temporal expressions:

- DATE expressions that refer to points[4] on a timeline, and contain *year*, *month*, *week* or *day* granularity temporal information. Examples: *eelmine kuu* 'last month', *1999. aastal* 'in the year 1999', or *22. veebruariks* '(for) the 22nd of February';

- TIME expressions that refer to points on a timeline, and could contain *day* granularity temporal information, but must contain *hour* or *minute* granularity, or part of a day (e.g. 'morning', 'afternoon') temporal information.

[3] In TIMEX2, such expressions were annotated as full-length phrases, see Ferro et al. (2005) for more details.

[4] One can also argue that expressions with a coarse granularity, such as year expressions (*2004*), refer to a period rather than a point on a timeline. However, for uniformity, we still considered such expressions as point-referring expressions, provided the referred temporal entity can be located on a timeline.

Examples: *reedel kell 13.45* ‘on Friday at 13:45’, *järgmise hommikuni* ‘(until) the next morning’;

- DURATION expressions that describe periods of time, e.g. *kolm päeva* ‘three days’, or *8 kuu jooksul* ‘(during) 8 months’. Duration expressions also include so-called directed duration expressions: expressions that make explicit the direction to which the duration unfolds, such as *eelneva kahe kuu jooksul* ‘during the previous two months’, or *järgmised kolm aastat* ‘the next three years’;

- SET expressions that refer to recurring time periods, i.e. sets of times. For example: *igal aastal* ‘every year’, or *kolm korda nädala jooksul* ‘three times a week’.

According to TimeML specification, temporal expressions in text are tagged with TIMEX3[5] tags. The most important (mandatory) attributes of a TIMEX3 tag are the *tid* (unique index of the expression), the *type* (type of the expression, from the aforementioned types), and the *value* (normalised semantics of the expression). An example of a TIMEX3 annotation is (1):

(1) 1929. aastal toodeti USAs 5,4 miljonit autot.
<TIMEX3 tid=”t1” type=”DATE” value=”1929”>1929. aastal </TIMEX3>
toodeti USAs 5,4 miljonit autot.
5.4 million cars were produced in the US in the year 1929.

The representations of the semantics of the expressions in the attribute *value* are based on the ISO-8601 standard representation of dates and times, and can be separated into three base formats:

- *Month-based date format*. For example, the expression *15. veebruar 2009* ‘15th of February 2009’ is normalised as ”2009-02-15”;

- *Week-based date format*. For example, if the speech time refers to the 35th week of the year 2010, the expression *järgmise nädala teisipäeval* ‘on next week’s Tuesday’ can be normalised as ”2010-W36-2”;

- *Duration format*. For example, the expression *kolm aastat* ‘three years’ is normalised as ”P3Y”;

Month-based and week-based date formats can be lengthened or shortened from the right side, depending the granularity of the temporal information contained in the expression. For example, the expression *2009. aasta maikuus* ‘in

[5]We use the tag name TIMEX3 in the examples, although it must be noted that our annotation format diverged from the TIMEX3 standard at some points, and therefore, our tool uses the name TIMEX instead. We will discuss the divergences in more detail in Subsection 2.2.3.

May 2009' contains only year and month granularity temporal information, and is normalised as "2009-05", while the expression *2009. aasta 20. mai hommikul kell kuus* '20th May, 2009, at 6.00 am' contains both date and time information, and can be normalised up to hour granularity as "2009-05-20T06".

Date and time formats are also extended with special tokens to allow representation of expressions denoting Before Current Era (BCE) dates, seasons, quarters, halves of years, weekends, and parts of days. For example, the expression *2009. aasta suvel* 'in summer 2009' is normalised as "2009-SU" (SU = *summer*), and the expression *2011, 6. mai õhtul* 'in the evening of the 6th of May 2011' is normalised as "2011-05-06TEV" (EV = *evening*).

The annotation of temporal intervals involves separately marking up the beginning and ending points of the interval, and the duration of the interval. One can distinguish expressions denoting intervals with explicit endpoints (2), and expressions denoting intervals with implicit endpoints (3):

(2) Maailma Kirikunõukogus töötas Tutu aastail 1972–1975.
Maailma Kirikunõukogus töötas Tutu
<TIMEX3 tid="t2" type="DATE" value="1972">aastail 1972 </TIMEX3>–
<TIMEX3 tid="t3" type="DATE" value="1975">1975. </TIMEX3>
<TIMEX3 tid="t4" type="DURATION" value="P3Y" beginPoint="t2"
endPoint="t3" />
Tutu worked at the World Council of Churches between 1972–1975.

(3) Flaami Bloki populaarsus on märkimisväärselt kasvanud viimase kümne aasta jooksul.
Flaami Bloki populaarsus on märkimisväärselt kasvanud
<TIMEX3 tid="t5" type="DURATION" value="P10Y" beginPoint="t6"
endPoint="t0" >viimase kümne aasta jooksul. </TIMEX3>
<TIMEX3 tid="t6" type="DATE" value="1990" />
The popularity of the Flemish Block has grown remarkably during the last ten years.

In Example 2, endpoints of the interval are explicit from the expression (*1972–1975*), and the corresponding (implicit) duration is added as an empty TIMEX3 tag (the tag with the index *t4*). Duration refers to the beginning and end points as the attributes *beginPoint* and *endPoint*. In Example 3, duration of the interval is explicit from the expression (*during the last ten years*), and attributes *beginPoint* and *endPoint* refer to (implicit) endpoints of the interval. The beginning point *t6* is added as an empty TIMEX3 tag, and endpoint *t0* refers to the document creation time (which was, for the given text, 2000-10-10).

There are three ways to represent fuzzy or unspecified temporal semantics. First, if the temporal expression contains a word or phrase that modify the temporal semantics (such as in the expressions *1990ndate alguses* 'at the beginning

of the 1990s', *rohkem kui 5 päeva* 'more than 5 days'), the modifier is specified in the attribute *mod*, as in the following example (4):

(4) 1999. aasta lõpul esilinastunud Macbeth on teeninud meedias erakordset tähelepanu.
<TIMEX3 tid="t7" type="DATE" value="1999" mod="END">1999. aasta lõpul </TIMEX3>
esilinastunud Macbeth on teeninud meedias erakordset tähelepanu.
Macbeth, which premiered at the end of 1999, has gained an exceptional attention in the media.

Second, if the temporal expression only refers to past, present, or future, without specifying any calendric information (e.g. *hiljuti* 'recently', or *praegu* 'now'), the attribute *value* is filled with a token exemplifying the reference (PAST_REF, PRESENT_REF, or FUTURE_REF). Third, if the temporal expression hints of temporal information (some specific granularity), but does not specify details, such as the date expression *ühel päeval* 'in one day', and the duration expression *aastateks* '(for) years', the placeholder X is used to indicate gaps in the attribute *value* (e.g. 'in one day' is normalised as "XXXX-XX-XX" and '(for) years' is normalised as "PXY").

The semantics of recurring times (SET expressions) are based on the duration format: the attribute *value* specifies a duration – the period covering the recurrence – and the attribute *freq* specifies the frequency of the recurrence: the number of recurrences during the period (an integer value); the frequency value is also augmented with a temporal granularity, or the placeholder X, if the granularity is unknown (Knippen et al., 2005). Example (5):

(5) Seda vahemaad läbib Heimonen vähemalt neli korda päevas.
Seda vahemaad läbib Heimonen
<TIMEX3 tid="t8" type="SET" value="P1D" freq="4X"
mod="EQUAL_OR_MORE">vähemalt neli korda päevas. </TIMEX3>
Heimonen passes this distance at least four times a day.

Note that the duration-based representation of recurrences in TIMEX3 has its limitations: as Mazur (2012) argues, "neither TIMEX2 nor TimeML express the semantics of set expressions sufficiently well to make these schemes applicable to all set expressions". At the end of the following subsection, we propose to augment the set representation possibilities of TIMEX3 with the possibilities of TIMEX2; however, the ambitious goal of having a wide coverage on all set expressions remains out of the scope of this work.

2.2.3 Specifics of Estonian Temporal Expression Annotation

In order to bring out the specifics of Estonian temporal expression annotation, we take The Grammar of Estonian (Erelt et al., 1993) as a starting point. The

Grammar of Estonian discusses temporal expressions (temporal adverbials) from the point of view of syntax[6]and this point of view can be contrasted against our TimeML-based view of temporal expressions.

The classification of temporal expressions in TimeML partially overlaps with the classification of temporal adverbials in The Grammar of Estonian (Erelt et al., 1993), which distinguishes the following types of temporal adverbials:

- *Occurrence times*, such as *eile* 'yesterday', *viimasel ajal* 'in recent times', *1967. aastal* 'in the year 1967', or *pingerikkal ajastul* 'in a stressful era';
- *Time boundaries*: start and end times of events/states, such as *kella kolmest* 'from three o'clock', *kongressist alates* 'from (the time of) the congress', or *1975. aastani* 'until the year 1975';
- *Durations*, such as *kaks päeva* 'two days', or *terve koolivaheaeg* 'the whole school holiday (period)';
- *Recurring times*, such as *kaks korda päevas* 'twice a day', *sageli* 'frequently', or *haruharva* 'very rarely'.

The most important differences between the two classifications are: 1) The Grammar of Estonian also considers expressions with no explicit calendric information (such as 'in a stressful era') and event-denoting expressions (such as 'from the congress') as temporal adverbials; 2) The Grammar of Estonian distinguishes *time boundaries* as a separate class of expressions, while TimeML merges time boundaries with point-referring expressions (DATEs and TIMEs); 3) The Grammar of Estonian does not make a distinction between different granularity point-referring temporal expressions (DATEs and TIMEs), all point-referring temporal expressions are considered as *occurrence times*.

In order to clarify the first difference between *temporal adverbials* in The Grammar of Estonian and *temporal expressions* in our research, and to support the practical concern that annotated expressions should also be normalisable in the given (TIMEX3-based) annotation format, we define two rough criteria for deciding whether a temporal expression is markable[7]:

[6]We do not consider *temporal expressions* and *temporal adverbials* to be exactly the same linguistic category, although the categories seem to have a large overlap. Temporal adverbials are defined from the point of view of syntax: they are sentence constituents that modify the meaning of the main verb or the sentence. Temporal expressions, however, are not restricted to the syntactic role of an adverbial, e.g. they can also modify the meaning of a single constituent (a word or a phrase) in the sentence, such as the expression *today* in the phrase *today's meeting*.

[7]An expression satisfying at least one of the criteria is considered markable.

1. A markable expression should contain temporal information of *year, month, week, day, hour* or *minute* granularity. This leaves out expressions with vague calendric semantics (such as *omal ajal* 'at its own time', or *sageli* 'frequently'), domain-specific expressions (such as *sel hooajal* 'in this season', or *terve koolivaheaeg* 'the whole school holiday (period)'), and event-denoting expressions (such as *pärast kooli lõpetamist* 'after graduation');

2. A markable expression should contain a reference to the *past, present* or *future*, anchored to the creation time of the document (like expressions *praegu* 'now' and *hiljuti* 'recently' in their prototypical usage) or to some other markable temporal expression (like expressions *varem* 'earlier' and *hiljem* 'later' in their prototypical usage). This leaves out expressions anchored to event mentions.

Considering the second difference between *temporal adverbials* and *temporal expressions*, we note that distinguishing *time boundaries* as a separate class of temporal adverbials is likely owning to Estonian morphology. Estonian has distinct morphological cases (semantic cases) that are regularly used for marking 'start and end times'. The elative case (word suffix *-st*) marks that the temporal noun refers to a 'starting time' (e.g. *kolmapäevast* 'from Wednesday'), and the terminative case (word suffix *-ni*) marks that the temporal noun refers to an 'ending time' (e.g. *neljapäevani* 'to Thursday'). In English, the time boundary information is expressed by prepositions.

TIMEX3 guidelines for English propose that temporal prepositions should be annotated separately from temporal expressions, and should be marked as temporal SIGNALs. For example, the phrase *from June 7, 2003* is annotated as[8] (6):

(6) <SIGNAL sid="s1">
from
</SIGNAL>
<TIMEX3 tid="t61" type="DATE" value="2003-06-07">
June 7, 2003
</TIMEX3>

The same specification applies to other temporal prepositions "indicating how temporal objects are to be related to each other" (Knippen et al., 2005), such as *at, on, in, until, before*, or *after*. We have chosen not to follow this part of the specification for Estonian, as it would require tagging at the morpheme level (marking word suffixes as SIGNALs). Tagging at the morpheme level would be problematic, as it would introduce a conflict with the TimeML principle that no nested

[8]This example is borrowed from Knippen et al. (2005).

annotations are allowed (as the SIGNALs would be nested inside the TIMEXes[9]).

Estonian counterparts for the frequently used English temporal prepositions *at, on*, and *in* are also morphological case markings, so in most cases we do not follow the idea of decomposing expressions into temporal signals and temporal expressions. This has an implication: expressions that are broken down to multiple expressions in English (as they are separated by temporal prepositions) are annotated as full-length expressions in Estonian. For example, consider the expression *2009. aasta 20. mai hommikul kell kuus* 'on the 20th of May 2009, at 6.00 am' annotated in English (7):

(7) <SIGNAL sid="s2">
on
</SIGNAL>
<TIMEX3 tid="t71" type="DATE" value="2009-05-20">
20th May 2009,
</TIMEX3>
<SIGNAL sid="s3">
at
</SIGNAL>
<TIMEX3 tid="t72" type="TIME" value="2009-05-20T06">
6.00 am
</TIMEX3>

and the Estonian annotation of the same expression (8):

(8) <TIMEX3 tid="t9" type="TIME" value="2009-05-20T06">
2009. aasta 20. mai hommikul kell kuus
</TIMEX3>

In terms of representing the semantics of temporal expressions, our annotation format mostly follows the TimeML TIMEX3 standard, though there are a few exceptions. First, we note that the ISO duration format used for expressing the semantics of SET expressions (as was described in Subsection 2.2.2) lacks the means of expressing the semantics of specific recurring time points, such as the expressions *teisipäeviti* 'on Tuesdays', and *talviti* 'in winters'. In such cases, we use the TIMEX2 (Ferro et al., 2005) way of expressing the semantics of recurrences: the attribute *value* contains a date, where the (finest) recurring granularity is marked with a concrete value, and all other granularities are marked with the placeholders X. Thus, 'on Tuesdays' is normalised as

[9]An alternative to using nested annotations would be resegmentation of the text, so that the temporal signal suffixes are separated from the nouns, and thus can be annotated separately. However, from an automatic annotation perspective, resegmentation would add additional complexity to interpreting the output of the temporal tagger, i.e. one would need to take into account that the text segmentations in the input and output of the tagger would not match. We wanted to avoid introducing such complexities.

"XXXX-WXX-2", 'in winters' is normalised as "XXXX-WI", and 'in Januaries' is normalised as "XXXX-01". Second, we note that although TIMEX2 and TIMEX3 standards have a means for expressing semantics of the first and the second half of a year (labels H1 and H2, e.g. *value*="1999-H2" stands for 'the 2nd half of 1999'), expressing the same semantics on other granularities, such as months and weeks, is not supported. We therefore add the temporal modifiers FIRST_HALF and SECOND_HALF to convey this meaning. For example, the expression 'the second half of April 2004' will be normalised to *value*="2004-04" and *mod*="SECOND_HALF".

2.3 Automatic Temporal Expression Extraction and Normalisation in Estonian

In this section, we describe our approach for automatic temporal tagging of Estonian. The creation and evaluation of the system has been described in Orasmaa (2010) and Orasmaa (2012), the former publication also gives a detailed technical description of the system. In this work, we describe the system at a general level, keeping the technical details to a minimum.

The first subsection gives a general overview of the system, the second discusses the extraction process in detail, and the third describes our temporal expression normalisation strategies. The last subsection gives an overview of the evaluation of the system.

2.3.1 General overview of the system

Temporal tagger is a program that takes a natural language text and the creation time of the text (i.e. the speech time) as an input, and outputs a text annotated with temporal expressions.

We took a language-specific approach to temporal tagging and designed a temporal tagger that addresses the characteristics of Estonian: i.e. rich morphology and flexible word order. The rich morphology of Estonian poses a challenge for describing temporal expression patterns: one needs to consider that nouns and adjectives decline in 14 morphological cases, and about 45 % of the word forms in Estonian texts have more than one morphological interpretation (Müürisep et al., 2003) (the ambiguity can depend on the text genre). The flexible word order adds an additional challenge: words inside the temporal expression can be reordered, e.g. the expression *täna kell 8 hommikul* 'today at 8 o'clock in the morning' has at least three plausible word orderings: *täna kell 8 hommikul*, *kell 8 täna hommikul*, and *täna hommikul kell 8*.

Our rule-based temporal expressions tagger uses two sets of rules: basic extraction rules and composition rules. A ***basic extraction rule*** consists of a *phrase pattern* that describes a set of temporal expression phrases[10], and a sequence of *normalisation instructions* that need to be applied in order to normalise the semantics of the expressions. A ***composition rule*** specifies how extracted consecutive temporal expression candidates are joined into longer temporal expressions candidates.[11]

Basic extraction rules are used to extract small phrases with an unchangeable word order, for example *järgmisel neljapäeval* 'on next Thursday', or *eelmisel kuul* 'in the last month'. Usually these phrases represent a temporal meaning in terms of a single time granularity (e.g *day*, or *month*), or act as modifiers of temporal expressions (e.g. *teisel poolel* 'on second half (of)'). Composition rules are used to direct how expressions extracted by basic extraction rules are joined into longer temporal expressions (e.g. 'on next Thursday' + 'at 10 am'). Such a decomposition of rules allows one to lessen the number of extraction rules used. Otherwise, we would have to define separate extraction rules for capturing multi-granularity expressions (such as *eelmise aasta 25. aprillil* 'on the 25th of April last year'), and for capturing varying word order within these expressions[12](e.g. *25. aprillil eelmine aasta*).

After the extraction phase, each extracted temporal expression candidate is associated with a sequence of ***normalisation instructions***. We distinguished three types of normalization instructions: 1) anchoring instructions; 2) calendar arithmetic instructions; and 3) markup changing instructions. By default, the input date of normalisation is the creation date of the text and relative temporal expressions like *täna* 'today', or *eelmisel aastal* 'in the last year' are solved according to this date. Anchoring instructions are used to override this default, allowing one to anchor a temporal expression to a preceding temporal expression in the text. Calendar arithmetic instructions are used to change the input date either by direct manipulation (e.g. setting values in calendar representation, or adding calendar units to the input date and calculating a new date), or by using heuristic calculation methods (discussed in more detail in Subsection 2.3.3 starting from page 43). Markup changing instructions were used to directly manipulate attribute values in the TIMEX annotation (e.g. to set value of the *mod* attribute).

Normalisation instructions also need to take account the extraction context:

[10]Here, the notion *phrase* is used in the sense of 'a sequence of words, consisting of one or more words, without any specification of syntactic relations between the words'.

[11]The overall design of the system (how to structure temporal expression extraction and normalisation into different types of rules) draws inspiration from Berglund's work (Berglund, 2004).

[12]However, our current decomposition of rules does not solve all word order problems, e.g. separate extraction rules still need to be defined if the word order varies within a single granularity expression, such as *viimased kolm kuud* and *kolm viimast kuud* 'the last three months'.

whether a single phrase detected by a basic extraction rule is normalized (e.g. *esmaspäeval* 'on Monday'), or whether a multi-phrase expression created by a composite rule is normalized (e.g. *järgmise nädala* + *esmaspäeval* 'on next week's' + 'Monday'). A context sensitive trigger can be specified for a normalization instruction, so the instruction is executed only in specific contexts.

A few notes on technical implementation. Compared to finite state transducer approaches to temporal tagging, which transform the input text (step by step) into the output text with inline timex annotations (Mani and Wilson, 2000b; Bittar, 2009), our temporal tagger does not directly transform the input text. Instead, an inner representation of the input text is created[13], the processing (extraction and normalisation) is made on this representation, and at the final stage, the input text is augmented with the processing results, producing an output with necessary annotations. This allows one to make the processing independent of the input format (e.g. currently, the program supports a morphologically analysed text input (the format exemplified by Kaalep and Vaino (2001)), and a JSON version of the same format), and allows one to make fixes in the input without altering the output (e.g. the tokenisation of punctuation is slightly normalised in the inner representation to improve matching of rules).

2.3.2 Extraction of temporal expressions

Before the extraction of temporal expressions, the input text must be preprocessed: segmented into sentences, words, and morphologically analysed and disambiguated (using Filosoft's morphological analysis tools (Kaalep and Vaino, 2001)). The program takes the preprocessed text as an input, and converts it to an inner representation. Then, in the first (preparatory) processing step, the text is analysed for features that are required in the following phases (e.g. locations and semantics of numeral phrases, and grammatical tenses of verbs).

For the initial extraction of temporal expression candidates, ***phrase patterns*** within basic extraction rules are used. A phrase pattern is implemented as a simplified finite state machine working at the token (word) level. Words of the text are fed to the phrase pattern one at a time, and if the pattern reaches a full phrase match at some text position, a ***temporal expression candidate*** is extracted.

More formally, a phrase pattern is a string consisting of substrings describing words (***word templates***), and operator symbols expressing relations between words. Only two operators are supported: the "concatenation" operator (whitespace, e.g. $A\ B$ indicates that a word matching the template A is consecutively

[13]More specifically: word tokens along with their morphological analyses (including multiple variants in cases of morphological ambiguities), and additional information (about sentence boundaries, grammatical tenses, and numeral phrases), are encapsulated as Java objects.

followed by a word matching the template B in the text), and the skipping operator (question mark, e.g. $A?$ indicates that at a given position, the template A can be matched or skipped). The pattern must have at least one word template without the skipping operator. From the string of the phrase pattern, a simplified non-deterministic finite automaton is built.

A word template can describe a group of words by: 1) a regular expression; 2) a word lemma; 3) a numeral phrase template; 4) a word class. We use regular expressions mostly to describe numerical parts of expressions or words with few alternative variants. Our primary way for describing words was by using lemma-based descriptions, which can capture all the possible morphological variants of a word. Note that this is in contrast to state-of-the-art temporal expressions tagging in English (Strötgen and Gertz, 2010), where mainly regular expressions are used for the description of words and phrases. A numeral phrase template is used to capture words and phrases denoting numbers (such as *kaksteist* 'twelve', or *kahekümne esimesel* 'at the twenty first'), and it can be restricted to match numbers of a specific type (cardinals, ordinals, and fractions) and numbers from a specific range.

A word class allows one to combine multiple other word templates into a list (allowing no recursion, so only word templates 1–3 can be used). A match on the word class is triggered if any of the listed templates matches the input word. Word classes can be used to describe paradigmatic elements of an expression, e.g. weekday names are represented as a word class *NADALAPAEV*, which is defined as a list of corresponding word lemmas: [*esmaspäev, teisipäev, kolmapäev, neljapäev, reede, laupäev, pühapäev*].[14]

After initial temporal expression candidates are extracted by phrase patterns, redundant candidates are detected and removed. This step consists of two phases: 1) removing the overlapping candidates; and 2) removing the candidates matching negative patterns. Removing the overlapping candidates is triggered if a longer candidate totally overlaps a shorter one. For example, from the expression *30. jaanuar* '30th of January' two candidates are extracted: the candidate covering the full expression (*30. jaanuar*) and the candidate covering only the month name (*jaanuar*), and the latter one gets deleted because it is totally overlapped by the longer candidate. A phrase pattern can be associated with a negative pattern (a sequence of regular expressions describing words), which outlines the context of a phrase in which no temporal expression candidate should be extracted. For example, negative patterns are used to restrict the extraction of month names (such as *August*) from person names (such as *Karl August Hermann*, or *August Mälk*).

After redundant temporal expression candidates have been removed, consecutive candidates are joined into longer phrases. The joining is performed at two

[14] [*Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday*]

levels: first at the phrase level, and then at the interval level. The **phrase level** joining is directed by the composition rules, which describe how expressions with different temporal granularities can be joined (e.g. *järgmise nädala* + *esmaspäeval* 'on next week's' + 'Monday'), or how phrases modifying the semantics of the expression can be added (e.g. *2009. aasta* + *lõpus* 'at the end of' + '2009'). The ***interval level*** joining is based on the results of two previous extraction steps (the extraction of the candidates and the phrase level joining), and is guided by built-in language-specific heuristics. The first heuristic joins two expressions into a temporal interval expression, if the first expression is in elative case and the second expression is in terminative case, e.g. *reede**st*** + *pühapäeva**ni*** 'from Friday' + 'to Sunday'. The second heuristic attempts to detect whether the interval expression can be formed if the numeric part of the expression is expanded from the left or right end of the phrase, e.g. *aastatel 2007* + ***kuni 2009*** ('in the years 2007' + 'to 2009'), ***1.-*** + *3. juunil* ('(from) the 1st to' + '3rd of June').

A few notes on technical implementation. Our implementation keeps the extraction process distinct from the normalisation process. While each temporal expression candidate becomes associated with a list of normalisation instructions during the extraction process, all of these instructions are executed later, during the normalisation. This is different from finite state transducer approaches, which transform the input text step-by-step, and add intermediate inline annotations to the text during the process (as described by Mazur (2012)). We will add annotations to the text only after the normalisation step, when all the processing is done. This allows us to keep the processing independent from the concrete input/output format.

Appendix B provides more detailed examples of the rule format, and the rules used for temporal expression extraction, phrase composition, and normalisation.

2.3.3 Normalisation of temporal expressions

Considering the normalisation of temporal expressions, date and time expressions can be divided into absolute and relative temporal expressions:

- ***Absolute expressions*** can be normalised independently from the context, and contain *year* granularity explicit temporal information, e.g. *12.03.2011*, *20. mai 2009* '20th of May 2009'. Normalisation of such expressions requires simply rewriting the explicit calendric information into the annotation format;

- ***Relative expressions*** are context-dependent expressions that need to be normalised based on some other time point, possibly by using calendar arithmetic. We distinguish two types of relative expressions:

– ***expressions that can be resolved based on the text creation time***, such as deictic expressions (*täna* 'today', *järgmisel nädalal* 'in the next week', *mullu* '(in) the last year'), and expressions lacking *year* granularity explicit temporal information (*24. aprillil*, 'on the 24th of April', or *mai lõpus* 'at the end of May');

– ***expressions that need to anchored to some other temporal expression*** for normalisation of semantics, e.g. in the sentence *1996. aastal oli tulu 5% väiksem kui aasta varem.* 'In 1996, the profit was 5 % smaller than the year before', the expression 'the year before' needs to be anchored to the expression '1996' in order to resolve the date correctly.

By default, creation time of the text (usually given as a full date, i.e. gives explicit *year, month* and *day* granularity information) is taken as the base date for normalising date and time expressions. The goal of the normalisation process is to modify the base date (by applying normalisation instructions) until a date corresponding to the semantics of the expression is obtained.

The simplest normalisation instruction is the ***set*** instruction, which is used to set a new value to a calendar field[15], rewriting the old value. Set instructions are mostly used for describing the semantics of absolute temporal expressions.

Add instructions change the value of the calendar field by adding or subtracting number of temporal units. These instructions are required for normalising date/time expressions with an explicit direction (e.g. *eelmisel nädalal* 'last week', *järgmisel nädalal* 'next week', *viis aastat tagasi* 'five years ago', or *viie aasta pärast* 'after five years'). However, all expressions with explicit direction cannot be described in terms of adding or subtracting instructions. For example, the expression *eelseisval reedel* 'on the forthcoming Friday' can refer to both 'this week's Friday' and 'next week's Friday', so one cannot give an exact rule for whether the calendar field *week* needs to be altered or not.

Schilder and Habel (2001) studied the task of temporal tagging on German and proposed an alternative strategy to complement adding and subtracting instructions: ***seek*** instructions.[16] A seek instruction takes a direction (past or future), and a required value of the calendar field as inputs, and finds a date that is: 1) nearest to the base date in the given direction; and 2) has the required value of the calendar field. For the previous example ('on the forthcoming Friday'), one can define a seek instruction that finds a future Friday nearest to the base date. We

[15]Calendar field: a subpart of the date/time representation, such as *year, month, week, day of week, day of month, hour, minute*. We adopt this term from the Joda Time library (`http://www.joda.org/joda-time/`, last visited 2015-05-23), which was our basis for implementing calendar arithmetic instructions.

[16]Schilder and Habel refer to the strategy as "the strategy of the gliding time window".

mostly use seek instructions to normalise semantics of phrases that express the direction with present participles, such as *eelneval reedel* 'on the preceding Friday', *eelseisval nädalavahetusel* 'on the upcoming weekend', or *tuleval kevadel* 'on the coming spring'.

Normalising temporal expressions with a direction can also be problematic when the expression is used at the beginning or at the end of the mentioned time cycle. For example, if the expression *viimasel nädalal* 'last week' is used at the end of the week cycle, i.e. on Sunday, it is not clear whether it refers to 'the week before this week' or to 'this week'. However, if the expression is used at the beginning of the week, on Monday, it likely refers to the 'the week before this week'. How semantics of temporal expressions are interpreted by humans on such border cases needs further investigation, which was outside the scope of this work.

Solving the semantics of expressions without explicit direction indication[17] can be considered as a separate subproblem of normalisation. Two strategies have been proposed for solving this problem. The first strategy involves finding a verb nearest to the temporal expression from the sentence, and deciding the direction based on the tense of the verb. This strategy has been employed for resolving English temporal expressions by Mani and Wilson (2000a) and Strötgen and Gertz (2010). An alternative strategy (Baldwin, 2002) involves creating a window of unique calendar field values, which is centred at a base date, and picking the sought value from this window. This strategy can be illustrated best by the example of resolving day of week expressions (9): given that the base date (the *speech time*) is 2010-03-28 (*Sunday*), and we want to resolve the semantics of the expression *teisipäeval* 'on Tuesday', we make a window of seven unique weekday names (around the base date, which is marked by *):

(9)

Thu	Fri	Sat	Sun	Mon	Tue	Wed
25	26	27	28	29	30	31
			*			

and pick Tuesday from the window (2010-03-30) as the solution. If the strategy is applied for resolving month or day of month expressions, one month will always be outside the window, so a fall-back strategy is required (e.g. one can pick the missing month as the corresponding month of the year of the base date).

[17] More specifically, we consider here relative *day of week, month,* and *day of month* expressions, which do not give any clue from which direction (past, present, or future) the mentioned day of week, month, or day of month needs to be sought. This group of expressions contains single-word expressions (such as *teisipäeval* 'on Tuesday', or *märtsis* 'in March') and extensions of these expressions that contain finer granularity temporal information (e.g. *teisipäeva õhtul* 'on Tuesday evening') or a temporal modifier (e.g. *märtsi lõpus* 'at the end of March').

These two strategies have been tested in resolving English day of week expressions by Mazur and Dale (2008). The authors found that on a corpus consisting of transcribed texts, newswire, newsgroup and weblog texts[18], Baldwin's 7-day window gave the correct solution in 94.28 % of cases, and the verb's grammatical tense gave the correct solution in 92.64 % of cases. Author of the current work has compared these strategies on a corpus of Estonian news articles (more specifically: on a corpus of daily news) (Orasmaa, 2010), and has found that for solving day of week expressions, the best results were obtained with the verb tense heuristic (accuracy of 90.2%), while for resolving month and day of month expressions, Baldwin's window showed the best performance (accuracy of 94.7% for month expressions, and 92.7% for day of month expressions). While current normalisation rules are configured based on these results, the problem itself requires further investigation on a more diverse corpora.

Although most of the date and time expressions can be resolved based on the creation date of the text (an assumption that seems to hold at least in the news domain), some words in the expression could indicate that a different anchoring strategy should be used. According to Negri and Marseglia (2004), the presence of the keywords 'following', 'previous', 'same', 'that', 'before', 'later' in an English temporal expression indicates that the expression needs to be anchored to a previously mentioned date expression for correct normalisation of the semantics. The authors applied a restriction on picking anchor expressions: the granularity of the anchor needed to be the same or finer than the granularity of the anchored expression. For example, the expression 'three days later' could be anchored to the expression 'on Monday' (because granularity is the same), but cannot be anchored to the expression 'on this month' (because the granularity is coarser).

In the current work, we also use heuristic strategies for anchoring Estonian date and time expressions. These strategies were based on the author's experience of studying temporal tagging in the news domain, but the effectiveness of these strategies has not yet been evaluated separately. Expressions containing the keywords *varem* 'earlier', *hiljem* 'later', *sama* 'same', *too* 'that' are anchored to a previous temporal expression, without any granularity constraints. Time expressions that do not have any coarser granularity (e.g. *day, week*) information are anchored to a previous *day* granularity temporal expression. In all cases, the preceding three sentences are examined for a suitable anchor candidate.

[18] Authors used the ACE 2005 Training Corpus: `https://catalog.ldc.upenn.edu/LDC2006T06`, last visited: 2015-05-27

2.3.4 Evaluation

The Corpus

In the previous work (Orasmaa, 2010), we developed and evaluated our system on a corpus of newspaper texts, but did not analyse how the system might work on other text genres or on different subgenres of news. In order to bridge this gap, we evaluated the system on a corpus where different text genres were distinguished, using data from the Reference Corpus of Estonian (Kaalep et al., 2010).

A subpart of the Reference Corpus–the corpus of Written Estonian–consists of approximately 250 million words: the newspaper texts make up 84% of the corpus, Estonian parliamentary transcripts 5.9%, legalese texts 4.7%, fiction texts 2.9% and scientific texts 2.5%. The size of the corpus and different text varieties it contains makes it rather difficult to choose a representative subcorpus for the evaluation. In this study, we focused on homogeneous texts, so we left out scientific and fiction corpora, which contain rather heterogeneous texts.

In case of newspaper texts, we choose two daily newspapers (*Postimees, Eesti Päevaleht*) and one weekly magazine (*Luup*) as a source for our corpora. This allowed us to further distinguish six subsections: *Local news, Foreign news, Opinions, Sport, Economics* and *Culture*. In order to evaluate the system on historical texts, we also chose articles focusing on historical topics (e.g. archaeology, and international relations in previous centuries) from the popular science magazine *Horisont*. From the domain of legalese texts, we only chose a subcorpus of Estonian laws.

We created an evaluation corpus of approximately 70,000 word tokens by choosing texts randomly from the corresponding subsections of the Reference Corpus of Estonian. Note that what was considered as a *text* varied in the different subgenres: in news, one news article was considered as a text, in parliamentary transcripts, a discussion of one item from the daily agenda was considered as a text, and in legalese texts a complete act of law was considered as a text.

Statistics of the final corpus are presented in Table 2.1. We applied automatic tagging of temporal expressions on the corpus and corrected manually the results of the automatic annotation (corrections were made by one person: the author). After the manual correction of the annotations, there were 1900 temporal expressions in the corpus. Note that this count excludes empty TIMEXes (implicit endpoints and implicit durations, like in Examples 2 and 3 in Section 2.2.2), which were left out in order to simplify the evaluation.

Evaluation Results

The system's performance regarding extraction was measured by standard information extraction measures (precision and recall):

Subcorpus	Texts	Tokens	% of tokens	Temporal expressions
Newspaper articles				
Local news	31	17271	24.6%	553
Foreign news	15	7724	11.0%	155
Opinions	15	7024	10.0%	130
Sport	16	6981	9.9%	160
Economics	12	5205	7.4%	177
Culture	12	5102	7.3%	142
Historical articles	6	7088	10.1%	229
Estonian parliamentary transcripts	3	6950	9.9%	109
Estonian law texts	3	6864	9.8%	245
Total	113	70209	100.0%	1900

Table. 2.1: Statistics of the evaluation corpus: text and token counts, proportions of tokens in each subcorpora, and temporal expression counts.

$$precision = \frac{\text{the number of correctly extracted expressions}}{\text{the number of all extracted expressions}} \tag{2.1}$$

$$recall = \frac{\text{the number of correctly extracted expressions}}{\text{the number of all expressions in the text}} \tag{2.2}$$

The performance of normalising semantics (determining *type* and *value* attributes) was measured by the precision:

$$norm\text{-}precision = \frac{\text{the number of correctly assigned attribute values}}{\text{the number of all assigned attribute values}} \tag{2.3}$$

The extraction performance was measured in two ways: in the *relaxed way* (only one character overlap between the gold standard and the system result was required for the match) and in the *strict way* (for the match, an exact overlap of two strings was required). The normalisation performance was measured only on correctly extracted expressions, following the relaxed evaluation scheme.

Table 2.2 gives an overview of the extraction performance. Values in the row *Total* were calculated as micro-averages: that is, numbers of correctly extracted expressions, automatically extracted expressions, and expressions of the gold standard were summarized over subcorpora and total measures were calculated from these sums.

Subcorpus	rec (relaxed)	prec (relaxed)	rec (strict)	prec (strict)
Newspaper articles				
Local news	80.8%	100.0%	76.1%	94.2%
Foreign news	87.1%	97.1%	83.9%	93.5%
Opinions	80%	97.2%	73.1%	88.8%
Sport	91.2%	98.0%	86.2%	92.6%
Economics	77.4%	97.2%	71.8%	90.1%
Culture	78.9%	96.6%	73.9%	90.5%
Historical articles	72.5%	98.8%	56.8%	77.4%
Estonian parliamentary transcripts	90.8%	96.1%	88.1%	93.2%
Estonian law texts	86.5%	99.1%	83.3%	95.3%
Total	82%	98.4%	76.1%	91.3%

Table. 2.2: The performance of the system at temporal expression extraction. Precision (*prec*) and recall (*rec*) are reported, following two evaluation schemes: relaxed (one character overlap was sufficient for the match), and strict (an exact overlap was required for the match).

Considering the extraction results, one can say that the current set of rules is biased towards precision: on the whole corpus, the precision was 98.4% (91.3% if the strict evaluation scheme is used), which contrasts with the relatively lower recall of 82% (76.1% in the strict scheme).

This low extraction recall can be explained by some of the choices made during the manual creation of the rules. The focus was on creating rules for expressions with clear semantics, and creating rules for expressions with ambiguous or vague semantics was lower in priority, and so these expressions were frequently missed. Missing ambiguous expressions included short/single-word expressions, such as *aasta* ('a year') and *kuus* ('in a month' or 'six'), single 4-digit years and short dates (e.g. *6.3.* referring to 'the 6th of March'). Missing vague expressions were mostly quantifier expressions with vague quantities, such as *mitu päeva* 'several days', *mõneaastane* 'a few years (old)', or *aasta-paari pärast* 'after a year or two'.

The lowest extraction performance (relaxed recall: 72.5%) was measured in the subcorpus of historical articles. Examining the results showed that the low recall was caused by missing 4-digit year expressions and by misinterpretation of *before common Era* expressions, which were only partially extracted. The rules developed on news texts did not manage to capture the variety of domain specific expressions used in historical texts.

Domain specific expressions were also noted in the subcorpus *Culture*, where short expressions referring to decades (e.g. *seitsmekümnendad*, 'the seventies') were relatively frequent. Missing these expressions seems to be the main cause of decreased recall (relaxed recall: 78.9%) on the given corpus.

The criteria set in 'Specifics of Estonian Temporal Expression Annotation' (Subsection 2.2.3) did not allow us to draw a clear-cut distinction between markable and non-markable expressions. For example, in the subcorpora of *Economics* and *Estonian laws*, domain specific expressions such as *viimase nelja börsipäevaga* ('during the last four market days'), or *eelarveaasta alguseks* ('by the beginning of the budget year'), were also considered as markable expressions according to criterion 1. However, domain specific expressions that frequently occurred in the *Sports* subcorpus, such as *teise poolaja alguses* 'at the beginning of the second half term' or *viimane hooaeg* 'last season' were considered as non-markable according to the criteria, and that can also be seen as contributing to the relatively high extraction recall (relaxed: 91.2%) on the given corpus.

Switching from the relaxed evaluation measure to the strict, recall and precision both dropped by typically 6% or more. A frequent source of errors on determining correct phrase boundaries was the inability to capture different variations of quantifier phrases, such as *kolm ja [pool aastat]*[19] 'three and half years' or *mitu [tuhat aastat]* 'several thousands of years'.

Table 2.3 gives an overview of the normalisation performance of the system. There were not many errors in determining the *type* of a temporal expression, as the relatively high precision (97.2%) indicates. The performance of determining *type* was lowest in the subcorpus of historical articles (87.3%), where it was a direct consequence of misinterpretation of phrase boundaries, e.g. *u. [7500 aastat] e.Kr.* 'circa [7500 years] BC'.

The average precision on normalising semantics of temporal expressions (determining the *value*) was 87.4%. A frequent source of errors was the system's inability to distinguish between general and concrete meanings of a temporal expression. For example, *täna* 'today' has a concrete meaning, which refers to the day of speech time, while its general meaning seems to refer to a broader period: to 'a contemporary time period' or 'nowadays'. Other frequent sources of error were errors caused by misinterpretation of phrase boundaries, and errors caused by using the wrong anchoring strategy. Relatively low normalisation precision (61.8%) in the subcorpus of historical articles was mainly caused by misinterpretation of phrase boundaries.

Highest precisions for *value* were measured in the subcorpora *Estonian parliamentary transcripts* (93.9%), *Foreign News* (93.3%), and *Estonian law texts* (92.9%). In the *Estonian parliamentary transcripts*, dominating temporal expres-

[19] Brackets mark the boundaries of an automatically extracted phrase.

Subcorpus	type (prec[21])	value (prec)
Newspaper articles		
Local news	98.7%	90.6%
Foreign news	97.8%	93.3%
Opinions	99.0%	85.6%
Sport	96.6%	86.2%
Economics	97.8%	90.5%
Culture	98.2%	88.2%
Historical articles	87.3%	61.8%
Estonian parliamentary transcripts	100.0%	93.9%
Estonian law texts	98.6%	92.9%
Total	97.2%	87.4%

Table. 2.3: The performance of the system at temporal expression normalisation. Precision (*prec*) on normalising the attributes *type* and *value* is reported.

sions were relatively easily solvable date expressions (date expressions consisting of a numeric day of the month and the month name, such as *1. juuli* '1st of July'): these expressions made up approximately 42% of all temporal expressions.

In the subcorpus of *Estonian laws*, almost all relative temporal expressions (such as *järgneva aasta 1. veebruariks* 'for the 1st of February next year') were not normalisable as concrete dates or times,[20]so normalisation strategies used for news texts did not work on them. However, because the majority of the expressions in the subcorpus were absolute date expressions (effective dates for laws, and dates associated with referred laws), the errors on relative temporal expressions had little effect on the overall precision.

In *Foreign News*, a relatively large amount of present references occurred (*now*-expressions, such as *praegu* and *nüüd*): approximately 20 % of all expressions (for comparison, in *Local News*, only approximately 12% of all expressions were present references); so relatively high precision could be attributed to the contribution of such easily normalisable expressions.

[20]This phenomenon was also observed by Schilder (2007), who noted that date expressions in laws (in statues and regulations) are more "concerned with normative legal concepts rather than with concrete events", and they "are linked to an event type as a temporal constraint", rather than "to an actual event".

[21]In some cases, the system produced temporal expressions with a missing *type* attribute. Considering these cases, one can also specify recall of assigning *type* in the following subcorpora: Culture 96.4%, History 86.7%, and Estonian law texts 98.1%. Total recall for *type* was 96.9%.

Contrasting with state-of-the-art in English. While our results cannot be directly compared with the state of the art in English temporal tagging, it is possible to draw some rough contrasts between the performance levels in the two languages. For this, we investigated the results reported in the TempEval-3[22] evaluation exercise (UzZaman et al., 2013), and contrasted the best results reported there with our aggregated results (namely, the *total* performances reported in Tables 2.3 and 2.2).

TempEval-3 used the F1-score for aggregating the extraction performance over precision and recall:

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2.4}$$

The extraction scores (*totals* of the scores) of the current study recalculated as F1-scores are 89.45% (relaxed evaluation) and 83.0% (strict evaluation). Our relaxed extraction F1-score was close to the best relaxed score reported for English (90.32%), and the strict evaluation score was even slightly higher than the best strict score for English (82.71%).

For calculating the performance on normalisation, TempEval-3 aggregated the extraction F1-score and the *value* normalisation accuracy, following the formula:

$$\text{value\_F1} = \frac{\text{extraction\_F1\_\%} \cdot \text{value\_accuracy\_\%}}{100} \tag{2.5}$$

The same score calculated for our results (using the relaxed extraction score as a basis) was 78.17%, and this result was also slightly higher than the best result reported for English (77.61%).

These contrasts indicate that the performance levels reported in this study can be comparable with the best results reported for English; however, one should refrain from drawing final conclusions, as the evaluation settings were not fully comparable (due to different domains, and differences in the annotation format).

2.4 Related Work

In this section, we aim to position our work on automatic temporal tagging with respect to other similar works. In the first subsection, we give a brief overview of the other works addressing automatic temporal expression detection and normalisation in Estonian. In the second subsection, we compare our work briefly with the finite state transducer approaches commonly used for rule-based information

[22] TempEval was a series of evaluation exercises (TempEval (Verhagen et al., 2009), TempEval-2 (Verhagen et al., 2010), and TempEval-3 (UzZaman et al., 2013)) that were organised with the goal of evaluating automatic TimeML-based text annotators.

extraction. In the third subsection, we give a short overview of the recent developments in English temporal expression tagging, which could be considered as state-of-the-art in the field.

2.4.1 Related research focusing on Estonian

Compared to English, there has been relatively little work on automatic temporal expression extraction and normalisation in Estonian.

Saue (2007) focuses exclusively on the task of extraction of temporal expressions. She uses a grammar to generate regular expression patterns which are later applied for annotating temporal expressions. The author evaluates the tagger on news texts and fiction texts, and reports precision and recall of 92% on the news texts, and a precision of 83% and recall of 77% on the fiction texts.[23] Compared to our work, the author does not use any language-specific preprocessing, and does not normalise the annotated expressions.

Treumuth (2008) considers temporal information extraction in a specific application domain: in information dialogue systems. He implements a module for a dialogue system that detects and normalises date expressions appearing in the used input. The module uses a rule based approach for temporal expression detection, and formalises the semantics of the expressions as SQL constraints. These constraints are then applied for querying a database of events, with the goal of finding information about events that the user might be interested in.

Treumuth's approach is similar to ours as the author also uses automatic morphological analysis for obtaining word lemmas and builds the extraction rules on the word lemmas, instead of the word forms.

While Treumuth's work is application specific, our work provides a general purpose temporal expression tagger, that could be integrated into different applications, including dialogue systems. Our coverage on temporal expressions is also larger, as in the context of information dialogues, Treumuth mostly focuses on detection of minute, hour, day, and month granularity temporal information (times/dates which lay in relatively close temporal proximity to the speech time), paying less focus to coarser granularity temporal information. He also does not consider durations, fuzzy/unspecified expressions and different normalisation strategies for relative temporal expressions.

The rule-based multi-lingual temporal tagger HeidelTime (Strötgen and Gertz, 2013) has recently also been extended with Estonian specific rules. HeidelTime's rules are grouped by the type of the rule (TimeML's types: *date, time, duration, set*), and each rule consist of three parts: the name of the rule, the extraction part describing the expression, and the normalisation part. While HeidelTime also uses

[23] However, the program is evaluated on a relatively small text corpus, consisting of less than 5000 tokens.

language specific preprocessing (the TreeTagger (Schmid, 1995) is employed for analysing the input text), it's extraction rules mainly employ regular expressions and do not rely much on the linguistic features (except for describing negative patterns of unwanted expressions). This contrasts to our work, where the extraction rules are lemma-based to a large extent. HeidelTime also does not use the composition of temporal expression phrases, so separate rules must be defined for expressions such as 'late Monday', 'morning', and 'late Monday morning'. Normalisation process in HeidelTime consists of several stages. The rules provide explicit part of the normalised value (e.g. the explicit year, like *2010*) and normalise implicit part of the semantics to an intermediate format (e.g. for weekday expressions, the weekday name is normalised to a common form). The full normalisation is achieved at a later stage (in the disambiguation stage), where domain-specific anchoring strategies are used for determining the reference time, and implicit parts of the semantics are resolved, using some linguistic cues (e.g. verb tenses) if necessary. This is different from our approach, where the rules describe the normalisation process to a full extent, including determining the type, applying the anchoring instructions, and performing calendric calculations.

At the time of this writing, HeidelTime's performance had not yet been evaluated on Estonian texts, so we cannot compare the performances of the two taggers.

2.4.2 Finite state transducer approaches

Our extraction approach is similar to approaches used in finite-state-transducer-based information extraction, where the short and most certain phrases are detected firstly, and the results produced by one transducer are used by a next one to create longer phrases, i.e. transducers are organised into cascades (Friburger and Maurel, 2004). Note that while cascades of finite state transducers (FSTs) typically transform the input text (they add intermediate inline annotations to the text (and also remove intermediate annotations in some cases) until the annotated output is reached, as described by Mazur (2012) and exemplified by Bittar (2009)), our approach does not directly transform the input text. Rather, it operates on an inner representation mirroring the input, and adds annotations only during the last stage of the work. This allows us to make the processing, to an extent, independent from the concrete input/output format (e.g. we can switch between the standard and the JSON format input/output). The downside is that our processing is likely slower than in finite state transducer approaches due to the work done on maintaining the inner representation.

Our motivation for using a custom rule format/engine, rather than building on existing finite-state-transducer-based information extraction frameworks, such as the GATE (Cunningham et al., 2002), was that we wanted to try an approach that directly builds on Estonian automatic morphological analysis. If Estonian

morphological analysis will be integrated into FST-based information extraction solutions in the future, lessons learned from our study could be used to build a revised and even more efficient FST-based temporal expression extractor.

2.4.3 Recent research in English

We do not intend to give a complete overview on temporal tagging in English; rather, we point out some of the research directions that are most related to our work, and discuss how our future work could be extended in these directions.

The TempEval-3 (UzZaman et al., 2013) evaluation exercise provides a recent evaluation of temporal tagging of English news domain texts. The results showed that statistical systems performed best at strict matching of temporal expressions (best F1-score: 82.71%), and the rule-based systems were best at relaxed matching (F1-score: 90.32%). Organisers of TempEval-3 also notice "a shift in the state of the art" of timex extraction, reporting that the performance of machine learning approaches has increased to a level similar to the performance of rule-based approaches (UzZaman et al., 2013). All normalisation was still done in a rule-based manner, and the best performance on the *value* F1-score was 77.61% (calculated using the formula 2.5).

The TIMEX annotated corpora created in the current work will also enable future experimentation with machine learning approaches on Estonian texts.

Recent research on English temporal tagging has broadened the scope of analysable texts by introducing new target domains, such as the medical[24] and encyclopaedic domains (Mazur, 2012).

Strötgen and Gertz (2012) argue that temporal tagging is a domain-sensitive task, and that "domain-specific strategies should be applied for extracting and normalising of temporal expressions". The authors analyse four types of texts – news-, narrative-, colloquial-, and scientific-style texts – and find that each text domain is characterised by different types of temporal expressions. They modify the state of the art English temporal tagger HeidelTime in a way that information about the document type can be provided to the tagger, so that the tagger can use domain-specific strategies. For example, in news, colloquial and scientific texts, HeidelTime chooses the document creation time as a reference time for resolving relative (deictic) temporal expressions (such as 'today' or 'in next week'). In narrative texts, such expressions are resolved by taking a previously mentioned expression (with required granularity) as the reference time.

The current work on Estonian temporal tagging has, so far, only considered news-style texts, although the results discussed in Subsection 2.3.4 also indicate the need for introducing domain-specific strategies. As most of the extraction and

[24] See Clinical TempEval `http://alt.qcri.org/semeval2015/task6/` (2015-06-11).

normalisation logic of our tagger is described in an input rules file (an XML file), domain-specific rule files can be created, taking the news domain rules as a basis and adapting these rules to a new domain. We consider this as future work.

Considering the latest research on the annotation format, Mazur suggests that in addition to annotating *global semantics* of the expression – "the temporal value obtained by interpreting the expression in the context of the document in which it is used" – temporal taggers should also annotate *local semantics*, that is, "the semantics of the expression with no context involved" (Mazur, 2012). The author proposes a representation of local semantics called LTIMEX, which "provides a vocabulary for capturing partially specified meaning". For example, the relative temporal expression 'January 3' is represented as 'xxxx-01-03', where the lowercase x marks the part of the semantics that is not explicit and needs to be determined based on the context. If the relative expression must be computed from the reference time, the representation indicates the operation (subtraction, addition) and the granularities changed, e.g. 'tomorrow' is represented as '+0000-00-01' ('a day added to the reference date') and 'last year' as '–0001' ('a year subtracted from the reference date'). The LTIMEX representation also goes beyond the normalisation possibilities of TIMEX3 and TIMEX2, providing, for example, means for normalising ordinally specified expressions, such as 'the third day' (normalised as '3D'), and 'the second year' (normalised as '2Y').

During the current research, we have also encountered the limits of TIMEX3 (and TIMEX2) on expressing full semantics of temporal expressions. For example, while we have created rules for extracting ordinal expressions, such as *kolmandal päeval* ('at the third day'), and relative expressions, such as *järgmisel päeval* ('on the next day')[25], our ways of expressing semantics of these expressions have so far been limited to the general underspecification of date XXXX-XX-XX. LTIMEX-based expression of semantics would allow us to make semantics of such expressions more specific. In addition to expressing semantics of the expressions that cannot be fully normalised based on the context, LTIMEX would probably contribute to overall transparency of the semantics (how was one or another calendric value arrived at?). We consider adapting LTIMEX as one possible future direction in developing our tagger.

2.5 Conclusions

In this chapter, we have introduced the task that can be seen as an integral part of (time-oriented) event analysis – the tagging of temporal expressions – and we

[25]Based on our experience on news corpora, Estonian 'on the next day'-expressions are frequently anchored to some event mentioned in text, and if the event date is unknown, the global semantics of such expressions can only be expressed partially.

have described an Estonian specific approach to solving the problem.

We have provided a brief overview of the development of TIMEX-based annotation formats, followed by a description of our Estonian specific adaptation of the latest commonly used version of the format: TIMEX3. We have contrasted the TIMEX3 temporal expression model with the concept of temporal adverbials in the Estonian grammatical tradition, proposing criteria for distinguishing between markable temporal expressions and the remaining temporal adverbials currently out of the scope of the automatic analysis. We have also outlined the differences between our annotation format and the English-specific TIMEX3 format, which were motivated by Estonian morphology, and concern the extent of expressions.

Considering the characteristics of Estonian, we have developed a rule-based language-specific temporal tagger for the language.[26] Compared to HeidelTime's multilingual approach (Strötgen and Gertz, 2013) to temporal tagging, which also addresses Estonian, our system allows free-word-order-aware composition of rules, and allows one to take advantage of Estonian morphological analysis, both in the extraction phase (allowing lemma-based extraction patterns) and in the normalisation phase (allowing one to consider verb tenses during the normalisation). The system comes with a set of rules developed for processing news texts, and as shown in the evaluation, it obtains relatively high performance on a variety of subgenres of formal written language, such as parliamentary transcripts, and law texts. The system can be used as a modular language analysis component in a range of NLP applications, from fine-grained event and temporal analysis, to analysing documents for temporal similarity (Alonso et al., 2010b).

In parallel with development and evaluation of the tagger, we have also created a manually corrected TIMEX-tagged corpora for Estonian, which can be used for future experimentation with different temporal tagging approaches (e.g. statistical approaches), and for fine-tuning existing taggers. The corpora have been made freely available at: `https://github.com/soras/EstTimexCorpora` (2016-01-04).

2.6 Philosophical Notes

Research on automatic temporal expression tagging shows that the state-of-the-art approaches can produce relatively accurate tagging, with the extraction performances (relaxed F1-scores) being close to 90%, and normalisation performances (the F1-scores on *value* attribute) close to 80% (UzZaman et al., 2013). These

[26]The source code of the tagger is available from `https://github.com/soras/Ajavt` (2016-01-04), and the tagger has also been integrated into the EstNLTK toolkit (Orasmaa et al., 2016).

results seem rather encouraging, and indicate that satisfactory practical performance levels (95% and above) may not be far from our reach.

However, while considering these results, it should be noted that the used TIMEX (TIMEX2, TIMEX3) annotation standards have been, to a large extent, "optimised" for capturing "calendric" temporal expressions, i.e. expressions whose semantics can be modelled in the calendar system (e.g. TIMEX2/TIMEX3 use the ISO 8601 standard which is based on the Gregorian calendar). Thus, the task is fairly focused on the part of human language usage that is already systematic, i.e. based on a well-defined conventional system of time-keeping. When comparing this focused view on temporal expressions with the view on temporal adverbials from The Grammar of Estonian (Erelt et al., 1993), we note that the latter view has a much wider perspective, also covering event expressions (e.g. 'until the congress', 'in the World Cup') at the positions of temporal adverbials. Naturally, the task would be much more difficult if one would aim to capture all temporal adverbials as temporal expressions, and normalisation of many of these expressions would also be beyond calendar based normalisation.

Considering the historical perspective, one could note that (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). Naturally, these interpretations rarely come to mind in contemporary everyday language use, where "days, months and years mostly function as time measuring units", rather than event descriptions (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. What about other, non-calendric events?

The TimeML framework (Pustejovsky et al., 2003a) proposes that a clear distinction between temporal expressions and event mentions in text can be drawn, and that event mentions can be captured (annotated) in a similar systematic manner as temporal expressions. Yet, it seems that the temporal expressions, with their semantics being rooted in the calendar system, represent a relatively unambiguous class of the two, while event mentions, however, are largely unsystematised. There is no apparent convention on how "events in general" are expressed in language, nor is there a larger philosophical/ontological level agreement on how events in general should be modelled. With these considerations in mind, we are approaching the next chapter, which gives an overview of our work on TimeML-based event mention annotation in Estonian.

CHAPTER 3

VERBAL AND NOMINAL PREDICATES AS EVENTS: A CASE STUDY OF TIMEML

3.1 TimeML-based Annotation Formats: An Overview

TimeML is an annotation framework for annotating temporal entities (event mentions and temporal expressions) and marking up temporal relations between these entities (e.g. annotating temporal relations between events and temporal expressions, and annotating the relative temporal ordering between events). The framework was initially created with the goal of improving "the performance of question answering systems over free text", more specifically, to provide support for answering temporal questions, such as *"Is Gates currently CEO of Microsoft?"* or *"When did Iraq finally pull out of Kuwait during the war in the 1990s?"* (Pustejovsky et al., 2003a). Subsequent studies have extended the list of possible applications with automatic summarization and information extraction (setting temporal bounds to the extracted facts) (Mani et al., 2005).

TimeML uses XML tags for expressing temporal semantics. The EVENT tag is used for annotating event mentions, the TIMEX tag is for temporal expressions, and the SIGNAL tag is for explicit relations between temporal objects. Event annotations in the TimeML specification 1.2.1 (Knippen et al., 2005) have two layers: EVENT tags are used for annotating event mentions in the text, and MAKEINSTANCE tags are used for evoking event instances, so that each annotated EVENT is associated with at least one event instance. Distinguishing event instances from mentions is motivated by cases where one event mention refers to multiple actual events, like the *teaching* event in the following example of a TimeML annotated English sentence (from Knippen et al. (2005)):

(10) *John taught on Monday and Tuesday.*

```
John
<EVENT eid="e1" class="OCCURRENCE">
taught
</EVENT>
<SIGNAL sid="s1">
on
</SIGNAL>
<TIMEX3 tid="t1" type="DATE" value="XXXX-WXX-1">
Monday
</TIMEX3>
and
<TIMEX3 tid="t2" type="DATE" value="XXXX-WXX-2">
Tuesday.
</TIMEX3>
<MAKEINSTANCE eiid="ei1" eventID="e1" tense="PAST"
aspect="NONE" polarity="POS" />
<MAKEINSTANCE eiid="ei2" eventID="e1" tense="PAST"
aspect="NONE" polarity="POS" />
<TLINK eventInstanceID="ei1" signalID="s1" relatedToTime="t1"
relType="IS_INCLUDED" />
<TLINK eventInstanceID="ei2" signalID="s1" relatedToTime="t2"
relType="IS_INCLUDED" />
```

Temporal relations between entities are marked with TLINK tags, e.g. in Example 10, the TLINKs convey that one instance of a *teaching* event takes place on (underspecified) *Monday* and another on (underspecified) *Tuesday*. In addition to TLINKs, the specification also contains two other link tags: SLINKs and ALINKs. SLINKs can be used to specify subordination relations between annotated entities (such as relations between perception verbs and their argument events, e.g. *John **saw** the accident*), and ALINKs can be used to specify aspectual relations between annotated entities (such as relations between aspectual verbs and their argument events, e.g. *John **started** to read*).[1]

As the TimeML framework has been widely adopted in research, efforts have been made on improving and extending it. Here, we briefly introduce four of these: ISO-TimeML (Pustejovsky et al., 2010), ISO-Space (Pustejovsky et al., 2011), TERENCE annotation format (Moens et al., 2011), and the GAF annotation format (Fokkens et al., 2013).

ISO-TimeML moves TimeML specification from a concrete, XML-based format, to a more abstract level, so that "rather different formats (e.g. UML diagrams) could be used to represent the TimeML model" (Pustejovsky and Stubbs,

[1] We will provide more detailed examples on the usage of SLINK and ALINK tags in Subsection 3.4.3.

2012). This specification also proposes to move from in-line annotations to stand-off annotations, which are built upon other ISO standardized linguistic annotations (e.g. sentence and word segmentations) (Pustejovsky et al., 2010). Other changes include removal of the MAKEINSTANCE tag, and introduction of a new type of TLINK (MLINK). MLINKs are used for expressing (possibly discontinuous) measurements of event length, such as the duration '4 hours' in the sentence 'John taught for 4 hours on this week'[2] (Pustejovsky and Stubbs, 2012). As TimeML became an international standard, annotated resources for languages other than English were also created, e.g. ISO-TimeML annotated corpora have been created for French (Bittar, 2010), Chinese (Xue and Zhou, 2010) and Italian (Caselli et al., 2011).

While ISO-TimeML focuses on the temporal information expressed in natural language, ISO-Space (Pustejovsky et al., 2011) addresses spatial information and its relation to event mentions. ISO-Space aims to capture both static spatial information (locations mentioned in texts, and their relations to other locations/regions), and spatiotemporal information (motion events indicating a change of location, and non-motion events that participate in some spatial relation). As the two standards have been designed to be interoperable (Pustejovsky et al., 2011), ISO-TimeML-based annotations can be extended with ISO-Space annotations to provide information on the spatial dimension of events.

The TERENCE project (Moens et al., 2011) employed TimeML framework in order to provide a machine-readable representation of the semantic structure of children's stories. The aim was to support the creation of story-based games that can "help to improve reading comprehension for low-literacy and deaf children" (Moens et al., 2011). In addition to TimeML-based temporal expression, event mention, and temporal relation annotations, the TERENCE format also provides means for annotating entity mentions (entities that play important roles in the development of the story, e.g. persons, animals, artifacts, and locations), event-participant relations, and coreference relations between story elements (both entity and event coreference).

The Grounded Annotation Framework (GAF) proposes a representation that can combine event information from different sources, both textual and extra-textual (e.g. web links, videos, or pictures), and "interconnect different ways of describing and registering events" (Fokkens et al., 2013). The motivation behind using multiple sources is that event information from a single source (e.g. text) is often incomplete, and might require complementing information from other sources. For annotating textual event mentions, GAF uses the TERENCE annotation format as a basis, which allows one to draw a distinction between textual event mentions and the formal representation of events (event instances). Event

[2]The *teaching* event might have been spread across different days on given week.

instances belong to the semantic layer, where an RDF[3] schema based model is used for "describing events, and related instances such as the place, time and event participants". The framework allows one to trace the provenance of the event information (from which source the information was retrieved), the temporal validity of the information (when the information was retrieved) and enables the storing of conflicting information (Fokkens et al., 2013).

In conclusion, TimeML is a framework widely adapted in developing general domain event analysis formalisms, and TimeML-based annotations can be, in principle, extended in several directions (e.g. by providing spatial annotations, or by linking events with extra-textual information). In the following section, we will discuss in more detail the theoretical considerations and practical problems related to TimeML.

3.2 The TimeML Event Model

3.2.1 Theoretical background of TimeML

Here, we discuss the theoretical work that is most often considered as having influenced / motivated the development of TimeML, and the event model within the framework. The list of works discussed here is by no means complete, but rather an extract covering works most related to the current study. A comprehensive listing of theoretical works related to TimeML can be found in the book "The Language of Time: A Reader" (Mani et al., 2005).

Essentially, events in TimeML are considered *temporal entities*: the main interest lays in associating events with time instants or time intervals, so that events can be compared temporally and ordered in a timeline. The focus on temporal semantics of events is also motivated by the theoretical well-foundedness of temporal relations: Allen's interval algebra (Allen, 1983, 1984) covers all possible relations between temporal intervals, and also provides a basis for inferring new relations from existing ones. This means that even if the temporal relation annotation is incomplete, automatic inference can be (up to a certain extent) applied to find missing relations. A similar mechanism can also be used to verify the consistency of temporal relations (e.g. to check for time loops in a set of temporal intervals).

However, the temporal inference requires that at least some of the temporal relations associated with the events have already been identified, based on available linguistic cues. These linguistic cues are often taken from the event expression's grammatical structure and/or lexical form, from syntactically related temporal ex-

[3] RDF (Resource Description Framework) "is a standard model for data interchange on the Web", see `http://www.w3.org/RDF/` for more details.

pressions, and from time relationship adverbials (such as *before, after, when*, or *while*) appearing in the context. While TimeML does not provide a linguistically well-grounded definition for the notion of *event*[4], the most influential theoretical accounts have considered verbs as prototypical events. This is also motivated by the fact that in many languages, verbs are marked with grammatical tense, which can be, following Reichenbach (1947), abstracted to the level of temporal relations. Furthermore, verbs can often be characterised by their inner temporal properties (the lexical aspect) as proposed by Vendler (1957), and by the grammatical aspect[5] (which "expresses whether the event is seen as finished, completed, or ongoing" (Mani et al., 2005)).

In the following subsections, we will provide more details on the theoretical accounts provided by Allen, Reichenbach, and Vendler.

Allen's theory on temporal relations

Most work on TimeML has used temporal intervals as temporal primitives for modelling the semantics of events and temporal expressions. It is argued that events mentioned in natural language sentences can usually be divided into subevents, e.g. "*John opened the door*" can be re-described with a sequence of subevents: "*John grabbed the door handle*", "*John pushed the door handle down*", and "*John pulled the door towards him*". In a similar way, times referred by temporal expressions can usually be divided into smaller times, e.g. "*yesterday*" can be divided into 24 hours that compose the deictically referred date. As time intervals have duration, but time instants (time points) do not, only intervals can be divided into subintervals, and so the intervals seem to suit more naturally for representing events (and times) mentioned in natural language utterances.

The interval algebra proposed by Allen (1983, 1984) states that the set of all possible temporal relations between two time intervals is covered by 13 mutually exclusive relations. These relations (listed in Table 3.1) are the identity relation (*equals*) and six symmetrical relations:

- *before/after* – interval A precedes interval B, and there is no overlap between the two;
- *meets/met by* – interval B begins where interval A ends, and there is no interval (time gap) between the two;

[4]Instead, it is stated that "events are generally expressed by means of tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases" (Pustejovsky et al., 2003a).

[5]However, not all languages have the grammatical aspect as a property of the verb, and this is also the case with Estonian. We will discuss this in more detail in Subsection 3.4.2.

- *overlaps/overlapped by* – an ending of interval A overlaps with the beginning of interval B, but neither of the two intervals overlap others at full extension;
- *during/contains* – interval A is fully contained within interval B;
- *starts/started by* – interval A fully overlaps with the beginning of interval B (and A is shorter than B);
- *finishes/finished by* – interval A fully overlaps with the ending of interval B (and A is shorter than B);

Relation	**Inverse relation**	**Pictorial Example**
A before B	B after A	AAAAA BBBBB
A meets B	B met by A	AAAAA BBBBB
A overlaps B	B overlapped by A	AAAAA BBBBB
A during B	B contains A	AAA BBBBBBBBB
A starts B	B started by A	AAA BBBBBBBB
A finished B	B finished by A	AAA BBBBBBBB
A equals B	B equals A	AAAAA BBBBB

Table. 3.1: Allen's 13 temporal relations between intervals.

Allen also proposed an algorithm which can be used to infer new temporal relations based on an existing set of relations. The main idea behind the algorithm is the iterative application of transitivity rules until all possible relations between intervals have been computed (the transitive closure of temporal relations has been found). Examples of transitivity rules are:

1. from $A\ before\ B$ and $B\ before\ C$, infer that $A\ before\ C$;
2. from $A\ during\ B$ and $B\ after\ C$, infer that $A\ after\ C$;

This technique can also be used to check for inconsistencies in the set of temporal relations. For example, given the set of relations $\{A < B, B < C, C < A\}$,

the application of the transitivity rule 1 will produce the relation $A < C$, which contradicts with the relation $C < A$, thus indicates a problem in the initial set of temporal relations (a time loop: $A < B < C < A$).

Reichenbach's theory of verb tense meanings

Reichenbach (1947) provides a theory for formalising grammatical tenses of verbs as configurations of temporal relations between three time points: the speech time S, the event time E, and the reference time R.

Point S refers to the time when the utterance/sentence was created, and point E refers to the time when the event mentioned in utterance/sentence took place. Using the relations $<$ (precedence) and $=$ (simultaneity) between S and E, the basic distinctions between past, present, and future can be represented. The simple past (e.g. *John opened the door*) is represented as $E < S$, the simple present (e.g. *John opens the door*) is represented as $E = S$, and the simple future (e.g. *John will open the door*) is represented as $S < E$.

In case of simple tenses, the reference time R – the "vantage point" from which events are being viewed (Mani et al., 2005) – is simultaneous with E.[6]The perfective aspect (in case of English) can be used to draw a distinction between R and E. For example, in case of past perfect (e.g. *John had opened the door*), the configuration of relations is $E < R < S$, and in case of future perfect (e.g. *John will have opened the door*), the configuration is $S < E < R$.

Following Mani et al. (2005), Table 3.2 lists English tenses along with their Reichenbachian relation configurations. Theoretically, there can be 13 different configurations of relations, but only 7 of these are actually realized in English. For example, there is no regular tense for expressing the relation $S < R < E$, although the construction *will be going to V* (e.g. *John will be going to open the door*) can be used to express the meaning (Mani et al., 2005).

Reichenbach also argued for two rules related to the usage of tenses in natural language sentences. The rule of *the permanence of the reference point* states that in a compound sentence, the reference point should be the same for all clauses. For example, in the sentence *When John had opened the door, the cat slipped out and ran away*, the first clause follows the configuration $E_1 < R_1 < S_1$, the second clause has the configuration $E_2 = R_2 < S_2$, the third has the configuration $E_3 = R_3 < S_3$, and reference times of all clauses coincide ($R_1 = R_2 = R_3$). The rule of *the positional use of the reference point* states that if there is an explicit temporal reference (e.g. a locative temporal expression such as *tomorrow, the last week, currently*) in the clause, it usually coincides with the reference time (e.g. *By*

[6]Except for the simple future, which is ambiguous in English. For example, Reichenbach (1947) notes that the utterance 'Now I shall go' has the interpretation $(S =)R < E$, while the utterance 'I shall go tomorrow' should be interpreted as $(S <)R = E$.

Relation	English Tense Name	Example
$E < R < S$	Past perfect	*John had opened the door*
$E = R < S$	Simple past	*John opened the door*
$E < S = R$	Present perfect	*John has opened the door*
$S = R = E$	Simple present	*John opens the door*
$S = R < E$	Simple future	*John will open the door*
$S < E < R$	Future perfect	*John will have opened the door*
$S < R = E$	Simple future	*John will open the door*

Table. 3.2: English tenses and their configurations of Reichenbachian relations.

nine o'clock, John had opened the door has the configuration $E < R < S$, and R is made explicit by the temporal expression *nine o'clock*). The explicit temporal reference in the form of a temporal relationship adverbial (e.g. *'before', 'after'*) can also override the rule of the permanence of the reference point, if indicating a relation other than simultaneity between the reference points. For example, in *John opened the door after he had knocked*, the first clause has the configuration $E_1 = R_1 < S_1$, the second clause has the configuration $E_2 < R_2 < S_2$, and the relation between the reference times is $R_1 > R_2$.

While the TimeML annotation scheme does not involve a direct annotation of Reichenbachian relations, these relations are often hypothesised as the mechanism underlying the verb tenses, which also affect temporal relations between (verbal) event mentions. A thread of work, initiated by Derczynski and Gaizauskas (2011), focuses on extending TimeML with the annotation of Reichenbachian verb tense structures.

Vendlerian classification of events

If event mentions are to be formalised as time intervals, there is a need for a formal categorisation of events based on their temporal properties. The most influential work on the subject is probably that of Vendler (1957), which proposes that natural language verbs can be divided into four classes: *activities, accomplishments, achievements*, and *states*.

Activities are events that proceed in time ("consist of successive phases following one another in time" (Vendler, 1957)), but do not have any culmination. Examples of verbs indicating activities are *run, walk*, and *hike*. An important property of activities is that they are homogeneous in time: if an activity occurs during a time period t, it also occurs during each subinterval of t. Although, as noted by Mani et al. (2005), one should also allow time gaps in t (e.g. if *John ran for an hour*, there might have been a few moments when John paused *running*

during that hour), and one can "zoom" into the subintervals only up to a certain degree, after which the activity could be no longer called *running* (e.g. "lifting a leg is not *running*").

Accomplishments are events that proceed in time (have a duration) and end with a culmination. Examples of verbs indicating accomplishments are *write, build*, and *destroy*. In contrast to activities, accomplishments are not homogeneous: if an accomplishment occurs during a time interval t, it does not imply that during a subinterval of t, the same accomplishment occurs. For example, when *John wrote a letter in an hour*, one could not choose a subinterval of that hour and say that *John wrote a letter* during that subinterval.

Achievements are accomplishments that do not have any apparent duration (or their duration is too short to be noted). Examples of verbs indicating accomplishments are *win, recognise*, and *indicate* (Mani et al., 2005).

States are homogeneous events that have a duration, but unlike activities, they do not "consist of phases succeeding one another in time" (Vendler, 1957), but rather "hold over a period of time" (Mani et al., 2005). Following Vendler, this distinction is usually traced to English language usage: verbs indicating states (such as *know, love*, or *be happy*) usually sound odd in progressive form (e.g. *John is knowing; John is loving; John is being happy*), while verbs indicating activities are natural to use in progressive (e.g. *John is running/hiking/walking*).[7]

In subsequent studies on the classification of events, Vendler's classes have been refined and extended, and linguistic tests for distinguishing between different classes have been developed further. However, it is still an open question to what degree the Vendlerian class is a lexical property of a verb (*the lexical aspect*), and to what degree it is determined by the context of the verb usage. This makes it difficult to apply Vendler's classification in practice. The TimeML framework only distinguishes between events (Vendler's activities, accomplishments, and achievements) and states in its event classification, because other distinctions "involve quite a bit more subtlety", and "the implications of the distinctions for inferring temporal relations are not at all straightforward" (Pustejovsky et al., 2005a). ISO-TimeML (ISO/TC 37/SC 4/WG 2., 2007) tries to bring Vendlerian classification back in a refined form, requiring that the event annotation should also specify the event type, which can be: *process* (Vendler's activity), *transition* (Vendler's accomplishment or achievement) or *state*. To our best knowledge, however, this typology has not yet been applied in practical TimeML corpus annotations.

[7] Thus, if we select a subinterval of *John knows*, it is odd to say that *John is knowing* during that subinterval; however, a subinterval of *John runs* can be described by *John is running* during that subinterval.

3.2.2 Event annotation in TimeML

According to the TimeML specification (Pustejovsky et al., 2003a), an event is "a cover term for situations that *happen* or *occur*" and for "predicates describing *states* or *circumstances* in which something obtains or holds true".

TimeML only requires the annotation of event mentions and their temporal circumstances (temporal expressions and relations), and leaves out other event components (the event's participants and circumstances other than time). Thus, it is seemingly a robust and lightweight annotation scheme, which does not require modelling the frame structure of events, nor creating complex event ontologies. However, it must be noted that the main aim of TimeML has not been "the event annotation", but the annotation of temporal semantics (temporal relations). Much of the TimeML-related research has focused on topics such as the role of temporal inference in manual (Setzer et al., 2003) and automatic temporal relation annotation (Mani et al., 2006), and how the decomposition of the TimeML annotation task affects the results of automatic annotation (Verhagen et al., 2010). The event annotation has mostly been regarded as an intermediate step towards temporal relations, and only recently, has it gained more attention as a problem on its own.[8]

In this work, we consider the event annotation as the main problem, which contrasts with many previous works that focused mainly on the temporal semantics side of TimeML. In the following subsections, we discuss the main questions of TimeML-based event annotation, and the principles that have so far been applied in addressing these questions.

Which linguistic units should be annotated as event mentions?

At the linguistic level, the spectre of linguistic units possibly being event-denoting expressions is considered wide, covering "tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases" (Pustejovsky et al., 2003a). Some examples (event mentions from referred categories are marked in bold face):

(11) a. tensed verbs: *A police officer **stumbled** and accidentally **fired** his weapon.*
b. untensed verbs: *He plans **to arrive** on Sunday **to attend** the talks.*
c. nominalizations: *The **talk** was followed by a **walk** in the cemetery.*
d. adjective: *He is becoming increasingly **successful** at his work.*
e. predicative clause: *I was told my package would **be delivered** today.*
f. prepositional phrase: *The entire crew was **on board** for a month.*

[8] We are referring to *The Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, which was organized for the first time in 2013, see `http://www.aclweb.org/anthology/W/W13/W13-12.pdf` (2015-08-12).

English annotation guidelines (Saurí et al., 2006; Saurí et al., 2009) also give detailed instructions on which grammatical / sentence level units should be annotated as event mentions, and we will discuss these instructions in more detail along with Estonian event annotation specifications in Subsection 3.4.1. However, one could summarise that the leading consideration behind the guidelines is that of the temporal relevance of events: according to Saurí et al. (2005), TimeML focuses on "event-denoting expressions that participate in the narrative of a given document and which can be temporally ordered". Considering the temporal relevance criteria, Saurí et al. (2005) also argue that "generics and most state-denoting expressions" should be left out of annotation.

Generics do not refer to "a specific event instance (or a set of instances)", but rather refer to "a class of events" (Saurí et al., 2006). For example, sentences such as *Running keeps you healthy and fit*, *The use of doping is a major problem in sports in general*, and *Car sales are usually indicative of the health of the economy* do not describe any particular instance of *running*, *use of doping*, or *car sale*, but rather give general statements about (classes of) such events. The genericity becomes more difficult to decide upon if concrete details (such as mentions of specific persons, locations, times) are added to the event description, e.g. compare the sentence *Running is good for you* with the sentences *John runs on a regular basis*, or *John usually runs in the mornings*.

The annotation of state-denoting expressions is also restricted. TimeML guidelines 1.2.1 (Saurí et al., 2006) specify four cases when state-denoting expressions should be tagged as EVENTs:

a. States that "are identifiably changed over the course of the document being marked up.". So, no event expression is tagged in *John's car is red* (assuming that the colour of John's car is mentioned nowhere else in the text), while in *John painted his car from red to blue*, both *red* and *blue* are marked as events;

b. "States that are directly related to a temporal expression" (such as *calm* in the sentence *The weather is calm today in Alabama*) are tagged as events;

c. States whose validity depends on the writing time are tagged (such as the quantity-denoting state *100,000 people* in the sentence *The population of Tartu is about 100,000 people*).

d. States that are, according to the document, governed by REPORTING, I_ACTION or I_STATE events[9]are to be annotated (e.g. *not guilty* in the sentence *John claimed that he is not guilty*).

[9]The TimeML event classes REPORTING, I_ACTION, and I_STATE will be introduced in Subsection 3.2.2 (starting on page 70).

What should be the extent of the annotation?

In the case of multiword event mentions, TimeML proposes that only the word that best represents the event should be annotated (Xue and Zhou, 2010). The general rule for complex syntactic structures is that only the head of a construction should be annotated as an event (Saurí et al., 2009). For example, in the construction *did not disclose* (in *Kaufman did not **disclose** details of the deal*), only the verbal head *disclose* is annotated as an event. This rule also applies to phrasal verbs (e.g. *John **set** up the sawmill in 1972*) and on idiomatic expressions (e.g. *Most of our suggestions have **kicked** the bucket*).

Robaldo et al. (2011) argues that the principle of annotating only the head (a "*minimal chunk*") makes it particularly feasible to use the dependency syntactic structures as a basis for deriving TimeML event annotations. Thus, the TimeML event annotation can also be seen as an extension of syntactic annotations, which aims to make a step towards (event) semantics, but still relies on the underlying syntactic structures (e.g. in the case when the full extent of the event expression needs to be derived). This is particularly the case, if one considers which features are characteristic to events annotated in TimeML.

Which characteristic features of an event should be annotated?

While not directly emphasised in the official English guidelines (Saurí et al., 2006; Saurí et al., 2009), works that have adopted TimeML to Korean (Im et al., 2009) and French (Bittar, 2010) have considered TimeML a "surface-based annotation language". According to Im et al. (2009), this means that instead of "encoding the actual interpretation of the annotated (event) constructions, only their grammatical features" are marked up. For example, if there is a present tense construction that can be interpreted as a future tense (e.g. *John is leaving tomorrow*), only the surface interpretation (the present tense) is encoded (Im et al., 2009).

In addition to *tense*, features encoded include *pos* (the part-of-speech of the word), *aspect* (the grammatical aspect of the verb), *polarity* (indicates whether the event is negated at the grammatical level) and *modality* (indicates which modal auxiliary word (e.g. *may, must, should*) modifies the interpretation of the event). Considering the strictly verb-specific features (*tense, aspect*), as well as features that mostly surface in verb phrases (*polarity, modality*), one can interpret the TimeML event model as one leaning towards the verb category.

How to classify events?

In classifying events, TimeML makes an exception to the surface-based annotation strategy, and requires that events are classified by their semantic level interpretations. Seven general classes are used, out of which five reflect the (semantic)

dependency relations between events.

Event classes carrying semantic relations with other events ("selecting for other events as arguments" (Saurí et al., 2009)) are:

- PERCEPTION – indicates an event involving a perception of another event. Typical verbs expressing events of this class are *see, hear, listen*. Example: *A bypasser **saw** the accident*.
- REPORTING – indicates an event communicating another event, e.g. narrating or informing about another event. Example verbs include *say, tell, state*, such as in *The police **said** that the criminal was arrested.*
- ASPECTUAL – an event indicating initiation, re-initiation, termination, culmination, or continuation of another event. Example verbs are *begin, stop, finish*, such as in *The police will **start** the investigation.*
- I_STATE – an intensional state[10]that introduces events belonging to "alternative or possible worlds" (Saurí et al., 2006). Example verbs include: *believe, doubt, feel*, as in *The inspector **believes** that the criminal will be caught soon.*
- I_ACTION – an intensional action that introduces another event "from which we can infer something given its relation with the I_ ACTION" (Saurí et al., 2006). Example verbs are *attempt, postpone, promise*, such as in *The criminal **tried** to escape.*

The remaining two classes, OCCURRENCE and STATE, represent a general distinction between events which "happen or occur in the world", and states which are "circumstances in which something obtains or holds true" (Saurí et al., 2006), which can be traced back to the Vendlerian distinction between states and events (activities, accomplishments, and achievements).

3.3 Creation of an Estonian TimeML Annotated Corpus

3.3.1 Goals of the research

In this section, we will introduce a manual annotation experiment, which involved the creation of a TimeML annotated corpus for Estonian. We used a corpus with gold standard morphological and dependency syntactic annotations as the basis, and manually added temporal annotations (event, temporal expression, and temporal relation annotations).

[10]ISO/TC 37/SC 4/WG 2. (2007) draws a distinction between "intentional" and "intensional" states, arguing that the latter is a broader class. In our view, this difference remains subtle, as it is not clearly pronounced in the examples listed in the guidelines and specifications.

- We interpreted the TimeML event mention annotations as an *extension* to (dependency) syntactic annotations, and aimed to achieve a relatively exhaustive event annotation, attempting to maximise the coverage in syntactic contexts that can be interpreted as "eventive". As a consequence, the current project set looser constraints on temporal semantics, e.g. many state-denoting expressions are also annotated if they appear in "eventive" syntactic contexts, regardless of their "temporal relevance" (e.g. whether they change "over the course of the text" or not);

- We aimed at *retrospective* analysis of annotation consistency: after annotators provided relatively exhaustive event and temporal relation annotation, we used alignment with syntactic annotations to extract meaningful subsets of these annotations (e.g. only event mentions denoted by syntactic predicates), and studied the annotation consistency (inter-annotator agreements) on these subsets;

- Our project had an *exploratory* nature, with the aim of charting initial consistencies (and inconsistencies) between event mentions and syntactic annotations; the results and conclusions drawn from this study can be used as a basis for future studies aimed at achieving higher levels of consistency between these layers of annotations;

Initial results of the project were described in Orasmaa (2014b) and Orasmaa (2014a). In this work, we describe a refined version of the work reported there. We also adjust some of the statements made and conclusions drawn there.

3.3.2 The corpus and its dependency syntactic annotations

We chose texts for the annotation experiment from the Estonian Dependency Treebank (Muischnek et al., 2014b). The focus was on the news genre, and articles belonging to that genre where chosen for annotation until a corpus with the size of approximately 22,000 tokens (including punctuation) was compiled. The final corpus consisted of 80 articles from three Estonian newspapers: Maaleht, Postimees, and SL Õhtuleht. The major subgenres of the corpus were Estonian news, Foreign news, Local news, Sports, Economics and Opinions.

The dependency syntactic annotations in the Estonian Dependency Treebank were initially generated with the syntactic analyser of Estonian (Müürisep et al., 2003), and then manually corrected following the procedure described in Muischnek et al. (2014b).

In the output of the syntactic analyser, three layers of annotation can be distinguished: morphological, surface-syntactic, and dependency annotations. The morphological annotations "contain information about lemma, part of speech and

grammatical categories (e.g. case and number for nominals; mood, tense, person and number for verbs) for every word-form in the text" (Muischnek et al., 2014b). The surface-syntactic annotations specify a syntactic function (e.g. a member of the verbal chain, subject, object or adverbial) for each word form. Dependency annotations "give information about the governor of every word form in the text" (Muischnek et al., 2014b).

The following is an example of syntactic annotations of the sentence *John lõpetas maja ehitamise* 'John finished building the house'[11]:

```
(12)  "<John>" % John
         "John" L0 S prop sg nom cap @SUBJ #1->2
      "<lõpetas>" % finished
         "lõpeta" Ls V main indic impf ps3 sg ps af @FMV #2->0
      "<maja>" % house
         "maja" L0 S com sg gen @NN> #3->4
      "<ehitamise>" % building
         "ehitamine" L0 S com sg gen @OBJ #4->2
      "<.>"
         "." Z Fst CLB #5->5
```

In Example 12, word forms are separated from syntactic analyses by line breaks. Each line of syntactic analysis begins with an indentation, which is followed by the word lemma in double quotes, inflectional affix of the word (starting with `L`), part of speech, morphological information[12], surface-syntactic label (starting with `@`), and dependency information.

Dependency relations can be read from tags `#x->y`, where `x` marks the number of current token and `y` its syntactic governor (e.g. in the previous example, `#4->2` marks that word 4 (*ehitamise* 'building') is governed by word 2 (*lõpetas* 'finished')).

3.3.3 The annotation process

Three guideline documents were created for Estonian TimeML annotation: one for event annotation, adapted largely from TempEval-2 event guidelines for English (Saurí et al., 2009), one for TLINK annotation, adapted from TempEval-2 temporal relation annotation guidelines (TimeML Working Group, 2009), and

[11] This illustrative example was automatically produced using the online version of the syntactic analyser of Estonian: `https://korpused.keeleressursid.ee/syntaks/index.php?keel=en` (2015-08-27)

[12] Appendix A describes the tagset used; a comprehensive description of morphological tags can be found at: `http://www.cl.ut.ee/korpused/morfliides/seletus.php?lang=en` (2015-08-27)

one for temporal expression annotation, which followed the annotation format described in Subsections 2.2.2 and 2.2.3.

The annotation project involved three annotators and one judge. The annotators had backgrounds in computational linguistics, but no previous experience with TimeML annotations. The judge (author of this work) had previous experience with adapting TimeML guidelines to Estonian and with doing some corpus annotation experiments in TimeML.

Before the main annotation process, a pilot annotation experiment was made, with the aim of getting the annotators acquainted with the task. All annotators were provided the guidelines and 5 newspaper articles for annotation. The results of the pilot study were then discussed among the annotators and the guidelines were elaborated.

During the main annotation process (the annotation of 80 newspaper articles), each text was annotated independently by two annotators and given to the judge for disagreement resolution. Double annotation was used because of the difficulty of the task and because double annotation provides a better basis for studies of inter-annotator agreement.

The main annotation process was split into 4 iterations. The work done in each iteration involved two stages: during the first stage, event mentions and temporal expressions were marked in the text, and during the second stage, temporal relations (TLINKs) were annotated between events and between events and temporal expressions. Note that due to the dependency between the first and the second stage, the time boundaries of the iterations overlapped. The following scheme illustrates how the iterations were organised[13] temporally (rows correspond to different iterations, `E`-s represent the timespan of the first stage, `J`-s represent the timespan of the judge checking the first stage, and `T`-s represent the timespan of the second stage[14]):

(13)
```
I     EEEEEEJJJTTTTTT
II          EEEEEEJJJTTTTTT
III                   EEEEEEJJJTTTTTT
IV                          EEEEEEJJJTTTTTT
      ------------  time  ---------------->
```

The first stage of an iteration was performed manually in a text file containing dependency syntactic annotations. Along with determining the extent of the event and time expressions, annotators were also asked to choose the class of the event

[13] The scheme (Example 13) is illustrative; in the actual annotation process, timespans had varying lengths.

[14] `T`-s only represent annotation, judge checking excluded. The annotations of the second stage were checked later, after the main annotation process.

and to determine the temporal expression's type and calendrical value. After the annotators had submitted their results, text files were processed with a script that checked initial validity of the annotation (e.g. detected typos and cases where the annotation did not follow the specified format) and then the annotations were manually checked by the judge.

The second stage of an iteration was performed using the Brandeis Annotation Tool (Verhagen, 2010). Following the TempEval-2 guidelines (TimeML Working Group, 2009), the annotation of TLINKs was further divided into 4 subtasks:

1. *event-timex*: determine relations between events and temporal expressions;
2. *event-dct*: determine relations between events and document creation time;
3. *main-events*: determine relations between the main events of two consecutive sentences;
4. *event-event*: determine relations between events in the same sentence (intrasentential relations);

In the subtasks *event-timex*, *main-events* and *event-event*, the relation annotation was guided by syntactic specifications. In the subtask *event-timex*, it was specified that relations should only be added between temporal expressions and event mentions that syntactically govern them (within a clause or a phrase). In the subtask *main-events*, main events chosen for relation annotation were required to be syntactically the most dominating event mentions in the sentence (usually root nodes of the syntactic tree). In the subtask *event-event*, sentence-internal temporal relations were only required to be marked between pairs of event mentions where one mention syntactically governed another.

As in TempEval-2, a simplified set of temporal relations was used: `BEFORE`, `BEFORE-OR-OVERLAP`, `SIMULTANEOUS`, `IS_INCLUDED` and `INCLUDES`, `OVERLAP-OR-AFTER`, `AFTER`, `VAGUE` and `IDENTITY`. The elaborate relations `SIMULTANEOUS`, `IS_INCLUDED` and `INCLUDES` were used instead of the general relation `OVERLAP` (which was used in TempEval-2), because in the pilot annotation experiment, annotators found that the general relation `OVERLAP` was confusing and needed elaboration.[15]

[15]Note that TimeML guidelines (Saurí et al., 2006) also propose using the special relation types `DURING` and `DURING_INV` for marking inclusion relations between events and times (denoted by duration temporal expressions). We simply used `IS_INCLUDED` and `INCLUDES` relations for these purposes.

3.4 Event Annotation

In this section, we give an overview of the event mention annotation in the Estonian TimeML annotated corpus: which guidelines were followed during mark-up, what were the initial event inter-annotator agreement results, and what were the results of retrospective analysis of inter-annotator agreement on different syntactic structures. The section ends with a discussion of event annotation.

3.4.1 Which linguistic units were annotated as event mentions?

We took TempEval-2 event annotation guidelines (Saurí et al., 2009) as the basis in creating event annotation guidelines for Estonian. To our best knowledge, this is the most detailed set of TimeML event annotation guidelines for English, containing richly exemplified instructions on which grammatical units should be annotated as events, what the extent of the annotation is, and how event attribute values should be annotated. As such, it suited to our goal, which was to explore event models that are rooted in grammatical structures. We also augmented our guidelines with information from other sources, such as information about different types of verbal predicates from "The Grammar of Estonian" (Erelt et al., 1993), and used text examples from Estonian corpora.

A distinctive feature of the guidelines was that they did not set rigid constraints for excluding state-denoting expressions that are not temporally located, such as is done in the English TimeML guidelines 1.2.1 (Saurí et al., 2006) (in these guidelines, state-denoting expressions were allowed to be annotated only in four cases, as listed in Subsection 3.2.2).

In general, the focus was on event mentions realised by verbs, nouns and adjectives.

Verbs as event mentions. A single verb functioning as the main verb (predicate) of the clause (such as *sulges* 'closed' in *John **sulges** ukse* 'John **closed** the door') was considered as an event mention. In case of event mentions realised as multiword expressions, we distinguished regular grammatical constructions, which were already captured in the automatic syntactic analysis of Estonian, and the constructions that were still an open area of research in the automatic syntactic analysis.

Regular grammatical constructions are *negation*, *compound tenses*, and *modal verb constructions*. In the case of grammatical negation (formed by using a negative particle and an infinite verb), only the infinite verb was annotated as an event mention. Only the content verb was also annotated in compound tenses, which consisted of a grammatical verb *olema* 'be' and the content-bearing past participle. In the case of modal verb constructions, we followed the examples of French (Bittar, 2010) and Italian (Caselli et al., 2011) annotation projects and

annotated both the modal verb and its argument (infinite) verb as (separate) event mentions.

The catenative verb constructions (consisting of a verb and an infinite verb, e.g. *pani tegutsema* 'forced to act', *plaanib lahkuda* 'plans to leave') represent a borderline case of regular constructions: while these constructions are covered by syntactic dependency structure, their automated classification has not yet been attempted (only one subclass – modal verbs – is distinguished in automated syntax, although theory (Erelt et al., 1993) suggests classes of phasal/aspectual, causative, and colourative constructions). In the case of catenative verb constructions, both verbs were annotated as separate event mentions, as in the case of the modal verb constructions.

Particle verb constructions, support verb constructions, and verb+noun multiword constructions in general can be considered as an open area of research in automatic syntactic analysis of Estonian[16]. In the case of particle verb constructions (consisting of a verb and a particle, such as *ette võtma* 'to take on (something)'), English TimeML guidelines were followed and only the verb was annotated as an event mention. In the case of support verb constructions (consisting of a semantically weak verb and event denoting noun/adjective/adverb, such as *kõnet pidama* 'hold a speech'), the annotators had three options: 1) annotate only the verb (if it is unclear whether there is a support verb construction or not); 2) annotate the verb and its argument as separate events (e.g. in aspectual constructions, such as *lõpetas kihluse* 'ended the betrothal'); and 3) annotate the verb and its argument as a single event (e.g. 'hold a speech'). In cases of general verb+noun constructions (aside from support verb constructions), it was also optional whether the only verb should be annotated as an event mention or whether a multiword expression consisting of both words should be annotated.

Deciding the extent of multiword expressions is also problematic in 'be' verb constructions. In constructions consisting of the verb 'be' and an infinite verb, such as the compound tense construction and the progressive construction (consisting of *olema* and *ma*-infinitive in the inessive, e.g. *on tulemas* 'is coming'), a general rule to annotate only the infinite verb as an event mention was used. However, if no other verb accompanied the verb 'be', it was also more difficult to decide the extent of the predicate. This difficulty is also reflected in the manual annotation of syntactic functions, where annotators often disagree on distinguishing the predicative and subject of the copular clause (Muischnek et al., 2014b). We tried to use a general rule that if the 'be' verb appears with a state denoting noun, adjective or adverb (such as *olema õnnelik* 'be happy'), the whole construction should be annotated as a single multiword expression, but in other cases, only

[16] Although, more recently, some of these constructions have also been addressed, e.g. see the work on particle verb constructions by Muischnek et al. (2013).

the ‘be’ verb should be annotated.

Infinite verbs that are not part of the main verb construction (e.g. catenative verb constructions or ‘be’ verb constructions) represent the last class of problematic verbal event mentions. Infinite verbs (especially past and present participles) that function as syntactic attributes or modifiers of other clause complements, can also be interpreted as “being in the background” with respect to the main event described in the clause, and can be more easily missed by annotators. For example, consider sentence *Medali võitnud sportlane võeti soojalt vastu* (syntax-changing translation: ‘The athlete, who had won the medal, was warmly welcomed’, and syntax-preserving translation: ‘The medal-winning athlete was warmly welcomed’):

(14) Medali võitnud sportlane võeti soojalt vastu
Medal had_won athlete was_taken warmly PARTICLE

‘The athlete, who had won the medal, was warmly welcomed.’

In Example 14, the past participle verb *võitnud* (‘(had) won’) functions as an attribute of the subject of the clause (‘athlete’) and thus can be considered as background information to the main event of the clause (‘(was) welcomed’). We considered that non-finite verbs “being in the background” should still be annotated as event mentions, although the pilot study hinted that annotators are more likely to miss such cases.

Nouns and adjectives as event mentions. Similarly to infinite verbs, event-denoting nouns and adjectives that are not part of the main verb construction (e.g. the support verb construction or “be” verb construction) can be more easily interpreted as “background events” and missed by annotators. For example, in the sentence *Esimeses geimis pääses meeskond ette ja hoidis edu tänu heale servile* ‘The team took the lead in the first game and maintained the lead because of the good serve’, it is relatively clear that main verbs *pääses* ‘took’ and *hoidis* ‘maintained’ should be annotated as events. However, it is difficult to decide whether the nouns *geimis* ‘game’ and *servile* ‘serve’ should be annotated as event mentions or whether they should be considered as “background information” that can be left unannotated. We decided that background noun and adjective event mentions should be annotated if: a) they are governing an annotated temporal expression; or b) they are directly governed by a verb annotated as an event and they appear more than once in the text, and thus are more likely to have an important relations with the annotated events.[17]

[17]The criterion b) can also be problematic, e.g. one needs to decide whether synonymous references to the “background event” should also be counted when counting its occurrences.

3.4.2 Grammatical features of Estonian verbal event mentions

Considering the surface grammatical features of event mentions, TimeML annotation projects have largely focused on the grammatical features of verbs: *tense*, *aspect*, *mood*, *modality* and *polarity* (Saurí et al., 2006; Bittar, 2010; Caselli et al., 2011). Here, we give an overview how these categories are represented in Estonian, and how they are annotated in our project.

Mood. In general terms, mood is a grammatical category that conveys a speaker's attitude towards a statement, e.g. whether the speaker is stating a fact, expressing a command or a request, or indicating that the statement only potentially holds. As such, mood is also important in the interpretation of "eventiveness", allowing one to distinguish between realis and irrealis statements. In Estonian, mood is expressed as a morphological category of verbs, and four moods are distinguished: indicative, conditional, imperative, and quotative.[18] The indicative is the most common mood, used generally for expressing realis events. The conditional mood indicates that the event is considered as irrealis by the speaker. The imperative mood conveys an order or a request. The quotative mood indicates that the statement is being mediated and the speaker takes no stance towards it's veracity, or is even doubting in it's veracity.

In manual annotations, we did not annotate the grammatical mood, because the information about mood is already provided in the underlying dependency syntactic annotations, from where it can be automatically added to event mention annotations.

Tense. The Estonian indicative mood has four grammatical tenses – simple past, present, perfect and pluperfect –, conditional, quotative and imperative moods can have past and present tense.[19] In the following discussion, we focus on tenses of the indicative mood. In terms of Reichenbachian relation configurations, three of the tenses (simple past, pluperfect and perfect) represent events occurring before the speech time, and one tense is used to express events taking place at the present moment. Table 3.3 lists configurations of Reichenbachian relations for Estonian tenses.

Arguably, the present tense is the most ambiguous of the four tenses. The tense is used for events culminating in the present (e.g. *Politsei saabub sündmuskohale* 'The police arrives at the scene'), for ongoing events (*Läbirääkimised kestavad edasi* 'Negotiations are continuing'), and for events taking place in the future (*Nad saavad töö kindlasti valmis* 'They (will) certainly finish the job'). In addition, recurring events and generic statements (such as *Tippsportlased on*

[18] According to The Grammar of Estonian (Erelt et al., 1993), it is also possible to distinguish a fifth mood – jussive – which is based on the imperative mood and expresses a mediated command.

[19] Although traditional grammars state that the imperative only has the present tense, Kaalep (2015) argues that the past tense can also be distinguished.

Relation	Estonian Tense Name	Example
$E < R < S$	Pluperfect	*John oli ukse avanud* 'John had opened the door'
$E = R < S$	Simple past	*John avas ukse* 'John opened the door'
$E < S = R$	Perfect	*John on ukse avanud* 'John has opened the door'
$S = R = E$	Present	*John avab ukse* 'John opens the door'

Table. 3.3: Estonian tenses and their configurations as Reichenbachian relations.

keerulised isiksused 'Top-athletes are complex personalities') are also mainly expressed in the present tense.

As Estonian does not have a grammatical future tense, most constructions expressing the future are based on the present tense. The future can be expressed by combining a present tense verb with a temporal expression (for example: *Nad saabuvad homme* 'They (will) arrive tomorrow'), by using a present tense verb in a perfective construction (e.g. using the particle *ära* to express the boundedness of a situation: *Nad teevad töö ära* 'They (will) complete the job'), or by using specific constructions involving an aspectual/modal verb in the present tense and an infinite verb representing a future event (e.g. by using *hakkama* 'begin' verb constructions: *Firma hakkab valmistama uusi tooteid* 'The company starts (=will start) making new products'). Additionally, if a future-pointing temporal expression is provided, the perfect tense can also be used to express events taking place in the future (e.g. *Homme õhtuks on nad lahkunud* 'By tomorrow evening, they will have left'), or events stretching through the present and ending in the future (e.g. *Pood on suletud homseni* 'The shop is (=will be) closed until tomorrow').

As information about grammatical tenses is already provided in the underlying dependency syntactic annotations, we considered that event mention annotations could be automatically augmented with tense information, so we did not ask the annotators to annotate *tense*.

Aspect. In languages such as English (Saurí et al., 2006), French (Bittar, 2010) and Italian (Caselli et al., 2011), aspect is a grammatical property of verbs, and it is annotated as an attribute of an event, as it can provide important cues for interpretation of temporal semantics. In general terms, grammatical aspect expresses the distinction between perfective situations (finished situations, such as conveyed by the verb *opened* in *John opened the door*) and imperfective situations (continuous or ongoing situations, such as conveyed by the phrase *was opening*

in *John was opening the door*). In Estonian, aspect "has not developed into a consistent grammatical category", although it can be expressed by "specific resultative or progressive constructions", by particle verb constructions, and by case alternations of direct object (Metslang, 2001). Following the principle of annotating only surface grammatical cues, we did not consider *aspect* as an attribute that should be annotated in Estonian event mentions.

Modality. The usage of modal verb constructions provides a periphrastic way for introducing irrealis statements. In English, modal verbs are considered as auxiliaries to the main verb and are not annotated as event mentions; the presence of the modal auxiliary modifying the event mention is marked with the attribute *modality*, which contains "the literal string of the modal auxiliary form" (Saurí et al., 2009). In Estonian, modal verbs can be conjugated in the same categories as the regular verbs, so they provide important surface cues for event interpretation and are therefore annotated with an EVENT tag. We used a special class value – MODAL – for marking modal verbs, and also used the attribute *modality* on infinite verb event mentions that are modified by the modal verb.

Polarity. In general terms, polarity conveys the basic distinction between truthful/affirmative statements, and negative statements. As grammatical negation in Estonian is expressed in a similar way as in English (using a negation auxiliary word), we followed the English example and excluded negation auxiliaries from event annotation, but indicated the presence of negation by setting *polarity="NEG"* on the event mention modified by a negation auxiliary.

3.4.3 Problems on mapping event-event argument relations to dependency syntactic structure

The TimeML event *class* provides an exception to its surface oriented annotation philosophy, providing semantic, rather than grammatical level classification. A part of this classification conveys intra-sentential semantic dependency relations between events, i.e. event mentions from the classes `REPORTING`, `I_ACTION`, `I_STATE`, `ASPECTUAL`, `PERCEPTION` and `MODAL` take other event mentions as arguments. As we interpreted event annotation as an extension upon dependency syntactic annotation, we were also interested in how well the TimeML event argument structure can be aligned with the underlying syntactic dependency structure. More specifically, we considered the problems that arose when dependency syntactic relations were used for finding arguments for events "selecting for an event-denoting argument".

First, in some cases **a dependency relation must be reversed** in order to find an argument for the event requiring argument in TimeML. In Estonian dependency annotations, the following cases can be distinguished:

1. Some aspectual and modal finite verbs are systematically annotated as dependents of the accompanying infinite verbs. For example, in *John **hakkas** maja ehitama* 'John began to build the house', the aspectual verb *hakkas* 'began' is a dependent of its argument *ehitama* 'to build'.

2. An event requiring argument in TimeML can function as a syntactic attribute of its TimeML argument. For example, in ***Taaspuhkenud** vägivald tuleb lõpetada.* 'Reinitiated violence must be stopped', the verb participle *Taas-puhkenud* 'reinitiated' (TimeML `ASPECTUAL` event) is syntactically governed by the event noun *vägivald* 'violence'.

3. Some `REPORTING`, `I_ACTION`, and `I_STATE` events are expressed by adverbials which are governed by their TimeML argument: the main verb of the clause. For example, in *Korraldaja **sõnul** toimub üritus detsembris* 'According to the organizer, the event will take place in December' the adverbial phrase *Korraldaja sõnul* 'According to the organizer' is governed by the main verb *toimub* '(will) take place'.

Second, an event requiring argument in TimeML can have **multiple dependent events**; however, it is possible that not all of the syntactic dependents are arguments according to the TimeML *class* of the event. For example, in *Eilsel valitsuse istungil **lubas** peaminister maksu vähendada* 'At yesterday's government meeting, the prime minister promised to reduce the tax' the main verb *lubas* 'promised' has two dependent events: *istungil* '(at the) meeting' and *vähendada* 'to reduce'; however, only *vähendada* is the actual argument according to the TimeML class `I_ACTION`.

Third, the required argument can be **indirect dependent** of the event requiring argument in TimeML, so it must be reached via a path of dependency relations. This mostly happens if the event requiring argument in TimeML is part of a periphrastic verb construction and its non-verb part is not annotated as EVENT; however, the non-verb part syntactically governs the TimeML argument event. For example, in *Nad **teevad** ettepaneku viimane otsus üle vaadata* 'They will make a proposal to reconsider the last decision', only the verb *teevad* 'make' is annotated as an EVENT in the periphrastic expression *teevad ettepaneku* 'will make a proposal'; however, it's EVENT argument *(üle) vaadata* 'to reconsider' is a dependent of the word *ettepaneku* 'proposal'.

The aforementioned problematic cases were still subject to manual event annotation and classification, despite the mismatches between dependency syntactic and TimeML event argument structures. Therefore, event annotations and dependency syntactic annotations in our corpus can be combined to further study these problematic cases in future.

The current annotation project did not involve the annotation of SLINKs and ALINKs, which, according to the TimeML specification, are generally used for marking relations between events "selecting for arguments" and their argument events. SLINK relations are used for introducing event-event subordination relations, such as *modal* relations between I_ACTION and I_STATE events and their arguments (e.g. *John **promised** Mary to clean the room.*), and *evidential* relations between REPORTING and PERCEPTION events and their arguments (e.g. *John **said** he cleaned the room.*). And ALINK relations are marked between ASPECTUAL events and their arguments, indicating the type of relation: e.g. whether the aspectual event marks the initiation of the event (e.g. *John **started** to read*), or its termination (e.g. *John **stopped** talking*) (Knippen et al., 2005).

As our inital analysis showed, such event-event argument relations cannot be straightforwardly generated based on Estonian dependency syntactic annotations. However, it is likely that a combination of event annotations and dependency syntactic annotations can be used to automatically provide a pre-annotation of SLINK and ALINK relations, which would help to reduce the human annotation effort on adding these relations.

3.4.4 Overall inter-annotator agreements on entities

In this subsection, we report and discuss overall inter-annotator agreements on marking temporal entities: on annotating event mentions and temporal expressions.

Table 3.4 lists inter-annotator agreements on deciding the entity extent, i.e. on deciding which phrases of tokens should be annotated as entities. A relaxed evaluation scheme was used in the calculations: one token overlap was considered as sufficient for a match.[20] The inter-annotator agreements were aggregated as F1-scores.

Layer	AB	AC	BC	JA	JB	JC
EVENT	0.88	0.76	0.79	0.94	0.93	0.79
TIMEX	0.78	0.71	0.80	0.89	0.86	0.75

Table. 3.4: Inter-annotator agreements (F1-scores) on entity extent. Agreements are reported over pairs of annotators: A,B,C refer to initial annotators and J refers to the judge. A relaxed evaluation scheme was used (one token overlap was sufficient for a match).

[20]This evaluation scheme is comparable with the relaxed scheme used in the evaluation of temporal tagging in Subsection 2.3.4. Note that there, the entities were indexed character-wise, while here, the entities are indexed token-wise, so we require one token rather than one character overlap here.

Table 3.5 reports inter-annotator agreements on assigning attributes for entities: annotating the *class* of the EVENT tag and *type* and *value* of the TIMEX tag. Only entity pairs with overlapping extents were used as a basis for calculating the agreement on attributes. As annotators sometimes missed attributes, both precision and recall were calculated for each attribute, and then aggregated as an F1-score.

Layer	AB	AC	BC	JA	JB	JC
EVENT.class	0.82	0.54	0.51	0.91	0.85	0.53
TIMEX.type	0.83	0.75	0.76	0.91	0.96	0.72
TIMEX.value	0.80	0.51	0.46	0.89	0.88	0.47

Table. 3.5: Inter-annotator agreements (F1-scores) on assigning entity attributes. Agreements are reported over pairs of annotators: A,B,C refer to initial annotators and J refers to the judge.

It can be noted in Tables 3.4 and 3.5 that not all agreements were on a similar level: agreements with the annotator C were frequently substantially lower than agreements among the other annotators (A, B and J). This indicates that though the guidelines provided sufficient basis for two of the three annotators for achieving a relatively consistent annotation, there was still room for improvement considering the third annotator.

Automatic temporal tagging performance compared to inter-annotator agreement on temporal tagging. If the evaluation results on automatic temporal tagging (in Tables 2.2 and 2.3) are compared to manual inter-annotator agreements on temporal tagging (in Tables 3.4 and 3.5), one can observe the surprising trend that the results of automatic tagging exceeded the results of human annotation in many cases. While we do not know the exact reasons for these trends, some hypotheses can be put forth. The experience levels of the annotators were different in the two evaluations: the evaluation in Chapter 2 was done by an annotator who had a considerable level of previous experience with temporal expression annotation, while annotators A, B and C in this evaluation had little previous experience. Thus, there is a possibility that annotators A, B and C would have needed more previous experience in order to achieve highly consistent temporal expression annotation. The manual temporal expression annotations in Chapter 2 were based on automatic annotation (they were corrections of the results of automatic tagging), in contrast to the temporal expression annotations here, which were fully manually laid on the text. Therefore, there is a possibility that the evaluation done in Chapter 2 was biased towards the automatic annotations. Note, however,

that the literature also reports positive effects of automatic pre-annotation. Bittar (2010) found that automatic pre-annotation has significant effects on improving the consistency of temporal expression annotation, so we can hypothesize that the automatic pre-annotation would have also improved consistency levels in our inter-annotator agreement experiments.

A trend similar to the one that appeared in the current work was also observed in the N2 corpus (Finlayson et al., 2014), an English TimeML corpus created without pre-annotation. The manual temporal expression annotation performance reported there was also suboptimal compared to the state-of-the-art automatic tagging performance reported for TempEval-3 (UzZaman et al., 2013), indicating the difficulty of the task when performed in a fully manual way.

The problem needs further investigations in the future.

Inter-annotator agreement on temporal tagging compared to agreement on event mention annotation. The results in Table 3.4 also show that the agreement on temporal expression extent was lower than the agreement on event mention extent. This trend might be surprising if one assumes that event mentions are a more complex phenomenon to annotate than temporal expressions. Possible reasons why this trend occurred include: a) temporal expressions occur more rarely in sentences than event mentions, so they can be more easily missed; b) temporal expressions appear more frequently as multiword phrases (compared to event mentions), and consistent annotation of multiword units is more difficult to achieve; c) no preprocessing (automatic pre-tagging of temporal expressions) was used. Considering reason c, a similar trend was also observed in the N2 corpus (Finlayson et al., 2014), where automatic pre-tagging was not used, and the inter-annotator agreements on temporal expression extent were lower than on event mention extent.

When comparing agreements on EVENT *class* to agreements on TIMEX *type* and *value* (in Table 3.5), the general trend was that there was higher agreement on TIMEX *type* than on EVENT *class*. This was an expected result, as EVENT classification can be considered as a more difficult task than TIMEX classification due to the general difficulties of classifying event phenomena. Agreements also showed that of the three attributes, TIMEX *value* was most difficult to annotate. Assigning TIMEX *value* can be considered as an error-prone task due to different normalisation possibilities and the requirement to follow a rigorous normalisation format.

3.4.5 A study of inter-annotator agreement on EVENT annotation

As we interpreted event mention annotation as an extension to syntactic annotations, it is important to ask, how are different syntactic categories and structures

covered by EVENT annotations, and how does the inter-annotator agreement vary for different syntactic contexts. Given that gold standard dependency syntactic annotations are available, we can group EVENT annotations based on the underlying syntactic annotations, and compare coverages and inter-annotator agreements of these groups.

Generalising from the annotation guidelines, we made two hypotheses about what constitutes a prototypical event mention: 1) a prototypical event mention is a verb; and 2) a prototypical event mention is a part of the syntactic predicate of the clause. We expected that on prototypical EVENTs inter-annotator agreement would be higher than on non-prototypical EVENTs.

In order to test these hypotheses, we made experiments where EVENT coverage and inter-annotator agreement on different subsets of EVENT annotations were measured. These subsets were obtained from the set of all EVENT annotations by filtering: annotations that did not meet the specified syntactic criteria were removed.[21] After the removal, inter-annotator agreements were measured on the remaining annotations. Only annotations of the three annotators were used in the experiments; annotations belonging to the judge were excluded, as these were highly dependent on the underlying annotations.

Table 3.6 shows the results of the experiments, reporting EVENT coverages and EVENT agreements on extent (average F1 scores over annotator pairs AB, AC, and BC) on different subsets of EVENT annotations. Model 0 is the initial annotation where no EVENT filtering was applied. Models 1a–1d explore, how part-of-speech affects inter-annotator agreement, and models 2a–2d explore how belonging to the syntactic predicate affects the agreement. Model groups 1 and 2 were constructed in the following way: a prototypical case was taken as the base model (1a = keep only EVENT verbs; 2a = keep only EVENTs in syntactic predicates) and other models (b–d) were created by extending the base model.

The results of models 1a–1d supported the hypothesis that verbs are prototypical candidates for EVENT: the highest inter-annotator agreement (0.943) was observed if only EVENT annotations on verbs are preserved. The results also showed that the most problematic part-of-speech for EVENT annotation is the noun: adding EVENT-noun annotations (model 1b) reduced the agreement to 0.832. Adjectives were less problematic than nouns and this can be explained by their lesser frequency and by Estonian-specific decisions in syntactic annotation. In Estonian, verbal participles are similar to adjectives and are systematically marked as adjectives when appearing in specific positions (Muischnek et al., 1999). We observed that the majority of the annotated adjective EVENTs were past particles functioning syntactically as attributes or predicatives.

[21] In cases of multiword EVENTs, an EVENT was deleted only if its header token (the token with the EVENT *class* attribute) did not meet the criteria. Typically, a verb was the header token.

Model	Description	EVENT coverage[22]	IAA on EVENT extent
0	initial (no EVENT filtering)	4561 (100.0%)	0.809
1a	verbs	2973 (65.18%)	0.943
1b	verbs and nouns	4273 (93.69%)	0.832
1c	verbs and adjectives	3201 (70.18%)	0.916
1d	verbs, adjectives and nouns	4499 (98.64%)	0.815
2a	EVENTs that are part of the predicate of a clause	2607 (57.16%)	0.982
2b	2a + direct verb dependents of the predicate	2888 (63.32%)	0.953
2c	2a + direct non-verb dependents of the predicate	3643 (79.87%)	0.882
2d	2a + clause members not directly dependent of the predicate	3244 (71.12%)	0.898

Table. 3.6: EVENT annotation coverages and inter-annotator agreements (F1-scores) on different syntactically constrained subsets of annotations. Subsets were obtained by *filtering* the set of all manually provided EVENT annotations: only EVENT annotations which met the criteria (in the model's description) were preserved, and all other EVENT annotations were removed. Only annotations from the three initial annotators (A,B,C) were used in the experiment. A detailed version of this table (listing all pairwise agreements) can be found in Appendix C.

Models 2a–2d required that the syntactic predicate was automatically detected for each clause, based on the syntactic tags of words. In Estonian, the syntactic predicate has the following structure: a finite verb is always part of the predicate and if the finite verb governs all other members of the clause, this is also the only member of the predicate. In cases when the finite verb had a grammatical function in the clause (e.g. in cases of modal verbs or compound tenses), members of the

[22]In 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.

clause are governed by the infinite verb, so the infinite verb was also included in the predicate. The infinite verb also forms the predicate in cases when there is no finite verb at all (e.g. in cases of negation constructions).

The results of the models 2a–2d (in Table 3.6) support the second hypothesis: the highest inter-annotator agreement (0.982) was achieved when only members of syntactic predicate were allowed to be annotated as EVENTs. The agreement remained relatively high (0.953) when verbs that were direct dependents of the predicate were additionally kept as EVENTs. This mostly indicates cases of a catenative verb, where an infinite verb or a gerundive verb is governed by the predicate. However, when non-verbs were allowed to be annotated as EVENTs in subject, object and adverbial positions, agreement decreased to 0.882. Indirect dependents of the predicate (model 2d) caused a slightly smaller decrease in agreement, which can be explained by their lower frequency amongst the EVENT annotations.

3.4.6 Discussion of the results of event annotation

The results of our annotation experiment showed that the highest inter-annotation agreement (F1-score of 0.982) on "identifying" events was obtained with syntactic predicates, and that the agreement decreased if the EVENT annotations were extended to include event-denoting words outside of the syntactic predicate. A similar trend occurred when we considered event mentions belonging to different part of speech categories: the highest agreement (F1-score of 0.943) was obtained with verbs, and the agreement dropped if event mentions from other part of speech categories (nouns, or adjectives) were included in the set of EVENT annotations.

The finding that event annotation can be consistently agreed upon verbs and especially on "main verbs of the clause" (syntactic predicates) is not new, and can, perhaps, be taken as a "common intuition". For example, some pre-TimeML research on temporal annotation has used light-weight event models where syntactic clauses (Filatova and Hovy, 2001) or verbs (Katz and Arosio, 2001) were considered as event mentions. However, what is new in the current study is that we have shown how inter-annotator agreement decreases when one tries to make more complex event models, which extend beyond syntactic predicates and verbs. While TimeML event annotation can possibly cover "tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases" (Pustejovsky et al., 2003a), the results of the current study provide evidence that linguistic categories/structures beyond syntactic predicates and verbs are more difficult to annotate consistently, and specialised annotation guidelines and projects are likely needed for addressing such event mentions.

An example of an event annotation subproblem requiring a specialised ap-

proach is the annotation of noun event mentions. Sprugnoli and Lenci (2014) compared agreements of non-expert (crowdsourced) annotators and expert annotators on the task of identifying nominal events in Italian, and they achieved highly consistent annotation only between the experts who were specially trained for the task. They concluded that the task is not an intuitive one, and cannot be easily approached by non-experts. Arnulphy et al. (2012) focused on the annotation of French event nouns that are part of named entity phrases, e.g. '2008 Cannes festival', or 'the nuclear catastrophe of Tchernobyl'. Despite the available TimeML resources and guidelines for French (Bittar, 2010), they argued for developing a specialised approach to the phenomenon.

The current study had several limitations. Firstly, because no rigid guidelines were used for excluding "non-temporal" state-denoting expressions, it appears that almost each syntactic predicate was interpreted by annotators as forming an event mention. From this point of view, it was rather trivial to achieve high agreement on the structure. Secondly, although we identified some contexts with high inter-annotator agreements on event mention extent, this result may not be of high practical usage without high agreements on event classification. However, event classification is a more complex task, and our current experiments did not reveal any syntactic contexts with significantly high agreements on this task. Thirdly, the decision to allow the annotators to annotate event mentions as multiword units may have caused additional disagreements, which may have been avoided if the focus had been set only on creating single-word event annotations.

In conclusion, the current study showed that while event mentions realised as verbal constructions can be annotated with relatively high inter-annotator agreement, the agreement is likely to decrease if other, "less intuitive", syntactic contexts of event mentions are targeted. This suggests that the annotation of event mentions should be divided into specific subtasks by the syntactic contexts of event mentions. This finding is also supported by other research, e.g. Sprugnoli and Lenci (2014) showed that separate guidelines and training are necessary for annotating event nominals, and Arnulphy et al. (2012) argue for the need to separately address event mentions in named entity phrases.

3.5 Temporal Relation Annotation

In this section, we give an overview of temporal relation annotation in the Estonian TimeML annotated corpus. First, we introduce a commonly followed hypothesis of how the temporal information is conveyed in natural language, then discuss how temporal relations have been annotated in previous studies, and how we annotated temporal relations, and provide initial results of inter-annotator agreements. In the previous to last subsection, we introduce the results of retrospective analysis

of inter-annotator agreements, and in the last subsection, provide a discussion on temporal relation annotation.

3.5.1 How are temporal relations conveyed in natural language?

Temporal information is conveyed in natural language both by **explicit** means, such as surface grammatical cues about temporality and explicitly time-referring expressions, and by **implicit** means, which require semantic level interpretation, often involving world knowledge (Mani et al., 2005; Marşic, 2012).

The main mechanism for expressing temporal relations at the surface grammatical level is the **verb tense**, which makes explicit the temporal relations between the event time, the speech time, and the reference time (following the Reichenbachian account introduced in Subsection 3.2.1). Explicit temporal cues are also provided by **temporal expressions**, which allow one to position events in time, indicate durations of events, or reveal their recurrent nature. **Time relationship adverbials** (e.g. *before, after, since*, or *until*) are frequently used to signal explicit temporal relations between events, and between events and times referred to by temporal expressions. Other grammatical devices, such as prepositions of location in English and locative cases in Estonian, can also indicate temporal meaning, but arguably their primary usage is not solely related to expressing time.

While explicit temporal cues provide a basis for initial or "default" temporal interpretations, it is not uncommon that temporal semantics obtained via explicit mechanisms are overridden by semantics derived from implicit ones.

Temporal information can be expressed implicitly by following the **narrative convention**, that is by describing the events in text in their chronological order. While the usage of past tense throughout the text can indicate that the narrative convention is being followed, this is not an unambiguous indicator, as other discourse relations can also hold between past tense verbs. For example, an event mention can explain the cause of the previously mentioned event, as in the text passage: *Max fell. John pushed him* (Lascarides and Asher, 1993). Arguably, one uses world knowledge to find the correct **causality** interpretation for the previous text passage.

In addition to temporal information derived from knowledge about causality, temporal information can also be obtained by relying on world knowledge about typical **subevent** relations (Marşic, 2012). For example, in the text passage *John had lunch in the restaurant. He sat at the corner table and ordered a pizza*, the events mentioned in the second sentence (*sat* and *ordered*) can be interpreted as subevents of the event mentioned in the first sentence (*had lunch*), and thus they are temporally included in their superevent.

Finally, implicit temporal information can also be gained by using the temporal **inference**. If some of the temporal relations are already known, then transit-

ivity rules can be used to infer new temporal relations based on existing relations, as discussed in Subsection 3.2.1.

3.5.2 How to annotate temporal relations?

The annotation of temporal relations – i.e. encoding temporal information in terms of relations between events, and between events and temporal expressions – is a complex task. Researchers have noted that the task is "very difficult for humans, let alone to computers, to perform reliably and consistently" (Marşic, 2012), and that the difficulties arise "due to rampant temporal vagueness in natural language" (Verhagen et al., 2009). In order to make the task more manageable, it has been split into subtasks, and subtasks have been addressed separately, both in terms of manual annotation and automatic annotation experiments.

In TempEval-1 and TempEval-2 (Verhagen et al., 2009, 2010) evaluation exercises, the task of annotating temporal relations was split into subtasks covering relations between events and temporal expressions, events and document creation time, and main events of two consecutive sentences. TempEval-2 added further syntactic restrictions on the contexts, requiring that relations between events and temporal expressions should be marked only if the event mention syntactically governs the temporal expression. TempEval-2 also introduced a fourth subtask of marking temporal relations between events syntactically governing other events in a sentence.

Marşic (2012) proposed to split intra-sentential temporal relation annotation into smaller syntactically motivated subtasks, and provided a detailed account of how to do this via relying on dependency syntactic relations. The author argued that temporal relations in the contexts of syntactic subordination between temporal entities can be annotated with relatively high consistency, as opposed to annotating relations between entities in arbitrary sentence contexts or in contexts involving syntactic coordination.

In this work, we also split the manual annotation process into subtasks, following the TempEval-2 proposal (Verhagen et al., 2010; TimeML Working Group, 2009). Figures 3.1, 3.2, 3.3, 3.4 illustrate these subtasks, exemplifying syntactic contexts from which event mentions are chosen for temporal relation annotation.

The motivation for employing TempEval-2 split was the following. TempEval-2 uses syntactic relations to specify the contexts in which temporal relations should be marked, so the temporal relation annotation can be aligned with the underlying dependency syntactic annotations in our corpus. Aligning temporal relations with syntactic relations enables a systematic analysis of temporal annotations, e.g. to find out in which syntactic contexts temporal relations are more explicit and in which contexts more vagueness occurs. While there is a possibility to use an even more detailed syntactic specification (e.g. consider separately

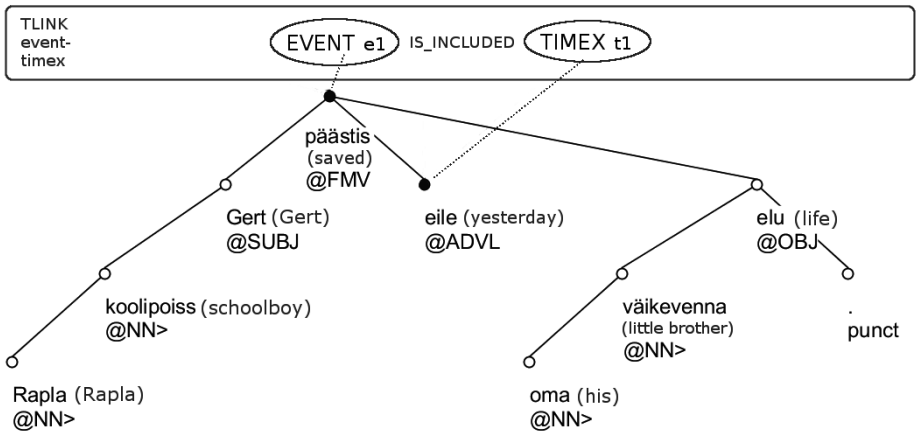

Figure 3.1: An illustrative example of TLINK annotation in the *event-timex* subtask: the temporal relation is determined between an event mention (*e1*) and its syntactic dependent temporal expression (*t1*).

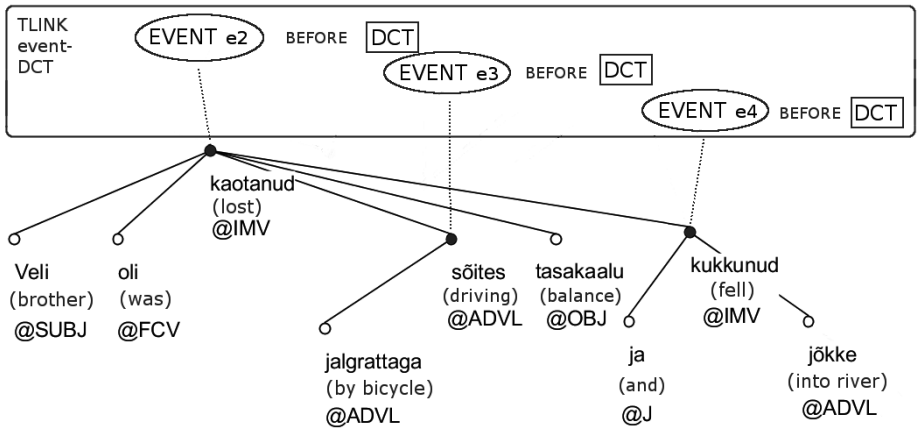

Figure 3.2: An illustrative example of TLINK annotation in the *event-dct* subtask: the temporal relation with DCT (document creation time) is determined for all event mentions marked in the sentence (*e2, e3, e4*).

relations between main verbs and their syntactic objects, subjects, or adverbials, as it is done in Marşic (2012)), we did not wish to go in such detail in the first

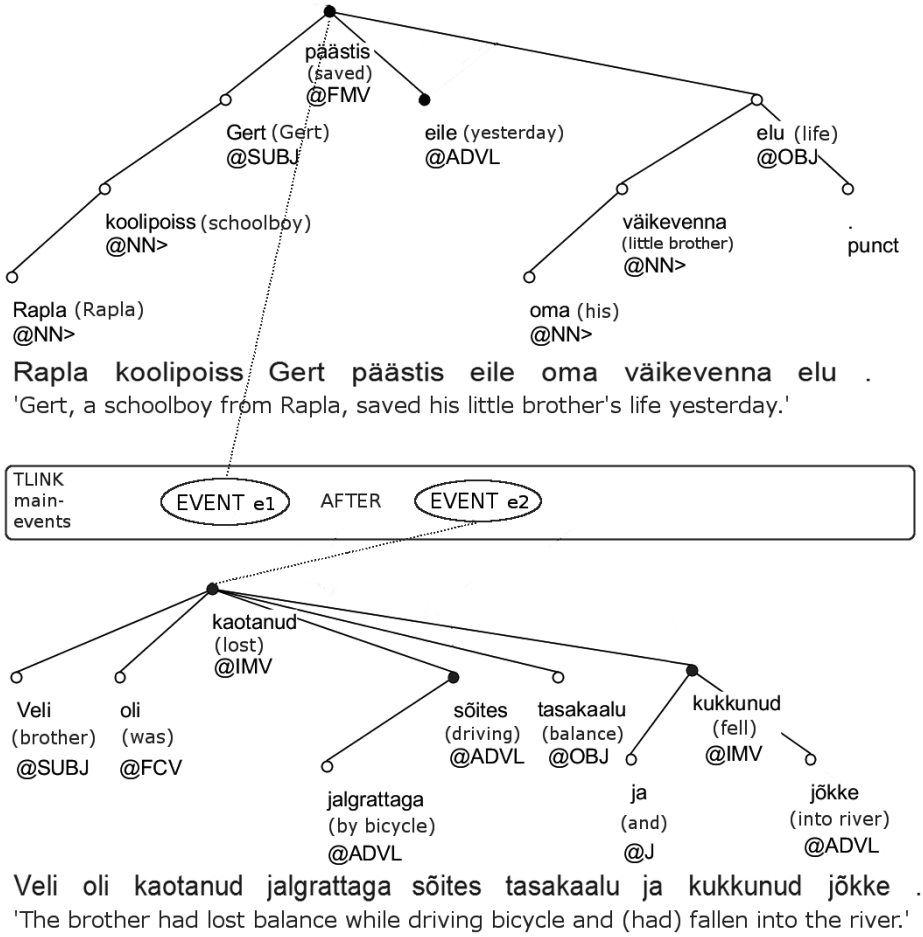

Figure 3.3: An illustrative example of TLINK annotation in the *main-events* subtask: the temporal relation is determined between the main events of two consecutive sentences. Event mentions covering root nodes of syntactic dependency trees (*e1, e2*) are chosen as *main events*, and a temporal relation is determined only between these events.

annotation experiment, and so we followed the more general split proposed by TempEval-2.

Evaluating the agreement between different temporal annotations. The first step in temporal relation annotation is choosing the pair of entities (event and timex, or event and event) between which the relation is to be specified. For this step, agreement can be calculated in terms of precision, recall, and F1-score.

The second step is specifying the type of the relation. A straightforward way for evaluating agreement at this step is to consider all pairs of entities mutually

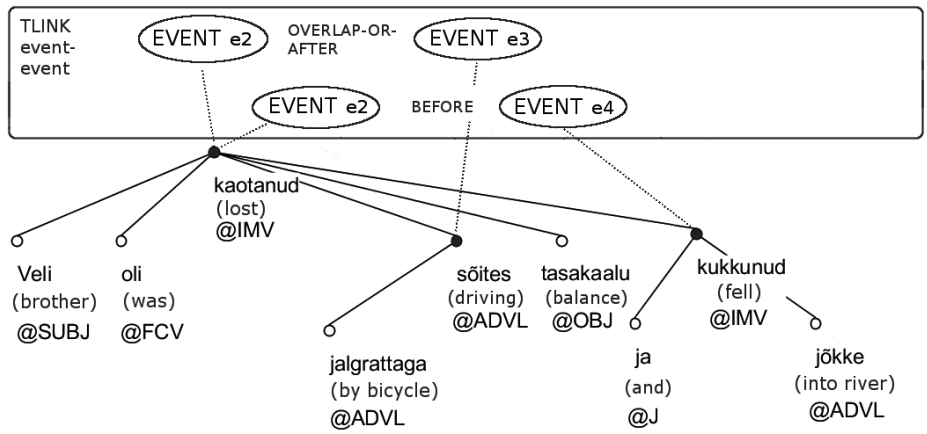

Figure 3.4: An illustrative example of TLINK annotation in the *event-event* subtask: temporal relations are determined for pairs of event mentions sharing a syntactic dependency relation (for pairs *(e2,e3)* and *(e2,e4)*).

chosen by both annotators, and to calculate the agreement as precision: the number of matching relations over all mutually annotated relations. This can be considered as evaluation of agreement at the surface level. However, the problem with this type of evaluation is that it does not take into account the fact that "the same temporal information can be captured by distinct annotations" (Mani et al., 2005), i.e. annotators may choose to annotate relations differently, but still convey the same semantics. Therefore, it is argued that temporal relation annotations need to be compared at the semantic level: by comparing temporal closures of annotations. A temporal closure is the set of all possible temporal relations that can be deduced from the initial set of temporal relations using inference rules. Two sets of annotations are considered as exactly matching if their temporal closures are equivalent (Mani et al., 2005).

In this work, we only employed the surface level agreement evaluation. The reasoning was as follows. Firstly, we were more interested in studying the agreement at the surface level: how consistently can temporal relations be aligned with syntactic dependency relations, and are there some syntactic contexts where higher levels of agreements can be observed? Secondly, we hypothesised that human annotators in general do not use temporal inference extensively while deciding the relations, but rather make decisions based on other types of implicit (and explicit) information available in a local context. Therefore, surface level evaluation would estimate the agreement with sufficient accuracy in most cases.

3.5.3 Overall inter-annotator agreements on temporal relations

In this subsection, we report overall inter-annotator agreements on marking temporal relations between entities (event mentions and temporal expressions), and discuss the results briefly. As the annotation process was split into four subtasks, we report the agreements separately for each subtask and also as an average agreement over all subtasks. For subtask results, we report the average agreement over three annotator pairs: AB, AC, BC.

Table 3.7 reports the results of the first step of annotation: the agreement on choosing pairs of entities for specifying a temporal relation.

event-timex	event-dct	main-events	event-event	*total avg F1-score*
0.771	0.999	0.708	0.605	0.771

Table. 3.7: Inter-annotator agreements (average F1-scores) on selecting pairs of entities for specifying temporal relations. Each of the first four columns reports a subtask average F1-score, which was calculated over three annotator pairs: AB, AC, BC. The last column reports a total average F1-score over all subtasks.

In the subtask *event-dct*, choosing the entities was actually trivial, as the annotators were asked to annotate the relationship with document creation time on all event mentions. Therefore, the only disagreement arose due to accidentally missed event mentions. Other subtasks left more room for disagreement. Although in subtasks *event-timex*, *main-events* and *event-event*, the annotators were asked to follow syntactic dependency relations on choosing entities, they did not have the syntactic dependency information available in the annotation tool, so the disagreements likely arose because they had come up with their own interpretations of syntactic dependencies.

After a pair of entities was chosen, the annotators had to specify the temporal relation. The agreement on specifying temporal relations was calculated as a precision and as Cohen's kappa (Cohen, 1960). The kappa agreement shows how much of the agreement was obtained beyond chance: i.e. beyond the agreement that could have been obtained if the annotators had made their choices randomly. $\kappa \leq 0$ indicates no agreement beyond chance, and $\kappa = 1$ indicates perfect agreement.

The (macro-averaged) agreements on specifying temporal relations are reported in Table 3.8; the detailed pairwise agreements that this table is based on can be found in Appendix D (Table D.1).

Following Landis and Koch (1977), the kappa values in Table 3.8 could be interpreted as showing mostly a "fair" agreement (kappas in the range 0.21–0.40),

measure	event-timex	event-dct	main-events	event-event	*total avg*
Precision	0.436	0.502	0.532	0.425	0.474
Kappa	0.296	0.383	0.42	0.32	0.355

Table. 3.8: Inter-annotator agreements on specifying temporal relations (choosing the relation type for each entity pair). Columns 2–5 report the average precision of a subtask and Cohen's kappa, calculated over three annotator pairs: AB, AC, BC. The last column reports total averages over all subtasks.

with the exception of "moderate" agreement (kappa in the range 0.41–0.60) obtained in the subtask *main-events*. However, considering the agreement thresholds commonly used in computational linguistics, our agreements fell on the lower side of the scale, e.g. Artstein and Poesio (2008) argue that kappas above 0.8 indicate a good agreement, although they also "doubt that a single threshold is appropriate for all purposes" and they leave room for reliability with lower agreements.

The low agreements reported here are roughly in line with agreements reported in similar settings by Verhagen et al. (2009). In their settings, annotators were asked to systematically annotate temporal relations in all contexts (e.g. relations between main events were determined in all consecutive sentences), without the possibility to skip problematic contexts. The "relatively low inter-annotator agreement" in their case, as in ours, seems to be the consequence of these settings: in problematic contexts, the annotators had no other possibility than to come up with their own idiosyncratic temporal interpretations, which likely caused the disagreement.

Some of the relatively low agreements could also have been the result of all mismatches being treated as equally problematic, regardless of the type of mismatch. We did some explorations on this issue. Firstly, a source of false disagreements could have been the usage of inverse relations, e.g. using *e1* BEFORE *e2* instead of *e2* AFTER *e1* was considered as a disagreement in our surface-based evaluation. We did a simple check for inverse relations: we allowed BEFORE/AFTER inversions and INCLUDES/IS_INCLUDED inversions, and noticed small improvements (total average precision increased up to 0.475). Secondly, we experimented with a special evaluation scheme which penalizes less for semantically close mismatches (e.g. BEFORE and BEFORE-OR-OVERLAP), and we also observed increased agreement levels, with the total average precision increasing to 0.59. It is possible that the agreements would further improve if one switched from surface-based evaluation to semantic level evaluation (comparison of temporal closures), however, this investigation remained out of the scope of this work.

3.5.4 A study on temporal relation inter-annotator agreement

As the overall inter-annotator agreements on temporal relations (reported in the previous subsection) were on the lower side of the scale, the next step was to investigate whether there were some subsets of annotations where higher consistency between annotators could be observed. The hypothesis we wanted to test was whether higher inter-annotator agreements could be observed in the contexts where explicit temporal information (information about verb tenses and temporal expressions) was available. As explicit temporal information is most characteristic to "main" verbs of the clause (as such verbs mostly have tense information,[23] and they are likely governing temporal expressions of the clause), our experiments focused on these verbs. More specifically, we only considered verbs that were part of syntactic predicates, adopting the model on which the highest inter-annotator agreement on event annotation was observed (see Section 3.4).

In each experiment, we selected a specific subset of "main verb" (syntactic predicate) events, along with all temporal relations associated with these events,[24] and measured inter-annotator agreements on specifying temporal relation (type) on the given subset. As in the previous evaluation, only temporal relation annotations provided by annotator pairs AB, AC, BC were considered.

Based on manual morphological and dependency syntactic annotations available in the corpus, we took out the following five subsets of annotations:

0. All syntactic predicate EVENTs;

1a. Event verbs in past simple tense;

1b. Event verbs in present tense;

2a. Event verbs governing an annotated temporal expression;

2b. Event verbs not governing any annotated temporal expression.

The set 0 is a superset of sets 1a, 1b, 2a, and 2b. The tense subsets 1a and 1b contrast each other: as we argued in Subsection 3.4.2, the past simple (1a) can be hypothesised as being relatively unambiguous compared to the present (2a).[25] Subsets 2a and 2b also contrast each other in terms of availability of an explicit temporal reference (a temporal expression).

[23] A syntactic predicate can also consist of a single infinite verb without any tense information; however, these cases were rare in our corpus.

[24] A temporal relation was included in the subset if it connected two events in the subset or if it connected an event in the subset with a temporal expression (or with a document creation time).

[25] We only took out subsets with two tenses (past simple and present), because there were too few annotations in the subsets of other tenses.

Before considering inter-annotator agreements, we investigated an alternative way of characterising temporal annotations in these subsets. As annotators were asked to use relations of type VAGUE in problematic/difficult contexts, we used the proportion of VAGUE relations as an indicator of how inexplicit the temporal information was in the given subsets. Table 3.9 shows counts of temporal relations (provided by annotator pairs AB, AC, BC) and proportions of VAGUE relations on the different subsets.

Event subset description	**Total relation count**	**VAGUE relation count**	**Proportion of VAGUE relations**
0. All syntactic predicate EVENTs	9,756	1,765	18.1%
1a. EVENTs in simple past tense	3,054	107	3.5%
1b. EVENTs in present tense	4,246	1,208	28.5%
2a. EVENTs governing TIMEX	1,558	63	4.04%
2b. EVENTs not governing TIMEX	7,218	1,521	21.1%

Table. 3.9: Counts of temporal relations, and proportions of VAGUE temporal relations of different EVENT subsets. Relation counts contain relations over entities that were commonly chosen for TLINK annotation by a pair of annotators (AB, AC, or BC).

As can be observed from Table 3.9, a relatively high proportion of relations were marked as VAGUE (18.1%) in the subset covering all syntactic predicate events (subset 0), which also indicates the overall vagueness of the task. However, a significant drop in vagueness (less than 5% of the relations were marked as VAGUE) can be observed in relatively unambiguous contexts: in contexts with past tense verbs (1a), and with verbs governing a temporal expression (2a). This lowering of vagueness can be further contrasted to increased levels of vagueness in subsets 1b and 2b, which represent relatively ambiguous contexts.

Table 3.10 shows inter-annotator agreements (Cohen's kappas) on annotating temporal relations on different EVENT subsets. Inter-annotator agreements were measured separately for each subtask (the average kappa was calculated over annotator pairs AB, AC, and BC), and then total average kappa was calculated as a macro-average over all the subtask averages. Detailed pairwise agreements are presented in Appendix D (Tables D.2, D.3, D.4).

According to the kappa interpretations provided by Landis and Koch (1977), most of the subtask inter-annotator agreements in Table 3.10 range from "fair" to "moderate", with the exception of "substantial" agreement (kappa in the range 0.61–0.80) obtained in the subset 2a *main-events*.[26]

[26]The "slight" agreement (kappa in the range 0.01–0.20) in the subset 2b *event-timex* should not be considered among the other agreements, because the agreement was measured on mistakenly

Event subset description	event-timex	event-dct	main-events	event-event	total avg κ
0. All syntactic predicate EVENTs	0.287	0.419	0.417	0.327	0.362
1a. EVENTs in simple past tense	0.238	0.279	0.432	0.382	0.333
1b. EVENTs in present tense	0.333	0.226	0.311	0.215	0.271
2a. EVENTs governing TIMEX	0.283	0.493	0.655	0.473	0.476
2b. EVENTs not governing TIMEX	0.089	0.402	0.378	0.296	0.291[27]

Table. 3.10: Inter-annotator agreements (Cohen's kappas) on specifying temporal relations on different EVENT subsets. Columns 2–5 report a subtask average Cohen's kappa calculated over three annotator pairs: AB, AC, BC. The last column reports total averages over all subtasks.

In general, the inter-annotator agreement results were in line with the vagueness measurements. There was higher agreement in the subsets containing relatively unambiguous temporal information (subsets 1a and 2a), in contrast to lower agreements in the subsets with ambiguous temporal information (subsets 1b and 2b). The results also indicated that the availability of an explicit temporal expression might be a more important factor in obtaining highly agreeable temporal relations than unambiguous tense information,[28] although this indication is not supported by the vagueness measures.

Some of the low kappa agreements were due to the sensitivity of the chance-corrected agreement to a highly skewed distribution. If data are highly skewed towards one category, then the measure tests agreement on rare categories, and low agreement on rare categories also leads to low overall agreements, even if there is high observable agreement on a frequent category (Artstein and Poesio, 2008). We observed this effect most in the subset 1a *event-dct*, where most of the specified relations were of the type BEFORE, and the observed agreement (precision) was actually relatively high (0.844), in contrast to the low chance-corrected agreement (0.279).

added TLINKs (if the guidelines had been followed rigorously, subset 2b *event-timex* would not contain any TLINK annotations).

[27] If the problematic subset 2b *event-timex* is excluded, the aggregate agreement is 0.358.

[28] Considering the type of temporal expression, we noted that the majority of temporal expressions falling into subset 2a were DATE expressions, so high agreements likely need to be associated with these expressions.

3.5.5 Discussion on temporal relation annotation

The experiments of this study showed that the presence of explicit temporal cues (verbs in the past simple and verbs governing a temporal expression) made the temporal relation annotation task less vague for annotators (less than 5% of the relations were marked as VAGUE), and supported higher inter-annotator agreements (e.g. a Kappa of 0.476 was measured for agreement on the relations associated with verbs governing a temporal expression). Our experiments also revealed opposite trends in contexts with limited/absent temporal cues (verbs in ambiguous present tense, and verbs not governing any temporal expression): there we observed lower levels of inter-annotator agreement (e.g. a Kappa agreement of 0.271 in contexts with a present tense verb) and higher degrees of vagueness (more than 20% of relations were marked as VAGUE).

Considering the assumptions on how the temporal information is conveyed in natural language (outlined in Subsection 3.5.1), the results of the current work confirmed the assumption that the presence of a temporal expression makes relations more explicit. Tense of the verb, however, provided more ambiguous results: indicating that only the past tense contributes to making temporal relations clearer, while the contexts of present tense were characterised by temporal vagueness. This can be explained by Estonian present tense conventionally being used to express closely related temporal semantics of present, future, recurrence, and genericity. Therefore, while in an ideal Reichenbachian interpretation, all tenses could be considered as making an equal contribution to temporal semantics, the findings indicated that the usage conventions of a tense pose additional challenges, and one could argue for distinguishing tenses with explicit and implicit meanings.

The findings support the general intuition that explicit textual/linguistic cues contribute to more consistent temporal annotation. To our best knowledge, this intuition has not yet been confirmed by a systematic study that shows which types of cues (e.g. temporal expressions, or verb tenses) support consistency on which types of temporal relations (e.g. intra-sentential or inter-sentential relations). Mani and Schiffman (2005) reported inter-annotator agreements on specifying temporal relations on pairs of English past tense verbs, but their work did not provide agreements on other verb tenses. Marşic (2012) decomposed the temporal annotation task into separate subtasks following the dependency syntactic structures, and while the author reported higher agreements on specifying *event-timex* relations (compared to *event-event* relations), no specific investigation was made on how the tense of the verb affects the results.

The current study had the following limitations. Firstly, the low inter-annotator agreement (and large amount of "vague" relations used) could have been the result of limitations in the annotation methodology. Guidelines on how to annotate temporal relations were not linguistically detailed, so in difficult cases the annotators

had no other guide than their own intuition. Also, a relatively large set of temporal relations was used (nine relations), which furthered the possibility to use idiosyncratic annotation strategies. It remains an open question whether similar results can be obtained using fewer temporal relations (only the relations *before, overlap, after*, as suggested by Marşic (2012)), and by using more detailed linguistic instructions on how to annotate temporal relations.

Secondly, it can also be argued that corpus composition (which genres of texts are annotated) plays an important role in temporal annotation (Bittar, 2010). Some texts, such as opinion articles, have a "less clear" temporal structure, and thus are more difficult to annotate. This study did not use any genre-specific criteria for choosing articles for annotation, as it was limited to using a corpus with available manual syntactic annotations. It is likely that some of the disagreements were due to opinion articles and editorials being included in the set of articles. It remains future work to investigate how text genre affects the consistency of annotation.

Concluding remarks. If the presence of explicit temporal cues (verbs in past simple and verbs governing a temporal expression) supports higher levels of inter-annotator agreement and lessens vagueness encountered during annotation, the question arises whether manual annotation efforts should be more targeted at contexts with explicit temporal cues? This contrasts to TempEval efforts, which have attempted to provide rather extensive coverage in manual annotations in order to approach the ultimate goal of the research ("to detect all temporal relations in a text") (Verhagen et al., 2009). The findings of the current work suggest that temporal expressions may be the most important device for revealing temporal relations in (a average) news domain text in Estonian. Our suggestion is that future research in Estonian temporal relation annotation should be more centred on contexts with temporal expressions.

3.6 Applications of the Corpus

In this section, we will discuss possible applications and future developments of the Estonian TimeML annotated corpus.

Although the inter-annotator agreements between the initial annotators were frequently on the lower side of the scale, annotations on all layers were also corrected by the judge, thus more consistent annotations are available in the final (gold standard) version of the corpus. We believe that this provides a sufficient basis for experimenting with different applications.

Experiments on end-user applications. The corpus can be used to perform experiments on how TimeML annotations support different end-user applications: automatic construction of chronologies, automatic question answering and summarisation. For these purposes, event annotations likely need to be developed fur-

ther by adding more event components. At minimum, annotations of participants and locations of the events could be added, but the set of annotations could also be extended with more sophisticated semantic role annotations, e.g. PropBank style annotations (Palmer et al., 2005) which focus on verb-specific argument structures.

Applications related to improving TimeML annotations. As the Estonian TimeML corpus is a subcorpus of the Estonian Dependency Treebank (Muischnek et al., 2014b), it can be extended with new texts from the treebank. An interesting question is, how can this extension be performed in a semi-automatic manner? On providing automatic pre-annotation, it can be investigated whether manually corrected syntactic annotation provides a better basis for machine learning of temporal annotations than the automatically provided syntactic annotation. Another sub-question is, how much the automatically provided pre-annotation improves the speed and consistency of manual annotation, especially when the annotation effort is focused on contexts with explicit temporal cues.

Tools developed in this work for evaluating inter-annotator agreements in different syntactic contexts can also be reused with new corpora. For example, the corpus can be extended with texts from other genres (such as fiction or science) available in the Estonian Dependency Treebank, so one can investigate whether the observations about inter-annotator agreements made on the newspaper texts also hold with other genres. Another possibility is to redesign the temporal relation annotation task (e.g. use a smaller set of relation types or provide more specific guidelines) and to investigate how this affects agreement.

Applications in linguistic research. The manually provided TimeML annotations also provide a basis for empirical investigation of some theoretical questions about time and aspect in Estonian. It can be investigated what is the interplay between TimeML event classes and Vendlerian event classes (e.g. whether TimeML classes can be taken as a basis in deciding Vendlerian classes of verbs), and how the Vendlerian event classes influence temporal semantics (temporal relation annotation). Texts in the corpus can also be analysed for two non-grammatical time-related categories of Estonian – aspect and future tense – to find whether temporal relation annotations support existing theories about the inner workings of these categories.

3.7 Conclusions

In this chapter, we have introduced the TimeML annotation format, its extensions, and the theoretical work that can be seen as among the important influencers of TimeML, namely Allen's interval algebra, Reichenbach's theory of verb tense meanings, and Vendler's event classification. We then gave an overview about

the event mention annotation in TimeML, focusing on questions such as which linguistic units are annotated as events, what is the extent of annotation, and how the annotated events are classified.

We have introduced a manual annotation experiment, during which a TimeML-annotated corpus of Estonian news texts was created. This considered TimeML event mention, temporal expression, and temporal relation annotations as *extensions* to dependency syntactic annotations, and the annotations were based on gold standard morphological and dependency syntactic annotations. We aimed at a relatively exhaustive event annotation, attempting to maximise the coverage in syntactic contexts that can be interpreted as "eventive".

We performed a retrospective analysis of annotation consistency: after the annotators provided relatively exhaustive event mention and temporal relation annotations, we used syntactic constraints to extract different subsets of these annotations, and to compare inter-annotator agreements in these subsets. We confirmed that there were relatively high inter-annotator agreements on prototypical event mentions: on verb event mentions, and more specifically, on event mentions covering syntactic predicates. While in principle the TimeML event model can cover a diversity of "eventive" linguistic contexts, our analysis showed that the agreement decreased when one tried to make more complex event models, which extended beyond syntactic predicates and verbs. Thus, specialised annotation projects are likely required to achieve high consistency in linguistic categories other than verbs. In the retrospective analysis of inter-annotator agreements on temporal relation annotations, syntactic contexts with explicit temporal cues (past tense verbs, and verbs governing temporal expressions) were compared with contexts characterised by limited/absent temporal cues. The analysis showed that in determining temporal relations, annotators perceived less vagueness in contexts with explicit temporal cues, and inter-annotator agreements were also higher in these contexts. The highest inter-annotator agreements were observed in contexts with temporal expressions, which suggests that the future research in Estonian temporal relation annotation should be more centred on such contexts. The findings also indicated that usage of the present tense in Estonian (news texts) is a rather ambiguous temporality cue (in comparison with the usage of the past tense).

The Estonian TimeML annotated corpus (containing both initial annotations and annotations corrected by the judge), and the tools for inter-annotator agreement experiments have been made freely available as a GitHub repository (`https://github.com/soras/EstTimeMLCorpus`, 2016-01-01).

3.8 Philosophical Notes

TimeML annotations essentially aim to extract (isolate) events from a story, and to recompose the story at a formal level, bringing out temporal relations between events. A compositional approach is assumed: first event mentions should be identified and then temporal relations drawn between these events. This perspective can also be reversed, by arguing that a perceivable presence of explicit temporal information is actually one important indicator of "eventiveness": that one can talk about "event mentions" with a high degree of certainty only in contexts where entities considered as events can be reliably placed on a time-line or temporally ordered with respect to each other. However, the results of the current manual annotation project indicate these contexts are scarce in news texts, and that higher than average consistency can be obtained only in certain syntactic contexts characterised by explicit temporal cues (such as explicit temporal expressions and past-indicating verb tenses).

One can argue that temporal annotation in TimeML is inherently a complex task, and that achieving consistency on this task requires an iterative annotation development process (called the MATTER[29] cycle). An iteration in this process involves modelling the phenomenon, annotating texts manually according to the model, testing the machine learnability of 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). If this framework is to be followed, the current work represents merely the first steps (modelling the phenomenon and performing annotation agreement experiments) of the first iteration, and other steps and iterations are required to achieve high consistency. Yet, it can also be argued that focusing on achieving high manual annotation consistency and high automatic annotation accuracy on the task may still not be sufficient to tackle the problem.

As it is argued by Zaenen (2006) that problems related to natural language understanding (such as event detection, and analysis of temporal relations) "have not been studied in linguistics nor anywhere else in the systematic way that is required to develop reliable annotation schemas". An optimisation-driven strategy that is supported by only a little discussion over the phenomenon being targeted has a risk of delivering models that do not satisfy the initial motivation, nor provide to be useful for applications. As a matter of fact, there has been little work on application-driven evaluation of TimeML annotations, and only recently have steps been made in this direction.[30]

It can also be argued that machine learning approaches have been most suc-

[29] MATTER stands for *model, annotate, train, test, evaluate, revise*

[30] We are referring to the QA TEMPEVAL task that was first proposed at SemEval-2015, see `http://alt.qcri.org/semeval2015/task5/` (2015-09-29) for more details.

cessful at the tasks where there is a large amount of data "available in the wild", e.g. in statistical machine translation and in speech recognition (Halevy et al., 2009). This large amount of data has been provided due to "natural tasks routinely done every day for a real human need" (such as translation and speech transcription), not due to skilled human annotation effort (Halevy et al., 2009). It is likely that the tasks of automatic event detection and temporal ordering also need to be reconsidered in work-flows where "a real human need" provides the motivation and supports the initial semi-organised data. The tasks of journalists, which involve making chronologies and summaries based on multiple textual sources, would be examples here.

We argue that the limitation of the current TimeML approach is the assumption that one can decide which units denote events and which temporal relations hold between events based solely on a single document (and frequently based on an even more narrow context, such as a sentence). This assumption may hold for texts that have been specifically composed with the goal of listing chronological facts or narrating a story in a chronological order (for example, children's stories (Bethard et al., 2012)), but based on the empirical experience gained during the current work, it does not hold for regular news text. As a matter of fact, we hypothesise that if one needs to make a chronology out of a regular news text, one likely needs to consider additional textual sources describing the same events in order to get better support for the facts, and to fill in the gaps in the information. In the next chapter, we will explore an alternative view on event analysis, by considering this as a multi-document task.

CHAPTER 4

BEYOND SINGLE DOCUMENT EVENT ANALYSIS: PRELIMINARY EXPERIMENTS ON CROSS-DOCUMENT EVENT COREFERENCE DETECTION

4.1 Different Perspectives on News Events

While grammatical level linguistic categories, such as verbs, verb nominalizations, and syntactic predicates, have relatively fixed linguistic scope, that of *event description* seems to be looser and easily extendible beyond a single word, phrase or sentence. When we consider a stream of daily news, it is often that a whole news article is built around one event, and important events are described in multiple articles and covered by multiple sources. Therefore, the task of event analysis can also be viewed in terms of detecting articles that discuss the same or directly related events, as in Topic Detection and Tracking research (Allan et al., 1998). In such settings, the notion of *event* seems to acquire a connotation of importance (e.g. Yang et al. (1999) define an event as a "(non-trivial) happening in a certain place at a certain time"), and it gradually merges with the notion of *topic* (as different instances of events are grouped under a topic, e.g. an earthquake happening at specific time and location is an *event*, while "earthquakes" form a *topic* (Yang et al., 1999)).

The open question is: what is the interplay between the fine-grained *event mention* perspective that assumes that a news article discusses many events, and the *topic/event* perspective that considers that a news article is focused on one or a few "seminal" events? From a fine-grained perspective, all event mentions

in a news article are informative, and should be analysed in order to gain an understanding of the story. From the topic/event perspective, a news article is most informative regarding a seminal event (the "focused event"), and we gain little from analysing all the event mentions in the story (they could only enhance our understanding of the seminal event). The topic/event perspective also suggests that events in news articles "are in the making": they are open to redescription/reinterpretation from other articles of the news stream, so we likely gain a better understanding of the story when we consider event descriptions from multiple articles/sources.

The argument that events are better "reconstructed" based on multiple sources is also supported by language comprehension research. Language comprehension research investigates how humans understand language, e.g. how they process "meanings" conveyed in text, and how they make inferences based on these "meanings". In this field, many researchers have argued that human understanding of text necessarily involves a mental representation of events (situations) discussed in text (Zwaan and Radvansky, 1998). However, it is also argued that much of the "text-based learning and reasoning" on some theme (e.g. on an historical event) actually "takes place when people integrate the information from multiple documents" into a common mental event model (Zwaan and Radvansky, 1998). As integrating knowledge from multiple sources seems to be the most efficient way for humans to learn about events, we believe that computer-based analysis of event mentions can also be advanced if multi-document settings are considered.

In this chapter, we will discuss cross-document event coreference detection, i.e. the task of finding event mentions referring to the same events across documents, and present our preliminary experiments on the task with Estonian news. The work considers a realistic information retrieval setting, where the user wants to find information about events related to a specific person from a corpus of daily news. In these settings, event coreference detection supports article browsing, as the list of coreferring mentions gives the user a glimpse into events related to the person searched for. We present a revised version of the work in Orasmaa (2015), and the revision includes remaking the experiment with different settings, and revision of the statements/conclusions based on new results.

4.2 The Problem of Event Coreference Detection in News

4.2.1 Theoretical and empirical concerns regarding event coreference detection

Event coreference detection is a task of automatically determining which fine-grained textual event mentions corefer. At the most basic level, the task involves a binary decision on whether two sentences or two words/phrases refer to the same event or not. For example, consider the following two sentences extracted from Estonian news:

(15) Soome sideminister Matti Aura andis eile hommikul lahkumispalve, sest tunnistas tehtud viga, kui ta hiljuti õigustas Vennamo käitumist.
'The Finnish Minister of Communications, Matti Aura, gave his resignation yesterday morning, as he admitted that his recent justification of Vennamo's behaviour was a mistake.' (source: Estonian Reference Corpus, Postimees 1999-01-05)

(16) Soome sideminister Matti Aura teatas eile oma tagasiastumisest seoses Sonera aktsiate müügi ümber puhkenud skandaaliga.
'The Finnish Minister of Communications, Matti Aura, announced his stepping down yesterday, related to the scandal of selling Sonera's shares.' (source: Estonian Reference Corpus, SL Õhtuleht 1999-01-05)

Both sentences 15 and 16 mention the same event, an announcement of a resignation, which is described by two event mentions: 'gave his resignation' and 'announced his stepping down'.

The task of event coreference detection can be divided into *within-document event coreference detection* (i.e. finding coreferring event mentions within the same document) and *cross-document event coreference detection* (i.e. finding coreferring event mentions from different documents) (Naughton, 2009).

Hovy et al. (2013) note that the difference between coreference and non-coreference is not clear-cut, and partial coreference of events can be distinguished. They studied partial coreference in detail, arguing that partial coreference could indicate either a membership or a subevent relation between two events. A membership relation holds, when one event mention describes a set of events of the same type, and the other event mention refers to a single member of that set. For example, in the sentence *"I had three meetings last week, but only the last one was constructive"*, the event referred to as *the last one* can be considered as a member of the set of events *three meetings*. In the case of subevent relation, one event is considered as a stereotypical sequence of events having a common goal (a script, such as *eating at a restaurant*, or *visiting a doctor*), and the other event (the subevent) is one of the actions/events executed as part of that script

(e.g. *ordering the meal* is part of *eating at a restaurant*, and *making an appointment to the doctor* is part of *visiting a doctor*). The authors also argued that time, location, and participants of events provide key information for distinguishing between different types of coreference (and non-coreference). However, their inter-annotator agreement results on distinguishing partial coreference relations indicate that these concepts are not yet well understood.

The idea that time, location, and participants could play an essential role in event coreference detection has also been explored by Glavaš and Šnajder (2013). These authors used a generic event model, where an event is defined by an event anchor (an event mention, usually a single word), and one or more arguments of four broad semantic types (agent, target, time, and location). They note and confirm by inter-annotator agreement experiments that if humans must decide on event coreference considering only mentions and four arguments, their agreement and certainty on the task is rather low, opposed to the case when they can examine the full context (both documents where the event mentions occurred) before making a decision. This suggests that event mentions are, in many cases, integrated tightly into stories, and one needs to consider the whole story before reasoning about an event mention, e.g. making a coreference judgement.

4.2.2 Event coreference detection in the context of limited linguistic resources

Most recent works on automatic event coreference detection have considered English or other languages well-equipped with linguistic resources, such as Spanish, Italian and Dutch (Glavaš and Šnajder, 2013; Cybulska and Vossen, 2013; Vossen et al., 2014b). The task at its full complexity seems to require that a number of advanced language analysis steps are implemented: 1) detection of event mentions; 2) named entity recognition and semantic role labelling for detection of event arguments; 3) normalization of temporal and locational expressions[1] and named entity coreference resolution for aligning event argument structures; 4) aligning event mentions along with their argument structures. However, it is not clear to what extent these fine-grained language analysis steps must be (and can be) solved, and some recent research suggests that light-weight approaches to event coreference detection are also worth trying.

First, it is likely that tasks of within-document event coreference and cross-document event coreference have different levels of difficulty. Naughton (2009) notes that two sentences mentioning the same event within the same document are likely to have a heterogenous vocabulary, as the factual information (e.g. location, participants) is rarely repeated on the second mention of the event. In contrast,

[1]Normalization of locational expressions would involve finding coordinates of the geographical regions that correspond to the expressions.

two cross-document sentences mentioning the same event likely have a more homogeneous vocabulary, because the important factual information (location, time, participants) is repeated in both documents. Thus, the task of cross-document event coreference can potentially be approached even with simple methods that rely on lexical similarity.

Second, if the set of documents under analysis can be narrowed down to the subset of documents discussing the same event(s) (such as in the Topic Detection and Tracking task (Allan et al., 1998)), even a simple method – matching event mentions by their lemmas – can yield relatively good results, which are roughly comparable to the results obtained with complex methods of matching event argument structures. This suggests that document clustering plays an important role in event coreference detection, and if one can obtain clusters of documents discussing the same events (e.g. by using lexical similarity and time constraints), one can achieve high precision mention level event coreference detection within these clusters using simple lexical similarity methods (Cybulska and Vossen, 2014b).

4.3 A Case Study: Finding Events Related to a Specific Person from Daily News

As a case study, we considered an exploratory search setting, where the user wants to find events related to a specific person from a corpus of daily newspapers.

The aim was not to provide an exhaustive listing of all related events, but rather to provide a summary-like overview, which gives the user a glimpse into the events most discussed, i.e. events that are mentioned in multiple articles.[2] The overview was to be provided in an extractive manner: sentences mentioning events are extracted from the articles, grouped by event coreference, and presented to the user.[3] From the provided text snippets, the user can proceed to reading the full articles if they spot something interesting.

In the current experiment, it was assumed that the scope of the search is the set of all articles where the person is mentioned by full name. From this set, the aim was to find cross-document pairs of sentences that mentioning the same event.[4] A

[2]The difference between the approach of the current work and the automatic summarisation approach is that the latter aims at reducing redundancy (e.g. picking only one sentence from the set of sentences mentioning the same event), while we aim to keep a moderate redundancy in order to give the user a better overview about the event discussed.

[3]We are aware that in addition to simply presenting the results to the user, there is a need for a more sophisticated sentence ranking or sentence simplification logic in order to improve the readability of the results. However, this remains a future investigation.

[4]An alternative would be to aim for detecting clusters of sentences, so that each cluster contains sentences referring to the same event. However, this approach would be more difficult to evaluate (considering partial coreference relations and the fact that one sentence can refer to multiple events),

time constraint was also imposed: the focus was on articles from a single week.

4.3.1 Document level considerations regarding event coreference

Before exploring sentence level event coreference, one likely wants to have some rough characterization of how much of the given set of articles is focused on describing "a narrow set of events".[5] Presumably, if only a narrow set of events is being described in the articles, there is also less room for errors in sentence level event coreference, and thus the task could be easier.

Lexical similarity within the set of articles can be taken as an automatically computable characteristic, which indicates whether the set of articles is: 1) likely describing a narrow/focused set of events (as in the case of lexically homogeneous articles); or 2) more likely describe loosely related or even different sets of events (in the case of lexically heterogeneous articles).

However, the relationship between high lexical similarity and concentration on a narrow set of events may not be straightforward. First, as exemplified by Bagga and Baldwin (1999), lexical homogeneity ("large number of overlapping words") can also indicate the presence of a complex subevent structure, where subevents frequently share same participants and locations.[6] In such settings, the task of event coreference resolution can actually be harder due to the ambiguities between full and partial coreference. Second, lexical heterogeneity does not necessarily imply that different events are being described, it may also be that the same events are being described using rather different vocabularies. Third, lexical heterogeneity is affected by the timespan chosen for analysis: longer timespans typically mean that there are more events reported, thus a lexically more heterogeneous set of articles is obtained. However, as only articles mentioning a specific person are considered in the current work, the increase of reportings also depends on how much media coverage a given person's activities receive, and how this coverage changes over time.

In the current work, the lexical similarity of a set of articles was measured as an average lexical similarity over all pairs of articles. For calculating similarity of a single pair of articles, the vector space model and calculation of cosine similarity over word lemmas were used. Tf-idf weighting[7] for lemmas was also

so it was chosen to explore the pair-wise detection approach instead.

[5] In Topic Detection and Tracking terminology: focused on a certain "seminal event" and events directly related to it.

[6] Although Bagga and Baldwin (1999) used slightly different settings: they compared sentences mentioning the event instead of comparing whole documents, we believe that their hypothesis (that a high word overlap in a set of sentences indicates high ambiguity of event coreference) could be generalizable to comparing documents.

[7] For calculating tf-idf scores and cosine similarities, standard tools provided by the scikit-learn library (`http://scikit-learn.org/`, 2015-11-10) were used.

used, ensuring that lemmas occurring many times in few articles had the highest discriminative power (and sharing such lemmas contributed most to the similarity between articles), while lemmas occurring in almost all articles of the corpus had the lowest discriminative power (thus sharing such lemmas contributed little to the similarity of documents) (Manning et al., 2008). Lemma counts and occurrences in the articles were calculated over the whole corpus (one week of news articles).

4.3.2 Sentence level considerations

As the aim was to give the user a glimpse into the events discussed, it was desirable to have the extracted sentences context independent: easily comprehensible outside of their document context. This can be problematic, because news texts often "concentrate on coherence of the narrative" (Glavaš et al., 2014), and thus sentences extracted from arbitrary locations in the story may be difficult to comprehend outside of their context due to coherence ties.

However, sentences from certain locations in an article may be more context independent. According to Bell (1999), the first paragraph of a news article (the abstract or lead) usually "summarizes the central action" of the story. It provides "the audience with the main point of a story", and based on that, the audience can decide whether to continue reading the article. While the title of the article can serve a similar function, the title can also aim at attracting the attention of the reader, thus it may be short and metaphorical ("catchy"), and not necessarily detailed enough to inform the reader about the central event(s). Therefore, it is the first paragraph, and frequently, the first sentence of the first paragraph, which guides the reader into the story. The first sentence also likely provides an independent description of central event(s): a description that can be comprehended outside of the context of the rest of the article.

There is also Estonian-specific evidence that first sentences are good candidates to be included in summaries. Müürisep and Mutso (2005) studied automatic extractive summary generation of Estonian news, and observed that the first sentence of an article was included in the summary in 100% of cases, and the second and third sentences were included in 65% of cases (observations were made on a training corpus).

In this work, a simplifying assumption that central events are likely described within the first three sentences of an article was made,[8] and in the following, these three sentences are referred to as "the summary sentences" of the article.

[8]However, this was a very rough approximation. Markings of the paragraph structure, title, and font could provide a better approximation, but the version of the corpus used in the current work did not have these markings available.

4.3.3 Methods for sentence level event coreference

In the experimental setup (details are discussed in the next section), four different sets of articles with varying degrees of lexical similarity were considered, and two different methods for sentence level event coreference detection tested: 1) a simple lexical similarity measure (measuring the number of lemmas overlapping between two sentences); and 2) an overlap of event argument structure components (location and time). We were interested in the following questions: a) how does the lexical similarity of a set of articles affect the results (can one assume more precise results in lexically more homogeneous sets?)?; and b) which of the methods returns more results from the summary sentences of the articles?

As a similarity measure, the Jaccard Similarity Coefficient (Jaccard, 1901) was used, which is defined as a similarity between two sets: the size of the intersection of two sets divided by the size of the union of two sets.

In order to find simple lexical similarity between two sentences (method 1), word lemmas were extracted from both sentences, converted to sets (i.e. removed duplicate lemmas) and the Jaccard Similarity Coefficient between these sets of lemmas was calculated. All pairs of sentences that had a similarity coefficient value greater than or equal to threshold k_1 were selected as pairs potentially containing corefering event mentions.

Note that one of the possible limitations of using the Jaccard Similarity Coefficient over whole sentences is that the measure is sensitive to contrast between sentence lengths (for example, if a sentence is less than half the length of the longer one, a coefficient value below 0.5 is obtained). This problem can be alleviated using a method less sensitive to the difference between sentence lengths: the Second Kulczynski Coefficient (Pecina, 2010). This method finds an average of two ratios: the size of the intersection divided by the cardinality of the first set, and the size of the intersection divided by the cardinality of the second set. However, as our experiments showed, this measure introduces another problem: a short sentence containing mainly non-content words (such as *Aga kes siis veel* 'But who else then') can match with many long sentences containing these non-content words, thus introducing an amount of noise to the results. Although filtering of non-content words could potentially alleviate this problem, we chose to stay with the Jaccard Similarity Coefficient in order to keep the models simple.

In the second similarity method, overlap of argument structure components of the event mentions was considered: temporal expressions and location names mentioned in sentences. As there were no syntactic structure nor semantic role annotations available, a very crude approximation was used that considered all temporal expressions and locations appearing in one sentence as belonging to one event argument structure. In the case of temporal expressions, normalized calendrical values of the expressions were used instead of lemmas, which al-

lowed matching lexically different date expressions. A strict matching scheme was used that required an exact match of the strings of `value` attributes of the temporal expressions. For example, the expressions *täna* 'today' and *neljapäeval* 'on Thursday' were considered as matching if their normalized `values` were matching (e.g. both had the `value` "1999-01-07"; see Chapter 2 for more on the annotation of temporal expressions). The focus was on temporal expressions of the type DATE, containing day granularity (e.g. *two days ago*), month granularity (e.g. *in April*), or year granularity (e.g. *last year*) temporal information. Expressions of the type TIME were also included, for which a looser matching scheme was allowed: only the date part of their `value` was used in calculating the match (e.g. *yesterday morning* was treated as a date, ignoring the part of day information, and thus it was possible to match it with the expression *yesterday* if both referred to the same date).[9] In the case of location names, only lemma matches were used, as the current work did not have means for normalizing location expressions to a standard format.

Similarly to the first method, the Jaccard Similarity Coefficient was applied and two scores were calculated: a coefficient for matching calendrical `values` of temporal expressions, and a coefficient for matching lemmas of location expressions. If both coefficient values were greater than or equal to threshold k_2, then the pair of sentences was considered as potentially containing coreffering event mentions.

4.4 Experiments

4.4.1 The corpus

For the experiment, all news articles from one week (from the period 1999-01-04 to 1999-01-10) from three Estonian daily newspapers (Postimees, Eesti Päevaleht, and SL Õhtuleht) were used, as they can be found in the Estonian Reference Corpus (Kaalep et al., 2010). The corpus has been automatically annotated for sentence boundaries and morphological information (word lemmas, part of speech tags, morphological case, and conjugation information). In addition, there are two layers of automatically added factual/semantic annotations: named entities (persons, organizations, locations, addresses, and quantities), and temporal expressions (using the temporal expression tagger introduced in Chapter 2).

[9] There is a possibility to develop even more fine-grained matching of TIMEX `values`: one could match `values` granularity-wise, e.g. consider that "2014-09-11" and "2014-09-13" match in *year* and *month* granularities, and only differ in *day of month* granularity. However, we anticipated that a granularity-wise matching would also introduce more mismatches/noise, and thus the stricter matching scheme was retained at this stage of research.

For the experiment, four persons were chosen – Jüri Mosin, Pekka Vennamo, Saddam Hussein and Mart Siimann – as article sets mentioning these persons were characterized by different lexical similarity levels (measured as an average of cosine similarities between all pairs of articles), and all the article sets contained articles from three different newspapers. Table 4.1 reports the statistics related to the four article sets.

Person	Articles	Sentences	Words	Avg. cosinus similarity (with tf-idf weighting)
Jüri Mosin	4	63	1079	0.57
Pekka Vennamo	9	142	2610	0.46
Saddam Hussein	7	105	1932	0.33
Mart Siimann	27	623	10506	0.17

Table. 4.1: Statistics of the article sets used in the experiment. Each article set was obtained by gathering all the articles where the given person was mentioned by full name.

Figure 4.1 gives an overview of how person mentions (i.e. articles mentioning the person) were distributed over the week. In the following, characteristics of these distributions in terms of event and person focus will be briefly described.

Articles mentioning Jüri Mosin were mostly focused on a single daily event – the criminal trial of Jüri Mosin and his accomplices – although they also discussed background events, such as the criminal history of the persons under trial, and possible future appeals. This focus was also reflected in relatively short coverage period (2 days), and high lexical homogeneity within the set of articles. A human reader could perceive this focus by observing that all summary sentences mentioned the central event (trial) and the given person.

Articles mentioning Pekka Vennamo were discussing a scandal related to the person: a possible abuse of his official position to purchase shares. The scandal had a longer development history, and on the given week, it culminated in the firing of Pekka Vennamo from his position, and the resignation of a Finnish minister related to the scandal. This culmination was also reflected in spiked media coverage (see Figure 4.1): on date 1999-01-05, the number of mentions suddenly "spiked" for a short period. In terms of events, the articles were quite focused on the two daily events (the firing and resignation of two high position persons), although, the history of the scandal, possible future developments, and impacts were more widely discussed than in the first article set. The focus was also reflected in the relatively high lexical similarity, and by the fact that the scandal or one of its subevents (firing and resignation) were mentioned in all summary sentences, as was the person.

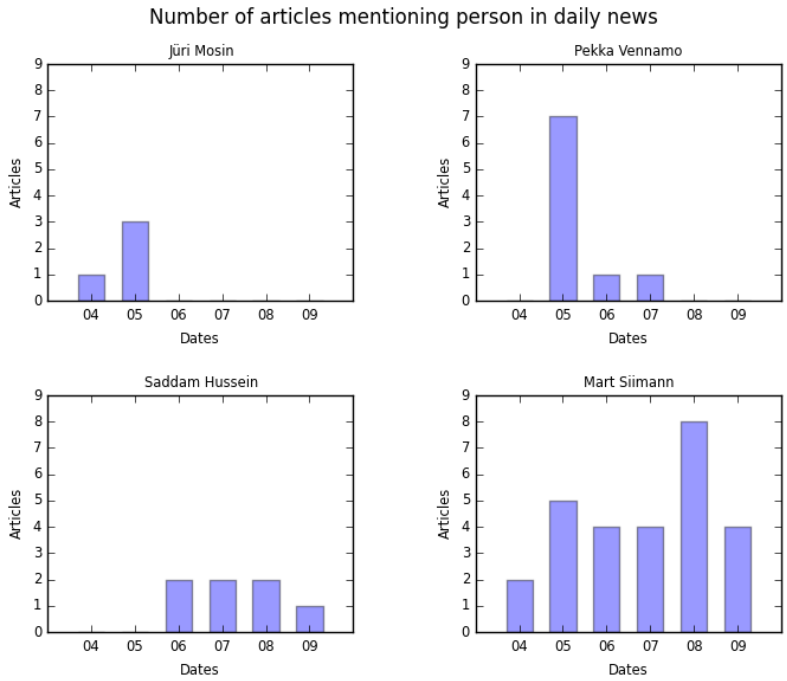

Figure 4.1: Newspaper media coverage of the given persons over the period 1999-01-04 to 1999-01-10. Each histogram shows how daily news articles mentioning the person (by full name) were distributed over the given period.

Articles mentioning Saddam Hussein were centred around two main events: an accusation that one of the sons of Saddam Hussein was involved in mass murder, and an allegation that a United Nations inspection team was involved in espionage for the United States in Iraq (where Hussein held presidency at that time). There was little in common between the two main events, which was also reflected in the relatively low lexical homogeneity. Also, in both events, the person did not play a central role, which is reflected by the fact that only 2 of 7 articles mentioned the person's name in the summary sentences.

Articles mentioning Mart Siimann, who was prime minister of Estonia at that time, were least focused on a narrow set of events, and had a mixed focus on the given person's actions. There were stories where his actions were more in the focus, such as meetings with prime ministers of neighbouring countries, and discussions related to coming elections. However, there were also articles just mentioning him once or twice, in a discussion of events not directly related to him (e.g. an article describing a sports event mentioned once that the prime minister was supporting the event). While Figure 4.1 shows a spiked increase of mentions

on date 1999-01-08, this was due to the cumulation of reports of different events, and not because of focus on a specific daily event. The overall lack of focus on a narrow set of events was reflected in the low lexical similarity between articles. The "mixed focus" on the given person's actions was reflected in the number of this person's name mentions in the summaries: 16 of 27 articles contained a mention of this person's name in the summary sentences.

4.4.2 Evaluation of sentence level event coreference

Evaluation principles

During this work, only the precision of finding pairs of coreferential sentences was evaluated. The reasons for this were:

- We assumed a browsing setting, where the relevance of the provided results is preferred over an exhaustive listing of all (potentially) relevant pairs. If more detailed information on the events is required, the user can access the full articles;
- Evaluating recall requires an exhaustive annotation of event coreference on all pairs of sentences. However, based on the empirical evidence on the manual annotation of event coreference in English (Hovy et al., 2013), we noted that a reliable annotation of event coreference is very difficult to establish, especially when considering partial coreference relations. Therefore, an exhaustive annotation would require a separate and focused annotation project, which was outside the scope of this work;

Next, we will discuss the principles used to manually decide upon event coreference.

Multiple event mentions. If two sentences under comparison contained multiple event mentions, it was required that only one pair of event mentions should corefer for the sentence pair to be correct. For example, consider the following pair of sentences:

(17) Läti peaminister Vilis Krishtopans lubas eile kohtumisel Eesti valitsusjuhi Mart Siimanniga ühtlustada kahe riigi vahelist viisarezhiimi.
'Latvian prime minister Vilis Krishtopans promised to homogenise the visa regime between the two countries at yesterday's meeting with Estonian prime minister Mart Siimann.' (source: Estonian Reference Corpus, Eesti Päevaleht 1999-01-08)

(18) /—/ Eesti-Läti sealihatüli oli peaminister Mart Siimanni ja tema Läti kolleegi, peaminister Vilis Krishtopansi eilse kohtumise üks põhiteemasid.
'/—/ an Estonian-Latvian dispute over pork trade was one of the main subjects

of yesterday's meeting between prime minister Mart Siimann and his Latvian counterpart Vilis Krishtopans.' [10](source: Estonian Reference Corpus, Postimees 1999-01-08)

Despite that event mentions such as *promised*, *to homogenise*, and *dispute* are not common between sentences 17 and 18, the pair of sentences was considered correct because the meeting (of two prime ministers) was mentioned in both sentences.

Possible coreference. Separately from correct/incorrect cases, cases of possible/partial co-reference were counted. However, the specification of possible coreference was less strict than the one proposed by Hovy et al. (2013). First, the type of possible coreference relation was allowed to be undetermined, e.g. due to the lack of information regarding the spatiotemporal positioning of the event. Second, it was considered problematic whether the subevent relation could be defined as a relation between a script and "one of the actions/events executed as a part of that script" (Hovy et al., 2013). For example, consider the following pair of sentences:

(19) Vennamo ise ütles eile Soome telekanalile MTV 3, et ei teinud midagi valesti ega ole andnud ka valeinfot Sonera aktsiatega sooritatud tehingute kohta.
'Vennamo himself told yesterday to the Finnish TV channel MTV 3 that he had not done anything wrong and he had not given any wrong information regarding the transactions with Sonera's shares.' (source: Estonian Reference Corpus, Eesti Päevaleht, 1999-01-05)

(20) Varem Vennamot toetanud Soome sideminister Matti Aura loobus skandaali tõttu teda usaldamast ja astus eile ka ise tagasi.
'A former supporter of Vennamo, Finnish Minister of Communications Matti Aura gave up his trust of Vennamo because of the scandal and also resigned yesterday from his position.' (source: Estonian Reference Corpus, Postimees, 1999-01-05)

The main event mentions of the sentences (i.e. *told*, *had not done* and *had not given* in sentence 19, and *gave up* and *resigned* in sentence 20) refer to actions performed by different persons and thus cannot corefer. However, the event mention *scandal* in sentence 20 refers to a rather general event (the event which was likely the main reason why both mentioned persons ended up in the spotlight of the Estonian media on the given week), and arguably one can consider the events *told*, *giving up trust* and *resignation* as subevents of the event *scandal*. Yet, it is difficult to consider *scandal* as a script, "a stereotypical sequence of events, performed by an agent in pursuit of a given goal" (Hovy et al., 2013). A scandal likely involves different agents with conflicting goals, and its outcome

[10]The beginning of the sentence was omitted for brevity.

can be rather unpredictable, as opposed to "stereotypical". Therefore, in the case of subevents, we only agree with the very general definition of Hovy et al. (2013): "subevent obtains when we have different events that occur at more or less the same place and time with the same cast of participants."

Results

We compared the methods in the settings where both methods yielded a similar amount of results: roughly 50 pairs of sentences in total. This required setting the threshold of method 1 (k_1) to 0.35 and the threshold of method 2 (k_2) to 0.50.

The results of manual evaluation of the simple lexical similarity method (method 1) are listed in Table 4.2. It can be observed that the method obtained high precision on most sets of articles, with the only exception on the lexically most heterogeneous set of articles (Mart Siimann). The method was also relatively easy to judge: high lexical overlap between words left little room for doubt on whether two sentences were mentioning the same events or not. This unambiguity of results was also reflected in low number of possible coreference relations among the results.

Person	Correct pairs	Pairs with possible coreference (subevents)[11]	Incorrect pairs	Precision
Jüri Mosin	11	1 (0)	0	91.7%
Pekka Vennamo	20	0 (0)	0	100%
Saddam Hussein	4	0 (0)	0	100%
Mart Siimann	8	3 (0)	4	53.3%
Total	43	4 (0)	4	84.3%

Table. 4.2: Results of similarity method 1 (Jaccard Similarity Coefficient for measuring lemma overlap between two sentences) on finding sentence pairs referring to the same event. In the case of possible coreference, numbers in parentheses indicate counts of pairs with a subevent relation among all the pairs with possible coreference.

Table 4.3 shows the results of the manual evaluation of method 2 (overlap of calendric values of temporal expressions and lemmas of location expression). Comparing the results in Tables 4.2 and 4.3, it can be noted that method 2 returned significantly more pairs with possible coreference relations than method 1. The

[11] From the two types of possible coreference relations (membership and subevent) (Hovy et al., 2013), we only report numbers of subevent relations, because we did not encounter any membership relations in the results.

results of method 2 were also more difficult to evaluate, due to the fact that besides overlaps in spatial and temporal cues, the sentences often had limited lexical overlap. In terms of finding sentences with full coreference, the overall precision was relatively low (50.0%); however, the actual number of incorrect pairs was also relatively small, and most of the errors were due to ambiguities between full and possible coreference.

Person	Correct pairs	Pairs with possible coreference (subevents)	Incorrect pairs	Precision
Jüri Mosin	3	0 (0)	0	100.0%
Pekka Vennamo	12	9 (8)	2	52.2%
Saddam Hussein	4	5 (4)	0	44.4%
Mart Siimann	10	7 (7)	6	43.5%
Total	29	21 (19)	8	50.0%

Table. 4.3: Results of similarity method 2 (Jaccard Similarity Coefficient for measuring overlap of temporal and locational expressions of two sentences) on finding sentence pairs referring to the same event. In the case of possible coreference, numbers in parentheses indicate counts of pairs with a subevent relation among all the pairs with possible coreference.

Effect of lexical homogeneity. The first question was about the effect of lexical homogeneity (of the set of articles) on the results of the methods. In particular, we hypothesised that higher homogeneity should support higher precision. The results in Tables 4.2 and 4.3 provide weak support for this hypothesis. There was a tendency that most errors (using both methods) were made on the lexically most heterogeneous set of articles (Mart Siimann), and that no incorrect pairs were returned from the lexically most homogeneous dataset (Jüri Mosin). However, considering that the most heterogeneous set of articles was also the largest one (almost ten times as large as the smallest), it might be that the size of the set, rather than its lexical heterogeneity, had an effect in the current experiment. In general, based on the current experiment, no hard conclusions could be drawn regarding the hypothesis.

Coverage on the summary sentences. The second question was, which of the methods returns most summary sentences. From the pairs found by the first method, only 12 pairs ($\sim$24% of all returned pairs) covered summary sentences (that is, at least one sentence from the pair was a summary sentence). In contrast,

from the pairs found by the second method, 44 pairs (∼75% of all pairs) covered summary sentences. This suggests that method 2 could yield sentence pairs with better summarisive qualities, assuming that the journalistic convention of summarising in the first paragraph is being followed, and that the approximation of the first paragraph (first 3 sentences as a first paragraph) holds. However, it needs to be further investigated, whether this was an artefact of the data or whether there was a general tendency (in the corpus) that important spatiotemporal cues, useful for linking articles discussing the same events, can mostly be found from the summarising parts of articles.

As Table 4.4 shows, there were also minor changes in the performances of the methods when switching from all sentences to summary sentences. The total precision of method 2 increased from 50.0% to 54.6%, and the total precision of method 1 decreased from 84.3% to 75%.

	Correct pairs	Pairs with possible coreference (subevents)	Incorrect pairs	Precision
Method 1	9	3 (0)	0	75.0%
Method 2	24	17 (16)	3	54.6%

Table. 4.4: Aggregated results of the two similarity methods on finding sentence pairs referring to same event in the summary sentences. For a pair of sentences to be included in these results, it was required that at least one sentence from the pair was a summary sentence.

Overlap in the results of the two methods. It was also investigated how large was the overlap between the results found by the two methods. From all sentence pairs returned by the two methods, only 4 pairs were overlapping (out of a total of 109 pairs). This shows that the two methods have the potential of complementing each other.

Comparing results to the previous work. The results reported here are slightly different from those reported in Orasmaa (2015), so a few clarifications are in order. Differences in the results of method 1 are due to different thresholds (in the previous work, k_1 was 0.5). Differences in the results of method 2 in the Pekka Vennamo subcorpus are due to the fact that in the previous work, we did not allow the matching of part-of-day date expressions (e.g. *yesterday morning*) to regular date expressions (e.g. *yesterday*). Differences in the results of method

2 in the Mart Siimann's subcorpus are due to different interpretations arrived at during different evaluations. Although the evaluator was the same in both cases, previous evaluation decisions were mostly forgotten because a year separated the two evaluations, and so a notable discrepancy emerged when the evaluation was redone (previous evaluation: 11 correct, 2 possible, 10 incorrect; this evaluation: 10 correct, 7 possible, 6 incorrect). This also hints at the low annotation reliability problems that seem to haunt such tasks in general (Hovy et al., 2013).

4.5 Discussion on Cross-document Event Coreference

The experiments of the current work showed that high precision cross-document event coreference can be obtained if model of the set of news articles is constrained by time and mention of a specific person. As the scale of the experiments (and manual evaluation extent) was relatively small, the experiments merely scratched the surface of the problem. However, a more systematic study also requires the creation of a corpus of Estonian news texts manually annotated for event coreference. Which leads to questions that have been little discussed in the literature: how does one create such a corpus in a way that the event phenomenon is modelled in balance between mention and topic levels?; and how does one achieve a reliable mention level event coreference annotation? We believe that the current experiments allow us to extend the discussion on these issues and provide a guide for future event coreference corpus design in Estonian.

Currently, the most commonly employed method for obtaining event coreference corpora is lexical similarity based news clustering. This seems to be practically the most straightforward approach due to the availability of news-clustering services, such as the European Media Monitor NewsBreif[12] and Google News[13]. A variant of lexical similarity based clustering was used for acquiring texts for the EventCorefBank (Bejan and Harabagiu, 2008), for the predicate argument alignment[14] corpus (Roth and Frank, 2012), and for the event coreference corpus created by Glavaš and Šnajder (2013). However, subsequent studies have also criticised this corpus creation method, arguing that the resulting corpora are characterised by low lexical diversity, which makes the task "artificially" easy, and supports high accuracies using lemma-based methods (Wolfe et al., 2013; Cybulska and Vossen, 2014b). A strong side of the method is that the high lexical similarity (combined with temporal proximity constraints) seems to assure that

[12] `http://emm.newsbrief.eu/NewsBrief/clusteredition/et/latest_et.html` (2015-11-25)

[13] `https://news.google.com` (2015-11-25)

[14] Predicate argument alignment: a task similar to cross-document event coreference resolution, where the goal is to align predicates along with their argument structures between two texts.

one or few seminal events are central to each of the acquired high lexical similarity clusters of news articles. Yet, to our best knowledge, none of the current event coreference corpus creation approaches has attempted to make this focus clearer by explicitly distinguishing annotations of seminal events from annotations of all other event mentions.[15]

One could anticipate that a reliable distinction between seminal and non-seminal events (in the manual annotation) would be difficult, because of the risk of falling into subjective grounds in making judgements. Yet, in the current experimental setup, we noticed that if the set of articles was focused on a few seminal events, these events were also mentioned in the summarising parts of the articles: in the first 3 sentences and (less frequently) in the headlines. This suggests future research questions: to what extent is the journalistic convention of producing summarising abstracts followed in practice (in Estonian newspapers), and can we use summary sentences as a basis for a reliable annotation of seminal events? Furthermore, one could investigate a more general question: to what extent is the information in news organised in the order of "the perceived decreasing importance" (Bell, 1999), that is, are the paragraphs organised by decreasing order of importance?

If there are news writing conventions of organising event information by importance, automatic analysis of events should also be aligned by these conventions, addressing the most important events first and then proceeding (gradually) to the less important ones, which could be more difficult to analyse (assuming less important events are described in less detail and repeated less often). The current research offered some, although preliminary, evidence on this: both methods showed differences in performance when tested first on all sentences and then only on summary sentences. This suggests that evaluation of automatic methods should also be organised in distinct phases, e.g. separate phases for measuring the accuracy of automatic event coreference detection in summarising parts and in content parts.

A distinctive part of the current study was that the articles were not grouped by a common topic/event, but by the mention of a specific person. This also showed that different persons have different media coverage patterns. The media coverage of Jüri Mosin and Pekka Vennamo on the given week was relatively scarce and focused on events directly related to these persons (the *trial* and the *scandal*). In contrast, the media coverage related to Saddam Hussein and Mart Siimann was less focused on a set of related seminal events and was more spread over the time

[15]For example, while Cybulska and Vossen (2014a) concentrate on event mention annotation in sentences that "describe seminal events", their annotation still does not distinguish explicitly the mentions of seminal events from all other event mentions in these sentences. Furthermore, their annotation seems to leave open the question, how much of the event coreference occurs outside "the sentences describing seminal events"?

period. In theory, the differences in media coverage could be explained by different news values behind the coverage. According to Galtung and Ruge (1965), elite people, such as prime ministers and presidents, get more media coverage, and this coverage is also more positive (or at least more diverse), and non-elite people tend to receive infrequent media coverage, which is also more likely to be associated with negative events. One could also try to characterise the media coverage of a person in terms of the types of reports: is the person frequently reported as a third person actor (person who is mostly valued for his/her actions)?; is the person more frequently being cited as a commentator (person who is valued for his/her opinions/comments, and perhaps less for direct actions)?; or is the person frequently just ”referred to”, but not particularly as an actor or as a commentator (person who is known or famous, but has no direct relationship to the current affairs)? Overall, one could try to distinguish different media coverage patterns, and try to balance the event coreference corpus in terms of different patterns, as these could indicate different challenges regarding the task.

4.6 Conclusions

In this chapter, we have introduced the problem of event coreference detection in news, described the theoretical and empirical concerns related to the problem, discussed the motivation for approaching the problem in a context of limited linguistic resources, and reported preliminary experiments on event coreference detection in Estonian news.

In the experimental part of the work, we studied cross-document event coreference detection in the context of finding events related to a specific person from a corpus of daily news. Rather than providing an exhaustive overview of all events related to the person, the goal was to give the user a glimpse into the events most discussed, i.e. to find events mentioned in multiple articles and sources. We discussed that high lexical similarity in a set of articles (mentioning a person) could indicate that the focus was on a narrow set of events, and thus higher performance event coreference detection could be achieved. We also discussed how events mentioned within the first sentences (of an article) could give the user a better overview about the whole story, due to the journalistic convention of summarising central events in the first paragraph (in the “summary sentences”). For the experiment, we chose four sets of articles with varying degrees of lexical similarity, with each set containing the weekly media coverage of a specific person. Two methods for finding coreferring event mentions were tested on these sets: a Jaccard Similarity-based method measuring lemma overlap between sentences; and a Jaccard Similarity-based method measuring overlap of temporal and spatial cues between sentences. The methods were found to complement each other: the

first method achieved higher precision, but captured less events mentioned in the summary sentences, and the second method returned more results from the summary sentences, but also delivered more ambiguous results (sentence pairs with uncertain event coreference).

Based on the literature and the experiments, we proposed that future research on event coreference detection in Estonian news should investigate two questions. First, the question of whether the trend of summarising central events in the first paragraphs of articles is consistent enough for distinguishing between two layers of event mention annotations: events mentioned in the summarising paragraph, and events mentioned in the content paragraphs. As the results suggest, method performances might differ on these layers. The second question is: does balancing the event coreference corpus in terms of different media coverage patterns, e.g. by distinguishing persons with higher and lower media coverages, reveal any differences in the difficulties of the task? Here also the results, although preliminary, suggest that such differences might be present.

4.7 Philosophical Notes

According to Pustejovsky and Rumshisky (2014), most annotation tasks can be broken down to *similarity judgements*. At each step, the annotator must decide whether a text segment belongs to one of the categories proposed by the annotation scheme, and making this decision involves a similarity judgement on whether the segment could be representative of the category or not. This task can be cognitively very demanding if category descriptions are rooted at an abstract level, such as in cases of events and event classes in the TimeML model. However, as Pustejovsky and Rumshisky (2014) argue, the cognitive load could be eased if the focus is shifted from abstract level similarity judgements to concrete level judgements, i.e. to comparing an annotatable text segment with a text segment representing a concrete example of the category. Following this idea, one can also hypothesise that the event coreference resolution task represents a cognitively less demanding alternative to mention level event annotation, which seems to be haunted by the vagueness of the event concept. Yet, empirically, this hypothesis has yet to be shown correct, as event coreference task involving concrete event mention comparisons has its own challenges, e.g. handling uncertainty (Glavaš and Šnajder, 2013) and deciding on partial coreference relations (Hovy et al., 2013).

Event coreference resolution introduces a shift from a single document perspective on events to a multi-document perspective, as repeated mentions of an event are more common to a collection of news, than to a single (news) document. This multi-document perspective should, in principle, be in line with the wider perspective of Topic Detection and Tracking (Allan et al., 1998), where

multiple documents are also analysed for events, although not by mention, but by the whole article standing as potential event description. On this matter, it is interesting to note how the definition of an event has changed during Topic Detection and Tracking research, following Makkonen's account (Makkonen, 2009).

According to Makkonen, the first definition of an event was closely tied to the physical world, defined as a unique occurrence with fixed temporal localisation. The problem with this definition seems to have been that it only covered the main or focused event (such as an airplane accident or an earthquake), which often had a clear spatio-temporal localisation in the story, and did not consider related events (e.g. in the case of an accident, descriptions of the rescue events, and the comments of rescue officials). So, the notion was extended to include "all necessary preconditions and unavoidable consequences" of the temporally specific main event. Still, the definition was problematic, as events with temporally scattered reports, e.g. long-lasting political campaigns, crises, or epidemics, did not meet the requirement of a specific temporal location. For covering such events, the notion of *activity* was introduced and defined as "a connected set of actions that have a common focus or purpose" (Makkonen, 2009).

As Makkonen notes, all of these definitions have a common problem: they seem to assume that news events are straightforward/factual reports of real world events. However, one can also argue for a clear distinction between real world events (generally outside the scope of the analysis) and news events (the actual target of the analysis). News events are, according to Makkonen (2009), "newsworthy real world events reported and packaged by the news media," and they are "literary products". If this perspective is taken, questions emerge about how events become news (what are the "news values" that determine the "newsworthiness"?(Galtung and Ruge, 1965; Harcup and O'Neill, 2001)) and how are the selected events "packaged by the news media" (what is the structure of a news story? (Bell, 1999)). In the context of automatic event analysis, these questions have received relatively little attention so far.

CHAPTER 5

REVISIONS AND SUGGESTIONS FOR FUTURE WORK

In this chapter, we will revise the general problems that have emerged in mention level event analysis, discuss different ways how these problems have been approached in other studies, describe the current state of solving these problems in Estonian event analysis, and provide a discussion on future work. We will centre the discussion on the TimeML approach, although we will also consider works building upon TimeML and works parallel to it.

5.1 Grounding Event Annotations in Grammatical Structures

Monahan and Brunson (2014) argue that "events are not a discrete linguistic phenomenon", and that natural language predicates and their usage contexts can convey "different degrees of eventiveness". Following Marşic (2012), we also believe that studying this complex phenomenon of "eventiveness" can be better guided on linguistically more familiar grammatical annotations. If event mention annotations are laid on grammatical level (syntactic and morphological) annotations, it allows one to decompose event annotations into linguistically structured subsets, upon which human understanding of the phenomenon, human annotation consistency, and machine learnability can be specifically studied and improved.[1]

In this work, we have started this grounding by laying event annotations on manually corrected dependency syntactic and morphological annotations. Next, we describe the current state and the remaining problems of such grounding, and also lay out directions on what should to be studied in the future.

[1]Marşic (2012) makes her argument, in particular, for grounding temporal annotations in syntax; however, we believe it is also applicable in the case of studying "eventiveness".

Event annotations on main verb constructions. Main verb / syntactic predicate constructions convey grammatical information that can be associated with "eventiveness" (e.g. tense, mood, polarity, and modality information in Estonian), and they also syntactically govern other clause constituents that can be interpreted as an event's arguments. Therefore, these constructions provide perhaps the most "natural" entry point for the grammatical grounding of event annotations. However, there are still areas where consistent event annotation centred on grammatical predicates is difficult to establish.

Copular verb constructions. While event mention annotation on "to be" + infinite verb constructions is relatively well-defined (only the infinite verb is annotated as an event mention), we encountered problems on establishing the annotation in contexts where no other verb accompanies the "to be" verb, and so a combination of "to be" and a non-verb clause member (or members) should convey "eventive" meaning (e.g. as in *John was ready for the debate.*)[2]

English TimeML guidelines propose that "in copular verb constructions, only the predicative complement is annotated as event" (e.g. *will not be ready*, or *may be ready*) (Saurí et al., 2005, 2006), and, in general, later works have followed this guideline, e.g. in annotating the EventCorefBank+ corpus (Cybulska and Vossen, 2014a). However, TempEval-2 guidelines (Saurí et al., 2009) diverge from this tradition and instruct one to mark up "both the verbal predicate and the predicative complement" as event mentions (e.g. *be ready*), and some of the non-English event annotation works have also considered the verb "to be" as a part of "the core predicate" (Matsuyoshi et al., 2010; Im et al., 2009).[3]

Our experience with Estonian annotation is that it is rather difficult to give syntax based guidelines for determining the extent of the predicate in such constructions. Words conveying the "eventive" meaning can be subject (*Tal on kahtlus, et asjad lähevad halvasti.* literally: 'A suspicion is being had by him (=he has a suspicion), that things are going wrong.'), predicative (*Suhtumine euroliitu on ebastabiilne.* 'The attitude towards the EU is unstable.'), or adverbial (*Nüüd on tuulik taas püsti.* literally: 'Now, the windmill is once again (standing) up.'). Thus, similar to English TempEval-2 guidelines (Saurí et al., 2009), we asked annotators to annotate both the "to be" verb and the accompanying non-verb, and in cases where the non-verb could not be decided, only the "to be" verb was to be annotated.

A more general problem here is that regular multi-word constructions in-

[2] Although copular verb constructions can also be interpreted as involving other verbs than *be* (such as the verbs *appear, seem*, or *look*), the discussion here is restricted only to *be* verbs as copular verbs.

[3] Although Im et al. (2009) only discuss "cancellation of the head-only rule", and annotation of verbal clusters as a whole, we assume that they are likely to also extend this idea to copulative constructions.

volving the grammatical "to be" verb (apart from the "to be" + infinite verb constructions) have not yet been the focus of a study in Estonian computational linguistics. For example, Kaalep and Muischnek (2009) create a database and an annotated corpus for verb-centric multi-word expressions in Estonian, and yet their work excludes "to be" centric expressions altogether. Our experience suggests that before more comprehensive guidelines can be formed, a systematic corpus-based study is required for the acquisition of regular multi-word "to be" constructions, and for an initial semantic classification of these constructions (e.g. based on TimeML classes).

Constructions involving negation and modality. While negation and modality appearing at the grammatical level (in specific main verb constructions) can be relatively consistently detected by automated syntax, these categories remain problematic in event analysis, especially when one aims to determine the relationships between events. For example, one can ask: should we draw a coreference relation between a mention confirming an event's occurrence and a mention negating it?; and should we aim to determine temporal relations between speculated (modality controlled) events?

TimeML guidelines for English suggest that "(main) verbs falling into the scope of a negative particle" should be annotated as event mentions (Saurí et al., 2006), and following works have largely adopted this principle (e.g. Bittar (2010), Caselli et al. (2011), Cybulska and Vossen (2014a)). However, the TERENCE annotation format that builds on TimeML has opted to skip the annotation of negated events, in order to simplify the assignment of temporal relations, and to improve annotator consistency (Moens et al., 2011; Bethard et al., 2012).

In a similar way, event mentions in the scope of the modal verb are considered as markables in TimeML (Saurí et al., 2006), though there is some disagreement on whether the modal verb itself should be annotated as well (Mani and Schiffman, 2005; Bittar, 2010; Caselli et al., 2011) or whether it should be excluded (Saurí et al., 2006; Cybulska and Vossen, 2014a). Some researchers have also opted to focus only on "realis"/"non-hypothetical events", and have therefore skipped entirely event annotation in modal contexts (Moens et al., 2011; Marovic et al., 2012; Bethard et al., 2012).

In our view, events in the scope of grammatical modality and negation should be annotated, provided these annotations can be easily discarded (e.g. removed based on underlying syntactic annotations) if there is a need for simplification of the event model. As for semantic level modality and negation, detection of these categories is largely unstudied in automated analysis of Estonian (and, as Moens et al. (2011) argue, this also seems to be the case for most languages), and we view such studies as a prerequisite for systematising event annotations in the contexts of semantic modality and negation.

Verb + verb constructions. While constructions of grammatical modality and negation can be interpreted as unambiguously referring to a single event, catenative verb constructions (finite verb + infinite verb) in general remain ambiguous in that aspect: depending on the construction, one can distinguish two distinct events (e.g. *John kavatseb lahkuda homme* 'John plans to leave tomorrow'), or a single event (e.g. *Parandused läksid maksma pool miljonit krooni.* 'Repairs cost (literally: went to cost) half a million kroons.').

Decisions on how catenative verbs should be segmented into events can be made by considering the possibilities of temporal relation annotation. Note that, so far, TimeML guidelines have left rather open whether and how TLINKs should be annotated in specific contexts, e.g. between an aspectual verb and its argument, or between an intention verb and its argument. To our best knowledge, the most comprehensive work (in English) that in addresses this gap is that of Marşic (2012), which attempts to systematise temporal relation annotations over different dependency syntactic contexts, and even in this work, the author does not go into detail regarding distinguishing annotations on different catenative verb contexts. In the case of Estonian, we suggest that temporal annotations in catenative verb contexts should be systematically studied over different lexical main verbs (e.g. catenative constructions headed by *kavatsema* 'to plan, to intend' and *suutma* 'to be able'), and over classes of main verbs (e.g. catenative constructions headed by verbs from the TimeML classes I_ACTION and I_STATE).[4] The Estonian TimeML annotated corpus provides a good starting point for such studies, as it allows one to calculate temporal relation coverages and inter-annotator agreements in specific catenative verb contexts, and therefore provides a basis for informed segmentation decisions (e.g. in the case of limited coverage or low agreement, one can choose not to interpret the construction as two event mentions).

Another problem in relation to catenative verb constructions is that the default syntactic attachment of the adverbial arguments can be "semantically misleading" in the contexts of these constructions. For example, in the sentence *Ta kavatseb tagasi jõuda tuleva aasta märtsis* 'He plans to return in March next year', the temporal expression *tuleva aasta märtsis* 'in March next year' was syntactically attached to the finite verb *kavatseb* 'plans', although semantically it should have been attached to the infinite verb *tagasi jõuda* 'to return' (Orasmaa, 2014b). The attachment of adverbial arguments (such as temporal and locational expressions) in the context of catenative verb constructions has gained relatively little attention in studies of automated syntax of Estonian; however, this is an important prerequisite for developing event analysis models, especially simple models ap-

[4]However, a prerequisite for a class-wise study is a consistent assignment of classes, which is yet to be achieved in Estonian.

proximating an event mention with the mention of a time and a location (as in Strötgen (2015)), and therefore requires a special study.

Verb + "eventive" noun constructions. Other classes of verbal constructions ambiguous between single event reading and two event reading are that of support verb constructions (a semantically weak verb + "eventive" noun, e.g. *made an offer*, or *get the appointment*) and verb + "eventive" noun constructions in general (e.g. *witnessed an accident*, or *passed the examination*).

Initial TimeML guidelines instructed one to annotate both the verb and the noun in a support verb construction as separate event mentions (for example, *take into consideration*) (Saurí et al., 2006). The TERENCE annotation format proposed to annotate only the noun, and to leave the support verb out of annotation, arguing that only the noun expresses the "main event" (Moens et al., 2011; Bethard et al., 2012). This principle was also adapted by Recasens (2011) and Lee et al. (2012) on their work on extending event mention annotations in the English EventCorefBank. However, in a later extension of the linguistic resource, EventCorefBank+, the annotators were asked to annotate both the support verb and the noun, because "part of the meaning would be lost" if either of the words were omitted (Cybulska and Vossen, 2014a).

We would like to point out two problems that we have encountered in relation to verb+noun constructions. We exemplify these problems with examples in English, but analogous examples can also be found for Estonian.

First, we have noticed that the list of TimeML's classes selecting for "event-denoting" arguments (e.g. PERCEPTION (*witnessed an accident*), ASPECTUAL (*began a journey*), or I_ACTION (*prevented the attack*)) could be incomplete: there are {verb} + {"eventive" noun} constructions where the verb seems to fall outside the classification. Examples of such constructions are: *join* {an event} (e.g. *joined the race*), *pass* {an event} (e.g. *passed the doping control*), *leave* {an event} (e.g. *left the game (at halftime)*).[5] Our suggestion is that while TimeML's event-selecting classes can be taken as a starting point for "sketching out" different verb+noun constructions, obtaining a more complete coverage could require a separate corpus-based study, likely focusing on the domain under analysis.

Second, we have noted that it is difficult to give a general rule on whether verb + "eventive" noun constructions should be segmented into two separate event mentions (or kept as a single mention), as this seems to depend on the task at hand and on the concrete construction (or class of constructions).

An example from event coreference detection. While we may choose to annotate *placed an order (for something)* as a single event mention in order to match it

[5] One could suggest the verbal events in these examples are classified as I_ACTIONs. However, Saurí et al. (2009) state that I_ACTIONs mostly "form a closed class", and they also provide a list of examples of representatives of the class that do not cover verbs similar to the ones in our examples.

with synonymous event mentions such as *ordered, preordered*, or *commissioned*, if we annotate the noun *order* as a separate event mention (interpreting it as "a state of commission"), it can also be matched over different perspectives. For example, we can interpret the sentences *Mary placed an order for cookies* and *John received an order for cookies* as referring to the same state (*"order (for cookies)"*), just by having different perspectives on it.

An example from temporal analysis. We can interpret the expression *held a conference* as referring to two events with temporal overlap: the act of holding/organising a conference, and the actual event – the conference – as the former event usually takes a longer time span. However, if the noun refers to an abstract state (such as in *gain a possibility*, *give a chance*, or *get an opportunity*), rather than to a concrete event, it would be more practical to annotate it as a single event mention in order to avoid possible confusions in determining the temporal relation (e.g. in the expression *gain a possibility*, does the event of *gaining* precede or overlap with the state *possibility*?).

In conclusion, we note that our knowledge on the verb + "eventive" noun constructions still needs improvement before consistent event annotation principles (and automatic annotation approaches) can be designed. For future work in Estonian, we suggest that (domain-specific) corpus-based studies should be conducted for finding possible patterns of these constructions, and then the patterns should be divided into categories, depending on the requirements of the task (e.g. event coreference resolution or temporal analysis).

Particle verb constructions and idiomatic verb expressions. The full "eventive" meaning of a verb expression can be a result of combining the verb with other words, such as in the case of particle verb constructions (e.g. *kokku saama* 'to get/come together') and idiomatic verb expressions in general (e.g. *jalga laskma* literally: 'to let the leg (=to run away)').

The TimeML guidelines for annotating particle verb constructions and idiomatic expressions stated that only the verbal part of the predicate should be annotated as an event mention (Saurí et al., 2006). However, when the focus shifted from temporal analysis to event coreference resolution, later guidelines instructed one to annotate all words in phrasal and idiomatic verb expressions, even in the case of a discontinuity in a phrase (e.g. *The actress **passed** yesterday **away** following a serious illness*) (Recasens, 2011; Lee et al., 2012; Cybulska and Vossen, 2014a).

We agree with the recent proposals that particle verb constructions and idiomatic verb expressions should be annotated as full length event expressions, as it supports coreference detection (e.g. enables matching synonymous mentions such as *jalga laskma* 'to run away' and *põgenema* 'to escape' – expressions which otherwise are rooted in lexically and semantically different main verbs). Detection of

particle verb expressions is also readily available for Estonian, integrated within syntactic parsing (Muischnek et al., 2013); however, automatic annotation of idiomatic verb expressions in general is still an open area of research.

In conclusion, we have argued that the understudied gap in main verb centric event annotation in Estonian is the event segmentation and event extent determination in multi-word main verb constructions. This also seems to be an open debate in TimeML-based approaches in general (as we observed from the divergences in event annotation guidelines). As multi-word constructions are relatively rare, the Estonian TimeML annotated corpus offers only limited coverage of such constructions, and can only provide a starting point for investigation. We propose that types of these constructions (e.g. catenative verb constructions and verb+noun constructions) require separate corpus-based domain-specific studies, which would use syntactic dependency relations and/or collocation extraction to find candidates of verbal multiword expressions, and then would cluster/classify these constructions manually or semi-automatically based on their segmentation possibilities and more generally based on their "eventive" properties. We also suggest that the segmentation may depend on the task (e.g. event coreference or temporal analysis), but this still needs to be confirmed by corpus-based studies. We envision that syntactically grounded automatic event mention annotation could be divided into three steps: 1) the detection of verb chains, which should capture regular multi-word verbal constructions neutral to their "eventive" properties; 2) filtering of the detected verb chains based on the desired event model (e.g. removal of modal or negation constructions); 3) segmentation of event annotations on the detected verb chains (i.e. deciding whether the construction should stand as a single word or multi-word event mention, possibly by considering the requirements of the task).

Annotation of non-verb event mentions. Considering that the Vendlerian and Reichenbachian conceptualisation of (temporal) events revolves around verbs, some of the TimeML related work has focused strictly on verbal event mentions and has excluded non-verb mentions altogether, e.g. Mani and Schiffman (2005); Puscasu and Mititelu (2008); Derczynski and Gaizauskas (2011). If we are to make a rough generalisation from English TimeML guidelines (Saurí et al., 2006; Saurí et al., 2009; Cybulska and Vossen, 2014a), with an admitted loss of some specific details, it appears that: 1) most of the annotation of non-verbs focuses on nouns, adjectives and prepositions; 2) out of the three parts-of-speech, only noun annotations cover a wide range of syntactic positions, as annotations of adjectives and prepositions are limited to predicative complement positions. In the following discussion, we will also focus on "eventive" nouns, considering that the annotation of other non-verb event mentions is a problem related to deciding the extent

of the main verb predicate.

The literature suggests that the annotation of "eventive" nouns is a difficult task, which likely requires a specialised annotation project and guidelines, as otherwise biased or inconsistent annotations can easily emerge. Caselli et al. (2008) compare English and Italian TimeML annotated corpora, and they argue that English annotations contain an "over-extension of the eventive reading to almost all occurrences of nominalization" (e.g. 100% of occurrences of the noun *agreement* were marked as events in English data, compared to only 43% of occurrences of the same meaning noun in Italian data), which they believe to be a result of annotator biases. Sprugnoli and Lenci (2014) observe that non-experts ("crowdsourcing contributors") obtain rather problematic inter-annotator agreements on the task of annotating event nouns, compared to the agreements between experts "specifically trained on the task".

The literature also suggests that the annotation of "eventive" nouns is more approachable as a domain-specific problem, rather than as a general domain problem. As Zhou and Hripcsak (2007) point out, "in general linguistics, events are often expressed by tensed and untensed verbs and nominalizations", while in the medical domain, "events are largely expressed by nouns or noun phrases". Galescu and Blaylock (2012) introduced an adoption of TimeML for the clinical domain, and they considered medical concepts (such as medical problems, treatments, and tests) as events, which effectively restricted their event mentions to noun phrases only. Though some of these concepts were not event mentions in strict TimeML interpretation, but "rather entities which participated in some event", annotators interpreted them as referring to "the event they were most closely associated with", e.g in sentence *Her blood pressure was measured at 240/120*, the noun phrase *her blood pressure* was annotated as a reference to the event of measuring (Galescu and Blaylock, 2012).

While the work on Estonian TimeML annotation included the annotation of event nouns, making the assumption that the annotation of event nouns is a task comparable to the annotation of event verbs, we now consider it (based on the literature and the annotation experience) as a problem requiring a separate study. We also suggest that it should be investigated whether the task should be approached in a domain-specific manner. For example, whether it helps to design a separate annotation approach for event nouns in sports news, considering the domain-specific vocabulary (e.g. *obstruction*, *centre pass*, or *half-time*) and ambiguities (e.g. deciding whether *half-time* refers to a time period or to an event).

5.2 Determining Event Relevance

Grammatical grounding of "eventiveness" is difficult, and a part of this difficulty seems to stem from different ***relevance perspectives*** motivating the analysis. Events in text can be analysed for their temporal relevance (whether they can be placed on the timeline or not), for their factual properties (whether they refer to the realis/actual real-world occurrences, rather than to hypothetical, speculated or suggested occurrences), and/or for their specificity (whether they refer to concrete event instances rather than to types or generic classes of events). However, focusing on one relevance perspective does possess a risk of arriving at a distinction that is incompatible with other relevance perspectives. To illustrate this problem, we'll briefly describe three relevance perspectives that mention level event analysis studies have accounted for,[6] and some of the disagreements on these perspectives that we have noted in the literature.

Temporality. TimeML annotations essentially aim to support the "temporal awareness" of event analysis (Mani et al., 2005), focusing on analysing events that "can be temporally ordered" (Saurí et al., 2005). However, it seems that TimeML's scope for temporality is too wide, and subsequent works have argued for more restrictive interpretations.

Saurí et al. (2005) introduced the open-domain text analysis tool EVITA, aimed at "locating and tagging all event-referring expressions in the input text that can be temporally ordered", following the TimeML event specification. This tool was later used by Chasin (2010) for creating timelines of historical Wikipedia articles, and the author argues for a need of additional, machine-learning based filtering of EVENT annotations, because not all of these are suitable for being placed on a timeline, despite the initial aims of the creators of EVITA.

The TERENCE annotation format attempts to simplify TimeML event mention annotations, excluding "events within direct speech" and "negated, modal and hypothetical events" altogether, as these events "can be quite difficult to place along a story timeline" (Moens et al., 2011; Bethard et al., 2012).

And a recent proposal of temporal relevance comes from the TimeLine task (Minard et al., 2015), where the focus is on events "in which target entities explicitly participate" and "that could be placed on a timeline". This focus also requires a simplification of the TimeML model, excluding "adjectival events, cognitive events, counter-factual events (which certainly did not happen), uncertain events (which might or might not have happened) and specific grammatical events" (Minard et al., 2015).

[6] This list is by no means a comprehensive overview of all relevance perspectives discussed in the literature, see Monahan and Brunson (2014) for a longer list of seven "qualities of eventiveness".

Specificity. *Specificity* can be defined as the degree to which an event is "well-individuated from others" (Monahan and Brunson, 2014). Having clear temporal boundaries does contribute to the specificity of event; for example, *The chicken laid an egg on Tuesday* can be considered as a specific event description, while *Chickens lay eggs when fertile* is a generic one (Monahan and Brunson, 2014). The problem of whether generic mentions should be annotated as events has been a persistent source of disagreement.

The English TimeML guidelines opt for specificity, and ask annotators to exclude generic event mentions, "even though capturing them could be of use in question answering" (Saurí et al., 2006). A Persian adaptation of TimeML choose to annotate generics ("for simplicity") (Yaghoobzadeh et al., 2012), and a Croatian adaptation choose to skip them (Marovic et al., 2012). Recent TimeML related event annotation guidelines for English – guidelines for EventCorefBank+ – suggest that mentions of "abstract and generic" events should be annotated, as should be the coreference relations involving them (Cybulska and Vossen, 2014a).

An interesting borderline case between generic and specific events is the class of mentions of recurring events ("habitual events", such as in *John taught on every Monday*). If the recurrence is expressed by a TimeML compatible temporal expression which allows one to place it on the timeline, one can argue that the mention should be annotated from the perspective of temporality, but should be left out from the perspective of specificity.

More recently, researchers have argued that in order to advance our understanding of genericity, separate annotation studies are required, and they have proposed a separate task for clause level annotation of event genericity, along with the annotation scheme and annotated corpora (Friedrich et al., 2015).

Factuality. Researchers have also outlined the need for distinguishing mentions of events that have actually happened or happen in the real world from other, hypothetical, speculated and/or negated ones. According to Monahan and Brunson (2014), event predicates possess the quality of *actuality*, which "refers to whether an action is realis or irrealis, that is, whether or not it actually occurs". According to Saurí (2008), the corresponding quality is *factuality*, which is (according to the author's model) a combination of an event's probability ("degree of certainty that the informant has about an event taking (or not taking) place in the world") and polarity ("whether the informant regards the event as referring to a situation that takes place in the world") at a specific time point, according to a specific source ("informant"). Saurí (2008) argues that *factuality* "is not one of the inherent features" of the event, but "a property relative to sources", so this seems to distinguish *factuality* from the concept of *actuality* proposed by Monahan and Brunson (2014).

Researchers have argued for improving TimeML event annotations by either adding a layer of information about the factuality to the TimeML annotations, (Saurí, 2008) or by straightforwardly focusing only on annotation of factual event mentions (Glavaš et al., 2014); a TimeML adaptation for Croatian started with a focus on realis events (on "events that are asserted to have already happened, are happening, or will happen") (Marovic et al., 2012), and then further improved the event annotations by adding a layer of factuality information (Glavaš et al., 2012).

Integrating the relevance perspectives of factuality and temporality does pose an interesting question regarding future events. From the perspective of temporality, future events that can be placed on a timeline are rightful representatives of the event category. However, as Monahan and Brunson (2014) note, future events can be considered as irrealis, as they have not yet happened, and thus could be discarded from the perspective of factuality; and some researchers have also done this in practice (Nothman et al., 2012; Nothman, 2013).

Considering the different relevance perspectives discussed in the literature, we agree, in general, with the claim of Monahan and Brunson (2014) that "eventiveness" can be (and should be) investigated along "several dimensions". The current work on Estonian mention level event annotation has only considered the temporality perspective outlined in TimeML, and thus can be improved by adding (at minimum) analyses from perspectives of factuality and specificity. That being said, we do note that the relevance perspectives proposed so far are intertwined to a degree, and the extent to which they can be disentangled remains an open question.

5.3 Towards Event Semantics

Taking TimeML-based mention level event analysis as a starting point, we will now discuss future research that could advance the automatic processing of "eventive" semantics in Estonian. We will distinguish four sub-problems, which can be, to an extent, independently approached: 1) developing four component light-weight event models; 2) improving event classification; 3) improving models of event-event relations; 4) improving models where events in text are considered as references.

Light-weight event models. In Chapter 1, we laid out the general goal of advancing four component light-weight event models, which combine annotations of event mentions, temporal expressions, location expressions, and participant mentions. In the practical part of the work, we focused on a detailed study of

event mention and temporal expression annotation, leaving development of other components as future work.

Although we have developed a manually annotated corpus of Estonian event mentions (introduced in Chapter 3), the inter-annotator agreement studies on the annotation and the literature review (in Sections 5.1 and 5.2) suggest that the annotation is still not reliable enough to be automated.[7] The studies made in Chapter 3 suggest an alternative, a trivial model – marking all members of the syntactic predicate as events – which achieves the highest agreement among annotators; however, this model covers approximately only 57% of all potential event mentions and can be implemented by simple rules; as discussed in Section 5.1, building a more advanced model has the prerequisite of further studying verbal multiword expressions and event nouns, possibly in a domain-specific manner.

We have developed an automatic temporal expression tagger for Estonian, and shown it obtains relatively high performance on various sub-genres of formal written language: on news texts, parliamentary transcripts, and law texts. As for future research on improving the tagger, we see several normalisation related questions that can be specifically studied, e.g. distinguishing between a general and concrete meaning of a temporal expression, and experimenting with different anchoring strategies. However, the primary research issue concerns domain adaptation: special sets of rules need to be developed for adapting the tagger to other types of texts, such as narrative texts (encyclopaedic and fiction) and biomedical texts.

Chapter 3 revealed a discrepancy between inter-annotator agreements on temporal tagging and the performance levels of automatic tagging, so future research should also investigate the consistency issues of manual tagging more closely, and re-evaluate the performance of the current automatic tagger (provide several independent expert evaluations).

As for the remaining event component annotations, the named entity recognizer of Estonian (Tkachenko et al., 2013) can be used for acquiring location and participant (person and organisation) annotations. To our knowledge, Estonian still does not have a NLP module available for normalisation of location mentions (e.g. in terms geographic coordinates), so this would be one important future development. Participant-denoting named entities also require disambiguation, which can be performed in terms of named entity coreference resolution (i.e. clustering named entities based on coreference), and at a more advanced level, in terms of linking entities to an external knowledge source, such as Wikipedia. Work on these tasks is still at the first stages in Estonian: while there are some experiments on syntax based anaphora resolution (Puolakainen, 2015), corefer-

[7] Although an event mention pre-annotator could be implemented, which would provide initial annotations to be post-corrected manually.

ence resolution between named entities still needs to be developed; and linking entities to Wikipedia likely has the general prerequisite that Estonian Wikipedia must increase in size.

In addition to named entity based location and participant models, syntax and WordNet based models can also be experimented with. Participants can be approximated as words in syntactic `SUBJECT` and `OBJECT` positions, and locations can be approximated as words carrying locative case markings and appearing in syntactic `ADVERBIAL` positions. WordNet based constraints can be used to the increase the precision of these models, e.g. to allow only participant denoting words that refer to human participants (as can be traced via hypernym relations), and to allow only location denoting words that refer to a physical location (or object).

Event classification. While TimeML introduces event classes that are arguably domain independent ("not restricted to a specific domain") and "relevant for characterizing the nature of the event as irrealis, factual, possible, reported, etc." (Saurí et al., 2005), subsequent studies have also offered some criticism on the class system, both at the theoretical and practical level.

As Kotsyba (2006) notes, it is unclear how classes distinguished by TimeML are specifically related to time (how they contribute to revealing the temporal semantics of the events). The author also suggests that instead of the current classification, event classes could be designed based on the distributions of temporal expressions appearing in event contexts. There is some research trying to give event annotations proper temporal grounding: research aiming to learn typical durations of events, e.g. that the duration of "war" typically ranges from months to years, and that the duration of "look" ranges from seconds to minutes (Gusev et al., 2011), but to our knowledge, the empirical results of this research have not yet been refined as event classification proposals.

The second problem with the TimeML's event classification is it is difficult to apply in practice. For example, Robaldo et al. (2011) reported an accuracy of approximately 70% for Italian event classification; in TempEval-2, best results were close to 80% for English and below 70% for Spanish (Verhagen et al., 2010). Investigating the problem more closely, Llorens et al. (2010) notes that high accuracy classification of events (on the English TimeBank corpus) is only achieved for the REPORTING class (F1-score of 90.51%), the next best-classified is the general class OCCURRENCE (F1-score close to 70%), and F1-scores for other classes remain below 70%. Similar observations were made on the manual classification of events in Croatian, where the highest agreements were on REPORTING and OCCURRENCE classes, F1-scores 82.1% and 65.4% respectively (Marovic et al., 2012). This does suggest that apart from the REPORTING class,

which is based on patterns of reported and direct speech, TimeML's event classes may not be well-aligned with distinctions made in conventional language usage, and this may also be the reason why they are difficult to distinguish at an empirical level.

Some researchers have proposed simplifications to TimeML's event classification, or have chosen alternative approaches altogether. Puscasu and Mititelu (2008) and Cybulska and Vossen (2014a) choose only five classes from the inventory of TimeML – REPORTING, PERCEPTION, ASPECTUAL, OCCURRENCE and STATE – excluding the classes I_ACTION and I_STATE. An alternative classification is employed in the NewsReader project, where only three broad semantic types of events are distinguished: 1) grammatical events (which "do not represent instances of events directly but express properties of events or relations between events", such as aspectual and causal relations); 2) speech acts or cognitive events (that "may be seen as provenance relations or as expressions of opinions"); 3) contextual events (all the remaining event mentions, which "usually describe the actual changes in the world") (Vossen et al., 2014a). Finally, arguments have been made for avoiding event classification altogether: Nothman (2013) argues that if the focus of the analysis is on an event coreference detection task (more specifically: linking event mentions to the sources that first reported them), there is no need for event classification.

Our experience with Estonian TimeML annotations also confirms that consistent event classification is difficult to achieve: agreements (F1-scores) between initial annotators ranged from 0.51–0.82, and agreements with the judge ranged from 0.53–0.91. This does suggest that an alternative, and more simple classification of Estonian event mentions should be explored in the future, possibly a classification in line with NewsReader's three class classification.

Event-event relations. Studies of narratology propose that the semantics of events have a lot to do with an events' relations to other events. If we view texts as narratives (which is, according to Bell (1999), a reasoned view in the case of news articles), we may consider Bal's perspective (Bal, 1997) that events "become meaningful" only "in series", and "it is pointless to consider whether or not an isolated fact is an event". This does suggest 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 TimeML does focus on temporal relations, rather than on events; however, because of the decomposition of the task, one employing the framework could easily get stuck with the problems of event annotation/analysis (e.g. how to ground the concept of event at the grammatical level), and may be

hindered from reaching temporal relation annotation. We see it also worthwhile experimenting with a simpler annotation model focusing directly on relations, without the decomposition of annotations into events and relations.[8] Focusing straightforwardly on the task (annotation of relations) could help to achieve higher inter-annotator agreement levels, and could enable simplified designs for machine learning set-ups, which, in turn, could foster more experimentation and hopefully improvements on current results.

The Estonian TimeML corpus created in our work contains both event mention and temporal relation annotations, thus, it can be used as a basis for experimenting with simplified temporal relation annotation models, and can also be used as a starting point for improving Estonian TimeML annotations.

A simplified model would approximate TimeML's events to verbs in syntactic predicates (a model that was supported by the highest inter-annotator agreement in the experiments), and at the extreme, would lose event annotation (event classification) altogether, considering syntactic roots as nodes to be connected with temporal relations. This model would aim at basic grounding between syntax and temporal semantics, keeping only TLINKs that are aligned with dependency syntactic relations, and TLINKs that connect syntactic root nodes of consecutive sentences. While this model does simplify the problem (e.g. by not considering non-verb event mentions), this model could be the first one to be evaluated domain-wise (e.g. considering text domains other than news available in the Estonian Dependency Treebank (Muischnek et al., 2014b)) before starting to develop more complex models.

As for improving Estonian TimeML annotations, we see future work branching in several directions: the corpus can be extended with aspectual and subordination (ALINK, and SLINK) relations between events (semiautomatic creation of these relations can be experimented with, based on syntactic dependency relations), temporal relation agreements could be evaluated in the case of a simplified set of relation types and, most importantly, the corpus should be extended with new annotated texts (including texts from other domains), possibly via the method of automatic pre-annotation, and manual post-correction.

A gap in Estonian temporal relation annotation that also requires a separate study is the annotation of explicit temporal signals (e.g. the adverbs *enne* 'before', or *hiljem* 'later'), and the temporal relations conveyed by them. Signal annotations can be added to the Estonian TimeML corpus, but because of the small size of the current corpus, a better approach would be to extend the corpus, and to annotate temporal signals in the larger Estonian Dependency Treebank. An

[8] We are aware of the pre-TimeML work proposing a similar idea: Katz and Arosio (2001) did not use event annotation and simply marked temporal relations on verbs. We think that this branch of research could also be advanced, in parallel to TimeML's.

ultimate goal would be the creation of a temporal relation annotator that specifically aims at high precision annotations in the contexts of explicit temporal cues (temporal expressions, temporal signal adverbs, and past tensed verbs).

Event mentions as references. An alternative to directly representing event mention semantics is to consider the mention as a reference, and to try to solve its target. This can be done either in the context of the event coreference resolution task, which aims to cluster together mentions that refer to the same event, or in a more general event grounding task, where the event reference is associated with the article describing the event (e.g. Wikipedia article, or the first-reporting news article) (Nothman, 2013), or with event information from extra-textual sources (e.g. videos, or pictures) (Fokkens et al., 2013).

In the context of analysing news articles, the idea to focus "on referent over semantics" has been recently promoted by Nothman et al. (2012); Nothman (2013). The author proposes a new event analysis task–*event linking*–where the goal is to link past event mentions in a news article to "the first story that reports the event in the news archive". Results of this task would be immediately useful to the corpus user (a news reader), allowing them to follow the link and to read the complete story behind the event mention (and to decide on the event's "semantics" based on the story). Arguably, this task also eases the problem of resolving event mention coreference: instead of finding an exact (mention level) co-referent, the co-referent is sought at the document level, diminishing the problems of partial coreference (Hovy et al., 2013) that appear at mention level coreference resolution.

Considering that the semantic resources supporting event analysis are limited in Estonian, we also see the benefit of focusing more on "solving reference rather than semantics". This could both advance our understanding of the event phenomenon (e.g. there is evidence that humans learn better about events when integrating information from multiple sources (Zwaan and Radvansky, 1998)), and could also provide data (e.g. sentences aligned by event coreference) for building more advanced semantic level analysis tools (e.g. semantic role labellers).

An available resource that supports cross-document event coreference detection studies is the Estonian Reference Corpus (Kaalep et al., 2010), which has a large newspaper section (with the total size of approximately 200 million word tokens). Detecting event coreference in this corpus could be driven by a practical information extraction/organisation task, such as the task of finding events related to a person as explored in Section 4.3. This would enable more extrinsic evaluation of the task, based on concrete user information needs, and such evaluation would also provide better knowledge about whether it is necessary to find "all coreferring event mentions", or whether it suffices to focus on some subsets of

events (e.g. events mentioned in article summaries). Another question that can be explored is how the media coverage patterns affect the results: e.g. does lower-/higher media coverage related to the person searched for affects the difficulty of the task?

From the perspective of event coreference models, we suggest that models with different textual granularity could be explored. In addition to the model matching two event mentions at the sentence level, which were explored in Section 4.3, it would also be worthwhile to explore more coarse-grained models, such as a model matching a sentence level mention to a paragraph level mention, a model matching a sentence to whole document (similarly to the event linking task in Nothman (2013)), or models matching a paragraph to a paragraph, or a paragraph to a document. In addition to potentially making the task easier, this also explores the intuition that events should be analysed in series (Bal, 1997), rather than as stand-alone units.

5.4 Final Notes on TimeML

In this chapter, we have pointed out several unstudied questions in relation to TimeML-based event annotation. We have argued that it is an open issue how "eventiveness" can be grounded on a range of syntactic structures (e.g. on support verb constructions); we have pointed out that in addition to the temporal relevance perspective, other relevance perspectives, such as factuality or specificity, should be taken into account; we have outlined future research directions in resolving "eventive" semantics: improving base annotations for four component event models, simplifying event classification, improving event-event relation annotations, and improving the models where an event is considered as a reference. Most of these issues seem to concern the compositional build-up strategy for semantics, where one first models semantics related to fine-grained event mentions (e.g. determines temporal ordering of consecutive main events), which then helps to infer something about the text in general (e.g. to induce that the events in text follow chronological order to a large extent). However, to our knowledge, the opposite direction of knowing something about the genre and in particular about the *type of text*, and then using this knowledge to make informed decisions about how and when to perform mention level event analysis, has not been explored much, or at least not systematically.

We noticed during our studies of news texts that there can be types of articles that tend to report more factual information and/or information in narrative-like form, such as weekly crime chronicles, reports on traffic issues/accidents, news briefs, short biographical summaries, and historical backgrounds of contemporary events. Rather than offering discussions on debates, opinions, plans, or intentions,

these texts seemed to be more oriented on what can be seen as the first function of news: giving an overview about the occurrences. These texts could also have relatively clear linguistic characteristics, such as a tendency to use more past tense, and temporal expressions, and a tendency to use less direct speech (which refers to discussions and commentaries, rather than reportings), intention verbs, and opinion words. Considering this, we suggest that TimeML-based text analysis could need a pre-step, which would classify a text based on its temporal "analysability": i.e. how suitable is the given unit for temporal fact extraction? While the high goal of analysing all fine-grained event mentions, both factual and debatable, is an admirable ideal, one should have a more firm grounding of the analysis at a coarser level: what are the types of texts that can be analysed with high confidence, and what types of texts likely remain temporally "vague" from the perspective of our current knowledge? This text typology is something that is not readily available, but needs to be established by corpus-based studies, with the help of temporal tagging tools, and the analysis of explicit temporal relations.

CHAPTER 6

CONCLUSIONS

Motivated by successes with automatic linguistic analysis at the grammatical level, we have explored the question whether fine-grained (word-, phrase- and sentence-level) event analysis can be considered a task similar to grammatical analysis: a task that can be approached in a broad-coverage and general domain manner.

Approaching the problem from a language specific–Estonian–perspective, we made detailed studies on time-oriented event analysis, focusing on TimeML-based temporal expression tagging, event mention and temporal relation annotation, and explored the possibilities of extending the approach as a generic fine-grained event analysis, which forms a basis for light-weight event representations (representations covering event, participant, time, and location expressions), and for event coreference detection.

In the study of ***temporal tagging***, we have contrasted TimeML's temporal expression model with the concept of temporal adverbials in Estonian grammatical tradition, showing that the TimeML's model is largely centred on calendric information, while Estonian grammatical tradition systematises temporal adverbials based on morphological and syntactic cues, and has a larger scope, which also includes event-denoting words as temporal adverbials. We have therefore proposed criteria for distinguishing between markable temporal expressions and the remaining temporal adverbials yet out of the scope of the automatic analysis. We have also argued that Estonian TIMEX format should apply a different phrase segmentation than the (English-specific) TIMEX3 format, as Estonian temporal expressions need not be segmented into smaller phrases due to intervening temporal signal words, and can be captured as full-length phrases.

Considering the characteristics of Estonian, we have developed a rule-based language-specific temporal expression tagger for the language, which allows a free-word-order-aware composition of rules, and takes advantage of morphological analysis both in the extraction phase (making lemma based extraction rules)

and in the normalisation phase (using verb tense information for normalisation). While we have developed the system mainly on news domain texts, obtaining a relatively high performance there, the evaluation shows that comparable performance levels are also obtainable on other types of formal written language texts, such as on law texts, and on parliamentary transcripts.

We conclude that implementing temporal tagging at the broad-coverage manner is currently limited to addressing calendar-based temporal expressions, though theoretically, the temporal adverbials cover both temporal and event expressions in Estonian, and a range of domain-specific expressions (such as *teisel poolajal* 'at the second half-time', or *viimasel hooajal* 'on the last season') lay on the borderline between timexes and events. Even when considering only calendar-based temporal expressions, the evaluation results along with the literature suggest that a general domain manner tagging of these expressions could be limited to news-style texts (and to formal written language texts similar to news), as other types of texts, such as historical articles or encyclopaedia texts, likely need alternative tagging strategies. The problem needs further investigations in the future.

The second part of this study focused on the creation of a **TimeML annotated corpus of Estonian** news articles. We have considered TimeML-based event mention, temporal expression, and temporal relation annotations as extensions to dependency syntactic annotations, aimed at a relatively exhaustive event annotation that maximises the coverage in syntactic contexts that can be interpreted as "eventive".

We have conducted a series of retrospective inter-annotator agreement experiments on **event mention** annotations, confirming that the agreements were higher on "prototypical" events: on verb event mentions, and more specifically, on event mentions covering syntactic predicates. Lower inter-annotator agreements occurred outside of syntactic predicates, and on nouns and adjectives, indicating that while in principle the TimeML event model can cover a diversity of "eventive" linguistic contexts, in practice, high agreement on non-verbs is difficult to obtain.

Based on the manual annotation experiment and on the literature review, we conclude that achieving a broad-coverage event mention analysis is currently limited due to difficulties of achieving consistent event mention annotation on multi-word verbal constructions (e.g. copular constructions and support verb constructions), and on non-verb event mentions; and we suggested that there should be a separate preprocessing step for detecting multi-word verbal constructions (independently from their "eventive" interpretations), upon which event mention annotations could be systematically added. The annotation of noun event mentions likely needs a specialised annotation project for achieving high consistency. The

literature also suggests that annotation of noun event mentions could be a problem that needs to be approached in a domain-specific, rather than in a general domain manner.

In the manual annotation of **temporal relations**, we have used a setup similar to TempEval-2, and confirmed the previous findings from English temporal relation annotation about the difficulties of achieving consistent (highly agreed) annotation. In the retrospective analysis of temporal relation annotations, we showed that syntactic contexts with explicit temporality cues (past tense, the presence of a temporal expression) were proportionally deemed as less vague by annotators and exhibited higher inter-annotator agreements, in contrast to contexts characterised by limited/absent temporal cues. These findings also show that usage of the present tense in Estonian (news texts) is a rather ambiguous indicator of temporality. This suggests that rather than considering tense as a uniform Reichenbachian mechanism of temporal semantics, one could distinguish explicit/implicit tenses (with the simple past being explicit, and the present being implicit) by their different usage conventions.

The Estonian TimeML annotated corpus developed in this work has a range of applications: it can be used as a basis for developing machine learned pre-annotators of temporal semantics, for application-driven research on improving TimeML annotations (focusing on applications such as question answering and summarisation), and for linguistic research on categories of future tense and aspect (categories which are not grammatical in Estonian). Being a subcorpus of the Estonian Dependency Treebank, the Estonian TimeML annotated corpus can be extended with other texts (including texts from genres other than news) from the treebank, and the tools developed for retrospective analysis of inter-annotator agreements can also be re-employed on extending the corpus.

In the last empirical part of this work, we changed the perspective from single-document event analysis to multi-document event analysis, namely to the task of **cross-document event coreference detection**. In contrast to the relatively exhaustive event mention annotation that was employed in the creation of the TimeML corpus, we now focused on events that were mentioned across documents, with the goal of finding events related to a specific person from a corpus of daily news. We tested two methods: a method deciding coreference based on overlapping lemmas, and a method deciding coreference based on overlapping temporal and spatial cues. We found preliminary evidence that the lexical homogeneity within the set of articles under analysis could influence the precision of the methods, and that the performance of the methods can also differ when restricted to the summary sentences (the first three sentences of a news article), and when unrestrictedly applied on all the sentences of the article. A large scale empirical

confirmation of these results remains a future work.

Based on the literature and the experiments, we proposed that future research on event coreference detection in Estonian news could investigate two questions. First, the question about whether the trend of summarising central events in the first paragraphs of articles is consistent enough for motivating a distinction of two layers of event mention annotations: events mentioned in the summarising paragraph, and events mentioned in the content paragraphs? Second: does balancing the event coreference corpus in terms of different media coverage patterns reveal any differences in task difficulties? For example, if the goal is to find events related to a specific person, can one expect the task of analysing the events related to a person with high media coverage more difficult than the task of analysing events related to a person with low media coverage?

In conclusion, based on the explorations on creating a broad coverage and general domain fine-grained event analysis tool for Estonian, we suggest that the current challenges of the research are: grounding event mentions on a range of syntactic structures (on multi-word constructions, and on non-verbs), exploring additional relevance perspectives as the basis for the event analysis (e.g. focusing on factuality instead of temporality), and finding more agreeable event classification (determining which event typology should be used, if any?). We also suggest that there is a need for a firmer grounding of event analysis at a coarser level: what are the types of texts or subsections of texts (e.g. summarizing sections) that can be analysed (both manually and automatically) with high confidence, and what texts/subsections likely remain problematic in their "eventive" interpretations?

ACKNOWLEDGEMENTS

During my studies, I received financial support from Estonian IT Academy program, from the European Regional Development Fund through the Estonian Centre of Excellence in Computer Science (EXCS), from Estonian Ministry of Education and Research (grant IUT 20-56 “Computational models for Estonian”), and from European Social Funds Doctoral Studies and Internationalisation Programme DoRa.

I wish to thank the pre-reviewers of this thesis, Associate Professor Dr. Agata Savary and Dr. Uuno Vallner, for the valuable feedback on this work.

I am very grateful to my supervisor Heiki-Jaan Kaalep for his guidance and patience on my work, and also for provoking me to consider the broader/philosophical context around the research problem. I thank Kadri Muischnek for providing me help with the annotation experiment, and for guiding me on the issues related to syntactic analysis, and Professor Haldur Õim for the discussions on event analysis. My special gratitude goes to the participants in the annotation project, and to my colleagues at the Research Group of Computational Linguistics at the University of Tartu.

And last, but not least, I am very grateful to friends, who have accompanied me on this journey, and to my parents and my brother for their constant support and faith in me.

KOKKUVÕTE (SUMMARY IN ESTONIAN)

Eesti keele üldvaldkonna tekstide laia kattuvusega automaatne sündmusanalüüs

Mitmete praktiliste keeletehnoloogia rakenduste – nt automaatsete küsimus-vastus süsteemide ja automaatsete sisukokkuvõtjate – kirjeldamisel ja kavandamisel on hea kasutada mõistet "sündmus" viitamaks (teatud määral) terviklikule infoüksusele, mis koondab endas vastuseid nt küsimustele *kes tegi mida? kus?* ja *millal?* Kuna paljud tekstid on narratiivse ülesehitusega, tõlgendatavad kui "sündmuste kirjeldused", siis viib sündmuskirjelduste tekstist ekstraheerimine ja formaalsel kujul esitamine meid ka lähemale "tekstide sisu/tähenduse" automaatanalüüsile.

Senised lähenemised sündmuste automaatanalüüsile eestikeelsetes tekstides on keskendunud probleemi uurimisele eeskätt teoreetilises plaanis ning eeldefineeritud sündmuste hulga puhul. Käesolevas töös keskendutakse probleemi empiirilisele poolele ning uuritakse, kuivõrd saab sündmusanalüüsi – sarnaselt tekstide grammatilisele (morfoloogilisele/süntaktilisele) analüüsile – käsitleda kui *laia kattuvusega* (st avatud sündmuste hulgale orienteeritud) ja *üldvaldkonna* tekste (siin: eeskätt ajakirjandustekste, lähtuvalt selle valdkonna heterogeensusest) hõlmavat keele automaatanalüüsi ülesannet.

Käesoleva töö (filosoofiliseks) lähte-eelduseks on, et sündmused on eelkõige ajas paigutuvad entiteedid ning sündmuste analüüsi tuleks alustada nende ajaliste omaduste (sündmusi hõlmavate ajaseoste) määratlemisest. Selleks otstarbeks kohandatakse eesti keelele TimeML märgendusraamistik ja luuakse raamistikule toetuv automaatne ajaväljendite tuvastaja ning ajasemantilise märgendusega (sündmusviidete, ajaväljendite ja ajaseoste märgendusega) tekstikorpus; analüüsitakse inimmärgendajate kooskõla sündmusviidete ja ajaseoste määramisel ning lõpuks uuritakse võimalusi ajasemantika-keskse sündmusanalüüsi laiendamiseks genereeliseks sündmusanalüüsiks sündmust väljendavate keelendite samaviitelisuse lahendamise näitel.

Esimeses peatükis antakse ülevaade "sündmuse" mõiste defineerimisraskustest filosoofias, tutvustatakse senist, info ekstraheerimise keskset lähenemist sündmusanalüüsile, püstitatakse eesti keele morfoloogilise ja süntaktilise analüüsi eeskujul üldvaldkonna tekstide laia kattuvusega sündmusanalüüsi probleem ning määratletakse eksperimentaalsed sündmusmudelid ja rakendused, mida sündmusanalüüs peaks toetama.

Teine peatükk keskendub automaatsele *ajaväljendite tuvastamisele* kui sündmusanalüüsi toetavale keeleanalüüsi ülesandele. Teoreetilises osas piiritletakse märgendamisele kuuluvad ajaväljendid ja märgendusviis. Pakutakse välja (väljendite kalendrilisusele tuginevad) kriteeriumid, mis võimaldavad eristada märgendatavaid ajaväljendeid eesti keele grammatikatraditsioonis eristatavatest ajamäärustest. Samuti võrreldakse eesti keele ajaväljendite märgendusviisi inglise keele spetsiifilise (TimeML TIMEX3) märgendusviisiga ning leitakse, et eesti keeles tuleks eelistada fraaside märgendamist terviklikena (vastandina inglise keele puhul rakendatavale fraaside segmenteerimisele).

Lähtuvalt eesti keele ajaväljendimärgenduse omapärast luuakse keelespetsiifiline automaatne ajaväljendite tuvastaja – programm, mis leiab tekstis üles ajaväljendid ning esitab nende semantika standardsel viisil. Tuvastaja on ülesehituselt reeglipõhine ning võimaldab tuvastamisreeglite ehitamist morfoloogilise analüüsi väljundile ning eesti keele vaba sõnajärje arvestamist tuvastamisreeglite koostamisel. Kuigi süsteemi on peamiselt arendatud ajakirjandustekstidel, näitas töös läbiviidud hindamine, et süsteem säilitab suhteliselt kõrge täpsuse ka teistel formaalse laadiga kirjakeele tekstidel: seadustekstidel ja parlamendi istungite transkribeeringutel.

Teise peatüki lõppjärelduseks on, et TimeML-põhise ajaväljendite tuvastamise ulatus hõlmab peamiselt kalendripõhiseid väljendeid, samas kui grammatikateooria järgi hõlmavad ajamäärused ka mitte-kalendrilisi väljendeid ning sündmusviiteid, mille süstemaatiline tuvastamine ja semantika esitamine üldvaldkonna viisil on aga suuresti läbiuurimata. Viimased uurimused inglise keelel on tõstatanud ka küsimuse, kuivõrd üldse saab TimeML-põhist ajaväljendite tuvastamist käsitleda üldvaldkonna ülesandena, kuna ajalehetekstidel väljatöötatud semantika leidmise strateegiad ei ole tulemuslikud muudel tekstiliikidel, nt teatmeteoste (entsüklopeediate) artiklitel ja meditsiinitekstidel. Probleem vajab tulevikus uurimist ka eesti keelel.

Kolmandas peatükis antakse ülevaade *TimeML ajasemantilise märgenduse* (märgendus, mis hõlmab lisaks ajaväljenditele ka sündmusviiteid ja ajaseoseid) kohandamisest eesti keelele. Töö käigus luuakse TimeML märgendusega ajalehetekstide korpus, võttes aluseks käsitsi parandatud sõltvussüntaktilise märgendusega tekstid ning püüdes maksimiseerida sündmusviidete märgenduse ulatust üle "sündmuslikena" tõlgendatavate süntaktiliste kontekstide. Sündmustevaheline

ning sündmuste ja ajaväljendite vaheline ajaseoste märgendus jagatakse alametappideks rahvusvahelise TempEval-2 märgendusvõistluse eeskujul.

Loodud märgenduse kvaliteedi hindamisel keskenduti inimmärgendajate vahel saavutatud märgenduse kooskõla mõõtmisele üle erinevate süntaktiliste kontekstide. Sündmusviidete märgendust analüüsides leiti, et suurema kooskõlaga oli just nn prototüüpsete sündmusviidete – verbide ja täpsemalt süntaktilisse predikaati kuuluvate verbide – märgendus, samas kui käändsõnaliste sündmusviidete ning väljaspool süntaktilist predikaati paiknevate sündmusviidete märgendamisel oli kooskõla madalam. Ajaseoste kooskõla hindamisel keskendutigi süntaktilisse predikaati kuuluvatele sündmusviidetele. Leiti, et lihtminevikus verbide ja ajaväljendite kontekstis oli ajaseoste märgendus kooskõlalisem ning ka selgemini määratletav (st märgendajad kasutasid vähem ebaselgusele viitavaid VAGUE märgendeid), võrreldуna kontekstidega, kus verbid olid olevikus ning ajaväljendid puudusid. Madalam kooskõla olevikuverbide kontekstis viitab ka sellele, et verbi grammatilised ajad (lihtminevik ja olevik) võivad küll teoorias olla võrdsed reichenbachilikud ajatähenduse väljendamise mehhanismid, ent praktikas näib ajatähenduse selgus sõltuvat kasutuskonventsioonidest: lihtminevikuaeg on ajalehetektsides sageli selgema/ilmutatud ajatähenduse kandjaks, samas kui olevikuaeg on raskemini tõlgendatav.

Kolmanda peatüki esimeseks lõppjärelduseks on, et kuigi TimeML sündmusmärgendus peaks (definitsiooni järgi) hõlmama "sündmusliku" tähendusega sõnu verbidest käändsõnade ja muutumatute sõnadeni, on praktikas probleemne kooskõlalise märgenduse saavutamine mitmesõnalistel üksustel (mis pole selgelt määratletavad "süntaktilise predikaadina") ning käändsõnalistel üksustel. Viimaste uurimuste järgi teistel keeltel tuleks käändsõnade (nimisõnade) sündmusmärgendust vaadelda ka kui eraldiseisvat uurimisprobleemi. Teiseks lõppjärelduseks on, et uurimustöö eestikeelsete tekstide ajasemantilise analüüsi vallas võiks tulevikus keskenduda ajaväljendeid sisaldavatele kontekstidele, mis pakuvad paremat lähtepunkti kooskõlalise märgenduse ülesehitamisel.

Käesolevas töös loodud eestikeelne TimeML märgendusega korpus omab mitmeid potentsiaalseid rakendusi edasises uurimustöös: seda saab kasutada masinõppepõhise ajasemantika eelmärgendaja treenimiseks, rakendustele-orienteeritud TimeML märgenduse edasiarendamiseks (rakenduste nagu ajaküsimustele vastamine ja automaatne sisukokkuvõtja katsetamiseks) ning ka lingvistilises uurimustöös grammatiliste aja väljendamise vahendite uurimisel. Kuna tegemist on eesti keele puudepanga alamkorpusega, siis on korpust võimalik laiendada uute tekstidega puudepangast (ka tekstidega teistest valdkondadest, nt ilukirjandusest) ning märgenduste kooskõlalisuse hindamiseks loodud tööriistu võib samuti taaskasutada uute tekstide märgendamisel.

Töö eelviimases osas tutvustati esmaseid katsetusi *sündmust väljendavate kee-*

lendite samaviitelisuse tuvastamisel üle mitme uudisteksti. Keskenduti infootsingu ülesandele, milles tuli leida samadele sündmustele viitavaid lauseid ühe konkreetse isikuga seotud artiklite hulgast, ning katsetati kaht meetodit: lemmade kattuvusel põhinev samaviitelisuse tuvastamine ning aja- ja kohaväljendite kattuvusel põhinev samaviitelisuse tuvastamine. Esialgsed tulemused näitasid, et tekstidevaheline leksikaalne sarnasus võib meetodite täpsust mõjutada ning et meetodite täpsus võib samuti erineda nende rakendamisel tervikartiklitel võrrelduna rakendamisega artiklite alguses paiknevatel nn sisukokkuvõttelausetel (juhtlõikudel). Eksperimentaalses osas tehtud tähelepanekute ning kirjanduse põhjal püstitati kaks küsimust tulevasteks uuringuteks. Esiteks, kas eesti ajakirjandustekstides jälgitakse piisavalt järjekindlalt artikli alguses (faktiliste) lühikokkuvõtete esitamise konventsiooni, et oleks põhjendatud sündmusanalüüsi meetodite eraldiseisev katsetamine ja hindamine juhtlõikudel ning sisu osades? Ning teiseks, kas oleks põhjendatud sündmusviidete samaviitelisuse uurimine erinevate meediakajastuse mustrite lõikes (nt konkreetse isikuga seotud sündmuste tuvastamisel vaadelda eraldiseisvate probleemidena sagedasti meedias mainitud isikuga seotud sündmuste analüüsi ning harva meedias mainitud isikuga seotud sündmuste analüüsi)?

Kokkuvõtvalt: uurimustöös kaardistati *laia kattuvusega* ja *üldvaldkonna* tekstidele orienteeritud automaatse sündmuste tuvastaja loomisega seotud probleeme eesti keelel. Leiti, et praeguse uurimisseisu juures on peamisteks väljakutseteks: sündmus(viide)te märgendamine spetsiifilistes süntaktilistes kontekstides (käändsõnadel ja mitmesõnalistel üksustel), analüüsi laiendamine uute relevantsusperspektiividega (nt lisaks ajasemantikale ka sündmuste faktilisuse arvestamine) ning kõrgema märgendajatevahelise kooskõlaga sündmuste klassifikatsiooni leidmine (sh võib olla ka põhjendatud klassifikatsioonist loobumine). Samuti leiti, et peenekoelise (sõna-, fraasi- ja lausetasemel) sündmusanalüüsi uurimisel tuleks tulevikus rohkem arvesse võtta lõigu ja tervikteksti tasandit: uurida, milliseid tekstitüüpe ja lõike (nt ajaleheartiklite juhtlõike) saab analüüsida (nii käsitsi kui automaatselt) kõrge kooskõlaga, ning millised tekstid / lõigud jäävad "sündmuslike" tõlgenduste osas problemaatiliseks.

Bibliography

D. Ahn. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8. Association for Computational Linguistics, 2006.

J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic Detection and Tracking Pilot Study. Final Report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

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

J. F. Allen. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154, 1984.

O. Alonso, K. Berberich, S. Bedathur, and G. Weikum. Time-based exploration of news archives. *HCIR 2010*, pages 12–15, 2010a.

O. Alonso, M. Gertz, and R. Baeza-Yates. Temporal analysis of document collections: framework and applications. In *String Processing and Information Retrieval*, pages 290–296. Springer, 2010b.

B. Arnulphy, X. Tannier, and A. Vilnat. Event Nominals: Annotation Guidelines and a Manually Annotated Corpus in French. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1505–1510, 2012.

R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

R. Baeza-Yates. Searching the future. In *Proceedings of ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval (MF/IR 2005)*, 2005.

A. Bagga and B. Baldwin. Cross-document event coreference: Annotations, experiments, and observations. In *Proceedings of the Workshop on Coreference*

and its Applications, pages 1–8. Association for Computational Linguistics, 1999.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley Framenet Project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

M. Bal. *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press, 1997. URL `https://archive.org/details/BalNarratologyIntroductionToTheTheoryOfNarrative`. (Date accessed: 2016-01-25).

J. A. Baldwin. Learning temporal annotation of French news. Master's thesis, Georgetown University, 2002.

C. A. Bejan and S. M. Harabagiu. A Linguistic Resource for Discovering Event Structures and Resolving Event Coreference. In *LREC*, 2008.

A. Bell. News stories as narratives. In A. Jaworski and N. Coupland, editors, *The Discourse Reader*, pages 236–251. Routledge, London and New York, 1999.

A. Berglund. Extracting temporal information and ordering events for Swedish. *Master's thesis report*, 2004. URL `http://fileadmin.cs.lth.se/cs/personal/pierre_nugues/memoires/anders/exjobbsrapport.pdf`. (Date accessed: 2016-10-21).

S. Bethard, O. Kolomiyets, and M.-F. Moens. Annotating Story Timelines as Temporal Dependency Structures. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

A. Bittar. Annotation of events and temporal expressions in French texts. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 48–51. Association for Computational Linguistics, 2009.

A. Bittar. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard*. PhD thesis, Université Paris Diderot, Paris, France, 2010.

R. Casati and A. Varzi. Events. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2014 edition, 2014. URL `http://plato.stanford.edu/archives/fall2014/entries/events/`. (Date accessed: 2015-04-01).

T. Caselli, N. Ide, and R. Bartolini. A Bilingual Corpus of Inter-linked Events. In *LREC*, 2008.

T. Caselli, V. B. Lenzi, R. Sprugnoli, E. Pianta, and I. Prodanof. 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, 2011. ISBN 978-1-932432-93-0.

R. Chasin. Event and Temporal Information Extraction towards Timelines of Wikipedia Articles. *Simile*, pages 1–9, 2010.

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

H. Cunningham. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics*, 5:665–677, 2005.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*, 2002.

A. Cybulska and P. Vossen. Semantic Relations between Events and their Time, Locations and Participants for Event Coreference Resolution. In *RANLP*, pages 156–163, 2013.

A. Cybulska and P. Vossen. Guidelines for ECB+ annotation of events and their coreference. Technical report, NWR-2014-1, VU University Amsterdam, 2014a. URL `http://www.newsreader-project.eu/files/2013/01/NWR-2014-1.pdf`. (Date accessed: 2015-12-20).

A. Cybulska and P. Vossen. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, pages 26–31, 2014b.

M. L. de la Calle, E. Laparra, and G. Rigau. First steps towards a predicate matrix. In *Proceedings of the Global WordNet Conference (GWC 2014), Tartu, Estonia, January. GWA*, 2014.

L. Derczynski and R. Gaizauskas. An Annotation Scheme for Reichenbach's Verbal Tense Structure. *Proceedings of the 6th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, 2011.

B. Dowden. Time Supplement. The Internet encyclopedia of philosophy, 2009. `http://www.iep.utm.edu/time-sup` (Date accessed: 2015-04-01).

T. Erelt, Ü. Viks, M. Erelt, R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, and S. Vare. *Eesti keele grammatika. 2., Süntaks (Grammar of Estonian: The syntax)*. Tallinn: Eesti TA Keele ja Kirjanduse Instituut, 1993.

L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. TIDES 2005 standard for the annotation of temporal expressions. 2005. URL `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-timex2-guidelines-v0.1.pdf`. (Date accessed: 2015-08-05).

E. Filatova and E. Hovy. Assigning time-stamps to event-clauses. In *Proceedings of the Workshop on Temporal and Spatial Information Processing*, volume 13, page 13. Association for Computational Linguistics, 2001.

M. A. Finlayson, J. R. Halverson, and S. R. Corman. The N2 corpus: A semantically annotated collection of Islamist extremist stories. In *LREC*, 2014.

W. R. Fisher. Narration as a human communication paradigm: The case of public moral argument. *Communications Monographs*, 51(1):1–22, 1984.

A. Fokkens, M. Van Erp, P. Vossen, S. Tonelli, W. R. Van Hage, B. SynerScope, L. Serafini, R. Sprugnoli, and J. Hoeksema. GAF: A grounded annotation framework for events. In *NAACL HLT*, volume 2013, pages 11–20. Citeseer, 2013.

N. Friburger and D. Maurel. Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1):93–104, 2004.

A. Friedrich, A. Palmer, M. P. Sørensen, and M. Pinkal. Annotating genericity: a survey, a scheme, and a corpus. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, pages 21–30, 2015.

L. Galescu and N. Blaylock. A corpus of clinical narratives annotated with temporal information. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 715–720. ACM, 2012.

J. Galtung and M. H. Ruge. The structure of foreign news. The presentation of the Congo, Cuba and Cyprus Crises in four Norwegian newspapers. *Journal of Peace Research*, 2(1):64–90, 1965.

G. Glavaš and J. Šnajder. Exploring coreference uncertainty of generically extracted event mentions. In *Computational Linguistics and Intelligent Text Processing*, pages 408–422. Springer, 2013.

G. Glavaš, J. Šnajder, and B. D. Bašic. Are You for Real? Learning Event Factuality in Croatian Texts. In *Conference on Data Mining and Data Warehouses (SiKDD 2012)*, 2012.

G. Glavaš, J. Šnajder, P. Kordjamshidi, and M.-F. Moens. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of 9th language resources and evaluation conference*, pages 3678–3683. ELRA, 2014.

A. Gusev, N. Chambers, P. Khaitan, D. Khilnani, S. Bethard, and D. Jurafsky. Using query patterns to learn the duration of events. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 145–154. Association for Computational Linguistics, 2011.

A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

T. Harcup and D. O'Neill. What is news? Galtung and Ruge revisited. *Journalism studies*, 2(2):261–280, 2001.

M. Haspelmath. *From space to time: Temporal adverbials in the world's languages*. Lincom Europa, 1997.

E. Hovy, T. Mitamura, F. Verdejo, J. Araki, and A. Philpot. Events are not simple: Identity, non-identity, and quasi-identity. In *NAACL HLT*, volume 2013, pages 21–28, 2013.

S. Im, H. You, H. Jang, S. Nam, and H. Shin. KTimeML: specification of temporal and event expressions in Korean text. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 115–122. Association for Computational Linguistics, 2009.

ISO/TC 37/SC 4/WG 2. *Language Resource Management - Semantic Annotation Framework (SemAF) - Part 1: Time and events*. 2007.

P. Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

A. Jatowt and C.-m. Au Yeung. Extracting collective expectations about the future from large text collections. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1259–1264. ACM, 2011.

H.-J. Kaalep. Eesti verbi vormistik (Estonian verb paradigm). *Keel ja Kirjandus*, (1):1–15, 2015.

H.-J. Kaalep and K. Muischnek. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas (Estonian multi-word expressions in computational linguistics: why and how). *Eesti Rakenduslingvistika Ühingu aastaraamat*, (5):157–172, 2009.

H.-J. Kaalep and T. Vaino. Complete morphological analysis in the linguist's toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pages 9–16, 2001.

H. J. Kaalep, K. Muischnek, K. Uiboaed, and K. Veskis. The Estonian Reference Corpus: Its Composition and Morphology-aware User Interface. In *Baltic HLT*, pages 143–146, 2010.

G. Katz and F. Arosio. 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, 2001.

R. Kessler, X. Tannier, C. Hagege, V. Moriceau, and A. Bittar. Finding salient dates for building thematic timelines. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 730–739. Association for Computational Linguistics, 2012.

B. Knippen, J. Littman, I. Mani, J. Pustejovsky, A. Sanfilippo, A. See, A. Setzer, R. Saurí, A. Stubbs, B. Sundheim, et al. TimeML 1.2.1. A Formal Specification Language for Events and Temporal Expressions. 2005. `http://www.timeml.org/site/publications/timeMLdocs/timeml_1.2.1.html` (Date accessed: 2015-05-15).

N. Kotsyba. Using Petri nets for temporal information visualization. *Études Cognitives/Studia Kognitywne*, 7:189–207, 2006.

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

A. Lascarides and N. Asher. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493, 1993.

H. Lee, M. Recasens, A. Chang, M. Surdeanu, and D. Jurafsky. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500. Association for Computational Linguistics, 2012.

K. Liin, K. Muischnek, K. Müürisep, and K. Vider. Eesti keel digiajastul – The Estonian Language in the Digital Age. *META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.*, 2012. URL `http://www.meta-net.eu/whitepapers/volumes/estonian`.
(Date accessed: 2016-10-20).

Linguistic Data Consortium et al. ACE (Automatic Content Extraction) English annotation guidelines for events, 2005.
`https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf` (Date accessed: 2015-03-27).

H. Llorens, E. Saquete, and B. Navarro-Colorado. TimeML events recognition and classification: learning CRF models with semantic roles. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 725–733. Association for Computational Linguistics, 2010.

J. Makkonen. Semantic classes in topic detection and tracking. 2009. URL `https://helda.helsinki.fi/handle/10138/21330`.
(Date accessed: 2015-12-08).

I. Mani and B. Schiffman. Temporally anchoring and ordering events in news. In J. Pustejovsky and R. Gaizauskas, editors, *Time and Event Recognition in Natural Language*. John Benjamins, 2005.

I. Mani and G. Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 69–76. Association for Computational Linguistics, 2000a.

I. Mani and G. Wilson. Temporal granularity and temporal tagging of text. In *AAAI-2000 Workshop on Spatial and Temporal Granularity, Austin*, 2000b.

I. Mani, J. Pustejovsky, and R. Gaizauskas. *The Language of Time: A Reader*, volume 126. Oxford University Press, 2005.

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

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press, 2008.

G. Marşic. 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, 2012.

M. Marovic, J. Šnajder, and G. Glavaš. Event and temporal relation extraction from Croatian newspaper texts. In *Proc. of the Eighth Language Technologies Conference. Slovenian Language Technologies Society*, 2012.

S. Matsuyoshi, M. Eguchi, C. Sao, K. Murakami, K. Inui, and Y. Matsumoto. Annotating Event Mentions in Text with Modality, Focus, and Source Information. In *LREC*, 2010.

P. Mazur and R. Dale. What's the date?: high accuracy interpretation of weekday names. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 553–560. Association for Computational Linguistics, 2008.

P. P. Mazur. *Broad-Coverage Rule-Based Processing of Temporal Expressions*. PhD thesis, Macquarie University (Australia) and Wrocław University of Technology (Poland), 2012.

H. Metslang. On the developments of the Estonian aspect: the verbal particle. *The Circum-Baltic Languages: Typology and Contact: Grammar and Typology. Amsterdam, Philadelphia: John Benjamins*, pages 443–479, 2001.

A.-L. Minard, M. Speranza, E. Agirre, I. Aldabe, M. van Erp, B. Magnini, G. Rigau, R. Urizar, and F. B. Kessler. SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering. 2015. URL `http://adimen.si.ehu.es/~rigau/publications/SemEval15-TimeLines-task.pdf`. (Date accessed: 2015-11-25).

M.-F. Moens, O. Kolomiyets, E. Pianta, S. Tonelli, and S. Bethard. 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, 2011. URL `http://www.terenceproject.eu/c/document_library/get_file?p_l_id=16136&folderId=12950&name=DLFE-1910.pdf`. (Date accessed: 2015-08-05).

S. Monahan and M. Brunson. Qualities of Eventiveness. *ACL 2014*, pages 59–67, June 2014.

K. Muischnek, K. Müürisep, and T. Puolakainen. Automatic Analysis of Adjectives in Estonian. In *Workshop in TALN99 (6eme Conference Annuelle sur le Traitment Automatiques des Langues Naturelles)*, TALN99, pages 108–114, 1999.

K. Muischnek, K. Müürisep, and T. Puolakainen. Estonian particle verbs and their syntactic analysis. In *Human Language Technologies as a Challenge for Computer Science and Linguistics: 6th Language & Technology Conference Proceedings*, pages 338–342, 2013.

K. Muischnek, K. Müürisep, and T. Puolakainen. Dependency Parsing of Estonian: Statistical and Rule-based Approaches. In *Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014*, volume 268, pages 111–118. IOS Press, 2014a.

K. Muischnek, K. Müürisep, T. Puolakainen, E. Aedmaa, R. Kirt, and D. Särg. Estonian Dependency Treebank and its annotation scheme. In *Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 285–291, 2014b.

K. Müürisep and P. Mutso. ESTSUM - Estonian newspaper texts summarizer. In *Proceedings of The Second Baltic Conference on Human Language Technologies*, pages 311–316. Citeseer, 2005.

K. Müürisep, T. Puolakainen, K. Muischnek, M. Koit, T. Roosmaa, and H. Uibo. A New Language for Constraint Grammar: Estonian. In *International Conference Recent Advances in Natural Language Processing RANLP 2003*, pages 304–310, Borovets, 2003.

K. Müürisep, H. Orav, H. Õim, K. Vider, N. Kahusk, and P. Taremaa. From Syntax Trees in Estonian to Frame Semantics. In *The Proceedings of the Third Baltic Conference on Human Language Technologies; Kaunas, Lithuania*, 2008.

D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

M. Naughton. *Sentence Level Event Detection and Coreference Resolution*. PhD thesis, National University of Ireland, Dublin, 2009.

M. Negri and L. Marseglia. Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004. *Rapport interne, ITC-irst, Trento*, 2004.

J. Nothman. *Grounding event references in news*. PhD thesis, The University of Sydney, 2013.

J. Nothman, M. Honnibal, B. Hachey, and J. R. Curran. Event linking: Grounding event reference in a news archive. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 228–232. Association for Computational Linguistics, 2012.

H. Õim, H. Orav, N. Kahusk, and P. Taremaa. Semantic analysis of sentences: the Estonian experience. In *Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, volume 219, page 208. IOS Press, 2010.

S. Orasmaa. Ajaväljendite tuvastamine eestikeelses tekstis *(Recognition and Resolution of Temporal Expressions in Estonian Texts)*. Master's thesis, University of Tartu, Estonia, 2010. (in Estonian).

S. Orasmaa. Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides *(Automatic Recognition and Normalization of Temporal Expressions in Estonian Language Texts)*. *Eesti Rakenduslingvistika Ühingu aastaraamat*, (8):153–169, 2012.

S. Orasmaa. 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, 2014a.

S. Orasmaa. Towards an Integration of Syntactic and Temporal Annotations in Estonian. In *LREC*, pages 1259–1266, 2014b.

S. Orasmaa. Event coreference detection in Estonian news articles: preliminary experiments. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 11:189–203, 2015.

S. Orasmaa, T. Petmanson, A. Tkachenko, S. Laur, and H.-J. Kaalep. EstNLTK - NLP Toolkit for Estonian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, 2016. European Language Resources Association (ELRA).

M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005.

P. Pecina. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):137–158, 2010.

T. Puolakainen. Anaphora resolution experiment with CG rules. *Proceedings of the Workshop on "Constraint Grammar-methods, tools and applications" at NODALIDA 2015*, pages 35–37, 2015.

G. Puscasu and V. B. Mititelu. Annotation of WordNet Verbs with TimeML Event Classes. In *LREC*, 2008.

J. Pustejovsky and J. Moszkowicz. The Role of Model Testing in Standards Development: The Case of ISO-Space. In *LREC*, pages 3060–3063, 2012.

J. Pustejovsky and A. Rumshisky. Deep Semantic Annotation with Shallow Methods. In *LREC 2014 tutorial*, 2014. URL `https://c2eb795ea057e29f68b4fcf554f0c4317d00be90.googledrive.com/host/0B-NS3rTq7KpcZzFPOXBlSUlscW8/index.html`.
(Date accessed: 2015-03-27).

J. Pustejovsky and A. Stubbs. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., 2012.

J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz. TimeML: Robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, 2003a.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. The TimeBank corpus. In *Corpus Linguistics*, volume 2003, pages 647–656, 2003b.

J. Pustejovsky, R. Knippen, J. Littman, and R. Saurí. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39(2-3): 123–164, 2005a.

J. Pustejovsky, A. Meyers, M. Palmer, and M. Poesio. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 5–12. Association for Computational Linguistics, 2005b.

J. Pustejovsky, K. Lee, H. Bunt, and L. Romary. ISO-TimeML: An International Standard for Semantic Annotation. In *LREC*, 2010.

J. Pustejovsky, J. L. Moszkowicz, and M. Verhagen. 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, 2011.

M. Recasens. Annotation Guidelines for Entity and Event Coreference, 2011. URL `http://nlp.stanford.edu/pubs/jcoref-corpus.zip`.
(Date accessed: 2016-01-15).

H. Reichenbach. *Elements of symbolic logic*. Macmillan Co., 1947.

L. Robaldo, T. Caselli, I. Russo, and M. Grella. From Italian text to TimeML document via dependency parsing. In *Computational Linguistics and Intelligent Text Processing*, pages 177–187. Springer, 2011.

M. Roth and A. Frank. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 218–227. Association for Computational Linguistics, 2012.

E. Saue. Eestikeelsete ajaväljendite automaatne eraldamine *(Automatic Extraction of Estonian Temporal Expressions)*. Bachelor's thesis, University of Tartu, Estonia, 2007.

R. Saurí. *A Factuality Profiler for Eventualities in Text*. PhD thesis, Brandeis University, 2008.

R. Saurí, R. Knippen, M. Verhagen, and J. Pustejovsky. 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, 2005.

R. Saurí, J. Littman, R. Gaizauskas, A. Setzer, and J. Pustejovsky. TimeML annotation guidelines, version 1.2.1. 2006. `http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf` (Date accessed: 2015-08-12).

R. Saurí, L. Goldberg, M. Verhagen, and J. Pustejovsky. *Annotating Events in English. TimeML Annotation Guidelines*. 2009. `http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/EventGuidelines-050409.pdf` (Date accessed: 2015-06-14).

F. Schilder. Event extraction and temporal reasoning in legal documents. In *Annotating, Extracting and Reasoning about Time and Events*, pages 59–71. Springer, 2007.

F. Schilder and C. Habel. From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of the Workshop on Temporal and Spatial Information Processing*, volume 65-72, page 9. Association for Computational Linguistics, 2001.

H. Schmid. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28, 1995.

S. Schneider. Events. The Internet encyclopedia of philosophy, 2005. `http://www.iep.utm.edu/events/` (Date accessed: 2015-04-01).

K. K. Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

A. Setzer, R. Gaizauskas, and M. Hepple. Using semantic inferences for temporal annotation comparison. In *Proceedings of the fourth international workshop on inference in computational semantics (ICOS-4)*, 2003.

R. Sprugnoli and A. Lenci. Crowdsourcing for the identification of event nominals: An experiment. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

J. Strötgen. *Domain-sensitive Temporal Tagging for Event-centric Information Retrieval*. PhD thesis, Heidelberg University, 2015.

J. Strötgen and M. Gertz. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics, 2010.

J. Strötgen and M. Gertz. Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards. In *LREC*, volume 12, pages 3746–3753, 2012.

J. Strötgen and M. Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.

TimeML Working Group. *TLINK Guidelines*. 2009. `http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/tlink-guidelines-081409.pdf` (Date accessed: 2015-06-14).

A. Tkachenko, T. Petmanson, and S. Laur. Named Entity Recognition in Estonian. *ACL 2013*, pages 78–83, 2013.

M. Treumuth. Normalization of temporal information in Estonian. In *Text, Speech and Dialogue*, pages 211–218. Springer, 2008.

H. Uibo. Experimental Two-Level Morphology of Estonian. In *LREC*, 2002.

N. UzZaman, H. Llorens, L. Derczynski, M. Verhagen, J. Allen, and J. Pustejovsky. SemEval-2013 Task 1: TEMPEVAL-3: Evaluating Time Expressions, Events, and Temporal Relations. 2013. URL http://derczynski.com/sheffield/papers/tempeval-3.pdf.
(Date accessed: 2015-03-27).

L. Vanderwende, M. Banko, and A. Menezes. Event-centric summary generation. *Working notes of the Document Understanding Conference*, 2004.

Z. Vendler. Verbs and times. *The philosophical review*, pages 143–160, 1957.

M. Verhagen. The Brandeis Annotation Tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association, 2010. ISBN 2-9517408-6-7.

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, J. Moszkowicz, and J. Pustejovsky. The TempEval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179, 2009.

M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics, 2010.

Ü. Viks. Tools for the Generation of Morphological Entries in Dictionaries. In *LREC*, 2000.

P. Vossen, A. Cybulska, E. Laparra, O. L. de Lacalle, E. Agirre, and G. Rigau. D5.1.1 Event Narrative Module, version 1 Deliverable D5.1.1. 2014a. URL http://www.newsreader-project.eu/files/2012/12/NewsReader-316404-D5.1.1.pdf. (Date accessed: 2015-12-28).

P. Vossen, G. Rigau, L. Serafini, P. Stouten, F. Irving, and W. R. V. Hage. Newsreader: recording history from daily news streams. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31 2014b. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/436_Paper.pdf. (Date accessed: 2016-10-21).

G. Wilson, I. Mani, B. Sundheim, and L. Ferro. A multilingual approach to annotating and extracting temporal information. In *Proceedings of the Workshop on Temporal and Spatial Information Processing*, volume 13, pages 1–7. Association for Computational Linguistics, 2001.

P. H. Winston. The Strong Story Hypothesis and the Directed Perception Hypothesis. In P. Langley, editor, *Technical Report FS-11-01, Papers from the AAAI Fall Symposium*, pages 345–352, Menlo Park, CA, 2011. AAAI Press.

T. Wolfe, B. Van Durme, M. Dredze, N. Andrews, C. Beller, C. Callison-Burch, J. DeYoung, J. Snyder, J. Weese, T. Xu, et al. PARMA: A Predicate Argument Aligner. In *ACL (2)*, pages 63–68, 2013.

N. Xue and Y. Zhou. 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, 2010. Association for Computational Linguistics.

Y. Yaghoobzadeh, G. Ghassem-Sani, S. A. Mirroshandel, and M. Eshaghzadeh. ISO-TimeML Event Extraction in Persian Text. In *COLING*, pages 2931–2944, 2012.

Y. Yang, J. G. Carbonell, R. D. Brown, T. Pierce, B. T. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.

A. Zaenen. Mark-up barking up the wrong tree. *Computational Linguistics*, 32 (4):577–580, 2006.

L. Zhou and G. Hripcsak. Temporal reasoning with medical data – a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40(2):183–202, 2007.

R. A. Zwaan and G. A. Radvansky. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162, 1998.

APPENDIX A

ESTONIAN MORPHOLOGICAL AND SYNTACTIC TAGS USED IN THE EXAMPLES

Table. A.1: Part of speech tags:

A	adjective	S	substantive
D	adverb	V	verb
H	proper name	P	pronoun
J	conjunction	Y	abbreviation
K	adposition	Z	punctuation
N	numeral	X	other

Table. A.2: Syntactic tags:

$@SUBJ$	subject	$@OBJ$	object
$@ADVL$	adverbial	$@PRD$	predicative
$@FMV$	finite main verb	$@IMV$	infinite main verb
$@FCV$	finite form of 'to be' or modal verb	$@ICV$	infinite form of 'to be' or modal verb
$@NN>$	noun as premodifier	$@AN>$	adjective as premodifier
$@J$	conjunct		

APPENDIX B

EXAMPLES OF RULES USED BY THE TEMPORAL TAGGER

In this Appendix, we will give some examples on how the rules used by the temporal tagger are defined in XML. Note that the examples presented here are simplified for brevity. A reader interested in more details should consult the description of the rule format, which is available at `https://github.com/soras/Ajavt/blob/master/doc/writingRules.txt` (last accessed: 2016-10-25).

An extraction rule using two word classes. In the following, we give an XML example of an extraction rule that can be used to extract Estonian "N years ago" expressions, e.g. *viis aastat tagasi* 'five years ago', *10 aasta tagune* '10 years ago/back'.

First, a **word class** (`SonaKlass`) named `TAGASI_EEST_TAGUNE` is defined, which can match words by lemmas (*tagasi*, *tagune*, *eest*), capturing 'ago'-like postpositions:

(21)
```
<SonaKlass nimi="TAGASI_EEST_TAGUNE">
  <Element tyyp="algv" vaartus="tagune" />
  <Element tyyp="algv" vaartus="tagasi" />
  <Element tyyp="algv" vaartus="eest"   />
</SonaKlass>
```

Each `Element` in the word class definition corresponds to a single word template that the class should match. The attribute `tyyp` specifies the type of the word template (`algv` refers to the lemma based template), and `vaartus` is the concrete value of the template (the format of the value depends on the type).

Second, a word class (named `ARV_LOENDA_VAIKE_A`) is defined for capturing numbers and numeral words representing a small range, from 0 to 599:

(22)
```
<SonaKlass nimi="ARV_LOENDA_VAIKE_A">
  <Element tyyp="reg" vaartus="([0-5]?[0-9]?[0-9])"
    semValue="REF:1" />
  <Element tyyp="arvSona" arvuPiirang="0-599"
    arvuLiik="_N_|_F_" semValue="REF:1" />
</SonaKlass>
```

Note that in Example 22, two different types of word templates are used in the word class: the template with `tyyp="reg"` is a regular expression based word template (where `vaartus` specifies the regular expression that the word should match), and the template with `tyyp="arvSona"` is a numeral phrase template.

The numeral phrase template matches numbers expressed in words, with the attribute `arvuPiirang` specifying the (integer) range of allowed numbers, and the attribute `arvuLiik` specifying the allowed numeral types (`_N_` = cardinal numeral, `_F_` = floating-point numeral);

In Example 22, word templates also contain pre-filled parts of the normalisation instructions (the `semValue` attributes), which are later carried over to the normalisation instructions. In the case of the regular expression template, `semValue="REF:1"` guides that the `semValue` should be taken from the first captured group of the regular expression match (i.e. the number between the parenthesis). In the case of the numeral phrase template, `semValue="REF:1"` guides that the `semValue` should be initialised with the number expressed by the numeral phrase.

Third, using the word classes defined above (Examples 21 and 22), a **basic extraction rule** for capturing "N years ago" expressions is defined:

(23)
```
<Reegel>
 <Muster>
   ARV_LOENDA_VAIKE_A? |aasta| TAGASI_EEST_TAGUNE
 </Muster>
 <Filter morfTunnused="_ {sg} _">
     <!-- semantics of singular #1: subtract the
             given number of years from the reference
             date -->
     <SemReegel priority="1"
        seotudMustriosa="ARV_LOENDA_VAIKE_A"
        op="SUBTRACT" semField="YEAR" />
     <!-- semantics of singular #2: if the number
             of years is unspecified, subtract one
             year from the reference date -->
     <SemReegel priority="1"
        seotudMustriosa="^0 1"
        op="SUBTRACT" semField="YEAR"
        semValue="1" />
 </Filter>
 <Filter morfTunnused="_ {pl} _">
```

```
      <!-- semantics of plural: a vague reference
           to the past  -->
      <SemReegel priority="1"
         seotudMustriosa="1" op="SET_attrib"
         attrib="value" semValue="PAST_REF"  />
 </Filter>
</Reegel>
```

A basic extraction rule is within `Reegel` tags, and in the above case, it consists of a phrase pattern (within `Muster` tags), normalisation instructions (`SemReegel` tags), and morphology-based filters (within `Filter` tags).

The **phrase pattern** (within `Muster` tags) describes the extractible temporal expression phrase. In Example 23, the phrase begins with a token matching the word class `ARV_LOENDA_VAIKE_A`, and this match can also be skipped, as `?` in the end indicates. The second token in the phrase (or the first one, if the previous match is skipped) should match with the lemma `aasta` (*year*), and the last token should match the word class `TAGASI_EEST_TAGUNE` (*ago*).

Morphology-based filters (`Filter` tags) restrict the execution of normalisation instructions: instructions within the `Filter` tags are executed only if the morphological constraints imposed by the `Filter` are satisfied. The morphological constraints are listed in the attribute `morfTunnused`. In Example 23, the first `Filter` enables the execution of its normalisation instructions only if the word matching the second word template (i.e. word matching the lemma `aasta`) is in the singular (`{sg}`). The second `Filter` requires the word at the same position to be in the plural (`{pl}`). This filtering servers the purpose of separating generic past reference expressions (marked with plural: *aastaid tagasi* 'years ago') from concrete past reference expressions (marked with singular: *aasta tagasi* 'a year ago' or *2 aastat tagasi* '2 years ago').

Normalisation instructions (`SemReegel` tags) guide how the reference time should be changed for arriving at the semantics of the expression.

The instructions are executed in the order specified in the `priority` attributes, and each following rule gets the output of the previous rule as an input. In Example 23, all normalisation instructions have the `priority="1"` as they are complementary to each other: each of them is executed in a different context.

The attribute `seotudMustriosa` specifies the phrase pattern matching condition under which the instruction will be executed. For example, `seotudMustriosa="^0 1"` specifies that the instruction is only executed if the extracted phrase has matched the second template (index `1` corresponds to matching the second template), and has skipped the first one (index `^0` corresponds to skipping the match on the first template). Rather than using numeric indices, `seotudMustriosa` can also refer to word class names in the pattern. Note that if the specified word class also has pre-filled parts of the normalisation instruction

(like the word class ARV_LOENDA_VAIKE_A in Example 22 has), then these pre-filled parts are carried over to complete the normalisation instruction. Thus, the following normalisation instruction (from Example 23):

(24)
```
<SemReegel priority="1"
  seotudMustriosa="ARV_LOENDA_VAIKE_A" op="SUBTRACT"
  semField="YEAR" />
```

subtracts the number of years from the reference time, with the exact number taken from the pre-filled semValue of the word class ARV_LOENDA_VAIKE_A.

The normalisation instruction:

(25)
```
<SemReegel priority="1" seotudMustriosa="^0 1"
   op="SUBTRACT" semField="YEAR" semValue="1" />
```

subtracts exactly one year from the reference time, and the normalisation instruction:

(26)
```
<SemReegel priority="1" seotudMustriosa="1" op="SET_attrib"
  attrib="value" semValue="PAST_REF" />
```

instructs to overwrite the TIMEX attribute *value*, so that it will be replaced with the string "PAST_REF".

A composition rule adding modifiers to phrases. In the following, we give an example of a composition rule for joining quantifier modifiers of type APPROX and temporal expressions containing a duration part, e.g. *umbes* + *aasta tagasi* 'approximately' + 'a year ago'.

First, a basic extraction rule for capturing quantifier modifiers that carry the APPROX semantics (words *umbes*, *orienteeruvalt*, *ligikaudselt* 'approximately') is defined:

(27)
```
<Reegel>
   <Muster>
      /(orienteeruvalt|umbes|ligikaudselt|circa)/
   </Muster>
   <SemReegel priority="1" seotudMustriosa="0"
      op="SET_attrib" attrib="mod" semValue="APPROX" />
   <MustriTahis poleEraldiSeisevAjav="1"
      seotudMustriosa="0" tahised="KVANT_EESLIIDE" />
</Reegel>
```

In Example 27, the phrase pattern consisting of a single regular expression template (between / and /) that captures the quantifier modifier words. The tag

MustriTahis guides that a label named "KVANT_EESLIIDE" should be attached to the extracted expression candidate. The attribute assignment poleEraldiSeisevAjav="1" declares that the extracted candidate cannot exist as a stand-alone temporal expression: if it is not joined by composition rules, it will be deleted and will not reach the normalisation phase. The semantics part of the rule (SemReegel) simply instructs it to set the TIMEX attribute *mod* to the value APPROX.

Second, we need to augment the basic extraction rule in Example 23 with the tag MustriTahis, specifying a label to be attached to the extracted expression candidate:

(28)
```
<MustriTahis seotudMustriosa="1"
             tahised="VOTAB_KVANT_EESLIITE" />
```

This label attachment can be added just after the Muster tags and before the Filter tags in Example 23.

Third, we can now define a **composition rule** for joining quantifier modifiers that carry the APPROX semantics and "N years ago" expressions into larger temporal expression candidates:

(29)
```
<LiitumisReegel fikseeritudJarjekord="1">
  KVANT_EESLIIDE  VOTAB_KVANT_EESLIITE
</LiitumisReegel>
```

The composition rule lists labels of the expression candidates that need to be joined into a longer phrase. The attribute assignment fikseeritudJarjekord ="1" specifies that the order of candidates is fixed: the quantifier modifier can only precede, but cannot succeed the "N years ago" expression.

APPENDIX C

PAIRWISE INTER-ANNOTATOR AGREEMENTS ON SPECIFYING EVENT EXTENT

Model	Description	AB	AC	BC
0	initial (no EVENT filtering)	0.875	0.762	0.790
1a	verbs	0.985	0.924	0.922
1b	verbs and nouns	0.900	0.795	0.802
1c	verbs and adjectives	0.951	0.884	0.914
1d	verbs, adjectives and nouns	0.878	0.770	0.797
2a	EVENTs that are part of the predicate of a clause	0.992	0.976	0.978
2b	2a + direct verb dependents of the predicate	0.987	0.937	0.934
2c	2a + direct non-verb dependents of the predicate	0.930	0.846	0.869
2d	2a + clause members that are not direct dependents of the predicate	0.908	0.881	0.906

Table. C.1: EVENT annotation inter-annotator agreements (F1-scores) on different syntactically constrained subsets, reported for annotator pairs AB, AC, and BC. Subsets were obtained by *filtering* the set of all manually provided EVENT annotations: only EVENT annotations which met the criteria (in the model's description) were preserved, and all other EVENT annotations were removed.

APPENDIX D

PAIRWISE INTER-ANNOTATOR AGREEMENTS ON SPECIFYING TLINK TYPES

measure	*pair*	event-timex	event-dct	main-events	event-event
Precision	AB	0.73	0.661	0.578	0.394
	AC	0.239	0.451	0.508	0.384
	BC	0.341	0.395	0.509	0.496
Kappa	AB	0.554	0.541	0.495	0.315
	AC	0.149	0.322	0.374	0.275
	BC	0.186	0.285	0.392	0.37

Table. D.1: Overall inter-annotator agreements (precisions and Cohen's kappas) between annotators A, B, and C on specifying temporal relations (choosing the relation type for each entity pair) on the subtasks *event-timex*, *event-dct*, *main-events*, and *event-event*.

Event subset description	event-timex	event-dct	main-events	event-event
0. All syntactic predicate EVENTs	0.517	0.553	0.499	0.327
1a. EVENTs in simple past tense	0.537	<0.1	0.507	0.357
1b. EVENTs in present tense	0.547	0.383	0.333	0.188
2a. EVENTs governing TIMEX	0.544	0.595	0.653	0.143
2b. EVENTs not governing TIMEX	<0.0	0.539	0.466	0.35

Table. D.2: Inter-annotator agreements (Cohen's kappas) between the pair AB on specifying temporal relations on different EVENT subsets.

Event subset description	event-timex	event-dct	main-events	event-event
0. All syntactic predicate EVENTs	0.158	0.372	0.368	0.257
1a. EVENTs in simple past tense	<0.1	0.624	0.366	0.405
1b. EVENTs in present tense	0.237	0.154	0.319	0.193
2a. EVENTs governing TIMEX	0.154	0.435	0.734	0.677
2b. EVENTs not governing TIMEX	0.0	0.358	0.342	0.216

Table. D.3: Inter-annotator agreements (Cohen's kappas) between the pair AC on specifying temporal relations on different EVENT subsets.

Event subset description	event-timex	event-dct	main-events	event-event
0. All syntactic predicate EVENTs	0.185	0.332	0.384	0.366
1a. EVENTs in simple past tense	0.169	0.133	0.422	0.431
1b. EVENTs in present tense	0.215	0.141	0.282	0.266
2a. EVENTs governing TIMEX	0.152	0.448	0.579	0.6
2b. EVENTs not governing TIMEX	0.268	0.31	0.326	0.323

Table. D.4: Inter-annotator agreements (Cohen's kappas) between the pair BC on specifying temporal relations on different EVENT subsets.

CURRICULUM VITAE

Personal data

Name	Siim Orasmaa
Date of Birth	11.06.1985
Citizenship	Estonian
Languages	Estonian, English, Russian
E-mail	siim.orasmaa@ut.ee

Education

2011 –	University of Tartu, PhD student, Computer Science
2008 – 2010	University of Tartu, MSc, Information Technology
2005 – 2008	University of Tartu, BSc, Computer Science
1992 – 2004	Tartu Forselius Gymnasium

Employment

2008 –	University of Tartu, Institute of Computer Science, programmer
2011 – 2012	Filosoft OÜ, programmer
2007	Webmedia AS, junior programmer
2004–2005	Military service

Scholarships

2014	DoRa travel scholarship for attending HLT Baltic conference
2014	DoRa travel scholarship for attending LREC conference
2013	IT Academy scholarship

2013	DoRa travel scholarship for attending TSD conference
2012	IT Academy scholarship
2010	DoRa travel scholarship for attending LREC conference

Publications

1. Orasmaa, Siim (2015). Event coreference detection in Estonian news articles: preliminary experiments. In: Eesti Rakenduslingvistika Ühingu aastaraamat (189-203). Tallinn: Eesti Rakenduslingvistika Ühing.

2. Orasmaa, Siim (2014). How Availability of Explicit Temporal Cues Affects Manual Temporal Relation Annotation. In: Human Language Technologies - The Baltic Perspective (215–218). IOS Press. (Frontiers in Artificial Intelligence and Applications).

3. Orasmaa, Siim (2014). Towards an Integration of Syntactic and Temporal Annotations in Estonian. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14): The International Conference on Language Resources and Evaluation; Reykjavik, Iceland; 2014. Ed. Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Hrafn Loftsson an. Reykjavik, Iceland: ELRA, 1259–1266.

4. Orasmaa, Siim (2012). Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides. Eesti Rakenduslingvistika Ühingu aastaraamat (153–169). Tallinn: Eesti Rakenduslingvistika Ühing.

5. Orasmaa, Siim (2013). Verb Subcategorisation Acquisition for Estonian Based on Morphological Information. In: Proceedings of 16th International Conference on Text, Speech and Dialogue (TSD 2013); Czech Republic, Plzen; Sep 1, 2013 – Sep 5, 2013. Springer-Verlag, 583–590. (Lecture Notes in Artificial Intelligence)

6. Orasmaa, Siim ; Käärik, Reina ; Vilo, Jaak ; Hennoste, Tiit (2010). Information Retrieval of Word Form Variants in Spoken Language Corpora Using Generalized Edit Distance. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10): The International Conference on Language Resources and Evaluation; Valletta, Malta; May 17–23, 2010. Ed. Calzolari, Nicoletta; Choukri, Khalid; Maegaard, Bente; Mariani, Joseph; Odjik, Jan. Valletta, Malta: ELRA, 623–629.

ELULOOKIRJELDUS

Üldandmed

Nimi	Siim Orasmaa
Sünniaeg	11.06.1985
Kodakondsus	Eesti
Keelteoskus	eesti, inglise, vene
E-post	siim.orasmaa@ut.ee

Haridus

2011 –	Tartu Ülikool, informaatika, doktoriõpe
2008 – 2010	Tartu Ülikool, MSc, infotehnoloogia
2005 – 2008	Tartu Ülikool, BSc, informaatika
1992 – 2004	Tartu Forseliuse Gümnaasium

Teenistuskäik

2008 –	Tartu Ülikool, arvutiteaduse instituut, programmeerija
2011 – 2012	Filosoft OÜ, programmeerija
2007	Webmedia AS, nooremprogrammeerija
2004–2005	kaitseväeteenistus

Saadud uurimistoetused ja stipendiumid

2014	DoRa välislähetuse toetus konverentsil HLT Baltic osalemiseks
2014	DoRa välislähetuse toetus konverentsil LREC osalemiseks
2013	IT Akadeemia stipendium

2013	DoRa välislähetuse toetus konverentsil TSD osalemiseks
2012	IT Akadeemia stipendium
2010	DoRa välislähetuse toetus konverentsil LREC osalemiseks

Publikatsioonid

1. Orasmaa, Siim (2015). Event coreference detection in Estonian news articles: preliminary experiments. In: Eesti Rakenduslingvistika Ühingu aastaraamat (189-203). Tallinn: Eesti Rakenduslingvistika Ühing.

2. Orasmaa, Siim (2014). How Availability of Explicit Temporal Cues Affects Manual Temporal Relation Annotation. In: Human Language Technologies - The Baltic Perspective (215–218). IOS Press. (Frontiers in Artificial Intelligence and Applications).

3. Orasmaa, Siim (2014). Towards an Integration of Syntactic and Temporal Annotations in Estonian. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14): The International Conference on Language Resources and Evaluation; Reykjavik, Iceland; 2014. Ed. Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Hrafn Loftsson an. Reykjavik, Iceland: ELRA, 1259–1266.

4. Orasmaa, Siim (2012). Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides. Eesti Rakenduslingvistika Ühingu aastaraamat (153–169). Tallinn: Eesti Rakenduslingvistika Ühing.

5. Orasmaa, Siim (2013). Verb Subcategorisation Acquisition for Estonian Based on Morphological Information. In: Proceedings of 16th International Conference on Text, Speech and Dialogue (TSD 2013); Czech Republic, Plzen; Sep 1, 2013 – Sep 5, 2013. Springer-Verlag, 583–590. (Lecture Notes in Artificial Intelligence)

6. Orasmaa, Siim ; Käärik, Reina ; Vilo, Jaak ; Hennoste, Tiit (2010). Information Retrieval of Word Form Variants in Spoken Language Corpora Using Generalized Edit Distance. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10): The International Conference on Language Resources and Evaluation; Valletta, Malta; May 17–23, 2010. Ed. Calzolari, Nicoletta; Choukri, Khalid; Maegaard, Bente; Mariani, Joseph; Odjik, Jan. Valletta, Malta: ELRA, 623–629.

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.
19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.

20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. Töö kaitsmata.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.
43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.

44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 p.

66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.
74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
75. **Nadežda Bazunova.** Differential calculus $d^3 = 0$ on binary and ternary associative algebras. Tartu 2011, 99 p.
76. **Natalja Lepik.** Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Tartu 2011, 133 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
80. **Marje Johanson.** $M(r, s)$-ideals of compact operators. Tartu 2012, 103 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
82. **Vitali Retšnoi.** Vector fields and Lie group representations. Tartu 2012, 108 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
85. **Erge Ideon**. Rational spline collocation for boundary value problems. Tartu, 2013, 111 p.
86. **Esta Kägo**. Natural vibrations of elastic stepped plates with cracks. Tartu, 2013, 114 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
88. **Boriss Vlassov.** Optimization of stepped plates in the case of smooth yield surfaces. Tartu, 2013, 104 p.

89. **Elina Safiulina.** Parallel and semiparallel space-like submanifolds of low dimension in pseudo-Euclidean space. Tartu, 2013, 85 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
93. **Kerli Orav-Puurand.** Central Part Interpolation Schemes for Weakly Singular Integral Equations. Tartu, 2014, 109 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
95. **Kaido Lätt.** Singular fractional differential equations and cordial Volterra integral operators. Tartu, 2015, 93 p.
96. **Oleg Košik.** Categorical equivalence in algebra. Tartu, 2015, 84 p.
97. **Kati Ain.** Compactness and null sequences defined by ℓ_p spaces. Tartu, 2015, 90 p.
98. **Helle Hallik.** Rational spline histopolation. Tartu, 2015, 100 p.
99. **Johann Langemets.** Geometrical structure in diameter 2 Banach spaces. Tartu, 2015, 132 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
105. **Md Raknuzzaman.** Noncommutative Galois Extension Approach to Ternary Grassmann Algebra and Graded q-Differential Algebra. Tartu, 2016, 110 p.
106. **Alexander Liyvapuu.** Natural vibrations of elastic stepped arches with cracks. Tartu, 2016, 110 p.
107. **Julia Polikarpus.** Elastic plastic analysis and optimization of axisymmetric plates. Tartu, 2016, 114 p.