EPiC Series in Language and Linguistics

Volume 4, 2019, Pages 84–94

Proceedings of Third Workshop ”Computational linguistics and language science”

EPiC Language and Linguistics

Formal Representation of Temporal Expressions [*]

Svetlana Timoshenko

IITP RAS,
Bolshoy Karetny per. 19, build.1, 127051
Moscow, Russia
timoshenko@iitp.ru

Abstract

In this paper we address the semantics of temporal expressions in natural language (such as *vchera*,’yesterday’, *shestnadcatogo maja*, ’on the 16th of May’, *tri dnja* ’three days’) and the way they interact with some other manifestations of temporality (such as functioning of prepositions and aspectual verb forms). A formal and constituent description of heterogeneous temporal expressions is proposed. We consider the interval algebra presented by James Allen to be the right basis for such a description. The new formal system is compared with the known TimeML project. The latter has weak spots - the meaning of some temporal expressions simply can not be represented in terms of TimeML. We discuss such cases and show how to analyze them in our formal system.

1 Introduction

This paper presents a new formal system for a uniform description of temporal phenomena. This system should provide representations that are linguistically sound and suitable for the automatic time reasoning. The structure of the paper is as follows. The second section puts the main topic into the linguistic context. The third section is dedicated to the description of the proposed formal system. This system borrows the main notions and concepts from Allen’s interval algebra, but is enriched by some additional concepts and relations. These additions make it more close to language perspective and natural quantifications of time as they involve the representation of calendar time. The resulting formal representations have much in common with the TimeML standard for temporal text markup. The brief description of TimeML scheme and the comparison of two formal systems can be found in the fourth section.
The development of the formal representation for time is a part of a broader and challenging task - building an automatic semantic text analyzer capable to produce a full semantic structure of a text in Russian [4], [5].This analyzer called SemEtap performs a transformation which resembles a translation from one language to another - but the target language is an artificial one. The resulting semantic structures are directed graphs satisfying the requirements of the specification of RDF Schema [8]. The formal representations given as the examples in sections 3 and 4 are graphs of the same type. They have all features of the artificial semantic language used in

[*] Supported by Russian National Foundation, grant 16-18-10422.

G. Wohlgenannt, R. von Waldenfels, S. Toldova, E. Rakhilina, D. Paperno, O. Lyashevskaya, N. Loukachevitch, S.O. Kuznetsov, O. Kultepina, D. Ilvovsky, B. Galitsky, E. Artemova and E. Bolshakova (eds.), CLLS 2018 (EPiC Series in Language and Linguistics, vol. 4), pp. 84–94

SemETAP. Every single semantic structure is built of individuals (nodes of the graph) and relations (edges of the graph). Every individual belongs to one or several classes. Classes are organized into the ontology with monotonic inheritance of properties. More information about SemETAP can be found in [4]. Although the ’dictionary’ coverage of the semantic language is far from complete, our analyzer is already functioning and all formal representations in the paper can be generated automatically from texts.

2 Linguistic Perspective

The linguistic research of temporal semantics has been conducted on to major topics. Since the information about the time is very important, the expressions that convey it are very frequent, and since they are frequent, they are often grammaticalized. So the first topic is the functioning of two grammatical categories dealing with time - tense and aspect. The grammatical markers of tense are usually added to the prototypical mean of event designation - the verb - and in this context serve to express various temporal relationships that occur between events.

The second topic is the subject of lexicography as it is focused on the study of specific lexical units referring to time - nouns, noun phrases, adverbs and adverbials.

The set of notions used to describe functioning of tense categories comprises such notions as time line, localization (on the time line) - absolute or relative, time points and time relations (*before*, *after* etc). It has been established as a standard since the 80s (see, f.ex.[6]), and was adopted by modern computational linguistics - mostly in the Reichenbachian version (f.ex.[7]). Linguists who work with lexical units do not reject the notion of the time line, but they focus mainly on the classification of language units and their temporal meanings. As the result, they propose a much more elaborated set of notions all corresponding in some way to the ”time point” of grammatical studies. This is especially true when it comes to computational linguistics. TimeML markup, which will be discussed in more details in section 4, distinguish 4 types of time entities: ’DATE’, ’TIME’, ’DURATION’ and ’SET’ - [9]. French researchers went event further and proposed an approach where there are no actual classes but the set of operators that can be combined in many different ways. So they got the classification in the form of calculus [2].

2.1 Interaction between Grammatical and Lexical Means

Consider the following example

(1)	Zavtra	utr-om	rod-it-sja	rebёnok-Ø.
	Tomorrow	morning-INS.SG	born-3.SG.NPST-REFL	child-NOM.SG

’The child will be born tomorrow morning’

The adverb *zavtra* ’tomorrow’, which means ’the day next to the current day’ and the noun *utrom* ’in the morning’ provide the full information about the time of the birth. The grammar also gives information about time: the marker of non-past tense is added to the perfective verb stem, so the wordform as a whole has a clear future time reference. From a semantic point of view the future can be described as a time interval that begins to exist after the moment of speech generation. Taking into account that ’the day next to the current day’ is ’the day that begins immediately after the current day’, i.e. the day called *tomorrow* forms part of the future because it is determined by the same temporal relation ’after’ and is anchored in the moment of the speech, grammatical marking of time in this phrase may seem redundant. The nominalization can keep the same time reference without any grammatical markers of tense:

(2) roždeni-e rebënk-a zavtra utr-om
birth-NOM.SG child-GEN.SG tomorrow morning-INS.SG
'The tomorrow morning birth of the child'

The contradictions between the verb tense and the meaning of adverb are forbidden:

(3) * Zavtra utr-om rodi-l-sja rebënok-Ø.
Tomorrow morning-INS.SG born-M.SG.PST-REFL child-NOM.SG
'The child was born tomorrow morning'

From the above examples the rule of temporal cohesion in clause can be deduced. But there are examples that show that cohesion on the clause level is not enough and that there are rules operating semantic elements in a more tricky way. For example, some verbs can syntactically subordinate *zavtra* even if they are in the past tense:

(4) Ja žda-l tebja zavtra.
I whait-M.SG.PST you tomorrow
'I expected you (to come) tomorrow'

There is no conflict between the verb tense and the meaning of the adverb, because the adverb does not characterize the time of the top predication (wait), but, rather, the subordinate predication that can be found in its semantic explication - the arrival of the addressee. This case was first described by Igor Boguslavsky, see [3], section 3.2.
These examples prove that the full description of the temporal semantics in the language is only possible when verbs and nominal clauses are described together, as the semantically interacting units, and both temporal relations and time designations are taken into consideration.

3 The Formal Model for Time Representation

We assume that interval temporal logic developed by James F. Allen [1] is powerful and flexible enough to model time as it is represented in natural language. But it still needs some adjustment, so this section begins with a brief overview of the main elements and assumptions of the interval temporal logic and after that gives an account of our additions. The example is quite voluminous and is given at the end of the section.

3.1 The Interval Temporal Logic

The interval temporal logic is built up from one primitive object, the time interval, and one primitive relation - ordering relation. From now on we will denote time intervals as `TimeInterval`, and this relation - as `MeetsTemorally`. `TimeInterval` is a continuous span of time that by definition must start somewhere and end somewhere. If two intervals start at the same point and end again at the same point, they are simply one and same interval. `MeetsTemorally` means that one interval exactly precedes the other, i.e. the final point of the first interval is the starting point of the other. In other words, no interval exists between them. Thus `MeetsTemorally` orders finite time intervals and forms an infinite time line which represents the universal time.

Time intervals can relate to each other in many different ways and `MeetsTemorally` is only one of them. Allen shows how other intuitive temporal relations can be derived from `MeetsTemorally` and proposes a list of six relations. As all these relations are directed, the full Allen's list comprises 12 relations (6 basic and 6 inverted, including `meetsTemporally` and its inverse). Not all inverted relations are equally useful for our task.

3.2 Necessary Additions

The original Allen's model does not say much about `TimeInterval`'s themselvs. They are supposed to be presented each as a pair of natural numbers (one for the beginning and one for the end) because the time line is just a sequence of natural numbers. But real world conventions about time marking are more complicate: we have days of the week, months and years that are nested inside one another. Thus the new properties should reflect two sorts of characteristics: a) characteristics of interval's duration calculated in terms of other intervals (f.ex., one week is as long as seven days) and b) calendar characteristics. The following durative characteristics has been added to the model: (**`hasSeconds, hasMinutes, hasHours, hasYears`** etc.). All properties from this list take integers as their values. We believe that 8 such properties are enough to process temporal expressions that are in common use, but it is clear that this set is open because of nanoseconds and other special measurement units. So intervals with known duration have a set of corresponding slots.

The second set of properties added to the model serves to link intervals to the calendar. Having in mind the possibility of temporal reasoning performed by third-party program, we want this link to be compatible with ISO 8601, an international standard covering the exchange of date- and time-related data. The time value expressed in this format has 7 characteristics maximum: three for date, three for time and one for UTC, for example: 2010-04-17T04:52:11+00:00. If temporal expression in some of its parts contains information corresponding to any of characteristics relevant for ISO 8601 label, the formal representation provide not only `TimeInterval`, but the entity of another type, `DateTimeDescription` with its own slots which should not be confused with slots expressing duration: **`inSecond, inMinute, inHour, inDay, inMonth, inYear, hasUTCTimeZone`**. This set of slots is closed.

3.3 Example

Consider the sentence

(5) V 2016 godu zavod rabotal tol'ko 120 dnej.
'In 2016 the industrial plant worked only 120 days'

The fragment of its semantic structure is represented on the fig.1, vizualising the rdf grapf. The formal representation of the phrase *120 dnej* ('120 days') is a `TimeInterval_1_1` with filled slot `hasDays`. Original Allen's relation `during` links it to another interval, the year, presented as `Year_1_1` (`Year` is a subclass of `TimeInterval` class). `Year_1_1` has a unique number that corresponds to one of ISO 8601 positions, that's why we have an individual of type `DateTimeDescription` here with the same number. The phrase *120 dnej* from witch we started does not exist on its own - it is the duration of plant working. That is denoted by the special relation *hasTime* that occurs only between events and time intervals. The verb tense contributes to the semantic structure the fact that `TimeInterval_1_1` existed in the past, before the moment of speech. For this element of meaning we have the third `TimeInterval` `- SpeechTimeInterval`, but it is related to `TimeInterval_1_1` not with the standard `before`-relation, but with `startsBefore`-relation. It is a disjunction of 4 standard relations:

$$startsBefore(i,j) := before(i,j) \wedge includes(i,j) \wedge meetsTemporally(i,j) \wedge overlaps(i,j)$$

`startsBefore` is a generalized representation of the temporal meaning conveyed by the past imperfective verb form in Russian.

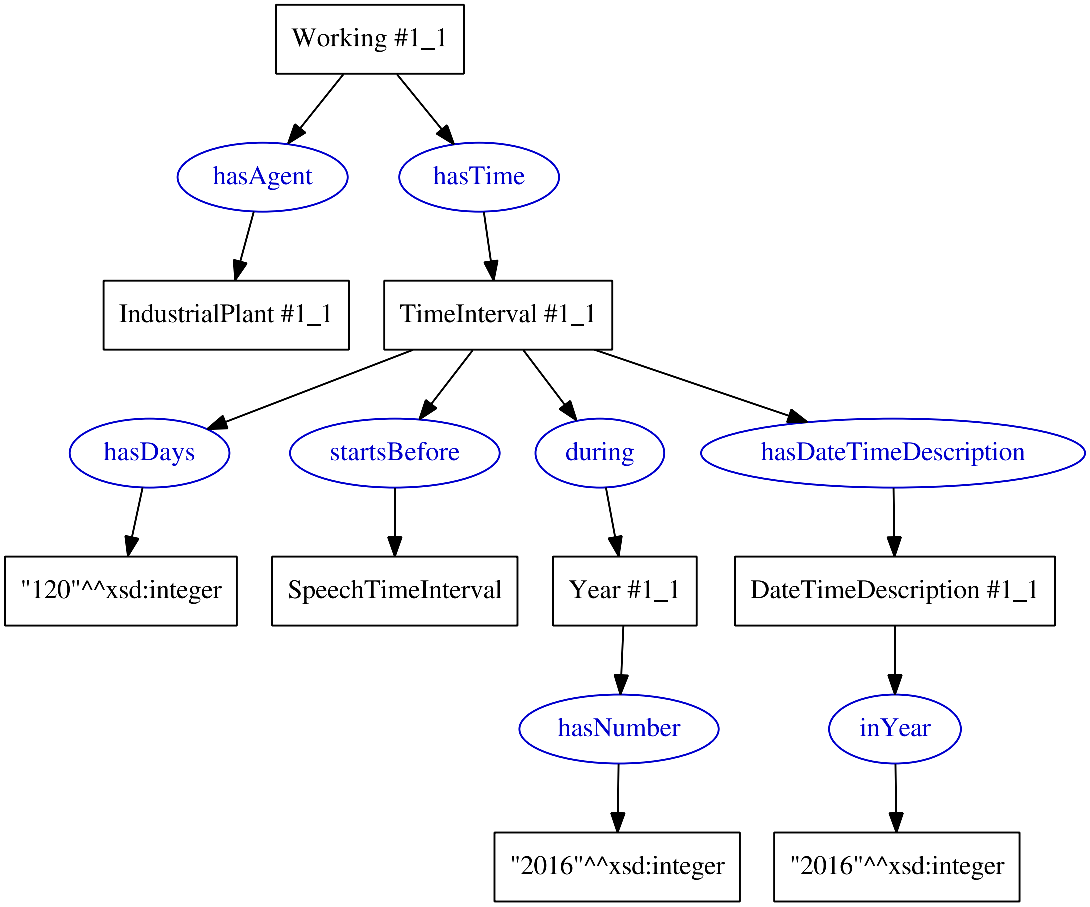

Figure 1: The fragment of the semantic structure which SemETAP generates for the sentence (5)
.

4 The Comparison of SemETAP Representation of Time with TilmeML

4.1 A Brief Description of TimeML

The proposed formal system has much in common with TimeML, a widespread annotation scheme dealing with temporal information. ML stands for Markup Language, because the goal of TimeML project is to create a standard markup language for temporal events in a document. This project emerged in the domain of automatic question answering as a preparatory stage in the development of a time reasoning tool. Texts marked with XML-compatible tags of TimeMl format form a corpus that allows training of statistical algorithms for detection of temporal information or can be used as a supplying material under construction of rule based systems solving the same task. For the full specification see [9].

As the main grounding reason for TimeML development was question answering, the format covers only events and time spans that can be clearly positioned in time. Time in generic statements (like ’*Jews are prohibited from killing one another*’) are beyond the scope of TimeML. The same is true for the descriptions of the typical patterns of activity (so called habitual situations - ’*Mr. Sununu has resorted to regular use of corporate planes for political travel*’) and also for some denotations of time like *soon*, *long time ago*, *recently*.

When the tag is assigned to a text fragment, a new entity emerges. This entity can be treated as an individual of the class whose name is the tag. Any further reference to individuals and classes in this section should be understood in this sense. TimeML distinguish three classes of entities. The entities can be linked to each other by links of three kinds.

There are following classes: EVENT, TIMEX3, SIGNAL. All situations that happen or occur are classified as EVENTs. All time designations (except those that are out of interest like *soon*) are TIMEX3s. The notion of SIGNAL covers ”auxiliary” text elements - this tag encloses words and phrases expressing the type of temporal link between events and / or temporal instances: prepositions, conjunctions etc. The MAKEINSTANCE tag creates instantiations of events. The instantiations and not the events themselves are linked to the TIMEX3 individuals.

There are following link types: TLINK, SLINK ALINK. ALINK is an abbreviation for Aspectual Link, it is used to mark verbal frases like *started to read*, *stops talking*. SLINK is an abbreviation for Subordination Link, it covers a large number of more or less grammatical meanings that in English happen to be expressed by the separate verb: modal meanings (wanted to buy), evidential meanings (John said he bought), factive meanings (managed to leave) etc. TLINK, the Temporal Link, represents the temporal relationship holding between events, times, or between an event and a time. TLINK has several parameters, and among them - relType. The values this parameter can accept, are: ’BEFORE’, ’AFTER’, ’INCLUDES’, ’IS_INCLUDED’, ’DURING’, ’DURING_INV’, ’SIMULTANEOUS’, ’IAFTER’, ’IBEFORE’, ’IDENTITY’, ’BEGINS’, ’ENDS’, ’BEGUN_BY’, ’ENDED_BY’. All these are the temporal relations from Allen’s model. This makes TimeML comparable with that one proposed here so we better give the corresponding example. See the sentence (5) in the TimeML markup below:

```
<SIGNAL sid="s1">
V
</SIGNAL>
<TIMEX3 tid="t1" type="DATE" value="2016">
2016 godu
</TIMEX3>
zavod
<EVENT eid="e1" class="OCCURRENCE">
rabotal
</EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1" tense="PAST" aspect="IMPERFECTIVE"/>
tol’ko
<TIMEX3 tid="t2" type="DURATION" value="P120D" temporalFunction="false">
120 dnej
</TIMEX3>.
<TLINK eventInstanceID="ei1" signalID="s1" relatedToTime="t1" relType="DURING"/>
<TLINK eventInstanceID="ei1" relatedToTime="t2" relType="HOLDS"/>
```

If we extract from this markup the graph with edges corresponding to temporal links, we get three main vertices joined by to edges in a triangular-like shape: the instance ei1 of working is linked with two different temporal units by the relation DURING and HOLDS. Some values could also be added to the graph (see fig.2). There are no TIMEX3 nodes on fig.2, because the

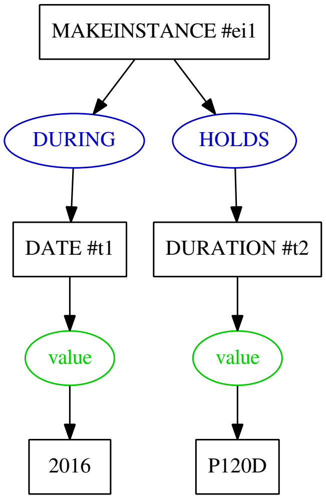

Figure 2: TimeML markup of the sentence (5) presented in a graph phorm
.

corresponding nodes are named according their types: DATE and DURATION. There are four distinct types of TIMEX's in TimeML: DATE, TIME, DURATION and SET. All expressions dealing with calendar time, i.e. comprising labeled units larger than hour, are DATEs (*Friday, October 1, 1999; next week*). All expressions that refer to a time of the day (*9 a.m.*; *ten minutes to three*; *last night*) are TIMEs. DURATIONs are obvious (*three weeks*; *all last night*; *20 days*; *3 hours*). SETs are also self-explaining - this class covers sets of times like *twice a week*.This distinction of types seems to be grounded semantically, but in fact it is formal: ”2016” will be classified as DATE even in the context suggesting that the event holds throughout the year, like in (6).

(6) Ves' 2016 god zavod prostaival.
'The plant stood idle for the year 2016'

The durative nature of this time reference must be expressed by the type of temporal link.

4.2 Temporal Expressions Beyond the Possibilities of TimeML

As we already mentioned, some temporal expressions like *soon* and *long time ago* are out of interest for the TimeML developers and users - they are not described in the specification, and the automatic taggers like HeidelTime not mark them either.

4.2.1 First Case: Soon and Similar Expressions

TimeML markup language does not provide markup decisions for such temporal adverbs and phrases like *soon*, *a long time (ago)* etc. However, it seems that such words play an important role in language functioning - they provide a temporal cohesion in texts. Therefore, a full-fledged time reasoning on texts is impossible without them. Consider the Russian word *skoro* ('soon') in a narrative context.

(7) Rjaboj pošël naverx i **skoro** vernulsja s Romanovym.
‘The man with the pockmarked face went upstairs and soon returned with Romanov’

In this context, the adverb *skoro* means ’after a short time period’. In fact, it reports that some event occurred after some other event. This part of the meaning can, therefore, be expressed with the help of any formal system that includes Allen’s temporary relations. But the exact meaning of the adverb contains the subjective estimate of the time interval that separates the events. A short time period is a period containing ’a little amount of time’ - in the speaker’s opinion. In the semantic representation of the SemETAP speaker’s evaluation of parametric expressions (*small weight*, *small length*, etc.) is reflected by the degree feature, which accept values from the closed set: `MaximalDegree`, `HighDegree`, `MediumDegree`, `LowDegree`, `MinimalDegree`. See fig. 3.

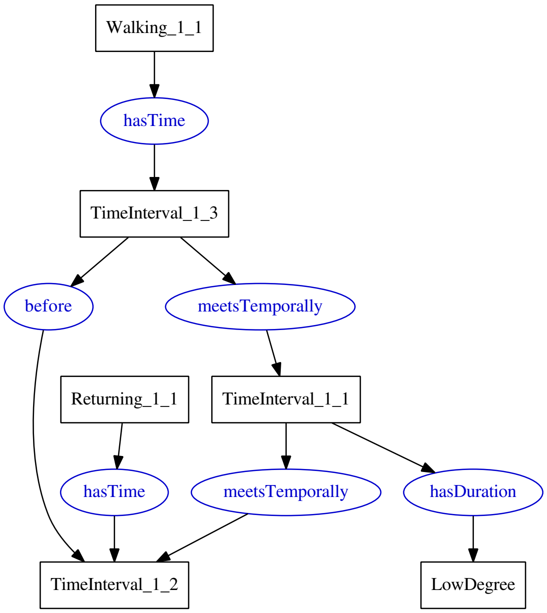

Figure 3: Semantic structure for the sentence (7) containing the adverb *skoro*
.

4.2.2 Second Case: Construction with Ordinal Numerals

. It has already been said that the TimeML distinction between DATE and TIME on one side and DURATION on the other, is purely formal. However, the opposition ’temporal localization’ vs ’temporal duration’ is semantic and intuitively very clear. It may be universal as well. The most common and easy way to distinguish localizations from durations is to ask the question. If the question corresponding to the temporal phrase is *when*, the phrase is classified as temporal localization; if the question is *how long*, the phrase is classified as duration. The sentence (5) contains information both about the temporal localization of the event and about its duration: the phrase *120 dnej* (’120 days’) describes duration, while the phrase *v 2016 godu* (’in 2016’) localizes the event on the time line. So in this example duration and localization have separate expressions, and this case is the most common one.

The meaning 'the time interval that lasts 120 days', is represented in the semantic structure as one branch:

```
hasDays(TimeInterval_1_1,"120"^^xsd:integer)
```

The meaning 'In the year number 2016', corresponding to the temporal localization in the exact year, is represented by the following subgraph.

```
during(TimeInterval_1_1,Year_1_1)
hasNumber(Year_1_1,"2016"^^xsd:integer)
```

The top position in both subgraphs is occupied by the same individual of the `TimeInterval` class, but the graphs themselves are different. Moreover, there is a third graph with the same top. It represents the meaning 'in the past with respect to the present moment (the moment of speech)'. In the sentence this meaning is expressed by the grammatical past tense of the verb.

Thus, the SemETAP representation of time at this point differs from TimeML because the information about duration and about temporal localization is not linked to the event directly, but to it's projection on the time line which projection is a time interval itself.

Some temporal expressions have a much more complex structure. The information about the event duration and localization can be expressed at once, with one synthetic expression. This is the case of Russian accusative construction with ordinal numerals.

(8) Ivan živët v Moskve **odinnadcatyj god**.
'Ivan is living in Moscow for ten years now'. (lit. for the eleventh year).

To represent the meaning of the temporal expression in (8) completely, 4 time intervals are required: the time interval corresponding to Ivan's life in Moscow, the time interval lasting ten years, the eleventh year and the moment of speech. All these intervals are connected: see fig. 4. TimeML allows time intervals, which lack lexical expression in the sentence, for example, the moment of speech. Such intervals could be added to the markup as usual TIMEX3s. Yet this Russian construction can not be marked in accordance with the specification [9], since the main interval characterizing the time of the event is neither DATE nor DURATION. The hybrid nature of the meaning of this expression is confirmed by a question test: it is impossible to ask when-question if we want to get an answer *the eleventh year*. But the how-long-question either does not form a good grammatical pair with this construction, although the answer to this question may be derived from the sentence.

4.2.3 Third Case: Temporal Expressions Characterizing 'Hidden' Events

The previous case is not the only case where the meaning of Russian constructions can not be correctly represented in the TimeML markup. Let's return to the sentence (4). The adverb *zavtra* ('tomorrow') in this sentence localizes on the time line not the expecting, but the arrival, which is not lexically expressed. By the way, in English such a construction is also possible, although it is seldom used (the example from British National Corpus):

(9) I expect them tomorrow morning.

TimeML allows adding the individualized events with the help of MAKEINSTANCE, but the assumption is that an added individual belongs to the same type of situation that the verb denotes. The individuals created with MAKEINSTANCE are designed to represent recurring events and event sets. The examples in question (4, 9) have a different semantic structure: the event that is not lexically expressed in the sentence is the event from subordinate clause. The correct linking of temporal adverbs in this case requires the detection of this 'omitted' event. Such a task can not be properly solved at the level of a shallow text markup and requires a deep semantic analysis.

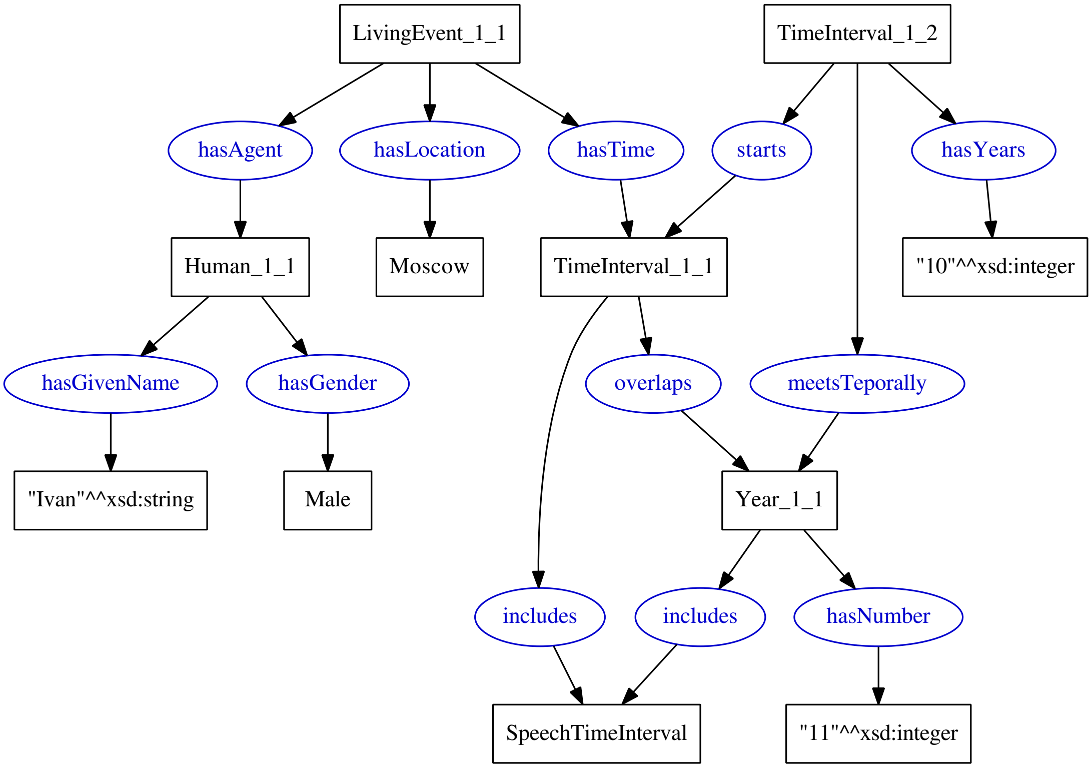

Figure 4: Semantic structure for the temporal construction with an ordinal numeral

.

5 Conclusion

The new formal representation of temporal information has been proposed. It is compatible with the TimeML standard, and more flexible. It covers even sophisticated temporal expressions relying on the temporal relations that occur in sentences ”beyond” the level of lexical units. This formal representation is implemented in the semantic text analyzer SemETAP, which provides an opportunity for applying it on a significant amount of language data and for creating a corpus for future linguistic research and for developing a tool for automatic event extraction and temporal reasoning.

References

[1] James F Allen and George Ferguson. Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5):531–579, 1994.

[2] Couto J. Minel J.-L. Battistelli, B. and Schwer S. Representing and visualizing calendar expressions in texts. *Proceedings of the 2008 Conference on Semantics in Text Processing, STEP ’08*, pages 365—-373, 2008.

[3] Igor Boguslavsky. *Sfera dejstviya leksicheskih edinic (The Scope of lexical units)*. Jazyki slav’anskoj kul’tury, 1996.

[4] Igor Boguslavsky, Vyacheslav Dikonov, Leonid Iomdin, Alexander Lazursky, Victor Sizov, and Svetlana Timoshenko. Semantic analysis and question answering: a system under development. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, pages 64–81, 2015. Available at http://www.dialog-21.ru/digests/dialog2015/materials/pdf/BoguslavskyIMetal.pdf.

[5] Igor Boguslavsky, Tatiana Frolova, Leonid Iomdin, Alexander Lazursky, Ivan Rygaev, and Svetlana Timoshenko. Semantic analysis with inference: High spots of the football match. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, pages 124–142, 2018. Available at http://www.dialog-21.ru/media/4290/boguslavskyim.pdf.

[6] Bernard Comrie. *Tense. Cambridge Textbooks in Linguistics*. Cambridge University Press, 1985.

[7] Leon Derczynski and Robert Gaizauskas. Temporal relation classification using a model of tense and aspect. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 118–122, 2015.

[8] Rdf specification. [online], 2004–2014. https://www.w3.org/TR/rdf-schema/, last accessed February 2019.

[9] Timeml 1.2.1 specification. [online], 2005. http://www.timeml.org/publications/timeMLdocs/timeml_1.2.1.html, last accessed February 2019.