Event Ordering. Temporal Annotation on Top of the BulTreeBank corpus

Laska Laskova
Institute for Parallel Processing, Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
Department of Bulgarian Language, Sofia University "St. Kl. Ohridski"
15 Tsar Osvoboditel Blvd., 1504 Sofia, Bulgaria
laska@bultreebank.org

Abstract

This paper describes the preliminary work on the project of extending the BulTreeBank with temporal information that will serve as a golden standard for Bulgarian language. We outline a flexible markup scheme that is based on a language-specific verb taxonomy and test its capabilities by implementing algorithms for temporal entities recognition in the CLaRK System tool.

Keywords
temporal expressions, temporal relations annotation, verb categories, boundedness

1. Introduction

Recently, an extensive work is being done on the automatic recognition and normalization of temporal expressions in natural languages (e.g. the MUC 6 and MUC 7 Named Entity Recognition Task, the Temporal Expression Recognition and Normalization Task). We propose a TimeML-based annotation scheme for temporal expressions in Bulgarian. The original scheme [9] was modified so that the annotation could benefit from the language-specific means for conveying temporal information: lexical aspectual type, Slavic Aspect (the so called *vid* category), tense and evidentiality. In Bulgarian, a language with rich verbal morphology, they play a crucial role in temporal order decoding.

Our final aim is to facilitate the creation of a gold standard by annotating automatically some of the temporal information. On structure level we focus on the interaction between verb phrases and temporal function words (conjunctions and prepositions). The technical part is carried out using the BulTreeBank, an HPSG syntactically annotated corpus of Bulgarian [11]. A rule-based algorithm for temporal relations detection is implemented in the XML-based CLaRK System [12]. Its performance proves that morphologically encoded aspectual data is important when analyzing temporal relations for Bulgarian.

2. Exploiting Bulgarian verb categories

Although when analyzing temporal relations (TRs) we would like to take into account world-knowledge information, especially causation and knowledge of language usage, at this stage of annotation we do not have the resources to complete such a task in a short time. We decided to calculate automatically temporal relations, which depend solely on sentential syntax, word order, morphological and limited lexical information. In order to achieve this goal we have systematized the information that can be found in the existing descriptive literature [2]. Our next step on this preliminary stage was to develop a taxonomy of lexical aspectual types, which proved to be relevant for encoding temporal ordering.

2.1 Aspectual verb classification

Verbal aspect category *vid* has two subcategories – namely, imperfective (IPF) and perfective (PF). Verbs are overtly marked for their *vid*, except for a relatively small group of biaspectual verbs in third declension. We accept that for Bulgarian language *vid* category encodes information about the boundedness of the eventuality denoted by the verb. This, of course, does not imply that the aspectual type of the verb is fixed, but we argue that this feature imposes some rigid limitations concerning the scope on the structure of the event, and hence some restrictions on the set of possible aspectual properties of the verb [6]. That is why we have decided to build our verb classification with respect to which nucleus element(s) verbs are related to. The well-known nucleus components (Figure 1) are described in the works of Moens and Steedman [8].

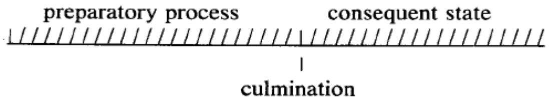

Figure 1. Nucleus structure.

Further subcategorization based on Vendlerian lexical aspectual classification is made with respect to affixation. For Slavic languages like Polish, Bulgarian, Russian and so on it has long been known that the aspectual type is marked by word-formational features and changed through derivational processes (just to mention a few recent studies: [3], [12], [6]). Verb classes whose differences proved to be relevant for TRs recognition are listed below.

2.1.1 Imperfective stem verbs

These are atelic verbs that focus on the unboundedness of the eventuality – states and activities, which are not related to any nucleus as its preparatory process.

2.1.2 Perfective verbs

Here we distinguish three groups. Telic stem verbs are typically achievements or accomplishments (culminated processes in Moens and Steedman terminology). The former focuses only on the culmination of the event structure and the latter both on the preparatory process and the culmination. The same holds for telic verbs derived by prefixes from imperfective base verbs.

Delimitatives derived by *po-* and *nad-* prefixation and expressing bounded but atelic eventualities are accomplishment verbs. In contrast, utterances with the so-called majorative-resultatives, which express activity that ends "beyond the proper limit" [5], e.g. *prejam* – "to have eaten too much", could equally receive the accomplishment as well as the achievement profile. Both classes focus on a process, but in the first case this process does not belong to a nucleus structure, and in the second it is identified as a preparatory process.

Verbs derived by -n_1- suffixation express punctual events with no internal structure. Only few of them denote points that are not incorporated in a nucleus structure. Most of these verbs could receive ingressive reading, focusing on the point which serves as the initial bound for the process. Either way, we treat all -n_1- perfectives as achievement verbs. This inconsistency is corrected on the level of TR annotation. When the perfective verb has a semelfactive reading, it is marked as *MOMENT*, but for an ingressive reading it receives *INITIATION* markup (see Table 1).

2.1.3 Secondary imperfective verbs

Verbs derived from perfectives by the *-a-* suffix or *-v-* suffix and its variants focus on the preparation process in the nucleus structure. For this reason, in many contexts the realization of the culminated process is implied, especially in a present historical tense, and on a number of occasions the nucleus component referred to by the utterance is not the process, but the culmination itself.

2.1.4 Ingressive and terminative verbs

Bulgarian perfective ingressive verbs, prefixed with *pro-* and *za-*, and terminative verbs, prefixed with *do-*, are derived from their imperfective counterparts: *zapeja* (PF) → *zapjavam* (IPF), "to start singing", *dopeja* (PF) → *dopjavam* (IPF), "to finish singing". Since perfectives focus on the process starting point, respectively culmination, they are assigned aspectual class achievement (that can be shifted to accomplishment). Again, for ingressive verbs this is obviously not the most adequate interpretation, but it suits us for the moment. On the other hand, imperfectives are assigned aspectual class activity (that can be shifted to achievement), because they focus on the beginning phase of a process, not necessarily culminated or otherwise limited, respectively the finishing phase of a culminated process that is implied to be interrupted.

2.1.5 Encoding aspectual class

Since on a token level verb forms in the BulTreeBank corpus are annotated with morphosyntactic tags providing information about *vid* category [10], we decided to use yet another attribute, *AspCat* (Aspectual Category). In accordance with the above classification, this tag receives one of the following five values: *state, act, ach, acc-ach, acc-act* (corresponding to Vendlerian types state, activity, achievement, accomplishment or achievement, accomplishment or activity). The ambiguity of the values is intended. The introduced attribute is not part of the tag set for temporal information mark-up. For the moment, the annotation is done manually but is computer-assisted[1]. Verbs that have only iterative readings are regarded as processes and their *AspCat* attribute receives *act* value, but on the level of TRs annotation they are further subcategorized as *SERIES*.

2.1.6 Encoding phase

Bulgarian verbs encode not only information about the type of eventuality expressed, but also about its phase. TimeML temporal annotation scheme provides a special mark-up for aspectual verbs and their complements, but we have to employ another attribute for ingressives and terminatives, namely, *@phase* (Table 1).

3. TimeML adopted for Bulgarian

TimeML emerged as a markup language for time, events and temporal links after the TERQAS workshop held in 2002 [9]. Temporal information should be represented via several tag types: EVENT – for event tokens, where event is any kind of situation that happens or occurs, MAKEINSTANCE for event instances (in contrast to event tokens), SIGNAL for textual elements that explicitly mark temporal or modal relations and quantification over events, TIMEX3 for temporal expressions, and LINK for relationships. The LINK tag is always one of the following types: TLINK (Temporal Link) for relations between two events or an event and a time, SLINK (Subordination Link) for relations between two events or an event and a signal, and ALINK (Aspectual Link) for relations between an aspectual event and its argument event.

The corpus annotated according to TimeML, TimeBank, comprises English newspaper articles marked for temporal information only, but our corpus is HPSG syntactically annotated on HPSG-based grounds, which, besides language specificity, calls for altering some of the TimeML tags.

[1] In the future this task will be accomplished by means of a regular grammar, based on the lexicon of Bulgarian verb Aktionsarten [5], revised accordingly to the classification presented in this paper (section 2.1.).

Table 1. EVENT attributes for Bulgarian

EVENT *ATTRIBUTE*	*VALUES*	*ACCOUNTS FOR*
aspect, altered	STATE, ACTIVITY, ACHIEVEMENT, ACCOMPLISHMENT, MOMENT, SERIES, NOT_SPECIFIED	Vendlerian aspectual class; note that two more classes are added – *series* for iterative "one episode" eventuality [4], and *point* for semelfactives [8]
class, altered	REPORTING, PERCEPTION, ASPECTUAL, INTENSIONAL, OTHER	lexical meaning. Events of type I_STATE or I_ACTIVITY (Intensional State, resp. Activity) are thus "decomposed" and signaled by two attribute values
conState, new	FACT, RESULT	lack or presence of culmination in the event structure for the related verbs in the present and past perfect tense; achievements, accomplishments *and* points are opposed to states and activities
evid, new	INDICATIVE, RENARRATIVE, CONCLUSIVE, DUBITATIVE	verbal evidentiality feature. It indicates both with the source of information and speaker's attitude about the statement validity
location, new	PAST, PRESENT, FUTURE, NONE	orientation of the event with regard to the perspective time – document creation time or other. BulTreeBank provides tense information derivable from the morphosyntactic tags [10]
persp Anchor, new	boolean	event potential to anchor shift of perspective. Its default value is true for perceptive and reporting verbs
phase, new	INITIATION, TERMINATION	beginning or end phase of the eventuality, encoded morphologically
status, new	NEGATIVE, POSITIVE	lack or presence of verb negation. When negated, verbs are treated as denoting moments or intervals where a particular situation does not hold

3.1 Adjustments of the annotation for the BulTreeBank corpus

All TimeML tags are represented as empty daughter elements with an appropriate attribute set. LINKs for intersentential relationships are embedded under the sentence node, TIMEX3, SIGNAL and EVENT elements are daughters in first position of the relevant lexical or phrasal node.

3.1.1 EVENT element

On this stage we annotate automatically only events expressed by means of verbs. In the BulTreeBank annotation scheme, verb complex, i.e. finite verb, accompanied by clitics, auxiliary particles (auxiliary verb forms and negative particles), participles and emphatic adverbs, is considered as a multi-token verb [10]. For this reason, some of the relations between event and signal, for example, are annotated not by means of LINK, but as a value for EVENT tag attribute.

The EVENT element introduced for the needs of the BulTreeBank temporal annotation is different from its TimeML counterpart. New optional attributes were added, and some of the values of the old attributes were altered. The differences are summarized in Table 1.

3.1.2 Other elements

There are some other changes in the scheme but due to the lack of space we cannot present them here. Since we focus on temporal ordering between events, we have to mention at least two of them. Originally, *RelType* attribute for TLINK has 13 possible values, based on James Allen's [1] 13 interval-interval and interval-moment relations: *BEFORE, AFTER, IBEFORE, IAFTER, INCLUDES, IS_INCLUDED, HOLDS, SIMULTANEOUS, IDENTITY, BEGINS, ENDS, BEGUN_BY, ENDED_BY.*

In our scheme @RelType is required, so we add the 14th value VAGUE, for temporal relations that are ambiguous or cannot by assigned automatically.

SLINK will not represent a relationship between a SIGNAL for negation particles and an EVENT when the verb is negated. Instead, this information will be encoded via the *status* attribute. As a consequence, ELINK (Entailment Link) is introduced to describe entailed TRs between an eventuality and a negation argument situation.

4. Experiment

Our aim was to test a rule-based approach for detecting TRs between events by employing information about sentential syntax, word order, temporal signal, tense, verb negation and sets of possible aspectual types.

The experiment was performed on a small set of 132 two-clause sentences extracted from the BulTreeBank corpus. 118 sentences of them are verbal head-adjunct phrases, 14 – coordinated phrases, in both cases clauses are connected by *dokato* conjunction. As a coordinating conjunction it corresponds to English "whereas". Subordinating *dokato* regarded as ambiguous: the two eventualities could be overlapping ("while"), or, one of the events, regardless of constituents' relation, is ended by the other ("until"). For instance:

(1) *Докато те чаках, гледах телевизия.*

while you.ACC wait.1sg.IPF.IMPERFECT watch.1sg.IPF.IMPERF TV

"**While** waiting for you, I was watching TV",

(2) *Гледах телевизия, докато (не) дойде сестра ти.*

watch.1sg.IPF.IMPERF TV until (not) come.3sg.PF.AORIST sister your

"I was watching TV **until** your sister came"

In the second example negating subordinated VP has no impact on the sentence meaning. We propose an interpretation that covers all cases illustrated above with the exception of coordinating *dokato* properties. As a subordinated constituent, *dokato* clause belongs to the frame adverbials class. The interval referred to is construed depending on the aspectual structure provided by the verb. If possible, the end point of the interval is anchored. When a subordinated verb expresses a moment-like eventuality (i.e. points or achievements) or an eventuality composed by process and culmination/termination (accomplishments), it serves as an endpoint (sentences (1) and (2), positive verb form variant). If not, the interval is identified by the activity/state, that is, when during the interval a particular positive or negative situation holds (sentence (2), negative variant).

Eventualities and temporal ordering annotation was implemented within the CLaRK System. We used Constraints, XPath Insert and Transformation tools. Our first step was to add information about aspectual class sets. Then a number of constraints and regular grammars were applied in a particular order: identification of EVENT and SIGNAL elements, the relevant attributes that receive "sure" value, TLINK insertion, ELINK insertion, and finally, establishment of TRs type. This simple algorithm ends with ascribing VAGUE value where more and different kinds of data are needed to calculate relType.

We create algorithms for assigning one of the 4 possible relType values for TLINK in coordinated sentences: SIMULTANEOUS, IS_INCLUDED, ENDED_BY and VAGUE in cases where the rule-based approach is insufficient, and 3 possible values for ELINK: IS_INCLUDED, INCLUDES and SIMULTANEOUS. For subordinated sentences the number of possible values increases, and even expert annotators have difficulties accessing relType.

The results we obtained are the following. 166 TLINK and ELINK elements are inserted automatically. Overall we achieve 63.7 % recall, 91.1 % precision, F1-score – 0.75. The worst performance is for ordering bounded-bounded eventuality, where @AspCat = "acc-ach" or "ach" and one of the verbs is in a non-perfective tense, while the other is in perfect (disregarding the type of the sentence). Best performance was for ordering bounded-unbounded eventualities (in complex sentences).

5. Conclusions and further work

The annotation of eventualities and temporal relationships is a subtask of a more general project – annotation of temporal information (first time for Bulgarian language) on top of the BulTreeBank. The CLaRK System, the system originally used for the creation of the BulTreeBank, will be further employed for implementing TIMEX annotation. As a preliminary step, we have created a verb classification and a refined annotation tagset, based on the TimeML standard, which was tested by implementing algorithms for automatic temporal entities recognition and markup in the CLaRK system.

6. Acknowledgements

The author wish to express her gratitude to her supervisor Ruselina Nitsolova, as well as to Petya Osenova and Kiril Simov for their encouragements and guidance. Special thanks are due to the anonymous reviewers whose criticism helped to improve this paper substantially. The work was granted partially by LTfLL project (FP7-212578).

7. References

[1] Allen, J. Maintaining knowledge about temporal intervals. *Communications of the ACM*. 26:832-843 November 26, 1983.

[2] *Bulgarian Academy Grammar*. Abagar, Sofia, 1983.

[3] Damova, Mariana. *Tense and Aspect in Discourse: A study of the interaction between aspect, discourse relations and* temporal reference within discourse representation theory with special attention to Bulgarian. PhD thesis, Stuttgart, 1998.

[4] Freed, A. The Semantics of English Aspectual Complementation. D. Reidel P.Company, Dordrecht, Holland, 1979.

[5] Ivanova, K. Nachini na glagolnoto dejstwie v syvremennija bylgarski ezik. Izdatelstvo na BAN, Sofia, 1974.

[6] Laskova, L. Taksisni otnoshenija v bipredikativni izrechenija za vreme. MA Thesis. SU St. Kliment Ochridski, Sofia, 2003.

[7] Młynarczyk, A. Aspectual Pairing in Polish. Utrecht: LOT, 2004. Available at: http://igitur-archive.library.uu.nl/dissertations/2004-0309-140804/inhoud.htm Last accessed Jun 08, 2009

[8] Moens, M. & Steedman, M. Temporal Ontology and Temporal Reference. Computational Linguistics, 14(2):15-28, June, 1988.

[9] Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., & Pustejovsky, J. 2002. TimeML Annotation Guidelines, Version 1.2.1. Available at: http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf Last accessed: Jul 13, 2009.

[10] Simov, K., Osenova, P. & Slavcheva, M. 2004. BTB-TR03: BulTreeBank Morphosyntactic Tagset. BTB-TS version 2.0. Available at: http://www.bultreebank.org/TechRep/BTB-TR03.pdf Last accessed: Feb 36, 2009.

[11] Simov, K. & Osenova, P. BTB-TR05: BulTreeBank Stylebook. BulTreeBank Version 1.0. Available at: http://www.bultreebank.org/TechRep/BTB-TR05.pdf Last accessed: Feb 36, 2009.

[12] Simov, K., Peev Z., Kouylekov M., Simov A., Dimitrov M. & Kiryakov, A. CLaRK - an XML-based System for Corpora Development. In: Proc. of the Corpus Linguistics 2001 Conference, pp 558-560, 2003.

[13] Petruhina, E. Aspektual'nye kategorii glagola v russkom jazyke: v sopostavlenii s c'eshskim, slovatckim, pol'skim i bolgarskimi jazykami. Izdatel'stvo Moskovskogo universiteta, Moskva, 2000