Annotation of the Syntax/Semantics interface as a Bridge between Deep Linguistic Parsing and TimeML

Mark-Matthias Zymla
University of Konstanz
Mark-Matthias.Zymla@uni-konstanz.de

Abstract

This paper presents the development of an annotation scheme for the syntax/semantics interface that may feed into the generation of (ISO-)TimeML style annotations. The annotation scheme accounts for compositionality and calculates the semantic contribution of tense and aspect. The annotation builds on output from syntactic parsers and links information from morphosyntactic cues to a representation grounded in formal semantics/pragmatics that may be used to automatize the process of annotating tense/aspect and temporal relations.

1 Credits

We gratefully acknowledge funding from the Nuance Foundation. We also thank collaborators from the *Infrastructure for the Exploration of Syntax and Semantics (INESS)* and the *ParGram* projects.

2 Introduction

In this paper we report on the progress of a project concerned with the development of a novel annotation scheme for tense/aspect. This annotation scheme is designed to interact with morphosyntactic information that is the result of deep parsing. It is also designed to be crosslinguistically applicable and was first introduced in (Zymla, 2017a; Zymla, 2017b).

The annotation scheme is designed to be applied to linguistically parsed input, i.e. syntactic treebanks. In particular, we work with analyses resulting from deep syntactic parsing within the ParGram effort (Butt et al., 2002), which includes a wide variety of different types of languages. In addition to working with data from the ParGramBank (Sulger et al., 2013), we adapted crosslinguistically applicable testsuites found in (Dahl, 1985). Furthermore, we began experimenting with application of the annotation scheme to a treebank based on the TempEval-3 TimeML corpus (UzZaman et al., 2013).[1] Our annotation scheme is also compatible with representations resulting from universal dependency grammars (section 4).

The annotation scheme goes beyond the effort presented by Ramm et al. (2017) in that it can interact with both deep linguistic parsers as well as the shallower dependency parsers solely utilized by Ramm et al. (2017). It is also generally cross-linguistically applicable, rather than being restricted to the closely related European languages English, German and French. Furthermore and most importantly it allows for the annotation and dynamic calculation of the semantics and pragmatics of tense/aspect that go beyond the individual morphosyntactic cues.

[1] The treebanks are available at http://clarino.uib.no/iness (Rosén et al., 2012)

However, our annotation scheme is not aimed at replacing existing standardized annotations schemes such as (ISO)-TimeML (Pustejovsky et al., 2003), but rather aims at bridging a gap between TimeML style annotations and the actual morphosyntactic cues about tense/aspect found crosslinguistically. The original TimeML corpora (TimeBank (Pustejovsky et al., 2003)) and TempEval3 (UzZaman et al., 2013)) provide semantic annotation in terms of temporal links. However, it is difficult to test predictions concerning the mapping from form to meaning based on the syntax/semantics interface provided in TimeML.

A notable innovation of the annotation scheme presented here is that it distinguishes between several layers of semantic (and pragmatic) reasoning by calculating annotations at three different tiers. In the first two tiers abstract meaning concepts are calculated directly from the morphosyntactic input. Tier 3 then relates the calculated semantic concepts to the actual (explicit or implicit) temporal context. This dynamic annotation scheme consists of two parts: the syntax/semantics interface, a rule-based system that maps syntactic information onto abstract semantic properties and, secondly, a layer that describes the temporal ordering of eventualities. This is explained in detail in section 3.

The subordinated goal in this paper is to explore the benefits of a bi-directional pipeline between TimeML corpora and our annotation scheme. For this pipeline we take inspiration from the Reichenbachian Tense Markup Model (RTMML) (Derczynski, 2016) and other work whose goal is to incorporate (neo-)Reichenbachian tense semantics, e.g., Gast et al. (2016). This means that we provide a semantic annotation that restricts the relations between speech time, reference time and run time of any given event expressed by a verb form (section 5).

The goals of our project are thus two-fold: The first goal, which has been in the focus of the project up to now, is to provide a cross-linguistically valid annotation scheme for the syntax/semantics interface that takes into account the current state of the art with respect to formal semantics (see Tonhauser (2015) for an overview). The current goal is to improve upon the linguistically motivated temporal relation typing and to thus contribute to the growing system of ISO-compliant temporal annotations.

3 Development of the Annotation Scheme

Our annotation scheme is loosely based on the Reichenbachian tradition (Reichenbach, 1947). In the Reichenbachian framework tenses are categorized by means of three different kinds of points/intervals in time. First, the speech time S corresponds to the moment of utterance. Second, the reference time R encodes a time a given utterance refers to. Third, the event time E describes the time during which an event takes place. Example (1) illustrates a subset of the tenses proposed in Reichenbach's system. In simple tenses the reference time and the event time overlap and are related to the speech time in a straightforward fashion via *anteriority*, *posteriority* or *overlap*. The need for the reference time especially arises due to constructions such as the *past perfect*, which is used when talking about events that happened before a certain point in the past. The Reichenbachian system treats the past perfect as one tense. To achieve a more flexible system, it can be translated into an interval calculus, where E,R and S are intervals $t, ...t^n$ and the relations $(\prec, -)$ between these points are what we define as tenses (Comrie, 1985). Thus, the annotation of the past perfect results in two separate tenses with a specific relative order – a result of the underlying semantic composition — as shown in the XML annotation of the semantic representation (2) in terms of temporal relations (trel).[2]

[2]TimeML utilizes TLINKs to express temporal relations between elements (events,temporal expressions). However, our `trels` are distinct from TLINKS. There are two main reasons: i) the `trels` may express sets of relations (see *future perfect*), ii) the `trels` may express a relation between variables and/or concrete temporal elements.

(1)

simple past:	$E - R \prec S$	past perfect:	$E \prec R \prec S$
simple present:	$S - E - R$	present perfect:	$E \prec S - R$
simple future:	$S \prec E - R$	future perfect:	$S \prec E \prec R, S - E \prec R, E \prec S \prec R$

(2) **Sample annotation – Representation**

a. (Yes), I had met Peter.

```
<doc time="now"/>
<timeref xml:id="t0" target="doc_time"/>
<timeref xml:id="t1" target="#token2"
tense abstract="past" semantics="r2"/>                    had

<event xml:id="e1" target="#token3">                      met
<!-- conceptual description -->
<concepts xml:id="c0">
<tense="past"/> ... </concept></event>
<!-- temporal realization -->
<trel xml:id="r1" relation="e1≺t1"/>                      E ≺ R
<trel xml:id="r2" relation="t1≺t0"/>                      R ≺ S
```

At the center of the annotation scheme is the mapping from syntax to semantics. The first layer of the annotation maps morphosyntactic elements onto abstract semantic exponents (Tier 1) as per the tense system introduced above. For example, the overt morphosyntactic cues express past in (3a), present in (3b) and non-past in (3c).[3] In a second layer, semantic components can be combined to construct new meanings not (necessarily) encoded in the parts they were derived from (mimicking compositionality, but more powerful) in a rule-based system (Tier 2). This is necessary for examples like (4b), which presupposes that the verbal predicate is directly or indirectly pluralized (`NUM pl`). In this case, (4b) is pluralized indirectly by virtue of the plural object.

Tiers 1 and 2 form a set of implication rules. These rules generate abstract meanings, in the sense that they are not yet anchored to a time line. (4b) is analyzed as a habitual imperfective in (4b), while (4c) is non-past with an explicit reference time introduced by the adverbial 'tomorrow' ((4c)). This reference time is integrated in Tier 3, which weaves the Tier 1 and 2 information into the explicit and implicit context. Tier 3 also serves as the locus for pragmatic reasoning and contains so-called compatibility rules since it relies on the compatibility of the overt morphosyntactic cues and the given context.

(3) a. John left. b. John build-s houses. c. John komm-t morgen an.
John arrive-Pres tomorrow

(4) a. `TENSE past` $\rightarrow$ `TEMP-REF 'past'` : $\lambda t.t \prec t_0$ Tier 1

b. `TENSE pres` : $\lambda t.t \circ t_0 \wedge$ `NUM 'pl'` : $\lambda e.pl(e) \rightarrow$
`ASPECT 'imperfective'` $\wedge$ `A-RESTR 'habitual'`[4] Tier 2

c. `TEMP-REF 'non-past'` : $\lambda t.t \not\prec t_0 \wedge$ `TIMEREF tomorrow(t1)` $\otimes$
`TEMP-REF 'tomorrow'` : $\lambda t.t \not\prec t_0 \wedge t \subset t1$ Tier 3

The annotation scheme was developed with two goals: First, to provide a representation of temporal semantic properties in compliance with formal linguistic research. Second, to provide a framework for the research on crosslinguistic semantic variation. We push the idea behind parallel treebanks as for example in the ParGram project (Butt et al., 2002) and the alignment of parallel corpora based on

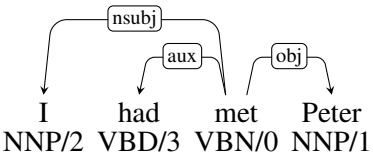
[3] All rules are language specific and illustrate cross-linguistic variation in the mapping from form to meaning. $\rightarrow$ = implication rule, $\otimes$ = compatibility rule; `TENSE` = morphological tense, `TEMP-REF` = semantic tense, `NUM` = semantic verbal number, `ASPECT` = Viewpoint, `A-RESTR` = viewpoint interpretation (e.g., progressive, habitual ...).

[4] We omit a formal representation for reasons of space. See (Ferreira, 2016).

their morphosyntax, semantics and pragmatics. As of now, parallel corpora are mostly aligned based on morphosyntactic properties. However, our annotation of the syntax/semantics and pragmatics interface allows for a more fine-grained alignment and thus provides a valuable way forward for cross-linguistic research based on semantically parallel corpora.

4 Normalization Across Syntactic Parsers

One of our main goals is to provide a pipeline between a semantic annotation and syntactic treebanks. For this purpose, we work with an explicit syntax and semantics. We exemplify this in terms of representations derived from the deep linguistic XLE parsers based on LFG (Lexical Functional Grammar), in particular the f(unctional)-structure, which is a syntactic representation that encodes grammatical functions (subject, object, complement) and functional information (e.g., tense, aspect, mood) (Crouch et al., 2017; Bresnan et al., 2015). We can also work with universal dependencies (UD)[5], where dependencies correspond to grammatical functions in LFG and UD features correspond to functional information. The two representations are illustrated in Figure 1 (UD on the left, LFG on the right). Both of these syntactic structures are mappable onto a conceptually equivalent computational structure. We work with a normalization whereby each token of the UD parse corresponds to an entry in a hashmap which contains a list of dependencies and/or features as value. Figure 1 illustrates the normalization of the UD structure adding UD morphosyntactic features.[6]



$$0\begin{bmatrix}\text{PRED 'met} < \text{I, Peter} > \\ \text{OBJ} \quad 1\left[\text{PRED 'Peter'}\right] \\ \text{SUBJ} \quad 2\left[\text{PRED 'I'}\right] \\ \text{TNS-ASP} \quad 3\left[\text{TENSE past, PERF +\_ MOOD indicative}\right]\end{bmatrix}$$

```
{0 = [(TOKEN, 'met'), (NSUBJ,1),(OBJ,2), (AUX,3)],
 1 = [(TOKEN, 'I')],
 2 = [(TOKEN, 'met'),(Tense,Past),(VerbForm,Part)],
 3 = [(TOKEN, 'had'),(Mood,Ind),(Tense,Past),(VerbForm,Fin)]
```

Figure 1: Universal dependency structure normalization

Sample annotation – Rules

```
#g AUX #h VerbForm Fin ∧ #h Tense Past →
(#h) TIMEREF past(t) : λt.t ≺ t0
#g VerbForm Part ∧ #g Tense Past ∧ #g AUX (#h) TIMEREF past(t) →
(#g) TEMP-REF 'past' : λt'.t' ≺ t
```

Figure 2: UD structure annotation

Based on this internal structure the INESS query language (Rosén et al., 2012) that is used to identify syntactic paths in treebanks has been modified to identify elements for annotation. The object of annotation are (sets of) syntactic properties which are expressed in terms of tree paths from the verbal spine

[5] http://universaldependencies.org/. We use the Stanford CoreNLP parser (Chen and Manning, 2014)

[6] The idea of the normalization is based on Unhammer (2010). It does not map f-structures and dependency structures on to a formally equal representation, but onto a representation that may be annotated in the same manner. The endeavor of translating f-structures faithfully to dependency structures is discussed in Meurer (2017).

(the root element) to the appropriate morphosyntactic feature. In Figure 2 partial syntactic structures (UD heads, f-structures) are identified via variables (`#g`–`#o`). For example the expression `#g AUX #h VerbForm Fin` refers to any UD head which stands in an `AUX` relation with another head with the UD feature `VerbForm Fin`. Inside a single rule query variables retain their instantiation, thus the second conjoint of the first rule in Figure 2 can be abbreviated. Semantic elements are associated with a specific head, e.g., `(#h) TIMEREF past(t)`:$\lambda t.t \prec t_0$ tells us that each distinct head that is bound by the variable `#h` (in this case only one if we only consider the sentence in Figure 1) introduces a temporal variable that is restricted to a time that is in the past of the speech time t_0. The second rule in Figure 2 introduces a second tense that is relative to the auxiliary tense.

5 Treebank Annotation — Worked Examples

Table 1 shows statistics on our material with respect to languages, number of sentences, rules (Tier 1 and Tier 2) and compatibilities (Tier 3) for our work concerning just the past tense. Overall 764 sentences in 10 languages were considered. Table 1 shows that the complexity of the syntax/semantics interface with regard to past tense is straightforward in terms of implication rules. However, most sentences are contextually further restricted explicitly (e.g., via temporal modifiers) or implicitly (via context).

The annotation scheme was first developed through sentence level annotations with the idea to provide a qualitatively sound and comprehensive semantic annotation of linguistic categories. For this purpose we produced a treebank based on the typological work by Dahl (1985).

	Total	**German**	**Italian**	**Polish**	**Urdu**	**Indonesian**
Sentences	196	56	50	45	47	48
Implication rules	9	3	3	1	3	2
Compatibilities	191	45	39	34	36	37

Table 1: Annotation of the Past Tense

The annotation consists of two sets of rules with different felicity conditions to distinguish semantic and pragmatic processes. Implication rules provide more general but robust meanings derived from the morpho-syntactic input and semantic construction rules (Tier 1 and 2), while compatibility rules anchor meanings in the implicit and explicit context (Tier 3). From another perspective, semantic (i.e. Tier 1 and Tier 2) rules cover meaning grammaticalized in linguistic categories and compatibility (i.e. Tier 3) rules define restrictions implicit in the context or stated by lexical elements (i.e. temporal modifiers). Consequently, Tier 1 and Tier 2 rules generate a context independent, abstract semantic representation that is mapped onto actual contexts by means of the third tier.

Our data shows that two main Reichenbachian relations are relevant: $E \prec S$ and $E \prec R \prec S$. These are a simple temporal backshift of an event (E) or an iterated backshift which situates the event in the past of a past reference point (R).[7] German and Italian express the simple past in two variants: past tense or present perfect morphology. Urdu distinguishes between perfective and past tense morphology (hence the difference in f-structural analysis). Indonesian usually does not specify tense morphologically and requires contextual inferences, but optionally uses perfect auxiliaries to express semantic past tense and iterated past tense.

In (5) a slice of the possible cross-linguistic variation in the expression of the iterated backshift $E \prec$

[7] $E, R \prec S$ (simple past) and $E \prec R, S$ (present perfect) are subsumed under under $E \prec S$.

$R \prec S$ is illustrated.[8] Both English (5a) and Urdu (5b) may be considered variants of the prototypical *past perfect*. In contrast, tenseless Indonesian (5c) only optionally employs iterated perfect markers.

(5) [Q: When you came to this place a year ago, did you know Peter?]

a. (Yes), I had met Peter. $\left[\text{TNS-ASP} \left[\text{TENSE past, PERF +}\right]\right]$

b. (hãã), mãĩ piter=se milaa thaa. $\left[\text{TNS-ASP} \left[\text{TENSE past, ASPECT prv}\right]\right]$
(yes), I Peter=with meet.Perf be.Past

c. (ya), saya sudah pernah ber-temu dengan Peter $\left[\text{TNS-ASP} \left[\text{PERF +}\right]\right]$
(yes), 1st already ever MID-meet with Peter

```
<timeref xml:id="t3" target="#token15"/>                    had

<event xml:id="e2" target="#token16">                       met
<!-- conceptual description ... --> </event>

<trel xml:id="r3" relation="e2≺t3"/>                        E ≺ R
<trel xml:id="r4" relation="t3≺t0"/>                        R_met ≺ S
<trel xml:id="r5" relation="t3=t2" />                       R_met = R_came
```

Our annotation system takes the possibility of cross-linguistic morphosyntactic variation into account by providing a combination of "translational" (Tiers 1 and 2) and inferencing rules (Tier 3). These rules calculate a formal semantic representation of temporal relations from information provided by deep linguistic parsers like the LFG systems or from UD representations. For Urdu and English, the morphosyntactic cues themselves provide a clear reading of iterated backshift (past perfect). In contrast, these temporal relations must be calculated via inferencing depending on the previous context for the tenseless language Indonesian. However, the end result is parallel in its meaning as expected.

6 Summary and Conclusion

Overall, our annotation scheme provides insights into the syntax/semantics interface by specifying rules as to how morphosyntactic material expresses semantic tense/aspect categories directly and indirectly across languages. This allows for an abstraction over distinct morphosyntactic material with respect to semantic analysis and is an important requirement for a crosslinguistically valid annotation of the syntax/semantics interface. Currently, there are two main approaches to integrating semantic annotations of tense/aspect into TimeML. First, encode semantic variables directly as TIMEX in the annotation (Gast et al., 2016). Second, use a tense/aspect annotation that is independent of the TimeML standard (but compliant with the relevant ISO norms) that serves as a preprocessing step (Derczynski, 2016). In the spirit of the latter, we have developed a neo-Reichenbachian annotation. The system consists of three tiers, whereby the first two tiers comprise of default (Tier 1) and constructed (Tier 2) meanings generated from a robust rule system. The role of the Tier 3 annotation is to resolve ambiguities to the point where explicit TimeML compliant temporal relations may be specified – a process that at this point still requires human assistance. In sum, we present a system that provides: i) crosslinguistically motivated insights into semantic properties of tense/aspect; ii) the possibility of systematically abstracting over crosslinguistic variation; iii) a bridge between deep linguistic parsing and interoperable semantic annotation schemes such as TimeML. This allows us to broaden the research perspectives for qualitative linguistic research by providing tools that allow for the quantitative testing of qualitative predictions.

[8] `PERF +` = perfect construction; `prv` = perfective; `MID` = middle voice

References

[Bresnan et al.2015] Joan Bresnan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2015. *Lexical-functional syntax*, volume 16. John Wiley & Sons.

[Butt et al.2002] Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In *Proceedings of the 2002 Workshop on Grammar Engineering and Evaluation*, volume 15, pages 1–7. Association for Computational Linguistics.

[Chen and Manning2014] Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser Using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 740–750.

[Comrie1985] Bernard Comrie. 1985. *Tense*, volume 17. Cambridge University Press.

[Crouch et al.2017] Dick Crouch, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula Newman, 2017. *XLE Documentation*. Palo Alto Research Center.

[Dahl1985] Östen Dahl. 1985. *Tense and Aspect Systems*. Oxford: Blackwell.

[Derczynski2016] Leon RA Derczynski. 2016. *Automatically Ordering Events and Times in Text*, volume 677. Springer.

[Ferreira2016] Marcelo Ferreira. 2016. The Semantic Ingredients of Imperfectivity in Progressives, Habituals, and Counterfactuals. *Natural Language Semantics*, 24(4):353–397.

[Gast et al.2016] Volker Gast, Lennart Bierkandt, Stephan Druskat, and Christoph Rzymski. 2016. Enriching TimeBank: Towards a More Precise Annotation of Temporal Relations in a Text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

[Meurer2017] Paul Meurer. 2017. From LFG Structures to Dependency Relations. *Bergen Language and Linguistics Studies*, 8(1).

[Pustejovsky et al.2003] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, and Lisa Ferro. 2003. The Timebank Corpus. In *Corpus Linguistics*, volume 2003, page 40. Lancaster, UK.

[Ramm et al.2017] Anita Ramm, Sharid Loáiciga, Annemarie Friedrich, and Alexander Fraser. 2017. Annotating Tense, Mood and Voice for English, French and German. *Proceedings of ACL 2017, System Demonstrations*, pages 1–6.

[Reichenbach1947] Hans Reichenbach. 1947. The Tenses of Verbs. *Elements of Symbolic Logic*, pages 287–298.

[Rosén et al.2012] Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An Open Infrastructure for Advanced Treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco, editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29.

[Sulger et al.2013] Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh M Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Özlem Çetinŏglu, I Wayan Arka, and Meladel Mistica. 2013. ParGramBank: The ParGram Parallel Treebank. In *ACL*, pages 550–560.

[Tonhauser2015] Judith Tonhauser. 2015. Cross-Linguistic Temporal Reference. *Linguistics*, 1:129–154.

[Unhammer2010] Kevin Brubeck Unhammer. 2010. LFG-based Constituent and Function Alignment for Parallel Treebanking.

[UzZaman et al.2013] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 Task 1: Tempeval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9.

[Zymla2017a] Mark-Matthias Zymla. 2017a. Comprehensive Annotation of Cross-Linguistic Variation in the Category of Tense. In *12th International Conference on Computational Semantics*.

[Zymla2017b] Mark-Matthias Zymla. 2017b. Cross-Linguistically Viable Treatment of Tense and Aspect in Parallel Grammar Development. In *Proceedings of the LFG17 Conference*. CSLI Publications.