Cross-linguistic annotation of tense and aspect syntax and semantics

Mark-Matthias Zymla

University of Konstanz

Universität
Konstanz

November 22nd, 2017

Outline

1. Introduction
2. Temporal annotation – A quick overview
3. Comprehensive annotation of the category tense

Outline

1. Introduction
2. Temporal annotation – A quick overview
3. Comprehensive annotation of the category tense

Tense and aspect in multilingual semantic construction

- Research project at the University of Konstanz
- Funded by the Nuance foundation
- Project goals:
  - Annotation of tense and aspect informed by formal semantics
  - Creating resources for NLP research and applications
  - Researching tense and aspect in under-resourced languages
  - Bringing together temporal annotation and deep linguistic parsing

NUANCE
FOUNDATION

ParTMA and INESS

- ParGram and ParTMA work in collaboration with the INESS infrastructure (Rosén et al. 2012)
  INESS website: `http://clarino.uib.no/iness`
- XLE parses are online and available to partners of the ParGram project
- Parses to be integrated into ParGramBank (Sulger et al. 2013)
- Working on visualization of semantic annotation for webpages

iness

In this talk ...

- We aim to present a comprehensive annotation scheme for the linguistic category of tense
    - We aim to bring together state-of-the-art formal semantic research and computational models of temporal mark-up
    - We address the semantic properties of tense within and across languages
    - Explicit annotation of its variation in terms of syntactic and semantic instantiation

Data

Primarily from ParGram ("Parallel Grammar"): NLP project based on Lexical Functional Grammar (LFG)

- Multilingual grammar development project
- International collaboration, with yearly meetings
- Large-scale, robust, parallel computational grammars
- So far:
  - Larger grammars for English, German, French, Norwegian, Chinese, Japanese, Polish
  - Smaller grammars for Indonesian, Malagasy, Turkish, Welsh, Wolof, Urdu, Georgian, Hungarian

Data II

- ParGramBank: parsebank/treebank for 11 languages, developed in INESS (Sulger et al. 2013)
- ParTMA treebank: Collection of treebanks expressing tense and aspect variation; steadily growing in collaboration with ParGram members
- **Currently:** 491 sentences in 13 treebanks from 11 languages. Parallel treebank for semantically past tense sentences (inspired by Dahl (1985))

Outline

1. Introduction
2. Temporal annotation – A quick overview
3. Comprehensive annotation of the category tense

Basics of temporal annotation

*"**Once** there **was** a scorpion **standing** by a river. The scorpion **was looking** for a way to **cross**, **when** he **noticed** a frog behind him. He **asked** the frog to **carry** him across the river."*

Basics of temporal annotation

*"**Once** there **was** a scorpion **standing** by a river. The scorpion **was looking** for a way to **cross**, **when** he **noticed** a frog behind him. He **asked** the frog to **carry** him across the river."*

a. **Eventualities:**
was standing(e_1), was looking(e_2)
noticed(e_3), asked(e_4)
cross(e_5), carry(e_6)

b. **Temporal variables:**
Speech time(t_0),
topic_time(e_1,t_1),
topic_time(e_2,t_2),
topic_time(e_3,t_3),
topic_time(e_4,t_4), once(t_5)

c. **Temporal relators:**
when(t_2,t_3)

a. **Eventualities:**
was standing(e_1),
was looking(e_2)
noticed(e_3),
asked(e_4)
cross(e_5), carry(e_6)

Tense and aspect annotation

b. **Temporal variables:**
Speech time(t_0),
topic_time(e_1,t_1),
topic_time(e_2,t_2),
topic_time(e_3,t_3),
topic_time(e_4,t_4),
once(t_5)

c. **Temporal relators:**
when(t_2,t_3)

Temporal annotation

A timeline

*"**Once** there **was** a scorpion **standing** by a river. The scorpion **was looking** for a way to **cross**, **when** he **noticed** a frog behind him. He **asked** the frog to **carry** him across the river."*

Table 1: Narrative time line

[w_0]	t_5 $t_1 \subset e_1$ $t_2 \subset e_2$ $t_3 \supseteq e_3$	$t_4 \supseteq e_4$		t_0	
[w_1]		e_5			
[w_2]			e_6		

→ Temporal progression →

TimeML

- Broadly accepted standard: TimeML Pustejovsky et al. (2003, 2002) and, more recently, ISO-TimeML(Pustejovsky et al. 2017, 2010)
- Gast et al. (2016) extend TimeML with topic time information allowing
  - Allows for formalization of viewpoint aspect
  - provides a finer granularity of temporal elements in general
- Has been applied in one way or an other to various languages, e.g. French, Italian, Korean, Chinese, Japanese

TimeML cross-linguistically

- The cross-linguistic adaption of TimeML has brought up various challenges
- Korean morphology → stand-off annotation (Im et al. 2009)
- Italian tense and aspect paradigma → annotation of contextual values (Caselli et al. 2011)
- Adaption to morphologically more rich languages, such as Chinese (Pustejovsky et al. 2017)

TimeML – desired improvements

- Several proposals for TimeML have been made, that argue for the independence of syntactic and semantic mark-up of tense categories, e.g.
    - Functional vs. Structural annotation (Gast et al. 2015)
    - Overhaul of ISO-TimeML tense values (Lefeuvre-Halftermeyer et al. 2016)
    - Our own annotation of syntactic and semantic variation of tense and aspect categories
    - **furthermore:** Mapping from (abstract) syntax to semantic representation (Bunt 2010)

Outline

1. Introduction

2. Temporal annotation – A quick overview

3. Comprehensive annotation of the category tense

Semantic construction of meaning

- Sometimes meaning is semantically or pragmatically constructed rather than syntactically marked
- This leads to semantic variation within a language but also distinguishes languages from one another
- **Our goal:** We want to mark up and explore these meaning shifts and test various possibilities of semantic construction

Three different tense and aspects systems

- *Once a scorpion was standing by a river.*

ENGLISH:	Once	a	scorpion	**was**	**standing**	by a river
	Once	a	scorpion	**be.Past**	**stand.Prog**	by a river

URDU:	Ek	tHA	biccHU,	jO	daryA=kE	kinArE	**kHaRA**	**tHA**
	one	Aux.Past	scorpion	Rel	river=Gen	bank.M.3Sg.Obl	**stand**	**Aux.Past**

INDONESIAN:	**Konon**[1]	**ada**	seekor	kalajengking	**berdiri**	di	pinggir	sungai
	Once	**there.is**	a	scorpion	**stand**	on	edge	river

[1] Can also be translated as: 'Supposedly, It is said, that ...'

Variation in the category of English past tense

(1) People kill.**ed** the king
People kill.**past** the king

(2) Tom said that Karen **was** dancing
Tom say.past COMP Karen be.**past** dance.prog

(3) If John **owned** a donkey, he would beat it
If John **kill.past** a donkey he will.past beat it

Annotation of semantic construction

- Analysis of semantic construction processes as exemplified above, comes with a theoretic load
    - Competing analyses available without a (clear) "winner"
        - pragmatic vs. co-indexing account in Sequence-of-tense
        - fake tense as proper past vs. as modal in counterfactuals
        - ....

→ Templatic analysis of secondary meanings

The ParTMA annotation scheme

- Consists of three modules:
- **Syntax**
  - The expressiveness of the ParTMA annotation scheme is directly linked to the richness of the syntactic representation
  - For a concrete implementation we refer to LFG

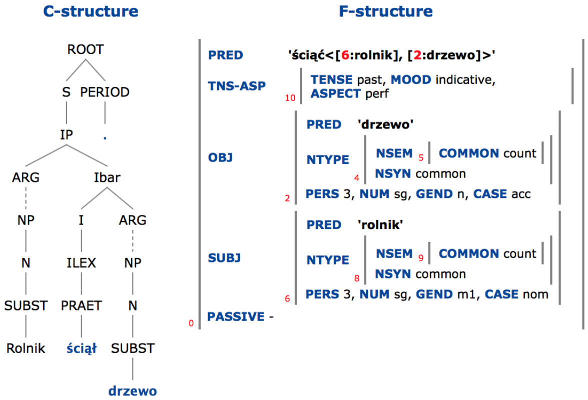
The ParTMA annotation scheme

- Consists of three modules:
- **Syntax**
  - The expressiveness of the ParTMA annotation scheme is directly linked to the richness of the syntactic representation
  - For a concrete implementation we refer to LFG
- **Semantics**
  - A set of cross-linguistically attested formally founded semantic features (represented as logic formulas)

The ParTMA annotation scheme

- Consists of three modules:
- **Syntax**
  - The expressiveness of the ParTMA annotation scheme is directly linked to the richness of the syntactic representation
  - For a concrete implementation we refer to LFG
- **Semantics**
  - A set of cross-linguistically attested formally founded semantic features (represented as logic formulas)
- **Syntax/Semantics interface**
  - A set of language-specific inference rules (or relations) that hold between syntactic and semantic features
  - Follow a set of cross-linguistically universal constraints to restrict variability

Lexical Functional Syntax



F-structure

PRED 'ściąć<[6:rolnik], [2:drzewo]>'

TNS-ASP 10 [**TENSE** past, **MOOD** indicative, **ASPECT** perf]

OBJ 2 [**PRED** 'drzewo'
NTYPE 4 [**NSEM** 5 [**COMMON** count] **NSYN** common]
PERS 3, **NUM** sg, **GEND** n, **CASE** acc]

SUBJ 6 [**PRED** 'rolnik'
NTYPE 8 [**NSEM** 9 [**COMMON** count] **NSYN** common]
PERS 3, **NUM** sg, **GEND** m1, **CASE** nom]

0 **PASSIVE** -

Figure 1: The farmer cut down the tree.

ParTMA semantics

- We propose a semantics with two types of objects:
  - Objects that are anchored to a <world,time> pair(for example situations, time intervals)
  - Abstract objects whose properties are not directly anchored to a <world,time> (for example time spans, events)
- **An example:**
  John climbed the wall for two hours last night.
  - ***last night*** defines a time interval that spans one specific night
  - ***two hours*** defines a time span which corresponds to the run-time of the climbing event
  - ***climb the wall*** describes the concept of climbing a wall

ParTMA semantics

- $[\![$John climbed the wall for two hours$]\!] =$
  $\lambda s.s \prec s_0 \wedge s \leq_p [\![\textit{last night}]\!]^{s_0} \wedge$
  s *exemplifies* $P =$
  $\iota x \exists e[\mathit{climb}(e) \wedge \mathit{ag}(e) = j \wedge \mathit{th}(e) = \mathit{wall}(x) \wedge \tau(e) = [\![\text{2hours}]\!]]$
- $[\![\text{PAST}]\!] = \lambda P.\lambda s.s \prec s_0 \wedge s$ *exemplifies* P
- **Simplification:**
  $[\![\text{PAST}]\!] = \lambda P.\lambda t.t \prec t_0 \wedge P(t)$
  existential closure => $\exists t[t \prec t_0 \wedge P(t)]$

The syntax/semantics interface

Crucial use of inference rules/relations between syntactic and semantic features

- α, β, γ are syntactic constraints in LFG, and ϕ and ψ are semantic features

The syntax/semantics interface

Crucial use of inference rules/relations between syntactic and semantic features

- α, β, γ are syntactic constraints in LFG, and ϕ and ψ are semantic features
- $\rightarrow$ describes the **implication** relation,
  s.t.: $\alpha \rightarrow \phi$ means, that ϕ obligatorily follows from α (morphosyntactically realized semantic features)

The syntax/semantics interface

Crucial use of inference rules/relations between syntactic and semantic features

- α, β, γ are syntactic constraints in LFG, and ϕ and ψ are semantic features
- $\rightarrow$ describes the **implication** relation,
  s.t.: $\alpha \rightarrow \phi$ means, that ϕ obligatorily follows from α (morphosyntactically realized semantic features)
- $\circ$ describes the **compatibility** relation,
  s.t.: $\alpha \circ \phi$ means, that ϕ is optionally available for α (implicatures, non-overtly realized(contextual) semantic features)

An actual example II

(4) Q: Do you know Peter?

(5) jeg møtte Peter på markedet i går
I meet.pst Peter at market yesterday
‘I met Peter at the market yesterday.’ Norwegian

F-Structure:

$$\begin{bmatrix} \text{TNS-ASP} & \begin{bmatrix} \text{TENSE 'past'} \\ \text{MOOD 'indicative'} \end{bmatrix} \end{bmatrix}$$

An actual example II

(6) Q: Do you know Peter?

(7) jeg møtte Peter på markedet i går
I meet.pst Peter at market yesterday
'I met Peter at the market yesterday.' Norwegian

F-Structure:

$$\left[\text{TNS-ASP} \quad \left[\begin{array}{ll}\text{TENSE} & \text{'past'} \\ \text{MOOD} & \text{'indicative'}\end{array}\right]\right]$$

ParTMA Temporal reference:

$$\left[\text{TEMP-REF} \quad \text{'past'} : t \prec t_0\right]$$

An actual example II

(8) Q: Do you know Peter?

(9)

jeg	møtte	Peter	på	markedet	i går
I	meet.pst	Peter	at	market	yesterday

‘I met Peter at the market yesterday.’ Norwegian

F-Structure:

$$\left[\text{TNS-ASP} \quad \begin{bmatrix}\text{TENSE 'past'} \\ \text{MOOD 'indicative'}\end{bmatrix}\right]$$

ParTMA Temporal reference:

$$\left[\text{TEMP-REF} \quad \text{'past'} : t \prec t_0\right]$$

- TENSE past → TEMP-REF 'past' : $t \prec t_0$
- $t \subseteq \mathit{yesterday} \wedge t \prec t_0$

ParTMA inference rules

- α, β, γ are syntactic constraints in LFG, and ϕ and ψ are semantic features (or time intervals, semantic links)

ParTMA inference rules

- α, β, γ are syntactic constraints in LFG, and ϕ and ψ are semantic features (or time intervals, semantic links)
- **Basic rules:**
  - $\alpha \rightarrow \phi$
  - $\phi \rightarrow \psi$

ParTMA inference rules

- α, β, γ are syntactic constraints in LFG, and ϕ and ψ are semantic features (or time intervals, semantic links)
- **Basic rules:**
  - $\alpha \rightarrow \phi$
  - $\phi \rightarrow \psi$
- **Complex rules:**
  - $\alpha \wedge \beta \wedge ... \wedge \gamma \rightarrow \phi$
  - $\alpha \wedge \phi \rightarrow \psi$

ParTMA inference rules

- α, β, γ are syntactic constraints in LFG, and ϕ and ψ are semantic features (or time intervals, semantic links)
- **Basic rules:**
  - $\alpha \to \phi$
  - $\phi \to \psi$
- **Complex rules:**
  - $\alpha \wedge \beta \wedge ... \wedge \gamma \to \phi$
  - $\alpha \wedge \phi \to \psi$
- **Contextual/higher level rules:**
  - $\mathit{ctx} \wedge \alpha ... \wedge \phi \circ \psi$
  - ✗ $\mathit{ctx} \to \phi$

Primary and secondary meaning

- **Primary meaning (tier-1):**
  - The primary meaning is denoted by the most simple rule that includes the respective syntactic exponent as premise and implies a certain meaning. Lexical semantics also belong to tier-1, ideally: $\alpha \rightarrow \phi$

Primary and secondary meaning

- **Primary meaning (tier-1):**
  - The primary meaning is denoted by the most simple rule that includes the respective syntactic exponent as premise and implies a certain meaning. Lexical semantics also belong to tier-1, ideally: $\alpha \rightarrow \phi$
- **Secondary meaning(tier-2):**
  - Meanings that arise from more complex, or contextual/compatibility rules.
    Consumes tier-1 meaning, e.g.
    $\alpha \rightarrow \phi$,
    $\phi \wedge \beta \wedge \gamma \wedge ... \rightarrow \phi'$

Semantic construction – Sequence of tense

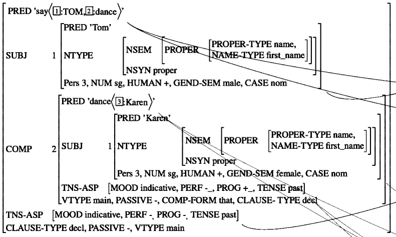
- The Sequence-of-tense phenomenon is a occurrence of tense deletion (or weakening) in embedded contexts:

(10)	Tom	said	that	Karen	**was**	dancing
	Tom	say.past	COMP	Karen	be.**past**	dance.prog

a. Tom said: "Karen is dancing."

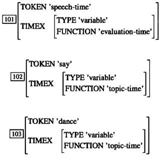
b. Tom said: "Karen was dancing."

Semantic construction – Sequence of tense



```
PRED 'say⟨1:TOM,2:dance⟩'
SUBJ   1  PRED 'Tom'
          NTYPE   NSEM   PROPER   PROPER-TYPE name,
                                  NAME-TYPE first_name
                  NSYN proper
          Pers 3, NUM sg, HUMAN +, GEND-SEM male, CASE nom
COMP   2  PRED 'dance⟨3:Karen⟩'
          SUBJ   1  PRED 'Karen'
                    NTYPE   NSEM   PROPER   PROPER-TYPE name,
                                            NAME-TYPE first_name
                            NSYN proper
                    Pers 3, NUM sg, HUMAN +, GEND-SEM female, CASE nom
          TNS-ASP  [MOOD indicative, PERF -_, PROG +_, TENSE past]
          VTYPE main, PASSIVE -, COMP-FORM that, CLAUSE- TYPE decl
TNS-ASP  [MOOD indicative, PERF -, PROG -, TENSE past]
CLAUSE-TYPE decl, PASSIVE -, VTYPE main
```

Figure 1: F-Structure

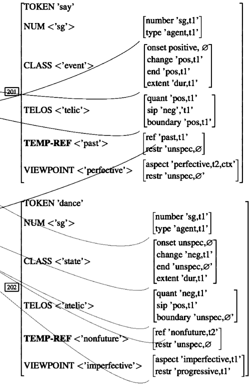


```
101  TOKEN 'speech-time'
     TIMEX   TYPE 'variable'
             FUNCTION 'evaluation-time'

102  TOKEN 'say'
     TIMEX   TYPE 'variable'
             FUNCTION 'topic-time'

103  TOKEN 'dance'
     TIMEX   TYPE 'variable'
             FUNCTION 'topic-time'
```

Figure 2: Relevant temporal variables as TIMEX



```
201  TOKEN 'say'
     NUM <'sg'>                    number 'sg,t1'
                                   type 'agent,t1'
     CLASS <'event'>               onset positive, ∅
                                   change 'pos,t1'
                                   end 'pos,t1'
                                   extent 'dur,t1'
     TELOS <'telic'>               quant 'pos,t1'
                                   sip 'neg','t1'
                                   boundary 'pos,t1'
     TEMP-REF <'past'>             ref 'past,t1'
                                   restr 'unspec,∅
     VIEWPOINT <'perfective'>      aspect 'perfective,t2,ctx'
                                   restr 'unspec,∅'

202  TOKEN 'dance'
     NUM <'sg'>                    number 'sg,t1'
                                   type 'agent,t1'
     CLASS <'state'>               onset unspec,∅
                                   change 'neg,t1'
                                   end 'unspec,∅'
                                   extent 'dur,t1'
     TELOS <'atelic'>              quant 'neg,t1'
                                   sip 'pos,t1'
                                   boundary 'unspec,∅'
     TEMP-REF <'nonfuture'>        ref 'nonfuture,t2'
                                   restr 'unspec,∅
     VIEWPOINT <'imperfective'>    aspect 'imperfective,t1'
                                   restr 'progressive,t1'
```

Figure 3: Annotation of eventuality predicates

Semantic Composition

- $[\![\text{PAST}]\!] = \lambda P.\lambda t.t \prec t_0 \wedge P(t)$
  $[\![\text{Tom said that Q}]\!]\ \lambda t.t \prec t_0 \wedge say(t, tom, Q)$
- $[\![\text{NON-FUT}]\!] =$
  $\{\lambda P.\lambda t'.\lambda t.t' \prec t \wedge P(t), \lambda P.\lambda t'.\lambda t.t' \circ t \wedge P(t)\}$
- $[\![\text{Karen was dancing}]\!] = [\![\text{Q}]\!] = \lambda t.t' \prec t \wedge dance(t', karen)$
  $[\![\text{Q'}]\!] = \lambda t.t' \circ t \wedge dance(t', karen)$
- $[\![\text{Tom said that Karen was dancing}]\!] =$
  $\lambda t.t \prec t_0 \wedge say(t, tom, \exists t'[t' \prec t \wedge dance(t', karen)]),$
  $\lambda t.t \prec t_0 \wedge say(t, tom, \exists t'[t' \circ t \wedge dance(t', karen)])$

Conclusion

- We presented a modular annotation scheme for tense and aspect
  - Allows for syntactic and semantic parallelism
  - captures cross-linguistic variation in the syntax/semantics interface
  - Expressive enough to model formal semantic intuitions

Conclusion

- We presented a modular annotation scheme for tense and aspect
    - Allows for syntactic and semantic parallelism
    - captures cross-linguistic variation in the syntax/semantics interface
    - Expressive enough to model formal semantic intuitions
- **Implementation**
    - Syntactically annotated treebanks for the category of past tense are available on INESS
    - Story-based treebank available offline (to be made public on INESS)
    - Coming soon: implementation of ParTMA annotation (and search) in INESS

References I

Bunt, Harry. 2010. A methodology for designing semantic annotation languages exploring semanticsyntactic iso-morphisms. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong*, Pages 29–46.

Caselli, Tommaso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in italian: the it-timeml experience for the ita-timebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, Pages 143–151. Association for Computational Linguistics.

Dahl, Östen. 1985. *Tense and aspect systems*. Oxford: Blackwell.

Gast, Volker, Lennart Bierkandt, Stephan Druskat, and Christoph Rzymski. 2016. Enriching timebank: Towards a more precise annotation of temporal relations in a text. In *LREC*.

Gast, Volker, Lennart Bierkandt, and Christoph Rzymski. 2015. Creating and retrieving tense and aspect annotations with GraphAnno, a lightweight tool for multi-level annotation. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, Page 23.

References II

Im, Seohyun, Hyunjo You, Hayun Jang, Seungho Nam, and Hyopil Shin. 2009. 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.

Lefeuvre-Halftermeyer, Anaïs, Jean-Yves Antoine, Alain Couillault, Emmanuel Schang, Lotfi Abouda, Agata Savary, Denis Maurel, Iris Eshkol-Taravella, and Delphine Battistelli. 2016. Covering various needs in temporal annotation: a proposal of extension of iso timeml that preserves upward compatibility. In *LREC 2016*.

Pustejovsky, James, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. In *Handbook of Linguistic Annotation*, Pages 21–72. Springer.

Pustejovsky, James, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering* 3:28–34.

Pustejovsky, James, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, Pages 394–397.

References III

Pustejovsky, James, Roser Saurí, Andrea Setzer, Rob Gaizauskas, and Bob Ingria. 2002. TimeML annotation guidelines. *TERQAS Annotation Working Group* 23.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An Open Infrastructure for Advanced Treebanking. In J. Hajič, K. de Smedt, M. Tadić, and A. Branco, editors., *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, Pages 22–29. Istanbul, Turkey.

Sulger, Sebastian, 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.

Thanks for listening