Cross-linguistic annotation of tense and aspect syntax and semantics

Mark-Matthias Zymla

University of 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
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
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
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
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
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
"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."
Introduction
Temporal annotation – A quick overview
Comprehensive annotation of the category tense
References

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(e₁), was looking(e₂)
   noticed(e₃), asked(e₄)
   cross(e₅), carry(e₆)

b. Temporal variables:
   Speech time(t₀),
   topic_time(e₁,t₁),
   topic_time(e₂,t₂),
   topic_time(e₃,t₃),
   topic_time(e₄,t₄), once(t₅)

c. Temporal relators:
   when(t₂,t₃)
Introduction

Temporal annotation – A quick overview

Comprehensive annotation of the category tense

References

a. Eventualities:
   was standing(e₁),
   was looking(e₂)
   noticed(e₃),
   asked(e₄)
   cross(e₅), carry(e₆)

Tense and aspect annotation

b. Temporal variables:
   Speech time(t₀),
   topic_time(e₁,t₁),
   topic_time(e₂,t₂),
   topic_time(e₃,t₃),
   topic_time(e₄,t₄),
   once(t₅)

c. Temporal relators:
   when(t₂,t₃)

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.

Table 1: Narrative time line

<table>
<thead>
<tr>
<th>[w₀]</th>
<th>t₅</th>
<th>t₁ ⊆ e₁</th>
<th>t₂ ⊆ e₂</th>
<th>t₃ ⊇ e₃</th>
<th>t₄ ⊇ e₄</th>
<th>t₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>[w₁]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>t₄ ⊇ e₄</td>
<td>e₅</td>
</tr>
<tr>
<td>[w₂]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>e₆</td>
</tr>
</tbody>
</table>

→ Temporal progression →
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 another to various languages, e.g. French, Italian, Korean, Chinese, Japanese
The cross-linguistic adaption of TimeML has brought up various challenges

- Korean morphology → stand-off annotation (Im et al. 2009)
- Italian tense and aspect paradigm → annotation of contextual values (Caselli et al. 2011)
- Adaption to morphologically more rich languages, such as Chinese (Pustejovsky et al. 2017)
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
one Aux.Past scorpion Rel river=Gen bank.M.3Sg.Obl
kHaRA tHA
stand Aux.Past

**INDONESIAN:** Konon^1^ ada seekor kalajengking berdiri di pinggir
Once there.is a scorpion stand on edge
sungai
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
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
**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 \(<\text{world}, \text{time}\) pair (for example situations, time intervals)
  - Abstract objects whose properties are not directly anchored to a \(<\text{world}, \text{time}\) (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

- $\llbracket$John climbed the wall for two hours$\rrbracket = \lambda s. s \prec s_0 \land s \leq P \llbracket$last night$\rrbracket^{s_0} \land s$ exemplifies $P =$ $\iota x \exists e[climb(e) \land ag(e) = j \land th(e) = wall(x) \land \tau(e) = \llbracket$2hours$\rrbracket]$  
- $\llbracket$PAST$\rrbracket = \lambda P. \lambda s. s \prec s_0 \land s$ exemplifies $P$

**Simplification:**

- $\llbracket$PAST$\rrbracket = \lambda P. \lambda t. t \prec t_0 \land P(t)$
- existential closure $\Rightarrow \exists t[t \prec t_0 \land P(t)]$
The syntax/semantics interface

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

- $\alpha, \beta, \gamma$ are syntactic constraints in LFG, and $\phi$ and $\psi$ are semantic features
The syntax/semantics interface

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

- $\alpha, \beta, \gamma$ are syntactic constraints in LFG, and $\phi$ and $\psi$ are semantic features
- $\rightarrow$ describes the implication relation, s.t.: $\alpha \rightarrow \phi$ means, that $\phi$ obligatorily follows from $\alpha$ (morphosyntactically realized semantic features)
The syntax/semantics interface

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

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

```
[ TNS-ASP
   [ TENSE 'past'
      [ MOOD 'indicative'
   ]
 ]
```
(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:
\[
\begin{bmatrix}
\text{TNS-ASP} & \left[\begin{bmatrix}
\text{MOOD} & \text{'indicative'}
\end{bmatrix}
\right]
\end{bmatrix}
\begin{bmatrix}
\text{TENSE} & \text{'past'}
\end{bmatrix}
\]

ParTMA Temporal reference:
\[
[\text{TEMP-REF} \quad \text{'past'} : t < t_0]
\]
(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:

\[
\begin{align*}
\text{TNS-ASP} & \quad \text{TENSE 'past'} \\
\text{MOOD 'indicative'} &
\end{align*}
\]

ParTMA Temporal reference:

\[
\begin{align*}
\text{TEMP-REF} & \quad 'past' : t \prec t_0 \\
\end{align*}
\]

- TENSE past $\rightarrow$ TEMP-REF 'past' : $t \prec t_0$
- $t \subseteq \text{yesterday} \land t \prec t_0$
ParTMA inference rules

• $\alpha$, $\beta$, $\gamma$ are syntactic constraints in LFG, and $\phi$ and $\psi$ are semantic features (or time intervals, semantic links)
ParTMA inference rules

\( \alpha, \beta, \gamma \) are syntactic constraints in LFG, and \( \phi \) and \( \psi \) are semantic features (or time intervals, semantic links)

Basic rules:

- \( \alpha \rightarrow \phi \)
- \( \phi \rightarrow \psi \)
ParTMA inference rules

- $\alpha$, $\beta$, $\gamma$ are syntactic constraints in LFG, and $\phi$ and $\psi$ are semantic features (or time intervals, semantic links)

**Basic rules:**
- $\alpha \rightarrow \phi$
- $\phi \rightarrow \psi$

**Complex rules:**
- $\alpha \land \beta \land ... \land \gamma \rightarrow \phi$
- $\alpha \land \phi \rightarrow \psi$
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:
- $\alpha \land \beta \land ... \land \gamma \rightarrow \phi$
- $\alpha \land \phi \rightarrow \psi$

Contextual/higher level rules:
- $ctx \land \alpha... \land \phi \circ \psi$
- $\chi ctx \rightarrow \phi$
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: $\alpha \rightarrow \phi$
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.
    $\alpha \rightarrow \phi,$
    $\phi \land \beta \land \gamma \land \ldots \rightarrow \phi'$
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

Figure 1: F-Structure

Figure 2: Relevant temporal variables as TIMEX

Figure 3: Annotation of eventuality predicates
Semantic Composition

- $\llbracket \text{PAST} \rrbracket = \lambda P. \lambda t. t \prec t_0 \land P(t)$
  $\llbracket \text{Tom said that } Q \rrbracket = \lambda t. t \prec t_0 \land \text{say}(t, \text{tom}, Q)$

- $\llbracket \text{NON-FUT} \rrbracket = \{ \lambda P. \lambda t'. \lambda t.t' \prec t \land P(t), \lambda P. \lambda t'.\lambda t.t' \circ t \land P(t) \}$

- $\llbracket \text{Karen was dancing} \rrbracket = \llbracket \text{Q} \rrbracket = \lambda t.t' \prec t \land \text{dance}(t', \text{karen})$
  $\llbracket \text{Q'} \rrbracket = \lambda t.t' \circ t \land \text{dance}(t', \text{karen})$

- $\llbracket \text{Tom said that Karen was dancing} \rrbracket =$
  $\lambda t.t \prec t_0 \land \text{say}(t, \text{tom}, \exists t'[t' \prec t \land \text{dance}(t', \text{karen})]),$
  $\lambda t.t \prec t_0 \land \text{say}(t, \text{tom}, \exists t'[t' \circ t \land \text{dance}(t', \text{karen})])$
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


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*.


Thanks for listening