Annotating Information Structure. The Case of Topic

Theme Session Beyond Semantics

Philippa Cook & Felix Bildhauer
Freie Universität Berlin
SFB 632 Informationsstruktur

24. Feb 2011
Why bother?

Insights in IS may lead to promising (re-)analyses of linguistic phenomena on the basis of information structure

- Using information structural constraints in describing phenomena previously accounted for in terms of syntax, e.g.
  - NP-PP split (De Kuthy 2002)
  - Quantifier scope (Cook and Payne 2006)
  - Extraction islands (Ambridge and Goldberg 2008)
  - Extraction from infinitival complements (Cook and Oersnes 2010)
  - German multiple fronting (Bildhauer and Cook 2010)
Why bother? (continued)

IS-annotated corpora are particularly valuable: test linguistic analyses that are based on notions such as “topic”, “focus” and “givenness”

- Definitions vary across different currents of research
- Definitions can cause considerable problems when applied to naturally occurring data

This talk: Report on a topic annotation task, discuss difficult cases
Aboutness Topic

- “Topic”: something to which the speaker draws the hearer’s attention, and about which s/he then provides further information (von der Gabelentz 1869)
- “Topic”: entity under which a proposition is stored in a structured “context set” (Reinhart 1981)

Building on this, Krifka (2007):

*The topic constituent identifies the entity or set of entities under which the information expressed in the comment constituent should be stored in the Common Ground content.*
Orthogonal IS-dimensions

<table>
<thead>
<tr>
<th>BACKGROUND</th>
<th>FOCUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOPIC</td>
<td>COMMENT</td>
</tr>
</tbody>
</table>

*text text text text text text text text*
Orthogonal IS-dimensions

<table>
<thead>
<tr>
<th>BACKGROUND</th>
<th>FOCUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOPIC</td>
<td>COMMENT</td>
</tr>
</tbody>
</table>

*text text text text text text text text*

Topic-Comment has sometimes been mixed up with Background-Focus and Given-New

- In practice: topics are prototypically given, foci are canonically new
- A merging of the dimensions is problematic: there are cases which deviate from the canonical alignment:
  1. There are topics which contain a focus
  2. There are new (i.e. non-given) topics
  3. The focus may only be a sub-part of the comment
  4. There are non-new foci
# Aboutness Topic: More Assumptions

<table>
<thead>
<tr>
<th>FOCUS</th>
<th>BACKGROUND</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOPIC</td>
<td>COMMENT</td>
</tr>
</tbody>
</table>

- Orthogonal dimensions of IS: does not rule out complete overlap of Aboutness Topic and focus
Aboutness Topic: More Assumptions

- Orthogonal dimensions of IS: does not rule out complete overlap of Aboutness Topic and focus
- Distinct from Vallduvi’s (1992) “Link”, which is defined positionally as the sentence-initial topic
- Further assumption: a sentence has at most one Aboutness Topic.
- Sentences which lack a Topic-Comment articulation are classed as thletic (cf. Krifka 2007, 43)
An NP X is the Aboutness Topic of a sentence S containing X if
a. S would be a natural continuation to the announcement
   *Let me tell you something about X*

b. S would be a good answer to the question
   *What about X?*

c. S could be naturally transformed into the sentence
   *Concerning X, S’*
   where S’ differs from S only insofar as X has been replaced by a suitable pronoun.
A corpus study

Purpose: Test a hypothesis about preferential topic realization (results were needed to substantiate a pragmatic explanation for a syntactic phenomenon)

- Four verbs selected (on the basis of a prior study)
- Only one argument frame per verb considered

<table>
<thead>
<tr>
<th>Argument frame</th>
<th>Verb</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP&lt;sub&gt;mit&lt;/sub&gt; XP&lt;sub&gt;LOC&lt;/sub&gt;</td>
<td>geraten</td>
<td>Er gerät [mit seiner Hose] [in die Kette].</td>
</tr>
<tr>
<td></td>
<td>‘to get (caught in)’</td>
<td>‘He got his trousers caught in the chain.’</td>
</tr>
<tr>
<td></td>
<td>reagieren</td>
<td>Sie reagiert [überrascht] [auf den Vorschlag].</td>
</tr>
<tr>
<td></td>
<td>‘to react’</td>
<td>‘She reacted surprisingly to this suggestion.’</td>
</tr>
<tr>
<td></td>
<td>profitieren</td>
<td>Sie profitieren [von den Steuersenkungen].</td>
</tr>
<tr>
<td></td>
<td>‘to profit’</td>
<td>‘They profit from the tax reductions.’</td>
</tr>
<tr>
<td></td>
<td>herrschen</td>
<td>Dort herrscht Ruhe.</td>
</tr>
<tr>
<td></td>
<td>‘to reign’</td>
<td>‘There reigns peace.’</td>
</tr>
</tbody>
</table>
Corpus study II

- Two independent coders
- Annotated 135 to 167 sentence tokens per verb (total 578)
- Task: Decide whether or not a given sentence has an aboutness topic (constituent)
  - if yes: select from among the NPs (and deictic expressions) in the sentence
  - if no: mark the sentence as non topic-comment
Inter-rater agreement (Cohen’s \( \kappa \))

<table>
<thead>
<tr>
<th>Verb</th>
<th>N</th>
<th>All categories matches</th>
<th>( \kappa )</th>
<th>Topic-Comment? matches</th>
<th>( \kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>profitieren</td>
<td>135</td>
<td>109</td>
<td>.57</td>
<td>129</td>
<td>-.01</td>
</tr>
<tr>
<td>herrschen</td>
<td>138</td>
<td>102</td>
<td>.55</td>
<td>106</td>
<td>.51</td>
</tr>
<tr>
<td>geraten</td>
<td>147</td>
<td>99</td>
<td>.33</td>
<td>100</td>
<td>.33</td>
</tr>
<tr>
<td>reagieren</td>
<td>167</td>
<td>106</td>
<td>.22</td>
<td>148</td>
<td>.25</td>
</tr>
</tbody>
</table>

- Agreement for all categories: highly variable across the four verbs
- Agreement for TC/non-TC: also variable, not better than chance for *profitieren*
- Disappointing result, given that the same guidelines were used
Controversial cases

Most of the controversial cases fall into two different types:

1. The annotators’ different interpretation of “Aboutness”
   ▶ Deciding “what the sentence is about” when there is more than one candidate expression
   ▶ Guidelines often do not yield a clear answer, or their application is not straightforward.
   ▶ Overlap of focus and a potential topic expression
Controversial cases

Most of the controversial cases fall into two different types:

1. The annotators’ different interpretation of “Aboutness”
   - Deciding “what the sentence is about” when there is more than one candidate expression
   - Guidelines often do not yield a clear answer, or their application is not straightforward.
   - Overlap of focus and a potential topic expression

2. Deciding whether there is an AT at all
   - Sentence construed as being “about” an entity, or description of situation?
   - Again: overlap of focus and a potential topic expression
More than one candidate expression: Example


‘I learned more in the area of public relations work in particular. I noticed what sort of reaction was caused by which actions and how to communicate certain events correctly.’


‘I will clearly be able to profit from this experience at the new place.’
More than one candidate expression: Commonalities

(7) b. Von \[\text{dieser Erfahrung}\]_{top?} kann \[\text{ich}\]_{top?} am neuen Ort selbstverständlich profitieren

In (7b) and many other cases:

- One candidate is a non-subject that is realized in prominent (initial) position (but: topics need not occupy that position in German)
- One candidate is the subject, and it is the topic of preceding utterances (“topic chain”; see Givón 1983).
- In the specific context, either of them can be viewed as AT: no obvious way to make an objective decision
Topic-Comment vs. Thetic

Different assumptions about possible interaction of Topic-Comment and Focus-Background

- Proposal: topic (theme) and the comment (rHEME) section of an utterance have their own focus-background structure (e.g. Krifka 1992; Steedman 2000)
- Overlap of topic and focus: only cases of contrastive topic discussed
- Is a complete overlap of (new-information) focus and AT a problem?

(9) Q: Who ate the apple?

A: Kim ate the apple.

[ ] _foc_ [ ] _background_

[ ] _top_ [ ] _comment_
Topic-Comment vs. Thetic: Example

(10) a. „Geh aus mein Herz, und suche Freud“, spielte der Posaunenchor zum Auftakt des Gemeindefestes der evangelischen Kirchengemeinde Edingen.

‘As a prelude to the Edingen parish’s celebration, the trombone choir performed Geh aus mein Herz, und suche Freud.’

b. Erwartungsfrohe Vorfreude herrschte auch bei [den zahlreichen expectant anticipation reigned also among the numerous Gottesdienstbesuchern, die auf das Freigelände am Amselweg service.visitors who on the outdoor.area at the Amselweg gekommen waren]top?
come had

‘There was an air of expectant anticipation amongst the numerous visitors to the service who had come to the outdoor area near Amselweg as well.’
Annotating “Aboutness Topic”

- Controversial cases

### Topic-Comment vs. Thetic: Commonalities

(10) b. Erwartungsfrohe Vorfreude herrschte auch bei [den zahlreichen Gottesdienstbesuchern, die auf das Freigelände am Amselweg gekommen waren]

**Topic = PP?**

- Subject NP not suitable as Aboutness Topic
- PP is focussed: (10b) is a suitable answer to the question „Wo herrschte noch Vorfreude?“
Topic-Comment vs. Thetic: Commonalities

(10) b. Erwartungsfrohe Vorfreude herrschte auch bei [den zahlreichen Gottesdienstbesuchern, die auf das Freigelände am Amselweg gekommen waren]

Topic = PP?

- Subject NP not suitable as Aboutness Topic
- PP is focussed: (10b) is a suitable answer to the question „Wo herrschte noch Vorfreude?“

Topicless sentence?

- Sentences which lack an AT: thetic (e.g. Lambrecht 1994; Krifka 2007)
- Thetic sentence: all-focus (e.g. Lambrecht 1994; Rosengren 1997)
- Thetic sentence in German: bear a single accent on the subject (e.g. Krifka 1984; Sasse 1987)
Interim summary

1. Choice between various candidates
   - Unclear how decision can be operationalised

2. Topic-Comment vs. Thetic distinction
   - Interaction/overlap of topic and focus is not discussed explicitly in the literature (with some exception)
   - Annotators might conclude that sentence S cannot have a topic
   - Analysing these sentences as “thetic” does not conform to cases discussed in the literature
Towards a resolution of the second type of controversial case

Definition of theticity which allows for a type of thetic utterance that introduces or presents an entity (rather than a situation or event).

- Fits in with the approach in Lambrecht (1994, 2000) 
  (*presentational* vs. *event-reporting*)

- Fits in with Sasse’s (1987) distinction 
  (*entity-central* vs. *event central*)
Subclassification of “thetic” structures

- All thetic sentences are ‘all-comment’
- Not all thetic sentences are ‘all-focus’
- Capture difference between entity-central and event-central by assigning different focus structures

- event-central: focus spreads across the whole sentence
- entity-central: only introduced referent is focussed
Illustration

The two options for example (10b) can be sketched thus (assuming that the main stress is on the subject-NP):

(10) i. **Topic-comment structure**  
[Erwartungsfrohe Vorfreude herrschte auch]_{comment} [bei den zahlreichen Gottesdienstbesuchern, die auf das Freigelände am Amselweg gekommen waren]_{foc}^{top}

ii. **Entity central thematic**  
[Erwartungsfrohe Vorfreude herrschte auch [bei den zahlreichen Gottesdienstbesuchern, die auf das Freigelände am Amselweg gekommen waren]_{foc}^{comment}
Conclusion

Practical

- Data of the type exemplified in (10) is frequent: the problem should be clarified beforehand in future annotation tasks
- Annotators should state (in rough terms) the accent pattern they assume for a given sentence
Conclusion

Practical

- Data of the type exemplified in (10) is frequent: the problem should be clarified beforehand in future annotation tasks
- Annotators should state (in rough terms) the accent pattern they assume for a given sentence

Theoretical

- Clarify possible interaction between orthogonal IS-dimensions (overlap of Aboutness Topic and Focus)
- Clarify the correct delineation of thetic utterances and the role that presentation may play there
Conclusion

Practical

- Data of the type exemplified in (10) is frequent: the problem should be clarified beforehand in future annotation tasks
- Annotators should state (in rough terms) the accent pattern they assume for a given sentence

Theoretical

- Clarify possible interaction between orthogonal IS-dimensions (overlap of Aboutness Topic and Focus)
- Clarify the correct delineation of thetic utterances and the role that presentation may play there

Thank you.
Conclusion

Practical

- Data of the type exemplified in (10) is frequent: the problem should be clarified beforehand in future annotation tasks
- Annotators should state (in rough terms) the accent pattern they assume for a given sentence

Theoretical

- Clarify possible interaction between orthogonal IS-dimensions (overlap of Aboutness Topic and Focus)
- Clarify the correct delineation of thetic utterances and the role that presentation may play there

Thank you.


Götze, Michael, Weskott, Thomas, Endriss, Cornelia, Fiedler, Ines, Hinterwimmer, Stefan, Petrova, Svetlana, Schwarz, Anne, Skopeteas, Stavros and Stoel, Ruben. 2007. Information Structure. In


Annotating “Aboutness Topic”

Conclusion

Publications.


