Principles of information-structure and discourse-structure analysis

Lisa Brunetti, Kordula De Kuthy, Arndt Riester

Motivation

Motivation

- Information structure has been intensively investigated in the theoretical linguistic literature.

  - Attention has shifted from the analysis of sentences preceded by short constructed contexts to that of utterances in *real discourse contexts*.

- How can we link the information structure of an utterance to the overall structure of the discourse?

In this talk:

- We introduce our methodology for an analysis of *authentic language data* in terms of discourse and information structure.

Motivation: Focus

Focus typically characterized as *the answer to an (explicit or implicit) question*:

- von Heusinger (1999) (discussing an observation by Paul (1880)): A sentence "can have different intonation centers on different constituents corresponding to an explicit or implicit question."

- Halliday (1967): "A specific question is derivable from any information unit except one with unmarked focus."

- Focus is the instantiation of the missing variable in an open proposition (Lambrecht 1994; Prince 1992)

- Jasinskaja et al. (2004): "One of the contexts most universally acknowledged as a diagnostic for focus are question-answer pairs."

Motivation: QUDs

Questions under Discussion (QUD)
(Roberts 1996, 2012; Büring 2003; Van Kuppevelt 1995)

- For any assertion in a discourse, there is an implicit QUD that determines which parts of the assertion are focused or backgrounded.

- Büring (2012): "A plausible analysis assumes that focus (...) is licensed by question/answer congruence (QAC), where 'question' is understood as the question under discussion, QUD, in a discourse model."

Our Analysis Approach (Riester, Brunetti and De Kuthy, submitted)

Our Analysis Approach (Riester, Brunetti and De Kuthy, submitted)

Our assumptions about discourse:

- Discourse is not linear but hierarchically organized in the form of a discourse tree.

- Text is built from *elementary discourse units* (cf. Hobbs 1985; Polanyi 1988; Mann and Thompson 1987; Asher and Lascarides 2003).

- Discourse structure is based on QUDs (or topics) (cf. Roberts 1996, 2012; Büring 2003; Van Kuppevelt 1995; Beaver and Clark 2009).

  - QUD stacks contain increasingly specific questions.

  - NB: question nodes of discourse trees can be **anaphorically dependent** on other nodes, and

  - only in specific cases they are connected by entailment relations.

Information Structure in Authentic Data

Information Structure in Authentic Data

- Complementing the information structure research based on *constructed examples*,

- some authors have tackled the analysis of *authentic data in corpora* (Dipper et al. 2007; Ritz et al. 2008; Calhoun et al. 2010).

Information Structure in Authentic Data

Focus Annotation in Switchboard (Calhoun et al. 2010)

- Definition:
  - *Kontrast* encodes whether a word has a salient alternative in the context (*kontrast*) or not (*background*).
- Annotation guidelines: annotators identified words which were
  - "salient with an implication that this salience is in comparison or contrast to other related words or NPs explicitly or implicitly evoked in the context"

Focus Annotation using LISA (Dipper et al. 2007)

- Definition:
  - "New-information focus (nf) is that part of the utterance providing the new and missing information which serves to develop the discourse."
- Annotation guidelines:
  - "Which part of the utterance reveals the new and most important information in discourse? Try to identify the domain by asking implicit questions!"

Information Structure in Authentic Data

- While Ritz et al. (2008) and Calhoun et al. (2010) recognized that annotating focus relies on determining a previous question,

- their analysis procedures do not spell out
  - when and where QUDs can be inserted in the discourse, and
  - what their exact relation is to the information structural categories.

Our Analysis Approach

Our Analysis Approach

To make the information structure explicit in authentic data, we

- reconstruct the QUDs of a text on the basis of **explicit constraints**

- link discourse and information structure by specific **linking rules**

- transform natural discourse into a **compact tree representation**:
  - whose non-terminal elements are questions and represent the current QUD at that position
  - whose terminal elements are the assertions contained in the text, and represent an answer to the specific QUD

$$Q_0$$

$$A_{0'} \quad Q_1 \quad A_{0''} \quad A_{0'''}$$

$$Q_{1.1} \quad Q_{1.2}$$

$$A_{1.1} \quad A_{1.2}$$
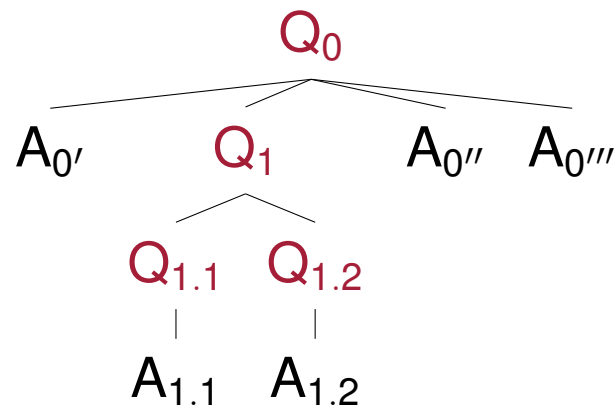
Building Discourse Trees & Formulating QUDs: 5 steps

1. Understanding and preparing the text

2. Formulating QUDs

3. Linking QUDs and Information Structure

4. QUDs of parallel structures

5. Identifying non-at-issue material

Understanding and Preparing the Text

- Read the text carefully and make sure you understand its content.

- Segment the text into separate assertions, namely:
    - sentences at sentence-level conjunctions;
    - coordinating phrases, even below the sentence level.

- Each separate assertion is marked by an *A*.

- Example from Snowden Interview (ARD TV, Jan 2014):

    (1) A: There was an article that came out in an online outlet called Buzz
             Feed

         A: where they interviewed officials from the Pentagon,

         A: from the National Security Agency,

         A: and they gave them anonymity to be able to say what they want

         A: and what they told the reporter was that they wanted to murder me.

QUD Principles I

Q-A-CONGRUENCE

QUDs must be answerable by the assertion(s) that they immediately dominate.

A QUD can in principle target any constituent of the assertion:

(2) Q: *What happened?*
     Q: *What about you?*
     Q: *When were you working for the NSA?*

(3) Q: *Who bought a bicycle?*   ⚡Q-A-CONGRUENCE

(4) A: **You were working until last summer for the NSA.**

If more context is introduced, it becomes clear that the questions in (2) are not all equally good.

QUD Principles II

Q-GIVENNESS

Implicit QUDs can only consist of given (or highly salient) material.

derived from Schwarzschild (1999)

(5) A_1: Edward Snowden is in the meantime a household name for the whistleblower in the age of the internet.

  Q: *What happened?*
  Q: *What about you?*
  Q: #*When were you working for the NSA?*   ⚡Q-GIVENNESS

  A_2: You were working until last summer for the NSA.

QUD Principles III

Maximize Q-Anaphoricity

Implicit QUDs should contain as much given material as possible.

derived from Schwarzschild (1999); Büring (2008)

(6) A$_1$: Edward Snowden is in the meantime a household name for the whistleblower in the age of the internet.

    Q: *#What happened?*   ⚡Max-Q-Anaphorizity

    Q: *What about you?*

    Q: *# When were you working for the NSA?*   ⚡Q-Givenness

    A$_2$: You were working until last summer for the NSA.

→ ensures discourse coherence

Linking QUDs and Information Structure

How can the information structure be derived from the QUDs?

- **Focus (F)**: that part of an assertion that answers the current QUD (obligatory)

- **Background (BG):** The non-focal part of an assertion which is already mentioned in the current QUD

- **Focus domain (~)**: combination of focus and background

- **Topic (T):** referential entity in the background

(7) Q_2: {*What about you?*}
     A_2: [[You$_T$]$_{BG}$ [were working until last summer for the NSA.]$_F$]~

Linking QUDs and Information Structure (Ex.)
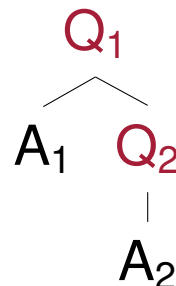
Compact tree representation:

- A question Q_n that depends on an assertion A_{n-1} occurs as the right sister of that assertion.

- Question-answer congruence is indicated by indices.

(8) Q_1: . . .

> A_1: Edward Snowden is in the meantime a household name for the whistleblower in the age of the internet.
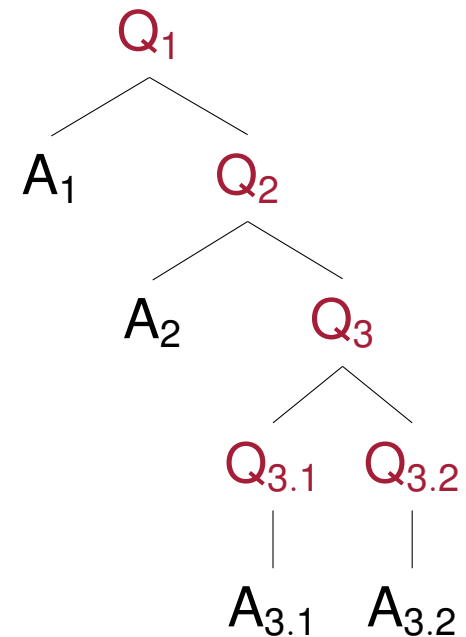
> Q_2: {*What about you?*}
> > A_2: [[You$_T$]$_{BG}$ [were working until last summer for the NSA.]$_F$]~

$$Q_1$$
$$A_1 \quad Q_2$$
$$A_2$$

Discourse tree with QUDs

Q_1: {... ?}
> A_1: ...
> Q_2: {... ?}
>   > A_2: ...
>   > Q_3: {... ?}
>   >   > $Q_{3.1}$: {... ?}
>   >   >   > $A_{3.1}$: ...
>   >   > $Q_{3.2}$: {... ?}
>   >   >   > $A_{3.2}$: ...

Textual analysis

Corresponding tree

Parallelism Constraint

PARALLELISM

The background of a QUD with two or more parallel answers consists of the (semantically) common material of the answers.

(9) > Q_{15}: *{Whom can you wire tap?}*

> > $A_{15'}$: [[You$_T$ can wire tap]$_{BG}$ [the president of the United States]$_F$]~

> > $A_{15''}$: [[you$_T$ can wire tap]$_{BG}$ [a Federal Judge]$_F$]~

- several parallel answers correspond to one QUD
- the constant material in the parallel answers reflects the background of the QUD
- the alternating parts of the assertions correspond to the wh-word in the QUD.

Parallelism

- PARALLELISM can override Q-GIVENNESS,
- since parallel, backgrounded material need not be salient already (narrowing down the focus).

(10) When you are on the inside and you go into work everyday and you sit down at the desk and the you realize the power you have.

Q$_{15}$: *{What power do you have?}* (by Q-GIVENNESS, MAX-Q-ANAPHORICITY)

Q$_{15}$: *{Whom can you wire tap?}* (by PARALLELISM)

> A$_{15'}$: [[You$_T$ can wire tap]$_{BG}$ [the president of the United States]$_F$]~

> A$_{15''}$: [[you$_T$ can wire tap]$_{BG}$ [a Federal Judge]$_F$]~

$$Q_{15}$$

$$A_{15'} \quad A_{15''}$$
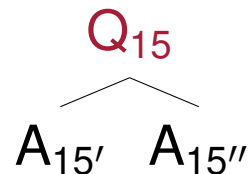
Complex Parallelism: Contrastive Topics

- Two or more assertions contrasted at two different positions:
    - QUD is not answered directly
    - broken down into partial answers about smaller parts,
      cf. Büring (2003)
- QUD structure consists of a super and several sub-questions:
    - Any answer to the sub-questions is also a (partial) answer to the super-question.
    - Super-question and sub-questions stand in an *entailment relation*.

(11) A_0: In many countries, as in America too the agencies like the NSA are not allowed to spy within their own borders on their own people.
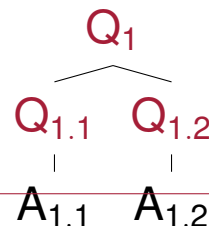
Q_1: {*Who can spy on whom?*}

> $Q_{1.1}$: {*Who can the Brits spy on?*}

> > $A_{1.1}$ So[[the Brits]$_{CT}$, for example, [they]$_T$ can spy on [everybody but the Brits]$_F$]~

> $Q_{1.2}$: {*Who can the NSA spy on?*}

> > $A_{1.2}$ but [[the NSA]$_{CT}$ can conduct surveillance [in England]$_F$]~

$$Q_1$$
$$Q_{1.1} \quad Q_{1.2}$$
$$A_{1.1} \quad A_{1.2}$$

Identifying non-at-issue material

NON-AT-ISSUE

An expression X whose denotation is discourse-new and which is contained in an assertion A is non-at-issue with respect to the current QUD Q

- iff X is optional with respect to Q,
- where optional means that under deletion of X, A is still an answer to Q.

- Typical non-at-issue content (Potts 2005; Simons 2007):
    - supplemental expressions (appositives, non-restrictive modifiers, parentheticals)
    - evidentials (*know/think that*, . . .)
    - expressives

Identifying non-at-issue material: Some examples

Evidential and apposition

(12)   Q$_5$:   {*What was the decisive moment for becoming a whistleblower?*}

     > A$_5$:   [I would say]$_{nai}$ [sort of the breaking point [is seeing the Director of National Intelligence, [James Clapper,]$_{nai}$ directly lie under oath to Congress.]$_F$]~

Factual Conditional

(13)   A$_{15}$: You can wire-tap a Federal judge.

       Q$_{16}$: {*What about it?*}

     > A$_{16}$: and [if you do it carefully]$_{nai}$ [[no one will ever know.]$_F$]~

Sentential adjunct

(14)   Q$_{20}$:   {*What did the political class do?*}

     > A$_{20}$: [Instead of circling around the public and protecting their rights]$_{nai}$ [[the political class]$_T$ [circled around the security state]$_F$]~

An example annotation

Q1: {*What was Snowden's role in the debate?*}

> A1': I: [[You]$_T$ [started]$_F$ [this debate]$_T$,]$_~$

> A1": [[Edward Snowden]$_T$ is [in the meantime]$_{nai}$ [a household name for the whistleblower in the age of the internet]$_F$.]$_~$

Q2: {*What did Snowden do until last summer?*}

> A2': [[You]$_T$ [were working]$_F$ [until last summer]$_T$ [for the NSA]$_F$]$_~$

> A2": and [[during this time]$_T$ [you]$_T$ [secretly]$_{nai}$ [collected thousands of confidential documents]$_F$]$_~$.

> Q3: What was the decisive moment {?}

> Q4: or was there a long period of time or something happening, {?}

> Q5: why did you do this?

> > Q5.1: {*What was the decisive moment for collecting thousands of confidential documents?*}

> > > A5.1: [I would say]$_{nai}$ [sort of [the breaking point]$_T$ is [seeing the Director of National Intelligence, [James Clapper,]$_{nai}$ directly lie under oath to Congress]$_F$]$_~$.

An example annotation

\> \> \> Q6: {*What about the intelligence community after this?*}

\> \> \> \> A6: [[There's no saving]*F* [an intelligence community that believes it can lie to the public and the legislators]*T*]~

\> \> \> \> Q7: {*What should the legislators be able to do with the intelligence community?*}

\> \> \> \> \> A7': [[who]~need to be able to [trust]*F* [it]*T*]~

\> \> \> \> \> A7": and [[regulate]*F* [its]*T* [actions]*F*].~

\> \> \> Q8: {*How did the experience effect Snowden?*}

\> \> \> \> A8: [[Seeing that]*T* really meant for [me]*T* [there was no going back]*F* .]~

\> \> \> Q5.1.1: {*The decisive moment for collecting thousands of confidential documents was the realization that WHO was going to do this?*}

\> \> \> \> A5.1.1: Beyond that, [[it]*T* was the creeping realisation that [no one else]*F* was going to do [this]*T* .]~

\> A5: [[The public had a right to know]*F* about [these programs]*T* .]~

An example annotation

> Q9: {*The public had a right to know that which the government is doing in WHAT relation to the public?*}

> > A9': [[The public]$_T$ had a right to know that which the government is doing [in]$_F$ [its]$_T$ [name]F,]$_\sim$

> > A9": and [that which the government is doing [against]$_F$ the public],$_\sim$

> > Q10: {*What permission did we have to do what with which of these things?*}
> > > Q10.1: {*What permission did we have to do what with neither of these things?*}
> > > > A10.1': but [[neither]$_{CT}$ of [these things]$_T$ [we]$_T$ were allowed [to discuss]F]$_\sim$,
> > > > A10.1": [[we]$_T$ were allowed [to know]$_F$]$_\sim$,

> > > Q11: {*Who was prohibited of doing what with these programs?*}
> > > Q11.1: {*What were the elected representatives prohibited of doing with these programs?*}
> > > > A11.1': even [[the wider body of our elected representatives]$_{CT}$ were prohibited from [knowing]$_F$]$_\sim$
> > > > A11.1": or [[discussing]$_F$ [these programs]$_T$]$_\sim$

> > > > Q12: {*What about these prohibitions?*}
> > > > > A12: and [[that's]$_T$ [a dangerous thing]$_F$]$_\sim$.

An example QUD tree
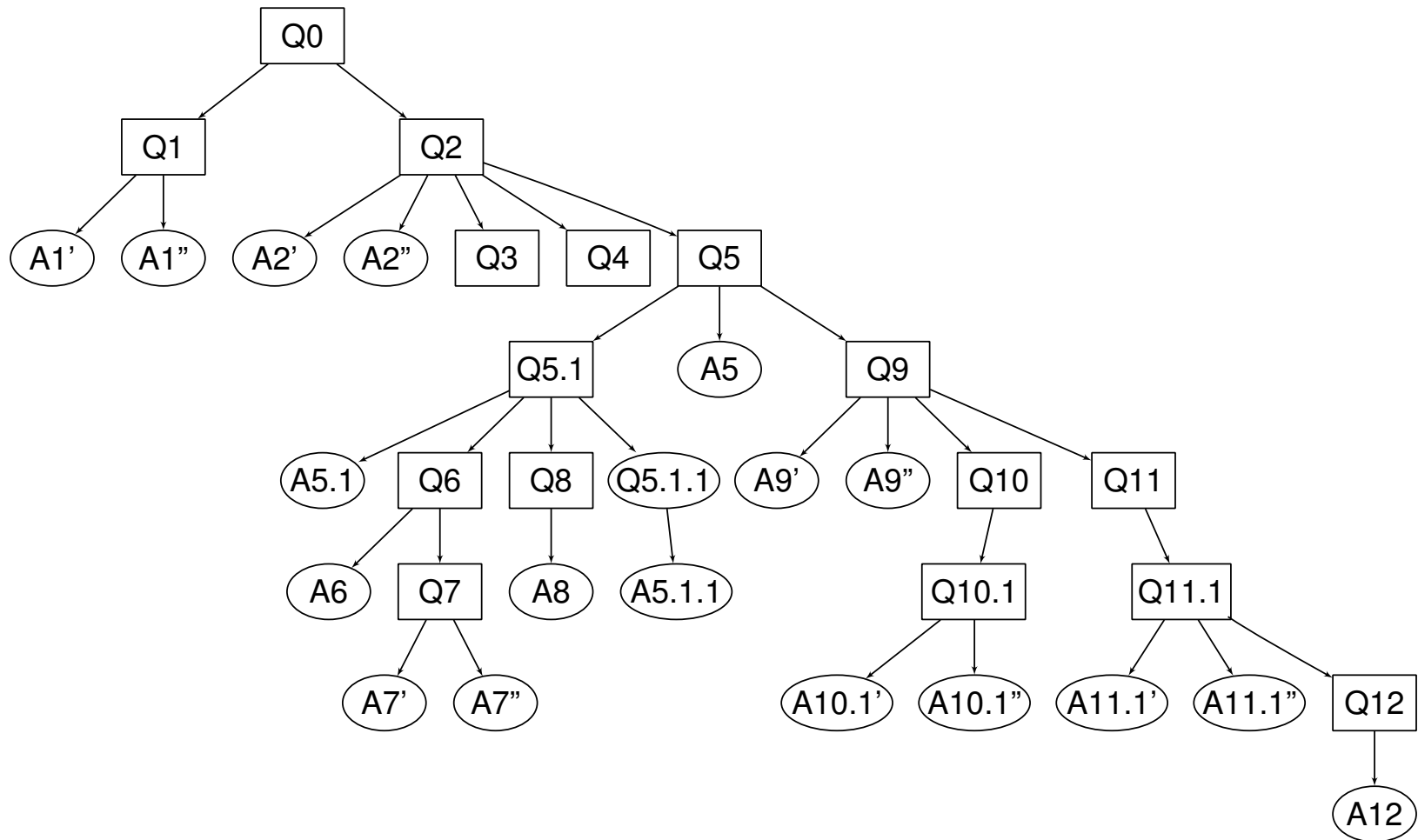
Conclusion

- We presented a combined analysis of discourse and information structure in naturally occurring data,
    - using QUDs as a hinge between the two.

- The formulation of QUDs is guided by pragmatic principles
    - building QUDs from the *given* (and only given) material
    - requiring QUDs to be addressed by the following utterance

- Making the QUDs explicit helps to
    - analyze the information structure of (complex) sentences
    - identify at-issue and non-at-issue material
    - make the discourse structure of a text transparent

- English, German, and French annotations support the applicability of the approach to naturally occuring data.

- Guidelines: `http://www.ims.uni-stuttgart.de/institut/mitarbeiter/arndt/`

Questions?

Contact:

Lisa Brunetti

`lisa.brunetti@linguist.univ-paris-diderot.fr`

Kordula De Kuthy

`kdk@sfs.uni-tuebingen.de`

Arndt Riester

`arndt.riester@ims.uni-stuttgart.de`

References

Asher, N. and Lascarides, A. (2003). *Logics of conversation: Studies in natural language processing*. Cambridge University Press, Cambridge.

Beaver, D. I. and Clark, B. Z. (2009). *Sense and sensitivity: How focus determines meaning*, volume 12. John Wiley & Sons.

Büring, D. (2003). On d-trees, beans, and b-accents. *Linguistics and Philosophy*, 26(5):511–545.

Büring, D. (2008). What's new (and what's given) in the theory of focus? In Berson, S., Bratkievich, A., Bruhn, D., Amy Campbell, R. E., Giovine, A., Newbold, L., Marilola Perez, M. P.-B., and Rhomieux, R., editors, *Proceedings of the Thirty-Fourth Annual Meeting of the Berkeley Linguistics Society, February 8–10 2008*, pages 403–424, Berkeley, CA. Berkeley Linguistics Society.

Büring, D. (2012). Focus and intonation. In Russell, G. and Fara, D. G., editors, *Routledge Companion to the Philosophy of Language*, Routledge Philosophy Companions, pages 103–115. Routlegde, London.

Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44:387–419.

Dipper, S., Götze, M., and Skopeteas, S., editors (2007). *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax,*

Semantics and Information Structure, volume 7 of *Interdisciplinary Studies on Information Structure*. Universitätsverlag Potsdam, Potsdam, Germany.

Halliday, M. (1967). Notes on transitivity and theme in English. part 1 and 2. *Journal of Linguistics*, 3:37–81, 199–244.

Hobbs, J. R. (1985). *On the coherence and structure of discourse*. CSLI.

Jasinskaja, E., Mayer, J., and Schlangen, D. (2004). Discourse structure and information structure: Interfaces and prosodic realization. In *Interdisciplinary Studies on Information Structure*, volume 1, pages 151 – 206. Universität Potsdam.

Lambrecht, K. (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge Studies in Linguistics. Cambridge University Press.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: A framework for the analysis of texts. Technical report, University of Southern California Marina del Rey, Information Science Institute.

Paul, H. (1880). *Prinzipien der Sprachgeschichte*. 8th edition, 1970. Tübingen: Max Niemeyer Verlag.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of pragmatics*, 12(5-6):601–638.

Potts, C. (2005). *The logic of conventional implicatures*. Oxford University Press, Oxford.

Prince, E. F. (1992). The ZPG letter: Subjects, definiteness, and information-status. In Mann, W. C. and Thompson, S. A., editors, *Discourse Description: Diverse linguistic analyses of a fund-raising text*, pages 295–325. John Benjamins, Philadelphia/Amsterdam.

Riester, A., Brunetti, L., and De Kuthy, K. (under review). Annotation guidelines for questions under discussion and information structure. In Adamou, E., Haude, K., , and Vanhove, M.,

editors, *Information structure in lesser-described languages: Studies in prosody and syntax*, Studies in Language Companion Series. John Benjamins.

Ritz, J., Dipper, S., and Götze, M. (2008). Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2137–2142, Marrakech, Morocco.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. In Yoon, J.-H. and Kathol, A., editors, *OSU Working Papers in Linguistics No. 49: Papers in Semantics*. The Ohio State University.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Schwarzschild, R. (1999). GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics*, 7(2):141–177.

Simons, M. (2007). Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6):1034–1056.

Van Kuppevelt, J. (1995). Discourse structure, topicality and questioning. *Journal of linguistics*, 31(01):109–147.

von Heusinger, K. (1999). *Intonation and Information Structure. The Representation of Focus in Phonology and Semantics*. Habilitationsschrift, Universität Konstanz, Konstanz, Germany.

Maximize Q-Anaphoricity (cont.)

More material from the previous discourse can be integrated:

(15) A_1: Edward Snowden is in the meantime a household name for the
             whistleblower in the age of the internet.

    Q: *What about you?*
    Q: *What did you do before becoming a whistleblower?*

    A_2: $[[You_T]_{BG}$ [were working until last summer for the NSA.$]_F$ $]_\sim$