Annotating Coreference and Information Status with RefLex

Janis Pagel
pageljs@ims.uni-stuttgart.de

University of Stuttgart
Institute for Natural Language Processing

May 5, 2016

Overview

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

The Problem of Definiteness

- There are different dimensions of grammatical definiteness in natural language

1. Definite expressions refer to unique entities,
   e.g. in a room with two cats and one dog, one may utter:
   The dog is black.
   But not: #The cat is black. (# = not felicitous)
2. Definite expressions refer to discourse-old entities:
   A dog sits in the room. This dog is black.

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Basic terms

- Phrases which refer to the same entity are **coreferent**
- Phrases which need another phrase to be interpretable are called **anaphors**
- The phrase an anaphor depends on to be interpretable is called **antecedent**

(1) Mary sees [**a crazy man**]$_1$ on campus. **He**$_1$ is a linguist.

Non-anaphorical Coreference

- Non-anaphorical coreference:

(2) [**Angela Merkel**]$_1$ is said to be the most powerful woman in the world.
[**Frau Merkel**]$_1$ is also the current German Chancellor.

Cataphors

- The antecedent could be specified after the anaphor
- In this case the anaphor is called **cataphor** (and the antecedent is called postcedent)

(3) Concerning **his**$_1$ age, **[Hugh Hefner]**$_1$ is still very vital!

Abstract Anaphora

- Some pronouns do not refer back to NPs, but to propositions or properties
- This behaviour is called **abstract anaphoricity**

(4) a. [**John slips on a banana peel**]$_1$. **That**$_1$ amuses Anna.

b. John [**sings the national anthem**]$_1$. Anna wants to do **it**$_1$ too.

Bridging

- In many cases, we perform inferences to resolve reference
- In linguistics this is often called **bridging**

(5) He entered a green room. The ceiling was very high.

- We infer that it is the ceiling of the green room (even though it is not explicitly stated!)

University of Stuttgart
Germany
Institut für Maschinelle Sprachverarbeitung

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

RefLex

- Guidelines for annotating information status and coreference

Arndt Riester and Stefan Baumann (in preparation). RefLex Scheme – Annotation Guidelines. Manuscript. URL: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/arndt/doc/RefLex-guidelines-01aug-2014.pdf.

RefLex

- The idea of information status that is implemented in RefLex goes back to papers by Ellen Prince
- It deals with different kind of knowledge a hearer has about referring expressions

Ellen F. Prince (1981). Toward a Taxonomy of Given-New Information. In P. Cole, editor, *Radical Pragmatics*, pages 233–255. Academic Press, New York.

Ellen F. Prince (1992). The ZPG Letter: Subjects, Definiteness and Information Status. In W. C. Mann and S. A. Thompson, editors, *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pages 295–325. Benjamins, Amsterdam.

R- and L-Level

There are two different layers of annotation in RefLex:

1. The **R-Level**
2. The **L-Level**

- The R-Level deals with referential information such as pronominal coreference or bridging
- The L-Level deals with non-referential, lexical information such as Metonymy, Synonymy or Hyponomy

R- and L-Level

(6) a. Lionel Messi is under pressure. [**The** [**football star**]$_{\text{l-new}}$]$_{\text{r-given}}$ seems to be involved in the Panama Papers Scandal.

b. A blue car drives on the street. [**A green** [**car**]$_{\text{l-given}}$]$_{\text{r-new}}$ is stuck in a traffic jam.

List of R-Level annotations

- R-New
- R-Given
- R-Environment
- R-Given-Sit
- R-Given-Displaced
- R-Bridging
- R-Bridging-Contained
- R-Cataphor
- R-Unused-Known
- R-Unused-Unknown
- $\pm$ generic
- R-Expletive
- R-Idiom
- Antecedent-of-Abstract-Anaphor
- $\pm$ predicate

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

List of R-Level annotations

- R-New
- R-Given
- R-Environment
- R-Given-Sit
- R-Given-Displaced
- R-Bridging
- R-Bridging-Contained
- R-Cataphor
- R-Unused-Known
- R-Unused-Unknown
- ± generic
- R-Expletive
- R-Idiom
- Antecedent-of-Abstract-Anaphor
- ± predicate

We will briefly look at these categories

R-New

- All indefinite expressions are labeled *r-new*

(7) I saw [a woman]$_{\text{r-new}}$ with [a hat]$_{\text{r-new}}$.

R-Given

- Standard (anaphoric) coreference
- For discourse-old information

(8) a. Mary sees [**a crazy man**]$_1$ on campus. **He**$_{1\text{[r-given]}}$ is a linguist.

b. [**Angela Merkel**]$_1$ is said to be the most powerful woman in the world.
[**Frau Merkel**]$_{1\text{[r-given]}}$ is also the current German Chancellor.

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

R-Given-Sit

- Sometimes, reference can only be resolved with respect to the place, time or speaker(s) of the discourse context
- In linguistics, this is described with the term *Deixis*

(9) a. **I**$_{\text{r-given-sit}}$ am talking right now.

b. **Yesterday**$_{\text{r-given-sit}}$ **we**$_{\text{r-given-sit}}$ saw a beautiful ballet.

c. She was standing **here**$_{\text{r-given-sit}}$ and looked at the sea.

R-Unused-Known

- Definite description that is known by the hearer (resp. by the annotator)
- Common knowledge

(10) a. **[The Pope]**$_{\text{r-unused-known}}$ lives in **[the Vatican]**$_{\text{r-unused-known}}$.

b. **[Angela Merkel]**$_{\text{r-unused-known}}$ is a frequent example in these slides.

R-Unused-Unknown

- Definite description that is not known by the hearer (resp. by the annotator)
- Typically complex NPs

(11) John saw [**the dog that barked in front of his house the whole night**]$_{\text{r-unused-unknown}}$.

University of Stuttgart
Germany

R-Bridging-Contained

- Bridging anaphors that include their bridging antecedent

(12) a. **[The construction of the new town hall]**$_{\text{r-bridging-contained}}$ is not finished yet.

R-Bridging-Contained

- Bridging anaphors that include their bridging antecedent

(13) a. **[The construction of the new town hall]**$_{\text{r-bridging-contained}}$ is not finished yet.

b. **[The construction [of the new town hall]**$_{\text{r-unused-unkown}}$**]**$_{\text{r-bridging-contained}}$ is not finished yet.

$\pm$ generic

- Generic expressions are labled with the feature $\pm$ generic

(14) a. [**The lion**]$_{+\text{generic}}$ is a carnivore.

b. He has [**no car**]$_{+\text{generic}}$.

$\pm$ generic

- Generic expressions are labled with the feature $\pm$ generic

(16) a. **[The lion]**$_{+\text{generic}}$ is a carnivore.

b. He has **[no car]**$_{+\text{generic}}$.

- The reason for this label: Generic expressions can be coreferent even though they are indefinite expressions

(17) **[The lion]**$_{1[\text{r-unused-known}|+\text{generic}]}$ is a mammal. **Lions**$_{1[\text{r-given}|+\text{generic}]}$ are also carnivores.

$\pm$ predicate

- Predicative constructions are **Subject + Copula + Predicate**

(18) Mrs. Clinton is [**the boss**]$_{+predicate}$.

± predicate

- Predicative constructions are **Subject + Copula + Predicate**

(20) Mrs. Clinton is [**the boss**]$_{+predicate}$.

- The reason for this label: Predicates of a predicative construction are not coreferent with their respective subject NP, but we want to establish some kind of relationship

(21) a. [The Euro]$_1$ rose to \$1.15. Now [the Euro]$_1$ is at \$1.12.

- Making the predicates coreferent with *Euro* would mean to make *\$1.15* coreferent with *\$1.12*

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

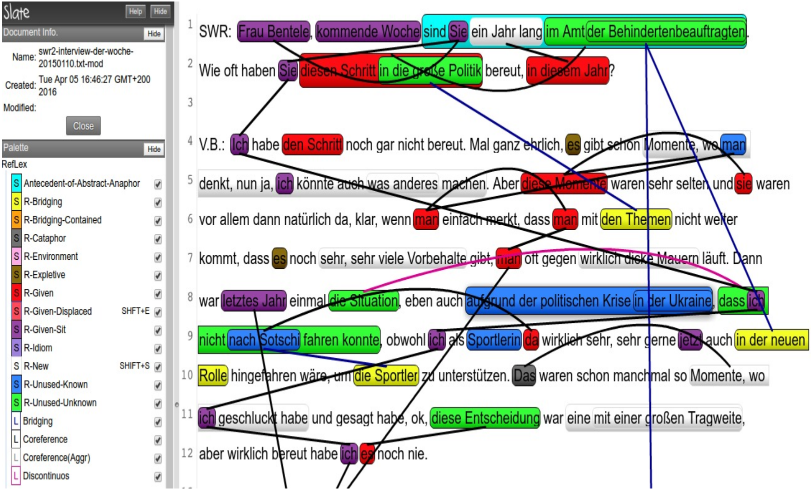
The Data

- German broadcast interviews by the SWR (Südwestrundfunk)
- All interviews are around 10 min of length
- Transcribed by the SWR and slightly normalized in terms of a suitable information status annotation
- i.e.
    1. keeping discourse particles, repairs and word order
    2. no phonetic transcription

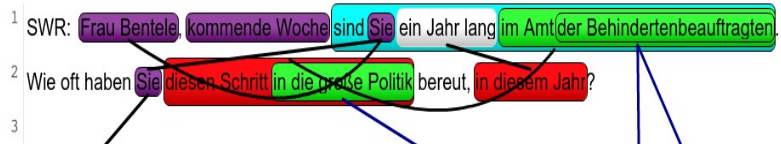
University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

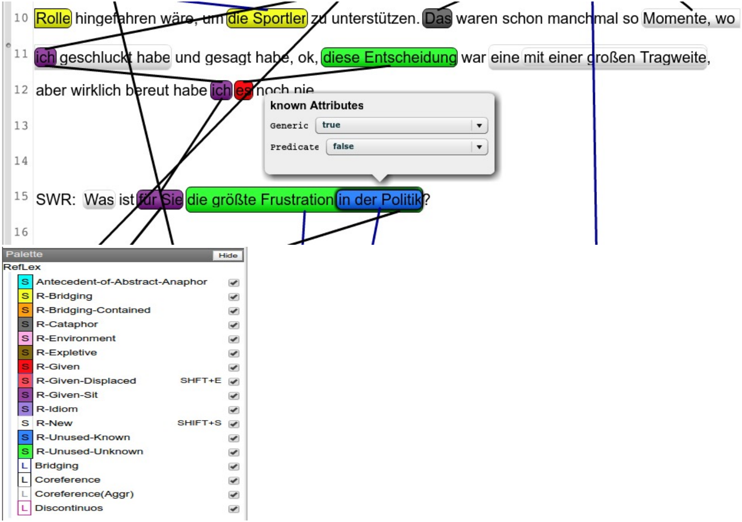
Verena Bentele - First lines of Interview



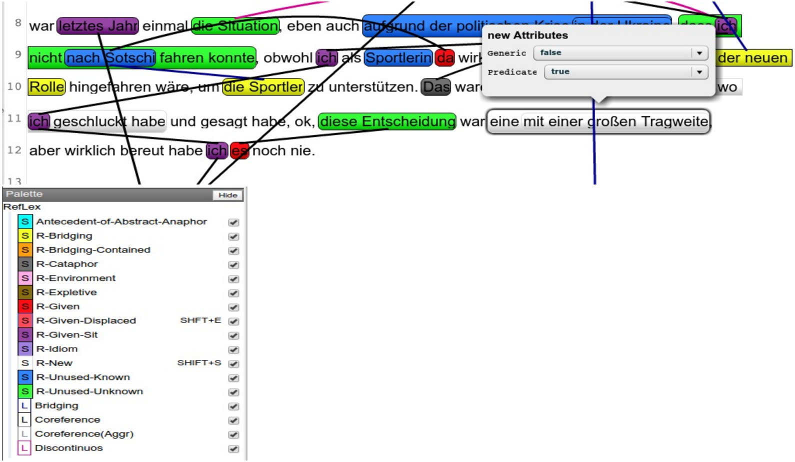
Verena Bentele - First lines of Interview



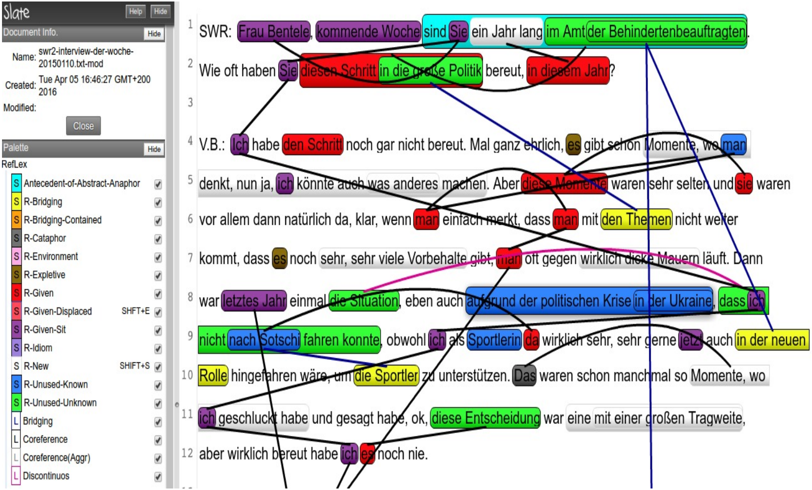
± generic


10 Rolle hingefahren wäre, um die Sportler zu unterstützen. Das waren schon manchmal so Momente, wo
11 ich geschluckt habe und gesagt habe, ok, diese Entscheidung war eine mit einer großen Tragweite,
12 aber wirklich bereut habe ich es noch nie.
known Attributes
Generic true
Predicate false
15 SWR: Was ist für Sie die größte Frustration in der Politik?
Palette
Hide
RefLex
Antecedent-of-Abstract-Anaphor
R-Bridging
R-Bridging-Contained
R-Cataphor
R-Environment
R-Expletive
R-Given
R-Given-Displaced SHFT+E
R-Given-Sit
R-Idiom
R-New SHIFT+S
R-Unused-Known
R-Unused-Unknown
Bridging
Coreference
Coreference(Aggr)
Discontinuos

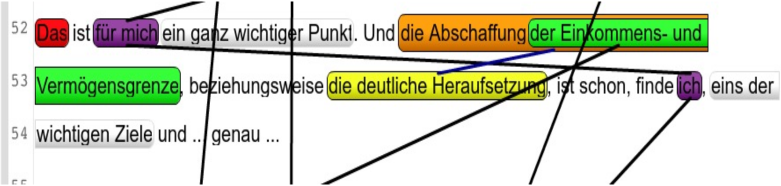
± predicate

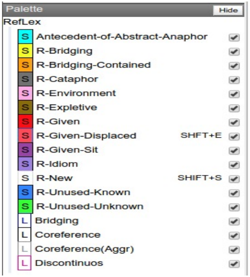

8 war letztes Jahr einmal die Situation, eben auch aufgrund der poli
9 nicht nach Sotschi fahren konnte, obwohl ich als Sportlerin da wir
10 Rolle hingefahren wäre, um die Sportler zu unterstützen. Das war
11 ich geschluckt habe und gesagt habe, ok, diese Entscheidung war eine mit einer großen Tragweite,
12 aber wirklich bereut habe ich es noch nie.
new Attributes
Generic false
Predicate true
der neuen
wo
Palette
Hide
RefLex
Antecedent-of-Abstract-Anaphor
R-Bridging
R-Bridging-Contained
R-Cataphor
R-Environment
R-Expletive
R-Given
R-Given-Displaced SHFT+E
R-Given-Sit
R-Idiom
R-New SHIFT+S
R-Unused-Known
R-Unused-Unknown
Bridging
Coreference
Coreference(Aggr)
Discontinuos

r-bridging

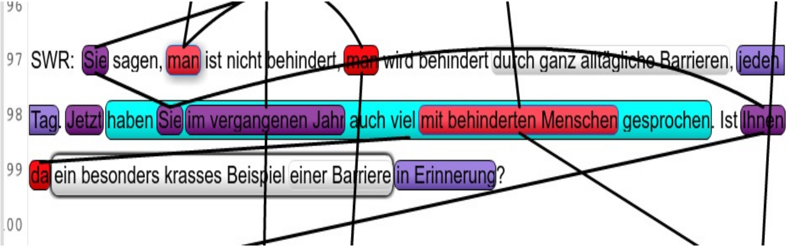

Slate
Help
Hide
Document Info.
Hide
Name: swr2-interview-der-woche-20150110.txt-mod
Created: Tue Apr 05 16:46:27 GMT+200 2016
Modified:
Close
Palette
Hide
RefLex
S Antecedent-of-Abstract-Anaphor
S R-Bridging
S R-Bridging-Contained
S R-Cataphor
S R-Environment
S R-Expletive
S R-Given
S R-Given-Displaced SHFT+E
S R-Given-Sit
S R-Idiom
S R-New SHIFT+S
S R-Unused-Known
S R-Unused-Unknown
L Bridging
L Coreference
L Coreference(Aggr)
L Discontinuos
1 SWR: Frau Bentele, kommende Woche sind Sie ein Jahr lang im Amt der Behindertenbeauftragten.
2 Wie oft haben Sie diesen Schritt in die große Politik bereut, in diesem Jahr?
3
4 V.B.: Ich habe den Schritt noch gar nicht bereut. Mal ganz ehrlich, es gibt schon Momente, wo man
5 denkt, nun ja, ich könnte auch was anderes machen. Aber diese Momente waren sehr selten und sie waren
6 vor allem dann natürlich da, klar, wenn man einfach merkt, dass man mit den Themen nicht weiter
7 kommt, dass es noch sehr, sehr viele Vorbehalte gibt, man oft gegen wirklich dicke Mauern läuft. Dann
8 war letztes Jahr einmal die Situation, eben auch aufgrund der politischen Krise in der Ukraine, dass ich
9 nicht nach Sotschi fahren konnte, obwohl ich als Sportlerin da wirklich sehr, sehr gerne jetzt auch in der neuen
10 Rolle hingefahren wäre, um die Sportler zu unterstützen. Das waren schon manchmal so Momente, wo
11 ich geschluckt habe und gesagt habe, ok, diese Entscheidung war eine mit einer großen Tragweite,
12 aber wirklich bereut habe ich es noch nie.

r-bridging-contained

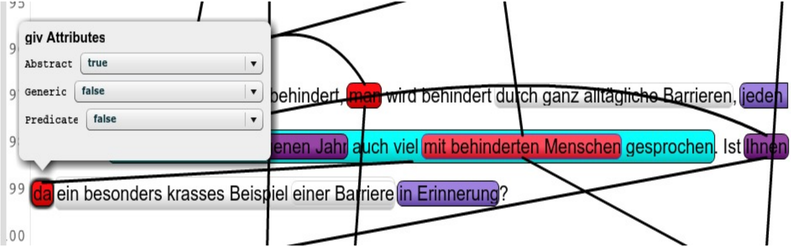




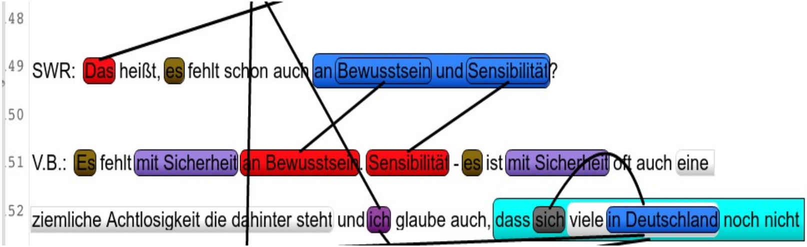
Abstract Anaphora



Abstract Anaphora



Conjunctions



University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Information Status Annotations in NLP Tasks

- A lot of Natural Language Processing (NLP) tasks require information about information status and coreference
- For example:
  1. Automatic coreference resolution
  2. Text generation/summarization
  3. (Statistical/Rule-based) Machine Translation
  4. Information extraction
  5. Question answering
  6. ...

Machine Translation

- Resolve coreference and choose correct pronoun

(22) a. **The bird**$_{1,\text{[sg-neut]}}$ sings. **It**$_{1,\text{[sg-neut]}}$ is black.
b. ***Der Vogel**$_{1,\text{[sg-masc]}}$ singt. **Es**$_{2,\text{[sg-neut]}}$ ist schwarz.
c. **Der Vogel**$_{1,\text{[sg-masc]}}$ singt. **Er**$_{1,\text{[sg-masc]}}$ ist schwarz.

- The system can only choose the correct pronoun when it knows about the coreference chain

Information extraction

- Easier to extract related information if the system knows which entities are coreferent
- Helps system to group related information together

Question answering

- System needs to know about coreferent entities in the questions
- Especially pronoun resolution can be helpful for such a system
- Coreference knowledge for the data helps anwsering the question (see Information Extraction)

Thanks for your attention! :)

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung