Annotation Guidelines: Annotating Information Status and Anaphoric Links in Scientific Text

Ina Rösiger

Version 4

March 2014

1 Information Status

1.1 Markables: Definite noun phrases (DPs)

1.1.1 Definite descriptions

- Definite descriptions: Nominal phrases (NPs) consisting of a determiner, an optional number of adjectives and a noun: `"the most efficient siRNAs"`, `"the siRNAs"`

1.1.2 Named entities

- Named entities/proper names: `"siRNAs"`
- Variables: `"x"`,`"y"`,`"z"` ...

1.1.3 Pronouns

- Personal pronouns: `"we"`, `"it"`, `"they"` ...
- Possessive pronouns: `"our"`, `"their"`, `"its"` ...
- Demonstrative pronouns: `"this"`, `"these"` ...

1.1.4 Quantifier phrases

- Universal quantifier phrases: `"all siRNAs"`

1.2 Non-markables:

1.2.1 Pronouns

- Relative pronouns
- Expletive and pleonastic "it":

  `It is interesting to see the different types of cells.`

- Do not mark "it" in:

  `Since it was discovered that ...`

1.2.2 Indefinite NPs

- Indefinite descriptions: `"a siRNA"`
- Existential quantifier phrases:

  `"some siRNAs", "most siRNAs", "15 siRNAs"`

- bare plurals: `"proteins"`

1.2.3 Bare singulars except...

when they corefer with another entity and are thus marked GIVEN, see Section 3.1.

1.2.4 Special cases

- Existential "there":

  `There are many amino acids that ...`

1.3 Categories

Table 1 shows the categories in the annotation scheme. The entity being marked is shown in bold face, referring expressions are marked by a box and their antecedent are shown in italics.

	Category	Example
Coreference links	Given (specific)	We present *the following experiment.* **It** deals with ...
	Given (generic)	We use *the Jaccard similarity coefficient* in our experiments. **The Jaccard similarity coefficient** is useful for ...
Associative links	Associative	Xe-Ar was found to be in *a layered structure* with Ar on **the surface**.
Categories without links	Associative (self-containing)	**The structure of the protein** ...
	Description	**The fact that the accuracy improves** ...
	Unused	**Noam Chomsky** introduced the notion of ...
	Deictic	**This experiment** deals with ...
	Predicative	Pepsin, **the enzyme, ...**
	Idiom	On **the one hand** ... on **the other hand** ...
	Unmarked	

Table 1: Categories and links in our classification scheme

Classify all definite noun phrases (DPs) according to the following annotation scheme.

1.3.1 Given

We consider a definite noun phrase GIVEN when the entity refers back to a discourse entity that has already been introduced in the discourse and is thereby known to the reader. This includes lexically new material, pronouns and repetitions or short forms of entities that have been referred to before. GIVEN entities include synonyms and are not limited to entities that have the same head. A pair of NPs is considered coreferent if the referring expression, the so-called *anaphor*, refers back to a previous expression it is associated with, the *antecedent*.

We present *a paper* that deals with the adaption to new domains.
It starts with an overview of the biomedial domain.

Sequences were mapped and deposited into *a database*.
The database comprises 10000 entries.

In 2006, Andrew Fire and Craig Mello shared the Nobel Prize in Physiology or Medicine for their work on RNA interference in the nematode worm *C. elegans*. **C. elegans** is unsegmented, vermiform, and bilaterally symmetrical.

The obligatory attribute GENERIC tells us whether the entity is generic or specific.

Attribute: generic

The attribute distinguishes between generic and specific expressions. If an entity is generic, the attribute has to be set to generic. No attribute means that the entity is specific. Generic expressions include reference to a kind or a general quantification whereas a specific reading has a fixed referent, i.e. we know which referent we select out of the set of entities that fulfil the description.

- Generic:
  Proteins perform a vast array of functions.
  siRNAs are considered ...

⇒ Bare plurals are always generic

- Specific:
  Pepsin is released by the chief cells in the stomach. **The protein** was discovered in 1836.

1.3.2 Associative

For ASSOCIATIVE DPs, the text presents an antecedent NP which does not stand in the relation of identity, but in some other form of relation to the associative phrase. The antecedent may be an associate in a typical relation such as part-of, is-a, or any kind of associate as long as there is a clear relation between the two phrases.

Xe-Ar was found to be in *a layered structure* with Ar on **the surface.**

We use *a classifer* to distinguish between the two categories.
The training data consists of ...

1.3.3 Associative (self-containing)

In some constructions, e.g. genitives or PP modifiers, we identify an associative relation between the head noun phrase and the modifier. We consider them ASSOCIATIVE SELF-CONTAINING and do not create a link.

The structure of the protein
the thoracic circuit stage in HK mutants
the giant fiber escape pathway of Drosophila

1.3.4 Deictic

All entities that refer to the current paper or aspects thereof are considered DEICTIC.

This experiment deals with ...
This paper talks about ...

1.3.5 Unused

If an entity is not mentioned before and is not related to some other entity in the text, but refers to something which is part of the common knowledge of the writer and the reader, it is called UNUSED. The entity can be interpreted due to world or domain knowledge. This is often the case for named entities.

We note that **the accuracy** has improved.
Noam Chomsky introduced the notion of ...
You can look it up in the **ACL Anthology**.

1.3.6 Description

The entity is either self-explanatory or given together with its own identification. This means the entity is not anaphoric, does not rely on information about the situation of utterance and is not associative of some trigger previously introduced in the discourse.

The fact that the protein has a layered structure is noncontentious.

The definite noun phrase refers to something new, but the syntactic construction makes the interpretation easier. DESCRIPTION entities are heavily related to the following syntactic patterns:

- NP complements: presence of a complement to the head noun

the fact that the accuracy has improved

- Relative clauses:

the protein that is essential

1.3.7 Idiom

Entities that include idiomatic expressions or metaphorical use are considered IDIOMS.

On **the one hand** [...] on **the other hand** [...]

1.3.8 Predicative

Any predicative expressions, including appositions, are annotated as PREDICATIVE.

Pepsin, **the enzyme**, ...
Pepsin is **the enzyme**
Short interferring RNAs (**RNAs**)

(1.3.9 Unmarked)

Coreference or associative links can refer to an entity that has not been marked before, e.g. because the entity is indefinite. As the annotation tool needs the entity to be classified to be able to create a link, we classify the entity as UNMARKED for practical reasons.

2 Associative Links

2.1 Markables

Same as in 1.1, with the exception of personal and possessive pronouns that are typically not associative, but rather GIVEN/coreferent.

2.2 Principle

The anaphor must be a definite noun phrase; the antecedent can be any type of nominal phrase (indefinite, definite, named entity ...). Verbal phrases or clauses cannot be the antecedent of an associative anaphor.

- 2 cases:

– antecedent already marked with information status:

Antecedent (DP, marked as an IS category)	Anaphor: DP, mark as ASSOCIATIVE

– antecedent has not been marked (for example because it is indefinite):

Antecedent (NP, UNMARKED)	Anaphor: DP, mark as ASSOCIATIVE

The links do not have to follow the chain principle: choose the best antecedent, not the last occurrence in the text. Associative antecedents can also have two antecedents (and two links), if this fits best.

3 Coreference Links

3.1 Markables

same as in 1.1, with two additions:

(1) **Pre-modifiers/Compounds**: Proper pre-modifiers (proper nouns, named entities) that can be coreferenced to other mentions should be added to the list of mentions. Adjectival and common noun pre-modifiers are not marked.

These pre-modifiers should be marked with the attribute PART-OF-COMPOUND.

```
The siRNA experiments ⇒ two markables (= the siRNA experiments, siRNA)
The protein experiment ⇒ one markable
he proteinaceous experiment ⇒ one markable
```

(2) **Bare singulars** if ...

... the insertion of a definite determiner is possible.

`The efficiency of RNAi is ... . RNAi efficiency can also be influenced by ...`

⇒ The RNAi efficiency can also be ...

3.2 Principle

The anaphor must be a definite noun phrase; the antecedent can be any type of nominal phrase (indefinite, definite, named entity ...). Verbal phrases or clauses cannot be the antecedent of a coreferent anaphor.

Analogous to the principle for associative links:

- 2 cases:
  - antecedent already marked with information status:

Antecedent (DP, marked as an IS category)	Anaphor: DP, mark as GIVEN

  - antecedent has not been marked (for example because it is indefinite):

Antecedent (NP, UNMARKED)	Anaphor: DP, mark as GIVEN

Coreference links follow the chain principle: always choose the last mention of the entity in the text as the antecedent. Multiple links are not allowed.

4 Special Issues

4.1 What if more than one category fits?

Whenever more than one classification fits, refer to the following rules to find priority:

- GIVEN overrules all other categories
- ASSOCIATIVE (SELF-CONTAINING) vs. DESCRIPTION: see test below

Test: Description vs. associative (sc)

The distinction between these two categories can be difficult, particularly for the scientific domain. The test that is described below is helpful to disambiguate between the two categories.

Let the genitive noun phrase or the PP modifier precede the head noun phrase. If the head NP is dependent on the modifier, it is considered ASSOCIATIVE (SC).

If it can be mentioned without the modifier and there is no prototypical relation between the two NPs, we call it DESCRIPTION.

The examples in Table 2 illustrate the difference:

Phrase	Test	Result
`the distribution of the data` →	`the data ... the distribution`	⇒ `associative (sc)`
`the structure of the protein` →	`the protein ... the structure`	⇒ `associative (sc)`
`the number of n-grams` →	`n-grams ... the number`	⇒ `description`
`the existence of a polynomial-time parsing algorithm` →	`... the existence`	⇒ `description`

Table 2: Test: description vs. associative

4.2 Span to be marked

- Always mark the longest DP span:

  `the number of different types of acids in the protein`

  - if there is a PP modifier, include the whole PP
    `the structure of the protein`,
    `the information in the right hand side of the table`
  - also include relative clauses
    `the protein that has the highest level of ...`

- Annotate embedded DPs (all DPs should be annotated)

  `the number of those siRNAs that ...`: two markables, "the number of those siRNAs that" and "those siRNAs that ..."

  BUT

- Do not split NPs in the form of ADJ+N

  `the efficient siRNAs`: do not annotate siRNAs as a separate entity
  `low thermodynamic`: one entity
  `high thermodynamic`: one entity

- Do not mark prepositions, e.g.

  in `on the other hand` – do not mark "on"

- Do not mark focus particles, such as `only, also, even`

4.3 Event reference

If the referent of an NP is a clause or a verb, do not annotate any kind of information status; just ignore the phrase

`We integrate this into the resolver. This integration` – do not mark "this integration"

4.4 Generic vs. specific entities

Generic and specific entities are part of the same chain. Add the attribute GENERIC for generic entities.

4.5 Possessive extents

Entites may include the possessive 's in the NP. This 's should be included in the markable: `Agent B's` strategy, `Sirna's` ability to ...

4.6 Gender and number

In case there is a disagreement in number or gender, but the two entities clearly refer to the same specific real-world entity, you can create a coreference link.

BUT: the default is that plural and singular entities are not coreferent:

`siRNAs` and `siRNA` form different coreference chains

4.7 Predicative bracket expressions

Often, specialist terms (particularly in genetics) are introduced by the term and followed by the abbreviation in brackets:

`short interferring RNAs (siRNAs)`

Both "short interferring RNAs" and "siRNAs" should be marked. The abbreviation in brackets gets assigned the type PREDICATIVE.

4.8 Compounds/multi-word expressions

Do not mark parts of compounds, unless it contains a named entity

⇒ annotate named entity as GIVEN (in case it corefers with another entity), but mark it with the special attribute PART-OF-COMPOUND.

The siRNA experiments – two markables
The protein experiments – one markable

4.9 We

The first occurrence of "We" is considered deictic, all others given/coreferent to the first occurrence. Occurrences of "our" are part of the same chain.

4.10 Conjunctions/Split antecedents

Default: mark as two entities
Special cases: additionally mark as one entity

See table below:

Default marking (underlined)	**Reference** to	
	only one	both
the A + the B	as usual	additonally: mark as one entity: the A + the B
A + the B	as usual	additonally: mark as one entity: A + the B
the A + B	as usual	additonally: mark as one entity: the A + B
A + B	as usual	additonally: mark as one entity: A + B
the A + B	create an associative link to the whole phrase	as usual

Split antecedents (last row in table):

(1) If it is possible to mark them as two entities, mark them both separately:

`the RNA experiments and the siRNA experiments` – mark as two entities

In case another entity refers to one part of the conjunction, you can then annotate a coreference link

(2) If it is not possible to mark them as two entities (because they are split), mark as one entity.

`the RNA and siRNA experiments` (one entity).

In case another entity (later in the text) is coreferent to only a part of the conjunction, mark it as associative.

`The RNA experiments` – mark this as associative

4.11 Multiple antecedents

In coreference chains, multiple antecedents cannot be marked. In this case, mark the anaphor as associative and create two associative links to both antecedents.

`Protein A has the same structure as protein B. Both proteins ...`

4.12 Titles, headings, acknowledgments and references

Anntotate DPs in

- the title of the paper
- headings of the paragraphs

Do not annotate

- DPs in the acknowledgements and references (at the end)
- authors of the paper (in the affiliation at the beginning)

4.13 Missing text

In some rare cases, numbers that are part of the text have been removed and are thus missing, e.g. in Section 5, in Table 2, in Figure 3, etc. Do not annotate these cases when the number is missing. If the text is complete (`Section 5`), we consider these NPs definite and therefore label them.

4.14 Cataphors

We do not annotate cataphors.

4.15 Citations

Annotate citations in the text, such as `Gale 2009`. Do not annotate the names in the citations as embedded ones, i.e. do not annotate `Gale` alone.

In the case that the bracket has been removed, such as in `Brown Brown 91`, mark the whole thing as one entity!