guidelines - **coref-annotation**

Introduction *corpus, target relations, goal setting, coref-annotation*

1. Entities

1.1. **Personal entities** may be referred to by their name, a noun or nominal phrase, or by a pronoun. Thus a single personal entity can be expressed in several ways:

origin	**Jon Bon Jovi**
name	Jon Bon Jovi, Mr. Bon Jovi, Jon,
nominal	the rocker, a member of that society, her father, the 40-year-old
pronoun	he, him, I

1.2. **Event entities** may be represented by their name, a noun or nominal phrase, by a pronoun, a verb or by an adjective phrase. Thus a single event entity can be expressed in several ways:

origin	**their marriage**
name	- ; often found for more specific, socio-historical events
nominal	marriage, wedding, newlyweds
pronoun	it, that, this
verb	to marry, wed, take vows
adjective	married

2. Entity Types and Subtypes

2.1. **Persons**

Each mention that represents a distinct person in the real world is annotated as type person. People may often be alternatively referred to by their **name, occupation, family relation, pronoun,** etc. or by some combinations of these.

We annotate as type person:	name mentions - nominal mentions - pronoun mentions religious figures - fictional characters - positions

We do not annotate as type person: animals - titles - honorifics

2.2.1 **Groups**

Groups of people are to be considered an entity of type person_group unless the group meets the requirements of an organization, which we state as a formal organizational structure within the group.

We annotate as type person_group:	name mentions	the Bon Jovis
	nominal mentions	the couple, the newlyweds, the three men
	pronoun mentions	they, their,

Whenever a group of people is introduced and it‘s members are named or counted, we build a member_of_group relation between all mentioned parts of the group and the group itself. We do this only once per group and document right with the first mention or beginning of the mention chain. If we can not find any name mention for the members, we follow the NearFirst rule (see 4.3) and add nominal or pronoun mentions to the existing relation.

> They were strangers on a plane, three men who probably were thinking of the business meetings that awaited **them** (group) in California. It was a fluke that **Jeremy Glick** (member)**, Tom Burnett Jr.** (member) and **Mark Bingham** (member) were all on the Boeing 757.
>
> *member_of_group: [them, Jeremy Glick, Tom Burnett Jr., Mark Bingham]*

2.3. **Events**

An event is considered to be something, that happens or takes place; a single specific occurrence either instantaneous or ongoing, that is unique and can be anchored at a point in time.[1]

An event mention in most cases consists of an event trigger and additional arguments. While the event trigger can be seen as the head word of the phrase, whether it is a noun, verb or adjective, additional arguments are not necessarily needed for an easy understanding of what is going on. Thus events can be referred to by, nouns, pronouns, verbs and adjectives. These are labelled as type event.

We annotate as type event:	name mentions	...
	nominal mentions	their marriage
	pronoun mentions	it, that, these
	verb mentions	to father, to marry, wed, took vows
	adjective	married, born
	additional arguments:	trigger type
		polarity
		tense
		modality (fictional, speculative, negated)

[1] Hasler, Laura, Orasan, Constantin, Naumann, Karin - NPs for Events: Experiments in Coreference Annotation, University of Tübingen, Computational Linguistics, URL: http://clg.wlv.ac.uk/projects/NP4E/539_pdf.pdf, p.1171

Additional arguments are temporarily not attached. We make one exception: We add an extra tag 'modality' to our target relation 'marriage' (see 6).

2.4 **Semantic Terms**

Every word which semantically expresses the core idea of our target relations and which is related to the semantic relation in some way is to be marked and labelled as „semantic term".

his sister	first entity	his sister
	second entity	his
	semantic term	**sister**
their marriage	first entity	their
	event	their marriage
	semantic term	**marriage**

Note the "term" in "semantic terms". Because of the focus on individual words or short compound nouns, the "Mention Extent and Mention Head" rule from 4.1 is not applicable here.

2.5. **Pre-annotated Entities of Type Coreference**

The entity type "coreference" should not be used any more. Existing entity mentions of this type are leftovers from prior annotation attempts and the corpus preprocessing, they can safely be ignored, or deleted.

3. **Coreference-Annotation**

3.1 **Source and target**

Any type of coreference relation consists of two parts, the "source" and the "target". While the target, which may be an anaphoric or cataphoric mention, is somehow related and referring to an antecedent within the text structure, the source is a name, noun or nominal phrase which directly represents the entity of the real world. With these two labels anaphoric and cataphoric relationships between two mentions can be shown.

Anaphoric expression:

Guy Ritchie **(source)** *followed, turning guest's heads as he* **(target)** *entered in a vibrant teal blazer...*

Cataphoric expression:

Madonna's 4-year-old-daughter **(target)**, *Lourdes* **(source)**, *....*

When establishing coreference relations between two or more mentions of type event, person or person_group *the rules given in 4.3 apply. We establish three types of relations through the entire corpus. A short description is given below.*

In early October, the annotation guidelines have been updated with a meaning of the "source" and "target" parts. For the sake of annotation simplicity and consistency, we continue the annotation process with the former naming convention, i.e., Guy Ritchie (target) & he (source) for the example above. In the postprocessing step, the annotation will then be fixed automatically.

3.2. **Identity Relations**

Referring expressions point back to the same entity.[2]

Every mention of type person, person_group or event which refers to the same entity in the real world and which follows each other within the text structure should be linked by an identity relation. Thus we build a coreference set for every single person, group or event entity mentioned in a text. We label the very first source as source and every other mentions which refer to the same entity as target. We do not establish more than one set per person, group or event, so that the sets include every single possible mention of that person, group or event.

- person **identity** [her husband, British film director, Guy, Guy Ritchie, Mr. Ritchie, dad, ...]
- person_group **identity** [they, the couple, the newlyweds, the parents, ...]
- event **identity** [married, marriage, it, that, took vows, what happened last summer, ...]

Coreference sets only consist of mentions that point back to the exact same entity. Therefore we build two separate identity chains for any divorce and marriage phrase mentioned in a text. When adding these mentions as arguments to our target relation 'marriage', we label every mention that refers to the ending of this relationship, e.g. divorce, as 'end_event'. Any other related event mention is labelled as 'event'.

3.3. **Inferred Relations**

The entity referred to by the source of the link is loosely connected to the entity referred to by the link's destination.[3]

Whenever an entity is mentioned or represented not in all it's parts, less complexly or more detailed and the semantic link between the two entities is loose we build an inferred relation. We distinguish the following subtypes:

- **synonymy** a word that means the same or nearly the same as another word
- **hypernym** a word whose meanings includes the meanings of other words
- **hyponym** a word used to designate a member of a class
- **part-of** a word that refers to an entity which is part of a more complex concept
- **other** every other possible inferred concept that does not fit the other subtypes

Subtypes are temporarily not attached.

[2] Komen, Erwin R. - Coreference Annotation Guidelines, Radboud University, Nijmegen, Netherlands, June 2009, p. 4
[3] see ibid., p.4

3.4. **Cross Speech Relations**

Referring expressions point to the same entity, but the link crosses a direct speech border.[4]

When we find an identity relation between two mentions which crosses a direct speech border, we do not establish an identity relation but a cross speech relation, though the two mentions refer to the same entity.

Every CrossSpeech relation is temporarily annotated as identity relation.

4. How to annotate

These are some basic rules we set for the annotation process:

4.1. **Mention Extent and Mention Head**

We annotate maximally long markables. This means the whole mention extent is to be marked.

> ... [a member of that small society of aging-but-blazing singers who‘ve learned how to stoke fans while keeping the home fires burning] ...

Additionally we annotate the mention heads:

> ... [[a member] of that small society of aging-but-blazing [singers] who‘ve learned how to stoke fans while keeping the home fires burning] ...

For coreference links we use the full mention extent.

4.2. **Nested Mentions**

Target relations we find in nested mentions like *her father, her mother, his dad, her sister, etc.* are to be annotated only once within the whole corpus.

4.3. **Hierarchy of Relation Types**

When annotating simple coreference links between groups, persons or event the following annotation rules apply:

> **NearFirst**: choose the nearest source
> A source should establish a coreference relation with nearest target possible.

[4] Komen, Erwin R. - Coreference Annotation Guidelines, Radboud University, Nijmegen, Netherlands, June 2009, p. 4

IdentityFirst: choose an „Identity“ type relation
If it is possible to refer back to a target with an Identity relationship, this should be done.

CrossSpeech >> Identity >> Inferred
When there seems to be a competition between NearFirst and IdentityFirst, it is IdentityFirst that wins, unless the coreference link crosses a direct speech boundary. In such a situation CrossSpeech is chosen.[5]

4.4. **Coref and Ucoref**

There is the option to use a coref or ucoref tag, which are used to show if there is no doubt that one entity corefers to another (**coref**) or if the annotator is relatively sure of coreference but there is an element of uncertainty (**ucoref**).

Coref and ucoref tags are temporarily not attached.

4.5. **Other**

Personal entities are given a name feature. We attach a complete name feature only to on of the mentions of that personal entity, which is then linked to the other mentions with coreference. Normally, the entity mention which gets attached the name feature should be the first one, but it could also be the first mention with the full name of the entity being spelled, or any other mention referring to this entity.

Objects that follow auxiliary verbs like „to be“ are seen as a markable mentions.

We do not exclude appositions, relative clauses or any other kind of attributive phrases.

5. **Entity class**

Each taggable entity must be assigned an entity class.

referential new entity or a definite descriptive term, name or anaphoric expression for a referential already occurred

- **generic** — mention refers to a class, kind or species of objects but does not point to any specific individual
- **non-generic** — mention refers to one or more individual entities of a class
  - if countable: **specific non-generic**
  - if not countable: **non-specific non-generic**

[5] adapted from Komen, Erwin R. - Coreference Annotation Guidelines, Radboud University, Nijmegen, Netherlands, June 2009, p. 4

attributive giving extra properties to an entity already referenced in the sentence (appositive) or qualifying an entity immediately within the same phrase

Entities often are quantified. We annotate also negatively quantified entities. For further help during class assigning in the annotation process see decision tree in 7. Entity class tags are temporarily not attached.

6. Target Relations

We specify our relations by adding additional arguments:

event — first event mention which states the relationship between two entities

related and/or — every entity that is not necessarily but in any way referred to the event / relation

expresses the core idea of the target relation (semantic term, see 2.4)

modality — fictional, speculative, negated
added as extra property to the target relation 'marriage'

marriage

argument	type	required	size
person (gender unknown)	person / person_group	x	=2
wife	person	-	<=2
husband	person	-	<=2
event	event	-	-
end_event	event	-	-
related	person/person_group/semantic term	-	-
from	date	-	-
to	date	-	-
ceremony	location	-	-

sibling_relationship

argument	type	required	size

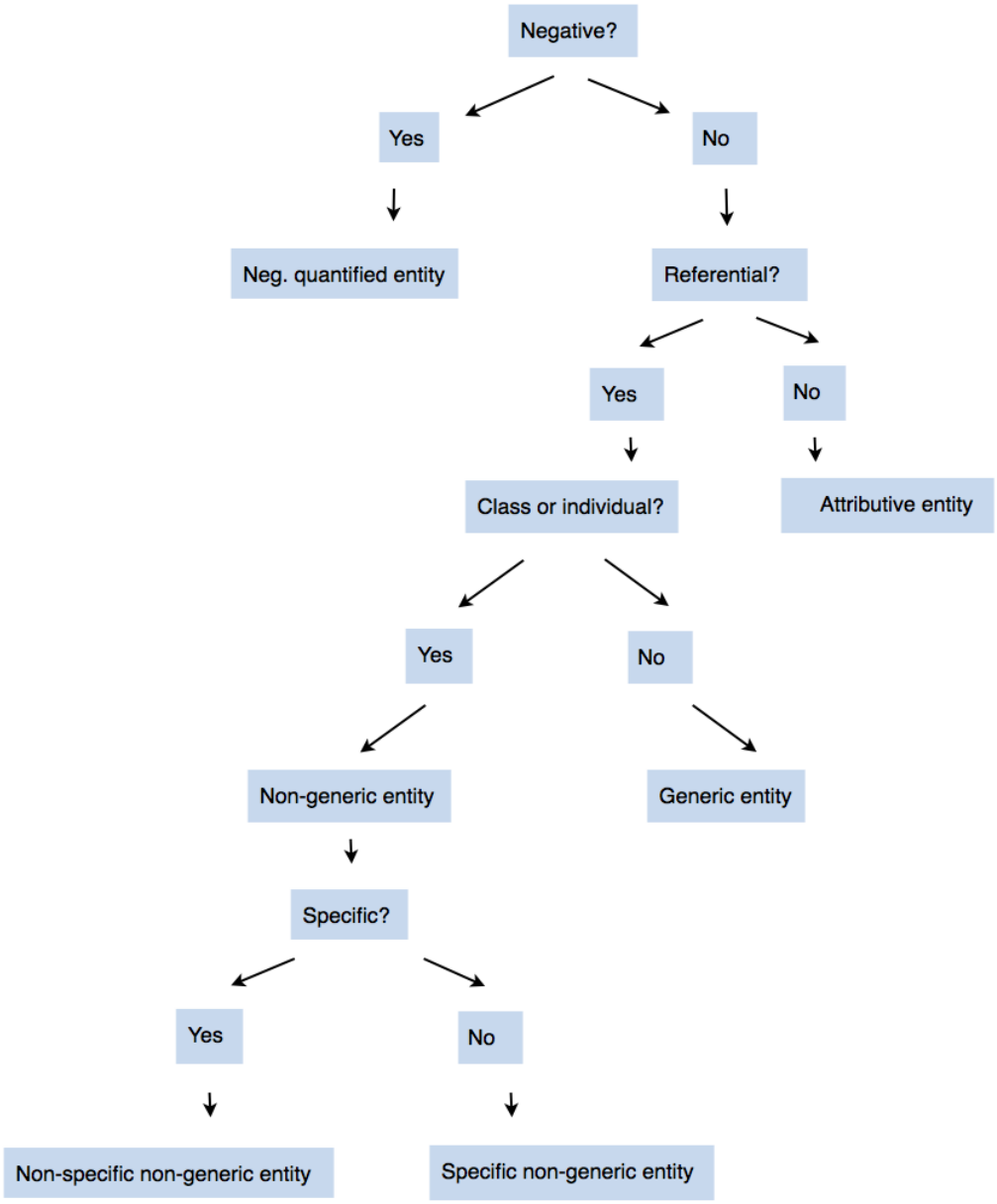
person	person/person_group	x	>=2
sister	person/person_group	-	-
brother	person/person_group	-	-
event	event	-	-
related	person/person_group/semantic term	-	-

person_parent

argument	type	required	size
person (gender unknown)	person/person_group	x	>=1
son	person/person_group		
daughter	person/person_group		
parent	person/person_group	x	[1,2]
mother	person		<=2
father	person		<=2
event	event	-	-
related	person/person_group/semantic term	-	-

7. Decision tree of entity class[6]

[6] adapted and modified f. Komen, Erwin R. - Coreference Annotation Guidelines, Radboud University, Nijmegen, Netherlands, June 2009, p. 14


Negative?
Yes
No
Neg. quantified entity
Referential?
Yes
No
Class or individual?
Attributive entity
Yes
No
Non-generic entity
Generic entity
Specific?
Yes
No
Non-specific non-generic entity
Specific non-generic entity