

Outstanding issues for Within-document NP Coreference Guidelines

(Laura Hasler: 30.01.2006)

During our discussions, there were several issues arising that we decided not to change now because they would make the annotation unfair and affect the inter-annotator agreement as some annotation had already been done. There were also some issues that we did not reach a definite decision about how to deal with. These are things that it would be a good idea to add/change/bear in mind in future guidelines and annotations.

General concepts: things like *violence, terror, terrorism, police, rebels, militants* etc. In several texts, these are used in a general sense and it is difficult to decide whether they should be annotated as coreferential and if so what with and how. I agree with Karin that there should be some recognition that a text is about *terrorism*, for example, and if it keeps mentioning *terrorism* we want to be able to encode somehow that this is the same concept under consideration even if it is general and encompasses many different things. However, this is difficult because coreference is about specifics. It is something to bear in mind for the future: we did not come up with a solution yet, but we can look at exactly how Tuebingen deal with it (they do annotate general concepts).

More than one feasible antecedent: there are several cases where there are two feasible indefinite antecedents in the text (not in the headline, this is dealt with differently) for the first definite mention of an entity. The definite mention is coreferential with both previous indefinite mentions, but the second indefinite mention cannot be marked as coreferential with the first because it is indefinite, even though all three mentions refer to the same thing. We need to decide how to deal with this: do we link the first definite mention and all subsequent mentions to i) the first indefinite mention, ii) the nearest (second) indefinite mention, or iii) it depends on the amount of detail/information given by the two mentions and the context/individual text, leaving us without a set rule. So in the following example, which antecedent should *the Boeing 757* and *its* be linked to? E.g. *Taiwan urged China on Tuesday to return the man detained for hijacking [a Taiwanese airliner] to the mainland... Liu, who doused himself with gasoline and forced [a Far East Air Transport domestic plane]... Beijing returned [the Boeing 757] and [its] 158 passengers and crew...*

Relative pronouns/clauses: in future we will mark relative pronouns, there is no reason not to – they are not as complicated to mark as first thought. Tuebingen just mark the pronoun and do not include the clause as a modifier within a larger NP, or as a markable itself, because they annotate at phrase level and therefore do not consider anything above that as a markable. They link the relative pronoun to the main clause only. We do not have this restriction: in our annotations so far we have taken as much information as possible within the NP, including relative clauses. So we could have the whole NP, including the relative clause, as a markable and then within that mark the relative pronoun as coreferential with the whole NP as this is the longest match/nearest mention. The nearest mention is the most appropriate to link to in this case because this type of construction is similar to appositives etc. in the sense that the relative clause and pronoun

fall within the larger NP but that the relative clause is not considered a markable itself. In the following example, *who* would be marked as coreferential with the whole NP. E.g. *Chinese officials were tightlipped whether [Liu Shan-chung, 45, [who] is in custody in China's southeastern city of Xiamen], would be prosecuted or repatriated to Taiwan.*

We in direct speech: we did not really agree on how this should be marked. We discussed whether they should be marked as coreferential with the organisation the speaker represents, as coreferential with the actual speaker, or not marked as coreferential at all because you cannot say with 100% certainty whose views are really being conveyed or who *we* refers to. It is currently annotated as coreferential with the organisation the speaker represents. This could be discussed further, or changed, in the future.

Problems with selecting ident/synonym/generalisation/specialisation tags: it is often difficult to assign one of these attribute tags to a coreference relation. Constantin added WordNet to PALinkA during Karin's annotation to make this easier. The first 94 texts annotated did not have use of WordNet in this way. If the selection of these tags still proves difficult in the future, it may be better to have just two attributes, **ident** and **not_ident**. This is something to bear in mind for future work.

New attribute tags: in future, it would be useful to have the following attribute tags added to PALinkA: **cataphoric**, **pronoun** and **dashes**.

We also discussed **which mention of an NP coreferential referents should be linked to** (i.e. first or nearest mention). As it stands, we differentiate between the two, having some referents linking to the nearest mention and others linking to the first mention, based on the trial annotation and intuition. It was decided that this is really an aesthetic/neatness issue and that providing all the mentions of an entity are annotated as coreferential to another entity within the same chain, regardless of whether it is the first mention or the nearest mention, it will still appear in that chain and be accessible. This is the most important thing.