2006 Guidelines for Annotation of Within-document NP Coreference
(L.Hasler in discussion with K. Naumann and C. Orasan: 30.01.2006)

General Strategy

- Prior to annotation, read the whole text to familiarise yourself with it.

- Make a note of all troublesome or ambiguous cases and discuss them with other annotators to decide upon the best solution to tackle them.

- Ensure that the annotation is done in one intensive period, as sporadically annotating a file can lead to the annotator having to re-read the document for familiarisation several times and to a lack of accuracy.

- Using the program you will perform a two-pass annotation process in which all markables (suitable NPs) are identified in the first pass and all coreferential links and attributes are assigned in the second pass.

- Having completed the annotation, check through it to see if there are any mistakes or additional problematic cases you may have missed before.

Markables

- Markables do not have to be annotated in linear order, i.e. it is possible to annotate a markable you may have missed in the first stage of annotation when assigning coreferential relations in the second stage. However, it is easier and more logical to annotate all markables in the first stage and in the order they appear in the text, marking first larger NPs then other NPs within them.

Newswire texts:
- As the files we are annotating are newswire texts, they usually contain two "headlines", one with a location followed by a colon and then the headline itself, usually appearing first, and one without this. DO NOT annotate the HEADLINE INCLUDING THIS LOCATION, annotate the one without it.

- Again, due to our newswire genre, the files usually contain a dateline, either at the beginning or end of the whole text, and the name of the author, sometimes with contact details. DO NOT annotate these parts of the text.

- Occasionally a document may contain text from partly deleted tables and captions. DO NOT annotate ELEMENTS THAT ARE IN INCOHERENT LISTS in the document.

- In short, markable elements should only appear in the "headline" and paragraphs.

NPs:

- Make sure you mark NPs at ALL levels, both DEFINITE AND INDEFINITE, from base to complex and co-ordinated, including all the noun phrases involved in events, regardless of whether they are coreferential or not. This is done in the first pass (see above). The next sections tells you when to annotate NPs as coreferential and when not.

- Include all modifiers of an NP in the markable. These include subordinate clauses (relative clauses and clauses introduced by gerunds/verbal participles), bracketed text, appositives, text between dashes, quantifiers (including negative quantifiers) and any other modifier types not listed here. Mark the whole NP including all modifiers and then mark appropriate modifiers within the larger markable as separate markables. Note that not all modifiers will be markables. DO NOT mark the NP being modified as a separate markable (in this case, *Three Israeli women*). E.g. [*Three Israeli women killed by* [*a suicide bomb in* [*a Tel Aviv café*]] *on* [*Friday*]]

- Mark all possible EMBEDDED NPs within a larger NP: e.g. [*passengers on* [*a flight from* [*Moscow*] *to* [*Nigeria*]]]

- Mark COORDINATED NPs as follows: [[***McVeigh***] **and** [***Nichols***]] *harboured* [*anti-government sentiment*]

- Annotate GERUNDS which are true NOMINALISATIONS of verbs (i.e. do not introduce a subordinate clause, are preceded by an article and are in the subject/object position): e.g. *Taleban information minister Amir Khan Mutaqi said* [***the fighting***] *took place in the Lolenj district…*

- Annotate numerals, dates and quantified NPs as markables.

- Annotate possessive pronouns and other possessors (e.g. [***Kabila's***] *forces*) as markables. Note that you cannot annotate the possessor without the *'s*. For possessive constructions containing *of*, annotate the whole construction as a markable and then any definite NP following the *of* within this as a separate markable: e.g. [*the latest victims **of*** [*the Palestinian-Israeli struggle*]]

- Annotate interrogative pronouns functioning as possessives, but not those functioning as relative pronouns: e.g. [*President Alberto Fujimori,* [[***whose***] *brother Pedro*] *is* [*one of* [*72 men still being held by* [*the armed Marxist guerrillas*]]]]*…*

- Annotate reciprocal pronouns (*each other*, *one another*) as markables.

- DO NOT mark RELATIVE PRONOUNS OR RELATIVE CLAUSES as markables in their own right. But remember that relative clauses can contain NPs which ARE markables and that relative clauses can modify other markable NPs.

- DO NOT annotate GERUNDS which function as a verb: e.g. ...[*Timothy McVeigh*] *accused of **blowing up** [a federal building]...*

- DO NOT annotate NPs which are part of fixed expressions, idioms, compounds and multi-word lexemes as markables: e.g. *in town*, *on board*, *came to power*, *bring to justice*.

- DO NOT mark *HERE* and *THERE* as markables, as they are not NPs.

Annotating coreference

Always link coreferential items back to the longest match of either the first mention or most recent mention (depending on the type of text under consideration – see below for more details).

'Do'…
- Annotate only NOMINAL IDENTITY-OF-REFERENCE DIRECT ANAPHORIC EXPRESSIONS as coreferential with their antecedents if they refer to the same entity in the real world. You should annotate NPs in the form of pronouns, definite descriptions and proper names: e.g. *it/he, the airport, Sam Rainsy* etc.

- Annotate INDEFINITE NPs which are COREFERENTIAL WITH DEFINITE NPs IN THE HEADLINE as coreferential. This goes against the standard rule of not annotating indefinite NPs as coreferential, but due to the nature of newswire texts (i.e. that NPs in the headline are usually definite, but then may be introduced again as indefinite in the first paragraph. As an example, in the following extract, *blast, an explosion* and *the blast* should be annotated as coreferential, with *blast* in the headline as the first mention: [***Blast***] *kills man* (headline)...[***an explosion***] *killed one man...*[***the blast***]*...* In cases such as this, ALWAYS INSERT A COMMENT that although the NP is indefinite, it is coreferential with a definite NP in the headline.

- Annotate DEFINITE DESCRIPTIONS which stand in the following relationship with the antecedent as coreferential:

  - identity (same head: *the blast - the 4 a.m. blast*; pronouns: *Hun Sen - he*)
  - synonymy (*the protest - the demonstration*)
  - generalisation (*survivors and family members of those killed - victims*)
  - specialisation (*victims - survivors and family members of those killed*)[1]

  For this type of coreferential link, the anaphor should be linked back to the **first mention** of the NP in the document. Assign the appropriate relation attribute (i.e.

---

[1] The terms *generalisation* and *specialisation* used here are concerned with lexical choice and detail rather than concept. This is an important distinction as generalisation and specialisation of concept (e.g. *the house...the door*) are used in indirect anaphora, which we do not consider to be coreferential for our purposes. We use the terms to denote the level of detail present in one NP in relation to another with which it corefers.

ident, **synonym**, **generalisation** or **specialisation**). The reference attribute **NP** should be assigned to the link. Note that it can be difficult to assign relation attributes. WordNet has been added to PALinkA to make this easier. There is the option to assign **other** if there is a problem/you really cannot decide.

- Annotate DEFINITE NPs IN COPULAR RELATION as coreferential: e.g. [*the blast*] *was* [***the*** *worst attack on* [*civilians*] *on* [*U.S. soil*]]. For this type of coreferential link, the anaphor should be linked back to the **nearest antecedent** in the document, i.e. the antecedent which appears closest to the anaphor. Assign the appropriate relation attribute from the list: **ident**, **synonym**, **generalisation**, **specialisation**. The reference attribute **copular** should be assigned to the link.

- Annotate DEFINITE APPOSITIVES (the description after the comma) as coreferential with the NP they apply to: e.g. [*Zaire Airlines,* [***the*** *main commercial airline in* [*Zaire*]]]. Here, mark the whole NP first, then mark the appositive statement after the comma, then mark this as coreferential with the whole NP. For this type of coreferential link, the anaphor should be linked back to the **nearest antecedent** in the document. Assign the appropriate relation attribute from the list: **ident**, **synonym**, **generalisation**, **specialisation**. The reference attribute **apposition** should be assigned to the link.

- Treat TEXT IN BRACKETS and TEXT BETWEEN DASHES after an NP as above (as long as it definitely refers to the NP, of course): e.g. [[*Hun Sen*]*'s Cambodian People's Party* [*(CPP)*]]. For this type of coreferential link, the anaphor should be linked back to the **nearest antecedent** in the document. Assign the appropriate relation attribute from the list: **ident**, **synonym**, **generalisation**, **specialisation**. The reference attribute **bracketed_text** should be assigned to the link.

- Annotate 1st and 2nd person pronouns *I*, *we* and *you* appearing in speech as coreferential with their antecedents. It is important that you DO NOT MARK these pronouns AS THE FIRST MENTION in a coreferential chain: the antecedent can appear BEFORE (anaphora) OR AFTER (cataphora) these pronouns, and the pronoun must be linked either backwards or forwards to its antecedent. So, in the sentence [[***IATA***]*'s director of security services*] *said, "*[***We***] *consider that* [[***Aeroflot***]*'s air security measures*] *correspond to* [*international standards*].*"*, annotate *We* as coreferential with *IATA*. For this type of coreferential link, the pronoun should be linked to the nearest antecedent, often the reporting clause. Assign the relation attribute **ident** from the list. The reference attribute **speech_pron** should be assigned to the link. In the case of *we*, the pronoun should be annotated as coreferential with the organisation/group etc. the person is speaking on behalf of as they are representing the views of that organisation/group.

- Annotate possessors (including possessive pronouns) as coreferential with their first mention in the text: e.g. [*Protest leader and apparent target Sam Rainsy*]...[***Sam Rainsy's***] *wife*...

- Annotate interrogative pronouns functioning as possessives as coreferential with their first mention in the text. In the following example, ***whose*** should be annotated as coreferential with the whole NP as this is the first mention and the longest match: [*President Alberto Fujimori,* [[***whose***] *brother Pedro*] *is* [*one of* [*72 men still being held by* [*the armed Marxist guerrillas*]]]]...

- Annotate reciprocal pronouns as coreferential with their first mention in the text: e.g. [*The two sides*] *have been in contact with* [***each other***] *since* [*Saturday*]

- It is possible to annotate cataphoric references as coreferential but there is no separate tag for this. If you find a cataphoric reference, link it forwards to its antecedent, select the attribute tags in the usual way and add a note in the COMMENT box that the reference is cataphoric. E.g. [***It***] *killed* [*168 people*] *and injured* [*hundreds more*]. [***The attack on*** [***the Alfred P. Murrah building in*** [***Oklahoma City***]]] *was* [*the worst attack on* [*civilians*] *in* [*US history*]]

- Annotate NPs at all levels, from base to complex and coordinated, including NPs embedded in larger NPs.

- Remember that definite noun phrases do not always have a definite article present, but should still be marked as coreferential with their antecedent: e.g. in the following sentence there is no article preceding *head*, but this NP (appositive) is still coreferential with *Mak Chito*: [*Mak Chito,* [***head*** *of* [[*the Phnom Penh police's*] *serious crimes office*]]]

How to annotate coreferential links

- In the first pass, all appropriate NPs should be marked as markables.

- In the second pass, coreferential links between NPs should be annotated using the **coref** relation and appropriate attributes (for relations and references) assigned for each link according to its type.

- If you are unsure whether one NP is coreferential with another, use the **ucoref** relation and then assign attributes as in the case of the **coref** relation. You should specify whether the uncertainty lies with the annotator or with the text (see **Tricky cases** near the end of the guidelines for more details).

- The following types of coreference should be linked back to the **first mention of the NP in the document**: NP (this includes pronouns although there is no separate label for this type).

- The following types of coreference should be linked back to the **nearest antecedent in the document** (i.e. the antecedent which appears closest to the anaphor): copular, appositive, brackets, speech pronouns.

- Sometimes you may need to mark a **zero element** in the document, for example, where the head or modifier of a noun phrase is not present (see **Tricky cases** for an example), so that a coreferential link can be annotated correctly. If this happens, you should annotate the zero element, using the zero tag, BEFORE marking the markable and assigning the coreferential link and link attribute. The zero element cannot be annotated after the markable has been marked. You should then indicate which element is "missing" by inserting it into the box. Zero elements should ONLY be inserted where not doing so would impede the annotation of coreferential links between two elements.

- There is ALWAYS the option to add a **comment** to each annotation, should you encounter any difficulties/wish to record any observations.

- There is the option to select **other** for any attributes you cannot assign a relation or reference type to.

'Do not'…
- Do not annotate any INDEFINITE NPs as coreferential, but remember that they ARE markables. However, bear in mind the one exception to this rule: indefinite NPs that corefer with a definite NP in the headline (see **'Do…'** for more details).

- Do not annotate INDEFINITE PREDICATE NOMINALS as coreferential with the subject they apply to: e.g *"…*[*the blast*] *doesn't look like* [***an accident***]*,"*. Here, mark the NP *the blast*, then mark the NP *an accident*, but don't mark this as coreferential with *the blast*.

- Do not annotate INDEFINITE NPs IN COPULAR RELATION as coreferential in the same way described for definite ones. Instead, mark the first whole NP, then mark the indefinite NP, but do not corefer this back to the whole NP: e.g. With [*the dead man*] *was* [***a retired employee of*** [*the state telecommunications company*]], mark the NP *the dead man*, then *a retired employee of the state telecommunications company*, but do not mark them as coreferential.

- Do not annotate INDEFINITE APPOSITIVES as coreferential, but do still mark them as NPs: e.g. do not corefer *The dead man* with *a retired Zairean telecommunications worker* in the phrase [*The dead man*]*,* [***a*** *retired* [*Zairean telecommunications*] *worker*]*…*

- Treat text in brackets and text between dashes after an indefinite NP as above.

- Do not annotate IDENTITY-OF-SENSE ANAPHORA (which includes one anaphora) as coreferential: e.g. [***The attack*** *on* [*Madrid*]] *killed more than* [***the one*** *on* [*London*]]

- Do not annotate INDIRECT ANAPHORA (also called bridging or associative anaphora), i.e. PART-OF and SET MEMBERSHIP (including SUBSET-SET) RELATIONS between the anaphor and the antecedent as coreferential: e.g. *Moscow - Russia*; *the bombers – McVeigh*; *four grenades – at least one*

- Do not annotate RELATIONS THAT CAN BE 'POTENTIALLY' REGARDED AS COREFERENTIAL, but do still mark them as NPs: consider 'definite' coreference only: e.g. in [*they*] *also put* [*the toll*] *at* [***up to 20***], do not corefer *the toll* with *up to 20* as this number is not definite.

- Do not annotate BOUND ANAPHORS as coreferential with their antecedent. A bound anaphor has a quantified NP as its antecedent. In the following example, *there is* not coreferential with *many Bukavu residents*: [*He*] *said* [***many Bukavu residents***] *were fleeing* [*the city*] *with* [*belongings*] *on* [[***their***] *heads*]...

- Do not annotate DISJOINED NPs, where two NPs are mentioned separately but then mentioned in the same NP later, as coreferential, e.g. if *McVeigh* and *Nichols* are mentioned separately, and then later as *McVeigh and Nichols*, do not corefer *McVeigh and Nichols* back to any of the previously mentioned single NPs *McVeigh* or *Nichols*.

- Do not annotate DIFFERENT READINGS OF AN NP as coreferential. The most common example of regular polysemy in our texts is the use of the name of a country as a geographical entity and as a government/authority. In the following example, the two mentions of *China* should NOT be annotated as coreferential as they refer to two different entities. [*A jobless Taiwanese journalist who commandeered* [*a Taiwan airliner*] *to* [***China***]]... [***China***] *ordered* [[*its*] *airports*] *to beef up* [*security*]...

- Do not annotate CROSS-DOCUMENT COREFERENCE.

Tricky cases…

- If there is a CONJOINED NP such as '[[*the fire*] *or* [*the ambulance service*]]' (2 NPs conjoined just 1 head present) and then later a reference to *the fire service*, first mark the whole conjoined NP, then mark the individual NPs within it, and then mark the later reference to *the fire service* as coreferential with the first element of the conjoined NP *the fire*. A **zero element** should be marked after *fire* (representing *service*) to enable the later reference *the fire service* to be linked correctly (see **How to annotate coreferential links** above).

- If there is an NP of the following kind, '[[[*the bomber*]*'s age*] *or* [*nationality*]]' (2 NPs conjoined - both heads present) then a subsequent mention of *the bomber's nationality*, treat the initial mention of *nationality* as a kind of ellipsis and corefer the second mention back to just the word *nationality*. A **zero element** should be marked

before *nationality* (representing *the bomber's*) to enable the later reference to be linked correctly.

- If there is a case where *his/her* or *his or her* is used to refer to an entity in the text, then treat it as one single NP and corefer it back to the relevant NP: e.g. in *If [**a citizen**] feels [[**his or her**] life] is in [danger]*, corefer *his or her* back to *citizen*.

- Sometimes the annotator may be unsure whether to annotate an NP as coreferential or not. This can be due to either uncertainty from the annotator themselves or from the text. For this purpose, the **ucoref** relation is available. As an example of uncertainty in the text, in the sentence [*The government*] *will argue that…* [[***McVeigh***] ***and*** [***Nichols***]] *were* [***the masterminds of*** [***the bombing plot***]], the verb *argue* may add uncertainty to the "objectivity" of the NP *the masterminds of the bombing plot*. The annotator should mark this NP as **ucoref** with *McVeigh and Nichols*, assign the relation and reference attributes as usual, and finally select whether the uncertainty was with the annotator or the text.