ParCor 1.0: Pronoun Coreference Annotation Guidelines

Liane Guillou[1], Christian Hardmeier[2], Aaron Smith[2], Jörg Tiedemann[2] and Bonnie Webber[1]

[1]University of Edinburgh
[2]University of Uppsala

March 21, 2014

Contents

1 Note

These guidelines were presented to the English and German annotators who worked on the annotation of the EU Bookshop and TED Talks texts in the ParCor 1.0 corpus. The document is split into three main sections: General guidelines (applicable to both text genres) and additional guidelines specific to the annotation of the EU Bookshop and TED Talks portions of the corpus respectively.

2 Pre-populated Markables

In order to assist the annotation process pre-populated MMAX-2 *markables* are provided as a starting point to the manual annotation. These markables represent:

- Pronouns: These will first appear as bold blue text with a magenta background and the background colour will disappear once you have determined a new *type* for them – e.g. anaphoric, pleonastic or event.
- Noun Phrases (NPs): A set of potential antecedents for pronouns. These will appear as normal blue text (as for any other markable in MMAX-2).

Please note the markables were produced by automated tools and may not be 100% accurate. You should look for errors such as:

- Pronouns that have not been identified (and therefore are not labelled as markables)
- Words that have been mis-labelled as pronouns
- Potential pronoun antecedents (the set of NPs) which may be missing and therefore need to be added (manually), or ones where their span may be too large or small and therefore needs adjusting (manually)

3 General Guidelines: What to Include

We wish to construct links between pronouns and their antecedent(s). These linked pronoun-antecedent pairs should exclude events (and references to the events) and pleonastic/dummy pronouns (see Section 4 on what to exclude).
The following set of guidelines has been condensed from the MUC-7 guidelines[1].
Pronoun *forms* to be annotated, include:

- Personal: First, second and third-person
- Possessive
- Demonstrative
- Relative
- Reflexive (TED Talks only)
- Pronominal adverbs (EU Bookshop only)
- Generic

The possessive forms of pronouns used as determiners are markable. Thus in:

> The company and [**its chairperson**]

there are two potentially markable elements: **its** and the entire NP, **its chairperson**.
First, second, and third-person pronouns are all markable, so in:

> "There is no business reason for [**my**] departure", [**he**] added.

my and **he** should be marked as coreferential.

3.1 Anaphoric and Cataphoric Pronouns

We are interested in marking pronouns (e.g. he, she, it, they, us,...) and their *antecedent* (the thing that the pronoun refers to). For example, in the following example:

> [**Alan Turing**]$_1$ was born at Paddington, London. [**His**]$_1$ father, [**Julius Mathison Turing**]$_2$, was a British member of the Indian Civil Service and [**he**]$_2$ was often abroad...

the pronoun **His** refers to **Alan Turing** - in other words the antecedent of **His** is **Alan Turing**· The pronoun **he** refers to **Julius Mathison Turing**, not to **Alan Turing**.

In some cases a pronoun can refer to more than one entity. Consider the following example in which the pronoun **They**, refers to two people: **John** and **Mary**, captured in the single conjoined NP antecedent **John and Mary**:

[1] `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html`

[**John and Mary**] went to the cinema. [**They**] saw a film about penguins.

When a pronoun appears after its antecedent/referent in a text we call this *anaphora* (the relationship is *anaphoric*). The pronouns in the above examples are anaphoric. When a pronoun appears before its referent in a text we call this *cataphora* (the relationship is *cataphoric*). The pronoun **she** in the example below is cataphoric:

If [**she**] is in town, [**Mary**] can join us for dinner.

We are only interested in cataphoric relations in which the pronoun and its referent occur in the same sentence. Also, consider the following rule for deciding if a pronoun is anaphoric/cataphoric: if the pronoun can be marked as anaphoric, mark it as such. If no possible antecedent appears before the pronoun, then consider linking it as cataphoric. (We will use the term *antecedent* to refer to the NP that either a cataphoric or anaphoric pronoun refers to.)

3.2 Speaker/Addressee Reference Pronouns

These are defined as:

- *Addressee reference*: Where the pronoun primarily refers to the addressee (person being addressed)
- *Speaker reference*: Where the pronoun primarily refers to the speaker or may not include the addressee

As a guideline:

- First-person pronouns normally refer to the speaker, in the case of the singular (e.g. the English "I"), or to the speaker and others, in the case of the plural (e.g. the English "we")
- Second-person pronouns normally refer to the person or persons being addressed (e.g. the English "you"); in the plural they may also refer to the person or persons being addressed together with third parties
- Third-person pronouns normally refer to third parties other than the speaker or the person being addressed (e.g. the English "he", "she", "it", "they")
- Plural pronouns like "us", "you" and "we" may be more difficult. "we", "us" and "our" will most likely be speaker reference, and instances of "you" and "your" will likely be addressee reference
- Be aware that it may be difficult to distinguish between *speaker reference* and *addressee reference* in some cases

3.3 Pleonastic Pronouns

These are pronouns that do not actually refer to an entity. In other words, the pronoun could not be replaced with an NP as with a regular pronoun. Often a *subject* is required by syntax i.e. something is required in that position. In some cases there will not be a *subject* so a "dummy" pronoun is required to fill the gap. For example in the following sentences the pronoun **it** does not refer to anything but is included as something is required by the syntax of the language in the subject position:

- **It** is raining
- **It** is well known that apples taste different from oranges

It is commonly used as a pleonastic pronoun in English. Other pronouns such as **they** and **you** may also be used in cases where they do not refer to a specific entity:

- In this country, if **you** own a house **you** have to pay taxes
- **They** say you should never mix business with pleasure

In the case of *pleonastic* pronouns we wish to make a partial annotation: Marking the pronoun as *pleonastic*, but not linking it to anything (because it does not refer to anything).

3.4 Identifying the Antecedent(s)

Once an *anaphoric* or *cataphoric* pronoun has been identified a pronoun, its *antecedent* needs to be determined. There are several cases. The pronoun may refer to:

- An entity (represented by a noun or NP)
- An event (see Section 4.1)
- Nothing (see Section 3.3)
- It may be possible to tell that a pronoun is anaphoric, but there is no specific antecedent in the text. For example the pronoun **these** in "Access to 0800 numbers...**these** calls" (see Section 3.7)
- A word may have been marked as a pronoun in error (i.e. the automated pre-processing pipeline made an incorrect choice)

In order to identify what a pronoun refers to, the pronoun itself should be used as a starting point. Look back earlier in the text (working backwards sentence by sentence) until the nearest non-pronominal antecedent is identified. For example, in:

> The details of [**Miyamoto Musashi**]'s early life are difficult to verify. [**Musashi**] simply states in Gorin no Sho that [**he**] was born in Harima Province

the pronoun **he** should be linked to **Musashi**, the nearest antecedent, and not to **Miyamoto Musashi** which appears earlier in the text.

3.5 Special Case: Pronoun has Multiple Antecedents

In cases like:

> [**John**] likes documentaries. [**Mary**] likes films about animals. The last time [**they**] went to the cinema [**they**] compromised and saw a film about penguins.

They refers to both **John** and **Mary**, who are mentioned in separate sentences so there is no NP span that covers both **John** and **Mary**. In cases like these, if all of the antecedents can be identified and it is clear from the texts what the antecedent are, the pronoun should be linked to each of the separate antecedent "parts". It is important to ensure that all "parts" are linked.
It is important to first ensure that no NP exists that covers all parts of the antecedent.

3.6 Special Case: They

When the pronoun **they** is used to refer anaphorically to a collective noun (such as **the government**), it should be considered a plural pronoun and marked as such.

3.7 Special Case: No Specific Antecedent

For example, in the sentence:

> There's a study called the streaming trials. **They** took 100 people and split them into two groups

There is no antecedent in the text tow which the pronoun **They** may be linked. In this case, the pronoun should be marked as "anaphoric but no specific antecedent".

3.8 Special Case: "he or she", "him or her", "his or her" and "his or hers"

English lacks ungendered person pronouns, and the former solution of just using "male pronouns" (e.g. "he") is now considered bad form. Therefore, you may come across instances of "he or she" in a text. For example:

> If your child is thinking about a gap year, [**he or she**] can get good advice from this website.

In such cases, **he or she** should be considered a single unit (or markable), just as if it had been written "s/he" (which is a common alternative). This solution will also make the phrase easier to resolve, since it can only be linked to a non-specific antecedent.
The same applies to instances of "him or her", "his or her" and "his or hers".

3.9 Special Case: "s/he"

Treat this as a complete unit (or markable) and as a pronoun.

3.10 Special case: The Pronoun Refers to a Modifier

In some cases, the pronoun may refer to a modifier in an NP. Consider the following example:

> The unionists used to be [**EU supporters**], but now they are questioning how [**it**] has developed...

Here, the pronoun **it** cannot be linked to the complete NP **EU supporters**, but **it** can be linked to **EU** (the modifier). If none of the automatically generated markables are suitable, the span of an existing markable should be adjusted or a new markable created. The resulting markable may or may not be an NP. However, with compounds like EU-supporters, these exist as a single unit and cannot be split any further (i.e. it is not possible to construct a markable that covers only the **EU** part). In such cases, it is necessary to search for a standalone instance of **EU** earlier in the text and link the pronoun **it** to that instance (assuming one can be found).

3.11 How Much of a Markable to Annotate

A markable is any pronoun, noun or NP that will be "marked" because it forms part of pronoun-antecedent pair, or a pronoun for which there is no antecedent to be marked. For pronouns, the markable will be a single word. For a pronoun's antecedent(s), the markable will be a noun or an NP. For noun or NP markables, the following rules apply. The markable must:

- Contain the head (main) noun
  - E.g. **task** is the head in **the coreference task**
  - If the head is name then the entire name (not just a part of it) should be marked. E.g. **Frederick F. Fernwhistle Jr.** in **the Honorable Frederick F. Fernwhistle Jr.**
- Also include all text which may be considered a modifier of the NP
  - E.g. **the Honorable Frederick F. Fernwhistle Jr.**
  - E.g. **Mr. Holland**
  - E.g. **the coreference task** (where **task** is the head) – this provides information about what the task is and separates it from **other coreference tasks**, **the scheduling task**, etc.
  - E.g. **the big black dog** (where **dog** is the head)
  - Determiners such as **the** should be included

N.B. The automatically generated set of markables may contain NPs that have incorrect spans. The spans may therefore require manual adjustment.

3.12 Relationships Between Markables

For a pronoun and its antecedent(s), the relationship between the elements is termed as *anaphoric/cataphoric*. For *pleonastic* and *event* pronouns there will not be a link to an antecedent markable.

4 General Guidelines: What to Exclude

4.1 The Events in Event Reference

The events in event reference — where pronouns are used to refer to an event that has happened or will happen, should not be marked. Event pronouns can refer back to whole sections of text or concepts evoked by the text. For example in:

> Ted [**arrived late**]. [**This**] annoyed Mary.

This refers to the event **arrived late**.
Another example:

> Vulnerable consumers in particular might need [**specific support**] to enable them to finance necessary investments to reduce energy consumption. [**This**] task...

Using deictics that vaguely refer to what the speaker is talking about (as in the above example) is bad writing, but examples like this exist in some of the texts. Here **this** should be treated as an instance of event reference.
In general, events should be easy to identify as they should contain verbs. As with the annotation of *pleonastic* pronouns a partial annotation is required: The pronoun is marked as *event*, but is not linked to the event itself.
Identifying pronouns that refer to events can be difficult, therefore the following simple rule is proposed:

- *English:* Try replacing the pronoun with a period and then start a new sentence *or* test if you can replace an instance of **which** with **this**
- *German:* Try replacing the pronoun with a period and then start a new sentence with **das**

If the resulting "new text" reads OK, then it is likely that the pronoun refers to an event. As an example of how this test would work, consider the following sentence:

> Ted arrived late, [**which**] annoyed Mary.

Question: Is **which**" an event pronoun?
Replace the pronoun **which** with a period and start the new sentence with **This**:

> Ted arrived late[**. This**] annoyed Mary.

Result: Mark **which** as an event pronoun as the "test" passed.
If two pronouns refer to the same event, each should be marked as an *event* pronoun (as opposed to marking the second as anaphoric to the first) and the two instances linked together.

5 Special Instructions for the Annotation of Written Text: EU Bookshop

The following instructions are specific to the annotation of written text and should be used when annotating the EU Bookshop documents.

5.1 Reflexive Pronouns

Reflexive Pronouns should not be marked.
In cases like "the man himself" we do not treat **himself** as a pronoun. Instead it should be considered an NP (the markable span can be amended in MMAX-2) if it has been automatically marked as a pronoun in error.

5.2 Indefinite Pronouns

An indefinite pronoun is a pronoun that refers to one or more unspecified beings, objects, or places. For example:

> [**Anyone**] can see that she was looking for trouble.

Here, **Anyone** is an indefinite pronoun as it does not refer to a specific person or group of people.
Indefinite pronouns should be marked as *pronoun*, to indicate that they have been "seen" in the text. As they will be marked as instances of the type *pronoun*, they will not be linked to anything, nor will any other features be recorded.

5.3 Numbers/Quantifiers Used as Pronouns

When deciding whether to link a pronoun to an *antecedent*, the following rules apply:

- many of **them** ...: **them** should be linked to its antecedent
- **one** of the fast growing economies: **one** should be marked as a pronoun but not linked to anything
- **others** ...: **others** is anaphoric and has an antecedent, but it is not coreferent with its antecedent. It should be marked as a pronoun but not linked to anything
- **both**: This is anaphoric, either to two individuals or two events or situations. If **both** here is a *bare* pronoun, it should be marked and linked. If it has a head (as in "both boys"), then it should be marked as a pronoun but not linked to anything
- **each**: This is anaphoric to a set. If **each** here is a *bare* pronoun, it should be marked and linked. If it has a head (as in "each boy"), then it should be marked as a pronoun but not linked to anything

5.4 Pronominal Adverbs

Pronominal Adverbs are a type of adverb occurring in both English and German (although they appear to be used more frequently in the German texts). They are formed by replacing a preposition and a pronoun. We wish to annotate these.
For example:

- For that → therefore
- In that → therein

- By this → hereby
- To this → hereto
- In which → wherein

5.5 Pronouns Within Quoted Text

Identifying whether a first-person or second-person pronoun within quoted text can become difficult. Furthermore, the focus is on translating coreference in *normal* text, not *quoted* text. We therefore simplify the annotation task using the following rules:

- First and second-person pronouns within quoted text should simply be marked as instances of type *pronoun*
- Third-person personal pronouns should be marked as normal
- In some cases, the text may read like an interview (with questions and answers) but with no quotes. In this case, the text is not to be treated as quoted text. Speaker/addressee reference pronouns should be annotated as normal.

5.6 Difficult Choices: Deciding Between Anaphoric or Event Categories

In some scenarios is it possible to read the text in more than one way and both readings appear to be equally likely. For example, it may be possible to mark the pronoun as either *event reference* (referring to a phrase with a verb) or *anaphoric* (referring to an NP), i.e. it is ambiguous. As an example, consider:

> In the framework of the North Seas Countries' Offshore Grid Initiative, ENTSO-E is already conducting grid studies for northwestern Europe with a 2030 horizon. [**This**] should feed into ENTSO-E's work for a modular development plan of a pan-European electricity highways system up to 2050.

In this example, the pronoun **This** could refer to:

- North Seas Countries' Offshore Grid Initiative (NP)
- conducting grid studies for northwestern Europe with a 2030 horizon (Verb Phrase)

In scenarios such as these, if multiple labels would be possible, select *anaphoric* and link the pronoun to the NP. This will provide more information when the data is used for training translation systems.
If it is *impossible* to tell what the pronoun refers to or if the text is very poorly written, the pronoun may be marked as *Not sure. Help!*. This will help to identify those scenarios that are very difficult for humans (and therefore even more difficult for machines) to determine.

6 Special Instructions for the Annotation of Spoken Text: TED Talks

The following instructions are specific to the annotation of transcribed spoken text and should be used when annotating the TED Talks documents.

6.1 Reflexive Pronouns

Reflexive pronouns should be annotated in English and German.
For English:

- Exclude instances of **myself** from the annotation as it is a singular first-person pronoun

For German:

- Include instances of **mich** even though it is a singular first-person pronoun as it can be reflexive *or* personal and it is important to make the distinction
- The pronouns **mich**, **dich**, **uns** and **euch** can all be used as either personal or reflexive pronouns. Mark whether they are personal or reflexive

6.2 First-person Pronouns

Singular first-person pronouns (I, me, etc.) do not need to be marked as they can be recovered automatically.

6.3 Speaker Reference

For pronouns that fall into the *speaker reference* category (used for all instances of **we**), the audience should be recorded. There is an *audience* attribute (in MMAX-2), which can be set as either:

- *Exclusive we*, meaning the speaker and his/her clique but not the audience
- *Co-present we*, meaning the speaker plus everyone physically present in the room
- *All-inclusive we*, incorporating everything else

6.4 Addressee Reference

For pronouns that fall into the *addressee reference* category, the audience should be recorded. There is an *audience* attribute (in MMAX-2), which can be set as either:

- *Deictic*, meaning that the speaker is really referring to the audience or a specific person
- *Generic*, as in phrases such as: In England, if [**you**] own a house [**you**] have to pay taxes

When a speaker uses deictic **you**, talking to the whole audience, it should always be marked as plural, even in cases like "Imagine **you**'re walking alone in the woods", where there is clearly a singular sense to the word.
For generic cases of **you**, it is not necessary to make a singular vs. plural distinction.
N.B. **you** should not be labelled as as pleonastic

6.5 Pronouns Within Quoted Text

These pronouns should be annotated strictly from the point of view of the quoted speaker, not of the speaker who quotes the utterance. In particular, this means:

- First-person pronouns are always speaker reference
- Second-person pronouns are always addressee reference
- A coreference relation is never marked between a first-person or second-person pronoun inside quoted speech with a pronoun outside the quoted speech passage (as in ' He said, "I do." ', where **he** and **I** could arguably be marked as coreferent)

In examples such as:

> I said, "[**Miguel**], what makes your fish taste so good?" [**He**] pointed at the algae.

Do not link the pronoun **He** (outside of the quote) to **Miguel** (inside of the quote). Instead, look for an earlier instance of the entity (i.e. Miguel) in the text that does not appear in quotes and link the pronoun (i.e. **He**) to that instance.

6.6 Extra-Textual Reference

For cases where the speaker refers to something such as a slide or prop, the pronoun should be marked as *extra-textual*. Two pronouns referring to the same object should both be marked as *extra-textual* and linked together as co-referents.
The *extra-textual* category can also be used within quoted text when a third-person is referred to such as the **he** in:

> People when they see me say "[**he**]'s a bit weird"

N.B. This is rarely required

6.7 No Explicit Antecedent

In cases like:

> There's a study called the streaming trials. [**They**] took 100 people and split them into two groups

where there is no explicit antecedent for **They**, the pronoun should be marked as *anaphoric* and the *no explicit antecedent* sub-category should also be selected. Do not mark **They** as pleonastic.

6.8 Split Antecedent

This should be marked if the pronoun has multiple antecedents. All components of the antecedent should be linked to the pronoun directly, and not to each other.

6.9 Simple Antecedent

For all cases except where there is *no specific antecedent* or there is a *split reference*.

6.10 Indefinite Pronouns, Pronominal Adverbs and Numbers/Quantifiers Used as Pronouns

Instances of these pronouns should not be marked.