Coreference Corpus Annotation Guidelines

Ekaterina Lapshinova-Koltunski, Christian Hardmeier

May 8, 2018

Contents

1 Introduction

These guidelines present instructions for the English and German annotators who work on the annotation of full coreference chains of the parallel English-German corpus. The parts of the corpora include texts from EU Bookshop and TED Talks texts and were included into the ParCor 1.0 corpus. These guidelines base on the guidelines for pronoun annotation described by [GHS+14], combined with those by [GS16]. Some categories are based on the descriptions by [Kun12] and [NM09].

2 Markables

2.1 Types of Markables: What to include

2.1.1 Full NPs

For example: *24 EWSA-Mitglieder* in (1-a) and der Aktivitäten in (1-b).

(1) a. *In diesem Buch berichten [24 EWSA-Mitglieder] über [ihren] Beitrag als Geschäftsleute und Gewerkschafter, Aktivisten und Freiwillige...*

b. *Eine Liste [der Aktivitäten], [die] man der Kategorie "aktive Bürgerschaft" zurechnen könnte, würde sehr umfangreich ausfallen – in [ihrer] Gesamtheit bilden [sie] das, was eine gesunde, partizipative Demokratie ausmacht.*

Definite article (from GECCo)

Definite articles signal a coreference relation, if the meaning of the nominal phrase (used with the articles) refers to an antecedent and triggers an identity relation. And the distance between the elements is not important here.

Here, we should differentiate between the case when the definite article is used with a nominal phrase but no coreference relation is triggered, as in (2-b). (2-a) represents the case when the definite NP is coreferent with the antecedent in the previous sentence.

(2) a. *This past spring, the U.S. Department of Education issued [a report, The Condition of Education 2000]. [The report] found that the benefits of attending college are greater today than ever before. [...]*

b. *Insgesamt stellen sich die ökonomischen Voraussetzungen zu Beginn des neuen Jahrhunderts für Deutschland recht gut dar. [Die Teilhabe] an den Weltmärkten hat sich verbessert – dies belegt die positive Dynamik der Exporte und Dienstleistungen.*

In example (3), all nominal phrases, no matter if they are expressed by the same word, refer to the same antecedent.

(3) *Unfortunately, the Bangladesh health system is unprepared for [a crisis of this magnitude]... [the arsenic problem]... [the crisis]... [the dilemma]... [the arsenic problem] .*

2.1.2 Proper names and Titles

Herr Almeida Freire in (4-b).

(4) a. *[Der EWSA] hat stets betont, dass aktive Bürgerschaft die gesellschaftliche Integration von Kindern und Jugendlichen unterstützen und ihnen das Gefühl vermitteln kann, ein Teil der Gemeinschaft zu sein.*

b. *In [seiner] EWSA-Stellungnahme zum im Januar 2011 veröffentlichten "Bericht der Kommission zur Beobachtung des Handelsmarktes" schreibt [Herr Almeida Freire]...*

2.1.3 NPs with quantifiers

Be careful when annotating NPs with quantifiers, e.g. *all people, two people, 105 Million euro*, etc. If you are not sure about the definiteness of an NP, apply the following test: try inserting a definite article or a demonstrative pronoun. If the meaning of the phrase is not changed, then the NP is definite and should be annotated as coreferent. Example: *all people* → *all these people* ⇒ definite NP [GS16, p. 3].

When deciding whether to link a pronoun to an antecedent, the following rules apply [GHS+14, p. 9]:

- *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. However, in the cases when one is used as a substitution, it is linked to its antecedent as in example (5), see Section 2.1.13

  (5) *Do you prefer the blue shirt or [the red shirt]? – I would like the red [one].*

- *others...*: *others* is anaphoric and has an antecedent, but it is not coreferent with its antecedent. It is a case of comparative reference, see Section 2.1.14. It should be marked as **mention** → **pronoun** → **comparative** and linked to the antecedent. In some cases, *others* can be repeated in a text and should be in this case marked as coreferent as in example (6).

  (6) *[Others] thought that I am stupid but I don't care what [the/these others] say.*

If *other* is a part of a nominal phrase (*other people*), it could also function as comparative reference. In this case, the whole nominal phrase should be marked as **mention → np → comparative**.

- *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 the whole np should be marked and linked to antecedents in cases like Toni and Marc ... both boys
- *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, as this is a relation of part-whole or set-subset or similar, but not coreference, see example (7), where *them* is coreferent with *Three girls*, but *each* is not, it would be a member of a set, but not identical.

  (7) *Three girls were walking on the street. Each of them was carrying a Döner.*

2.1.4 Nominal premodifiers

In case of English nominal premodifiers, we only annotate a nominal premodifier if it can refer to a named entity (like [[*EU*] *supporter*] in example (8)) or it is an independent noun in the genitive form ([[*creditor's*] *choice*]); in all other cases, nominal premodifiers are not annotated as separate markables (*bank account*) [GS16, p. 4].

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

In example (8), the pronoun *it* cannot be linked to the complete NP *EU supporters*, but to the *EU* (the modifier). However, with compounds like *EU-supporters* or *EWSA* in German in example (4-a) above, there is a single unit which cannot be split any further. In such cases, it is necessary to search for a standalone instance of *EU* or *EWSA* earlier in the text and link the pronoun *it* to that instance (assuming one can be found) [GHS+14, p. 7]. The same is applied in German compounds that cannot be split.

2.1.5 Generic reference

Generic nouns can co-refer with definite full NPs or pronouns, but not with other generic nouns (no repetitions).

(9) a. *[Computers] are expensive. But [they] are really useful. Computers cost a lot of money.*
b. *[Computer] sind teuer. Aber [sie] sind richtig nützlich. Computer kosten viel Geld.*

In example (9), only the anaphoric pronoun *they* should be linked to its antecedent in the first sentence (*computers*), but we do not annotate the generic noun computers in the third sentence.

2.1.6 Indefinite pronouns

An indefinite pronoun is a pronoun that refers to one or more unspecified beings, objects, or places, such as *anybody, anyone, anything, everybody, everyone, everything, nobody, nothing, one, somebody, something, someone.*

(10) *[Anyone] can see that she was looking for trouble.*

In (10), *Anyone* is an indefinite pronoun as it does not refer to a specific person or group of people. Indefinite pronouns should be marked as pronouns 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 [GHS+14, p. 9].

2.1.7 Personal pronouns

We only annotate personal pronouns if they have a specific referent in the text like *Tad Williams* – *he* in (11). They include: *he, she, they, wir, er, sie* and their forms in different cases (*him, her,* *them, ihnen*, etc.). In the first release of ParCorFull, first and second person pronouns are not annotated.

(11) *[Tad Williams] is one of the most famous writers of modern times. In addition to Memory, Sorrow and Thorn [he] has written the acclaimed Otherland series.*

Personal pronouns can also serve as **pleonastic pronoun**. These do not actually refer to an entity [GHS+14, p. 5]. 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, as in example (12).

(12) a. *[It] is raining. – [Es] regnet.*
b. *[It] is well known that apples taste different from oranges – [Es] ist bekannt, dass...*
c. *Der Europäische Rat hat einen ständigen Präsidenten, dessen Aufgabe [es] ist, die Arbeit des Europäischen Rates zu koordinieren und seine Kontinuität zu gewährleisten.*

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

(13) a. *In this country, if [you] own a house you have to pay taxes.*
b. *[They] say you should never mix business with pleasure.*

In German, the indefinite pronoun *man* is often used.

In the case of pleonastic pronouns we wish to make a partial annotation: Marking these pronouns as pleonastic, but not linking them to anything (because they do not refer to anything). These are the cases, where there is no antecedent in the text to which the pronoun *they* may be linked to, as in example (14). In this case, the pronoun should be marked as "anaphoric but no specific antecedent".

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

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 *he* in example (15).

(15) *People when they see me say "[he]'s a bit weird".*

N.B. This is rarely required.

English lacks ungendered person pronouns. Therefore, there are instances of *he or she* in a text, see example (16).

(16) *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 the instances of *he or she* in dative and accusative, e.g. *him or her*, *his or her* and *his or hers. S/he* should be treated as a complete unit (or markable) and as a pronoun.

2.1.8 Demonstrative pronouns

(17) *During the November/December ministerial, we created consensus in [some very important areas]. [These] include ...*

In example (17), the demonstrative pronoun *these* points back to its antecedent *some very important areas* mentioned in the previous sentence and must be annotated.

In German, the forms *der, die, das* can also serve as demonstratives, see (18).

(18) *Hast Du [Martin] gesehen? [Der] war heute nicht im Büro.*

All the forms of demonstrative pronouns can also be exophoric, i.e. refer directly to the situational context and thus, do not have any links in the text as in examples in (19).

(19) a. *Siehst Du [den] [da] drüben?*
b. *[That man] over [there], he is strange.*

These cases should be marked as **extratextual reference**. In [GS16], predicative constructions are annotated in the following way: *[This] is a bank, but [it] is not very well-known*. But in fact, *bank* and *it* corefer. So, we annotate *bank* and *it* as coreferent.

There are cases of usage of demonstrative pronouns *those, jene, der, die, das*, when they function as indefinite pronouns and do not have coreference relation (should not be linked), see examples in (20). They are similar to the cases of 'Korrelat' described in Section 2.1.9 below.

(20) a. *[Those], who hesitate, are lost.*
b. *längst hat in diesem Feld die Wirklichkeit [jene], die es besser wussten, eingeholt.*
c. *Für [die], die lieber von einer festen Unterkunft Tagestouren unternehmen wollen, bietet die Insel Sylt z.B. rund 200 km Radwege.*

Temporal adverbs Temporal expressions are to be annotated if they co-refer.

(21) a. *[In my young days] we took these things more seriously. We had different ideas [then].*
b. *Annette ist endlich fertig. Sie ist ein bisschen bummelig und unordentlich, wie ich [als Kind] gewesen sein muss. [Damals] hätte ich nie geglaubt, dass ich meine Kinder zurechtweisen würde, wie meine Eltern mich zurechtwiesen.*

The temporal adverbial *then* can trigger either coreference or conjunction (discourse relation). For disambiguation, we can use a rule that if the English *then* can be translated as *damals*, then it triggers a coreference relation as in example (22-a). In other cases, it is likely to be a discourse marker.

(22) a. *Well, though dollars were just as hard to come by in [the 1920s]. But you could have paid for the land with livestock [then] because there was still, you know, fifty-cent-an-acre-land – an entire acre.*
b. *To Add an Exception to the AutoCorrect Exceptions List 1. Choose Tools – AutoCorrect, and [then] click the Exceptions tab.*
c. *Having lectures, we probably had about 20 hours a week, so we went obviously we were so close we could stay in the University to do work and [then] we could come home and we were near enough to the town center to just, you know, get a cab in to it and go out.*

Local adverbs Local adverbs as *here* and *there* or *da, dort* can also trigger coreference and should be annotated. In example (23-a), *there* refers to *to the doctor*, in (23-b), *here* corefers with *your new home*.

(23) a. *Is Lina going [to the doctor] today? – No, she went [there] yesterday.*
b. *. How do you like [your new home]? - Oh, it's really wonderful [here].*

Event Reference Demonstrative pronouns also refer to verb phrases or to bigger discourse units, as in example (24).

(24) a. *The London G-20 meeting recognized that [the world's poorest countries and people should not be penalized by a crisis for which they are not responsible]. With [this] in mind, the G-20 leaders set out an ambitious agenda for an inclusive and wide-ranging response.*
b. *Die Europäische Union wurde gegründet, um [politische Ziele zu verwirklichen]; erreicht werden sollte [dies] auf dem Weg der wirtschaftlichen Zusammenarbeit.*

In (24-a), *this* does not have a specific referent, but refers to the whole subordinate clause of the previous sentence (fact sentence). In (24-b), *dies* refers to the event-vp *politische Ziele verwirklichen*. Neither [GS16], nor [GHS+14] mark these cases. Event anaphors can refer back to whole sections of text or concepts evoked by the text [GHS+14, p. 8].

Demonstrative *da*, which often has a temporal or a local meaning may also refer to bigger segments as in example (25).

(25) *Das ist eine schwierige Frage, weil es natürlich immer darum geht, [wer ist denn hier eigentlich der wichtigere, die die das umsetzen, technisch, oder die, die das journalistisch erdenken]. Ich möchte [da] keinen Unterschied machen.*

Using deictics that vaguely refer to what the speaker is talking about (as in example (26)) exist in some text registers/genres under analysis.

(26) *[Ein immer größerer Teil dieser Brennstoffe wird aus Ländern außerhalb der EU eingeführt. Gegenwärtig importieren wir 50 % unseres Erdgasund Erdölbedarfs]; [diese Abhängigkeit] könnte sich bis 2030 auf 70% erhöhen.*

Here *diese Abhängigkeit* should be treated as an instance of event reference that refers to a segment consisting of two sentences.

In general, events should be easy to identify as they should contain verbs. However, in some cases, the decision on how big the span of relation is could be difficult. For instance, in (26), *diese Abhängigkeit* could refer to the sentence *Gegenwärtig importieren wir 50 % unseres Erdgasund Erdölbedarfs* only. 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.

In some scenarios, it is 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, see example (27).

(27) *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 such scenarios, 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.

2.1.9 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, and in English, they are rather archaic). They are formed by replacing a preposition and a pronoun, like *gegen+das* → *dagegen* in example (28).

(28) *Viele Amerikaner haben Probleme mit [Rassismus]; doch wir sind [dagegen] immun.*

Dagegen refers to *Rassismus* and should be annotated and linked with it.

Some pronominal adverbs in German are used as 'Korrelat' and are not coreferential. In (29-a), there is no coreference; in (29-b), *davor* refers to the event *allein im Wald joggen*.

(29) a. *Ich habe Angst [davor], alleine im Wald zu joggen*
b. *[Joggst Du denn auch alleine im Wald]? Also ich habe Angst [davor].*

Some pronominal adverbs in German can be used both as reference (as in (30-b)) and discourse markers (as in (30-a)).

(30) a. *In den vergangenen beiden Jahren haben die 200 größten deutschen Firmen insgesamt weit über 50 000 Jobs abgebaut. Mittelständische Betriebe [dagegen] haben alleine im Jahr 2000 unterm Strich 350 000 Jobs zusätzlich geschaffen.*
b. *dann werden [diese speziellen Aspekte der Komplexe] deutlich miterlebt, und die Jugendlichen müssen [dagegen] anarbeiten.*

2.1.10 Relative pronouns

Relative pronouns include such cases as *who, whom, whose, which, that*, etc.

In example (31), *which/die* is linked to *The Army/die Arme*.

(31) a. *[The Army], [which] recruits heavily in the Punjab, will not use [their] force there in the way [it] is doing in the tribal areas.*
b. *[Die Armee], [die] einen Großteil [ihrer] Soldaten im Punjab rekrutiert, wird dort nicht mit Gewalt vorgehen, so wie [sie] es in den Stammesgebieten tut.*

Keep in mind that pronouns can be ambiguous. In example (32-a), *where* is a relative pronoun and refers to *Kashmir* (to show this, one can substitute where by in which). Conversely, in (32-b), *where* is not a relative pronoun and should not be annotated.

(32) a. *For both India and Pakistan, Afghanistan risks turning into a new disputed territory, like [Kashmir], [where] the conflict has damaged both countries for more than 50 years.*
b. *Daisy managed to discover where Mr. Baccini's dishonest partner was now living and was anxiously expecting her cheque.*

This type of coreference is one of the cases of grammatical coreference – a kind of coreference in which it is possible to identify the antecedent on the basis of grammatical rules. In German, the relative pronoun agrees in its gender, number and case with the antecedent, as *die* in (31-b) – female singular nominative.

2.1.11 Reflexive pronouns

We annotate reflexives as in examples (33).

(33) a. *It's beginning to rain! – [Daisy] exclaimed to [herself].*
b. *Es fängt an zu regnen! – sagte [Daisy] zu [sich] [selbst].*

[GS16, p. 3]: For German, reflexive pronouns must be annotated only if they are independent constituents, but not part of a reflexive verb. The following test should be applied: if the position of the reflexive pronoun can be changed, then the pronoun is an independent unit (example (34-b)), otherwise it belongs to the verb (example (34-a)).

(34) a. *Ich habe mich gestern gewundert. (*Mich habe ich gestern gewundert).*
b. *Ich habe [mich] 1 gestern gesehen. (Mich habe ich gestern gesehen).*

Similar to relatives, coreference with reflexives belongs to grammatical coreference. Reflexives agree in their number and case with the antecedent.

2.1.12 Groups

[GS16, p. 4]: If all elements from a group are referred to by an anaphoric pronoun, create a group markable consisting of the set elements and then link the anaphoric pronoun to it.

(35) a. *Did [your husband] buy Lorna, [Mrs. Humphries]? – No, [we] bought her together.*
b. *So wurden 2004 [Estland], [Lettland], [Litauen], [Polen], [die Slowakei], [Slowenien], [die Tschechische Republik] und [Ungarn] zusammen mit den Mittelmeerinseln [Malta] und [Zypern], Mitgliedstaaten der EU. [Bulgarien] und [Rumänien] folgten im Jahr 2007. Heute sind [sie] alle als Partner an dem großartigen Projekt beteiligt, das die Gründerväter der EU ersonnen haben.*

[GHS+14, p. 6]: In (35-a), *we* refers to *your husband* and *Mrs. Humphries*, and in (35-b), *sie* refers to all the countries mentioned in separate sentences, so there is no NP span that covers all of them. In cases like these, if all of the antecedents can be identified and it is clear from the texts what the antecedents are, the pronoun should be linked to each of the separate antecedent "parts". It is important to ensure that all "parts" are linked. All components of the antecedent should be linked to the pronoun directly, and not to each other.

It is important to first ensure that no NP exists that covers all parts of the antecedent. So, if a sentence like (36) precedes the first sentence in (35-b), then *sie* would refer to *Europäische Länder* and not to these concrete countries mentioned in the following sentences.

(36) *Europäische Länder, die jahrzehntelang keine demokratischen Freiheiten genossen hatten, kehrten endlich zur Familie der demokratischen europäischen Nationen zurück.*

2.1.13 Substitution and Ellipsis

Coreference expresses the relation of identity, whereas substitution triggers type reference relation between referents belonging to the same class [KS13, DBD81]. In this sense, substitution is similar to ellipsis, whereas the latter use elided elements instead of substitutes.

Ellipsis and substitution are subdivided into their structural types, according to the omitted/substituted element: nominal (e.g. nominal ellipsis in example (37-a)), verbal (see a case of verbal substitution in example (37-b)) and clausal (we illustrate clausal substitution in example (37-c)).

(37) a. *You might have to come up afterwards to count but if I take any one of these balls in the middle and I count how many [neighboring balls] that there are around it, the answer's always twelve [].*
b. *You'll see that it had [to accommodate] an incredible range of functions much more elaborate than any temple or palace in the past would have [done].*
c. *[Does everybody have a handout, for today]? If [not] Aaron's got handouts.*
d. *[So, well, any more questions]? – [no], okay, ...*
e. *[How many slices do you want]? - "[Two]", I said.*

Following [Men17], we also define two additional classes for ellipsis: yes-no type as in (37-d) and mixed type (a combination of nominal and verbal or clausal) as illustrated in (37-e). Non-repetition of the constituents from a question or statement is seen as an ellipsis of the whole clause by [HH76]. Like in case of coreference, substitution and ellipsis also form chains. The types of the antecedents (marked with curly brackets) in these chains are reflected in the types of their substituting or elliptical elements that we define: for instance, the elliptical element in (37-a) establishes a relation to a nominal phrase.

2.1.14 Comparative reference

Comparative reference does not trigger the relation of identity, co-reference in the strict sense. Together with other cases (substitution and ellipsis) it rather involves type-reference, co-classification or "sloppy identity", see [KS12].

The linguistic means signaling comparative reference include the following words:

- general:

  - EN: *same, equal, identical, identically, similar, such, so, corresponding, similarly, likewise, other, different, else, differently, otherwise*
  - DE: *Derselbe (+Nomen), gleich, identisch, Ähnlich (adj+adv), genauso, gleichermaßen, Anders, unterschiedlich, gegensätzlich, andersartig*

- particular:
  - EN: *More, fewer, less, further, additional, so/as/equally*+Quantifier, e.g. *so many*
  - DE: *weniger, mehr*
  - BOTH: comparative degree of adjectives such as *better, faster; so/as/equally* + adjective, e.g. *equally good* and comparative adverbs

Not all cases should be annotated. Grammatical comparison should not be annotated, as in examples (38-a-b).

(38) a. *Paul is [bigger] than Jim.*
b. *Ihr Appartment ist fast [so groß] wie meines.*

Sometimes, cases of comparison can be generic, and don't have any link to other text elements, as in examples (39-a-b).

(39) a. *Most people have [the same] breakfast everyday.*
b. *Alle Befragten beschrieben [ähnliche] Situationen.*

2.2 Span of Markables

[GS16, GHS+14]:

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 or a VP and sentence(s) in case of event reference. For noun or NP markables, the following rules apply. The markable must:

- contain the syntactic head of the NP
  - *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: *Frederick F. Fernwhistle Jr.* in *the Honorable Frederick F. Fernwhistle Jr.*
- determiners and adjectives (if any) that modify the NP
  - *the Honorable Frederick F. Fernwhistle Jr.*
  - *Mr. Holland*
  - *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.
  - *the big black dog* (where *dog* is the head)
  - Determiners such as *the* should be included
- deverbal modifiers (participial constructions, regardless whether in pre- or postposition) that can be substituted by a subordinate clause as in example *[Regional conflict, involving all of the region's states and increasing numbers of non-state actors], has produced large numbers of [trained fighters, waiting for the call to glory]*. In this case, both *regional conflict, involving all of the region's states and increasing numbers of non-state actors* and *trained fighters, waiting for the call to glory* are markables.
- dependent prepositional phrases (for example, *Queen of England*).
- appositions, i.e., additive material that is not syntactically integrated, are included into the markable span, but are not annotated separately: *[JuD, Party of Proselytizing,] was founded in 1972. / [Jud, Partei der Missionierung,] wurde 1972 gegründet.*

However, full clauses, in particular relative clauses, are not taken as parts of the markable rooted in the NP head. Therefore we annotate relative pronouns separately.

2.3 Anaphoric and Cataphoric Reference

[GHS+14, p. 4]:

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 example (40) is cataphoric.

(40) *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 Antecedents

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 lexical word, e.g. proper noun (example (41-a)) or common noun or a nominal phrase(example (41-b)), but also a pronoun

  (41) a. *[Tad Williams]? I just read a novel of [his].*
  b. $[EineverantwortungsbewusstePolitik]_1$ *kann* $[diesenProzess, derzudemvonobjektivenFaktorendeterminiertwird,]_2$ *nicht nur flankieren.* $[Sie]_1$ *muss* $[ihn]_2$ *vielmehr formen.*

- a verbal phrase (event-vp)

  (42) a. *just to remind you, for project number three which is due on Wednesday ... you have to basically [combine everything you learned from project one and project two]. ultimately [that]'s the goal .*
  b. *Also das Alphabet dieses Prozesses wäre jetzt [anzünden, ausgehen]. [Das] sind die beiden Aktionen, die das Ding machen kann.*
  c. *Hier geht es darum, [eine praktische Lösung zu einer aktuellen Fraestellung zu bearbeiten]. [Das] macht man üblicherweise auch in einer Kleingruppe.*

- a fact-sentence

  (43) a. *[We work for prosperity and opportunity]1 because they're right.* $[It]_1$ *'s the right thing to do.*
  b. *[Wir arbeiten für Wohlstand und Chancen]1, weil [das]1 richtig ist. Wir tun [damit]1 das Richtige.*

- Nothing (see Section ...)

- no explicit antecedent 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 page 5, ex. 14).

- 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.

- A word may have been marked as a pronoun in error

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. In example (44), the pronoun *he* should be linked to *Musashi*, the nearest antecedent, and not to *Miyamoto Musashi* which appears earlier in the text.

(44) *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.*

4 Annotation Process

4.1 Colours in MMAX

- All marked structures are in blue font.
- if something is in italics, then it is a member of a chain (and no chain means no italics:)
- if a pronoun is of type 'pronoun' (indefinite or not yet assigned a type), then it is bold and coloured in magenta
- all events are marked in pink (dirty pink) – events were marked in the previous framework but were not linked to any antecedents. Now, you need to find their antecedents and change the annotation accordingly.
- All the pronouns marked as 'Not sure. HELP!' are in bold font.
- mentions that have the value 'none' are marked with green colour.
- Green colour is also given to all cases of 'speaker reference' when the audience is marked as 'none'
- the cases of 'addressee reference' with the audience marked as 'none', we have red background colour
- all anaphoric pronouns have bold font
- all pleonastic pronouns are in bold font and coloured with cyan background
- 'speaker reference' and 'addressee reference' are in bold

4.2 Structures and their values in MMAX and in the annotated Markable files

Representation of the guidelines in the scheme.

1. Coreference chains: every chain has the same ID represented by the following tag in the annotated data: coref_class="set_230".
2. Mentions: mentions are marked as mention="VALUE", and the VALUE can be
   - "pronoun" for pronominal mentions and non-referring pronouns
   - "nps" for nominal mentions (and also comparative)
   - "vp" for verbal mentions (antecedents, substitution and ellipsis only)
   - "clause" for clausal mentions (antecedents, substitution and ellipsis only)
   - "none" for those elements that did not fit into one of the above categories
3. If the mention is pronominal then it has:
   - an attribute type="VALUE" and the VALUE can be:
     - "antecedent" for pronominal antecedents
     - "anaphoric" for anaphora
     - "cataphoric" for cataphora
     - "comparative" for comparative reference
     - "nom substitution" nominal substitution
     - "addressee reference" (not used in ParCorFull)
     - "speaker reference" (not used in ParCorFull)
     - "extratextual reference" for those cases of non-textual relation (world reference)

    - "pleonastic"
    - "pronoun" for indefinite pronouns without antecedents
    - "none"
    - "Not sure. HELP!"

  - an attribute agreement="VALUE" and the VALUE can be:
    - "none", where the agreement is not applicable
    - "they (sg.)" for all singular pronouns (in both languages)
    - "they (pl.)" for all plural pronouns (in both languages)
  - an attribute position="VALUE" and the VALUE can be:
    - "none"
    - "it (subject)" for all pronouns in subject position (in both languages)
    - "it (non-subject)" for all pronouns in non-subject position (in both languages)
  - an attribute type_of_pronoun="VALUE" and the VALUE can be:
    - "personal"
    - "possessive"
    - "demonstrative" (quantifiers like both boys are also marked as demonstratives)
    - "reflexive"
    - "relative"
    - "none"

4. If the mention is pronominal or nominal (not antecedents) it has:

   - an attribute antetype="VALUE", where value can be:
     - "entity" for the reference to entities represented by nominal groups
     - "event" for the reference to non-entities represented either by verbal phrases or by clauses (also longer text passages)
     - "generic" for generic nouns
   - an attribute split="VALUE", where value can be:
     - "simple antecedent" for antecedents representing one entity or one event
     - "split reference" representing reference to two or more entities / events
     - "no explicit antecedent" when the antecedent is difficult to identify
   - an attribute comparative="VALUE" and the VALUE can be:
     - "part" for particular comparative reference
     - "gen" for general comparative reference

5. If the mention is nominal, then it has:

   - an attribute nptype="VALUE" and the VALUE can be:
     - "antecedent" for nominal antecedents
     - "np" for anaphoric or cataphoric nominal phrases
     - "comparative" for comparative reference
     - "nom-ellipsis" for nominal ellipsis
     - "apposition" for elements further identifying the preceding NP
     - "Not sure. HELP!"
   - an attribute anacata="VALUE" and the VALUE can be:
     - "anaphoric"
     - "cataphoric"
   - an attribute npmod="VALUE" and the VALUE can be:
     - "possessive" for possessive pronouns
     - "demonstrative" for demonstrative pronouns
     - "def-article" for definite articles

  - "indefinite" for indefinite pronouns like each, every/jede(e), alle
  - "none" for bare nouns

- nominal mentions of nptype="apposition" can have attribute sub_apposition="VALUE" and the VALUE can be
  - "none" (not used in ParCor)
  - "head" for (not used in ParCor)
  - "attribute" for all appositions

6. If the mention is verbal, then it has:

   - an attribute vptype="VALUE" and the VALUE can be:
     - "antecedent" for verbal antecedents
     - "verb-substitution"
     - "verb-ellipsis"
     - "not sure. help!"

7. If the mention is clausal, then it has:

   - an attribute clausetype="VALUE" and the VALUE can be:
     - "antecedent" for clausal antecedents
     - "clause-substitution"
     - "clause-ellipsis"
     - "not sure. help!"

References

[DBD81] R.-A. De Beaugrande and W. U. Dressler. *Einführung in die Textlinguistik.* Niemeyer, Tübingen, 1981.

[GHS+14] Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. *ParCor 1.0: Pronoun Coreference Annotation Guidelines.* Edinburgh, Uppsala, March 21 2014.

[GS16] Yulia Grishina and Manfred Stede. *Parallel coreference annotation guidelines.*, November 2016.

[HH76] M.A.K. Halliday and Ruqaiya Hasan. *Cohesion in English.* Longman, London, New York, 1976.

[KS12] K. Kunz and E. Steiner. Towards a comparison of cohesive reference in english and german: System and text. In M. Taboada, S. Doval Suárez, and E. González Álvarez, editors, *Contrastive Discourse Analysis. Functional and Corpus Perspectives.* Equinox, London, 2012.

[KS13] K. Kunz and E. Steiner. Cohesive substitution in english and german: A contrastive and corpus-based perspectivet. In Karin Aijmer and Bengt Altenberg, editors, *Advances in Corpus-Based Contrastive Linguistics. Studies in honour of Stig Johansson*, pages 201–232. John Benjamins, Amsterdam, 2013.

[Kun12] Kerstin Kunz. *Richtlinien für die Korrektur von kohäsiven Referenzmitteln*, December 2012.

[Men17] K. Menzel. *Understanding English-German contrasts: a corpus-based comparative analysis of ellipses as cohesive devices.* PhD thesis, Universität des Saarlandes, Saarbrücken, 2017.

[NM09] Anna Nedoluzhko and Jiri Mirovsky. *Annotating extended textual coreference and bridging relations in the Prague Dependency Treebank.* Prague, 2009.