Annotation guidelines for dutchcoref

Andreas van Cranenburgh

1 How to annotate?

- Read the text from start to finish, make and correct annotations as you go.
- Identify mentions by asking yourself whether a span of text describes a specific identifiable object or person.
- When the same entity is referred to again, ensure that both mentions are in the same coreference cluster. Conversely, remove any incorrect links.

2 Mentions

A mention is a span of text that refers to an entity or person in the real or mental world. All mentions referring to objects or persons are annotated, including entities that are not referred to again (singletons). Mentions have been automatically identified, but they may need to be corrected.

The following subsections list the types of mentions that should be annotated. The span of a mention is indicated here with square brackets [and]; a span that should not be annotated as mention is indicated with [red brackets].

2.1 Pronouns

- Personal pronouns (*zij*, *hun*, ...). Includes *het* when used as pronoun.
- Possessive pronouns (*mijn*, *zijn*, ...)
- Demonstrative pronouns (*die*, *dat*, *deze*, *dit*, *daar*)
- Relative pronouns (*die*, *dat*, *wie*, *wat*)
- Reflexive/reciprocal pronouns (*zich*, *zichzelf*, *elkaar*). Both obligatory and normal reflexives are annotated.
- Indefinite/generic pronouns (*men*, *je*, *ze*, *iedereen*, *iemand*, ...) when the same unspecified person/thing can be referred to again. This excludes e.g. *niemand* or wh-pronouns in questions (*wie*, *wat*, *welke*, ...).
- Pronominal adverbs of location: *er*, *hier*, *daar*, *waar*, *waarin*, ...

Exclude non-referential, pleonastic pronouns:

- *[Het] regent.*
- *Daar moeten we [het] over hebben.*
- *[Er] zit niets anders op.*

2.2 Proper nouns (named entities)

- One-word names: *[Jan]*, *[Amerika]*.

- Multiword names form a single mention: *[Jan de Vries]*, *[de Verenigde Staten]*.

2.3 Noun phrases (NPs)

Always annotate the longest, most specific continuous span describing a mention. What to include:

- Determiners: *[het huis]*
  A possessive pronoun is a determiner, and is also its own mention:
  [[mijn] fiets]
- Adjectives, nouns: *[een warme kop thee]*
- Prepositional phrases modifying the noun: *[kandidaat voor [de coalitie]]*.
- Noun phrases within noun phrases. See previous example. Since *kandidaat* and *coalitie* describe different entities, they are both annotated. On the other hand, there is no need to mark *kandidaat* twice:
  [[kandidaat] voor [de coalitie]].

Special cases:

- Conjunctions (*Jan en Marie*). Include the whole conjunction as mention only when it functions as a unit in the text; e.g., when referred to again as a single group bij a plural pronoun "*ze*". By default, only the individual conjuncts *Jan* and *Marie* are considered as separate mentions.
- NPs with commas. Except in special cases, a comma indicates the end of a mention:
  [De nieuwste iPhone], *[een revolutionaire nieuwe smartphone]*.
  Special cases:
  - Geographical: *[Los Angeles, California]*
  - Adjective: *[Een mooie, rode roos]*
  - Conjunction functions as group (see above)
    [[Jan], [Marie] en [Joost]]
- Discontinuous NPs
  [[een belediging] /zijn/ van onze gastvrijheid]
  Mentions must be continuous, uninterrupted spans in the text. Since the verb "*zijn*" is not part of the noun phrase, it should also not be part of the mention. In this case only "*een belediging*" is marked as a mention (i.e., the part with the head of the constituent *belediging*).
- Relative clauses. The relative pronoun indicates the end of the mention:
  [[De burgemeester]$_1$ [die]$_1$ de vergadering opende] was behoorlijk nors.

What to exclude:

- Time-related NPs: *[gisteren]* , *[de langste dag van de zomer]*
- Actions, verb phrases: *[het verzamelen van liquide middelen]*
- Quantities, measurements: *[20 graden]* , *[100 MB]* , *[ongeveer 10 euro]*
  However, not every NP with a quantity is excluded, because the NP may describe a specific object that is referred to again:
  'En wij kregen als speciale missie om [vijf miljoen Nederlandse guldens]$_1$ uit de kluizen van de Nederlandsche Bank in Middelburg via Duinkerken naar Londen te brengen. De koers waartegen [ze]$_1$ in Whitehall konden worden

ingewisseld tegen Engelse ponden, was [. . .]. [Het geld]$_1$ zat in twee zwarte koffers, verdeeld over achthonderd linnen zakjes.

- Idioms: *Wat is er aan [de hand]?*
  [Hij] zag [Esmée] bij [het hoofdeinde] in [gesprek] met [een familielid]. (*gesprek* has no determiner, it is not a specific identifiable conversation that can be referred to again)
- Material, substances, and other non-specific mass nouns:
  [het deksel van [blank hout]]

3 Coreference links

Only a single type of coreference is annotated, indicating that mentions refer to the same entity. There is no annotation of the specific antecedent for an anaphor; by linking mentions, they become part of the same cluster and are considered equivalent. For example, given a cluster {John, he} and a new mention "him", linking the new mention to "John" or "he" makes no difference. Mentions that belong to the same cluster are indicated with subscripts. The following kinds of coreference are recognized:

- Identity, strict coreference
  [Jan]$_1$ ziet [Marie]$_2$. [Hij]$_1$ zwaait naar [haar]$_2$j .
- Predicate nominals
  [Jan]$_1$ is [een schrijver]$_1$.
- Relative clauses
  [De burgemeester]$_1$ [die]$_1$ de vergadering opende was behoorlijk nors.
  [Het huis]$_2$ [waar]$_2$ ik ben geboren.
- Appositions. If the first part is a name, mark separately:
  [Hu Jintao]$_1$, [de president van China]$_1$, hield een toespraak voor de VN.
  But a modifier followed by a name is a single mention:
  [zeilster Carolijn Brouwer]
- Type-token coreference:
  [The man]$_1$ who gave [[his]$_1$ paycheck]$_2$ to his wife was wiser than [the man]$_3$ who gave [it]$_2$ to [[his]$_3$ mistress]$_4$.
  The mentions in cluster 2 are not identical, but are tokens of the same type.
- Time-indexed coreference:
  [Bert Degraeve]$_1$, tot voor kort [gedelegeerd bestuurder]$_1$, gaat aan de slag als [chief financial and administration officer]$_1$.
  Cluster 1 contains mentions whose coreference is only valid at specific times, but we do not annotate this distinction.
- Bound anaphora:
  [Iedere man]$_1$ steekt wel eens [zijn]$_1$ nek uit.

Special cases:

- Always annotate the intended referent. In case of nicknames or jokes, you may have to distinguish mentions of the real referent, and nicknames or jokes that refer to someone else.
- Metonymy:

De VS heeft meerdere doelen gebombardeerd. Moskou heeft woedend gereageerd.
"*Moskou*" refers here not to the city, but to the government of Russia. We annotate the intended referent, not the literal meaning.

- Use/mention distinction:[1]
  [Jan]$_1$ is rijk, [hij]$_1$ heeft [een Ferarri]. [Jan]$_2$ is [een gangbare naam]$_2$.
  The second instance of *Jan* refers to the name/word itself, not the person. This is sometimes indicated with quotation marks.
  Maar verdomd, op [pagina vier] wordt [de aankomst in [de Hauptstadt]] gemeld van [een 'prominenter, unabhängiger politischer Publizist aus den Niederlanden']$_1$. [Politischer Publizist]$_2$! [Dat etiket]$_2$ zal [ik]$_1$ tijdens [dit bezoek] zeker niet meer kwijtraken.
  The first mention refers to the protagonist, but the second mention refers to the label.

Several more complex phenomena are excluded:

- VP coreference:
  [Mijn fiets was gestolen] . [Dat] vond ik jammer .
  [Heeft u ook een nieuwsbericht] , dan vernemen wij [dat] graag .
  Note that in addition to not annotating a link, these are not mentions because they do not refer to objects or persons.
- Part/whole, subset/superset relations (bridging relations):
  In de Raadsvergadering is het vertrouwen opgezegd in [het college]$_1$. In een motie is gevraagd aan [alle wethouders]$_2$ hun ontslag in te dienen .
  While the entities of *het college* and *alle wethouders* are related, they are distinct entities, and we do not annotate such a bridging relation between entities.
- Modality/negation:
  [Een partij als de CD&V] is nou niet echt [het toonbeeld van sociale betrokkenheid]

4 Differences with other annotation schemes

4.1 Differences with the Corea annotation scheme

Cf. Bouma et al. (2007)

- Only a single type of coreference relation is annotated, corresponding to the types IDENT, PRED, BOUND. The BRIDGE relation (part/whole, subset/superset relation) is not annotated.
- Mentions belong to coreference clusters which are equivalence classes; the specific antecedent of an anaphor is not annotated. The type of entity, the head of a mention, and the type of coreference relation are not part of the annotation.
- Mentions are manually corrected: all mentions that refer to an entity are annotated, non-referential spans are not included as mentions.
- Relative pronouns are considered mentions and coreferent.

[1] https://en.wikipedia.org/wiki/Use%E2%80%93mention_distinction

Corea: *[President Alejandro Toledo]$_1$ reisde dit weekend naar Seattle voor een gesprek met [Microsoft topman Bill Gates]$_2$. [Gates, die al jaren bevriend is met [Toledo]$_1$]$_2$, investeerde onlangs zo'n 550.000 Dollar in Peru .*
These guidelines: *[President Alejandro Toledo]$_1$ reisde dit weekend naar Seattle voor een gesprek met [Microsoft topman Bill Gates]$_2$. [Gates]$_2$, [die]$_2$ al jaren bevriend is met [Toledo]$_1$, investeerde onlangs zo'n 550.000 Dollar in Peru.*
Motivation: it can be difficult to identify the complete relative clause, due to discontinuity or long parenthetical remarks. Annotating the NP before the relative pronoun avoids a lot of difficult cases. Such cases are both difficult for annotators as well as for automatic parsers. For example:

  - Relative clauses can be discontinuous:
    Ik kan in elk geval getrouw [de indrukken]$_1$ weergeven [die]$_1$ deze feiten hebben achtergelaten .
  - Relative clause can be arbitrarily long:
    En dit was [de Perry]$_1$ [die]$_1$ vroeg op die ochtend in mei , voordat de zon te hoog stond om nog te kunnen spelen , op de beste tennisbaan in het beste door de recessie getroffen vakantieoord in Antigua stond , met de Russische Dima aan de ene kant van het net en Perry aan de andere .

- Obligatory reflexives are annotated:
  [Jan]$_1$ scheert [zich]$_1$

4.2 Differences with the Newsreader annotation scheme

Cf. Schoen et al. (2014)

- Entities are not restricted to a set of predefined types (person, organization, location, product, . . .)
- Relative pronouns, discontinuous NPs, and appositions are annotated differently.

References

Bouma, G., Daelemans, W., Hendrickx, I., Hoste, V., and Mineur, A. (2007). The corea-project, manual for the annotation of coreference in dutch texts. Technical report, University of Groningen. https://www.clips.uantwerpen.be/~iris/corea/publications/manual_corea.aug07.pdf.

Schoen, A., van Son, C., van Erp, M., and van Vliet, H. (2014). Newsreader document-level annotation guidelines-dutch. Technical report, VU University. http://www.newsreader-project.eu/files/2013/01/8-AnnotationGuidelinesDutch.pdf.