Annotation guidelines for dutchcoref

Andreas van Cranenburgh

March 27, 2019

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 identifiable 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 correct span for mentions is indicated with square brackets [and]; a span that should not be annotated as mention is indicated with [red brackets].

Due to technical limitations (i.e., the CoNLL 2012 format), mentions must follow word boundaries. Therefore we must annotate:[1]

[hoek [Groot Hertoginnelaan-Laan] [van Meerdervoort]]
and not: *[hoek [Groot Hertoginnelaan]-[Laan van Meerdervoort]]*
The following subsections list the types of mentions that should be annotated.

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/object can be referred to again. This excludes e.g., negations *niemand* or wh-pronouns in questions (*wie, wat, welke, ...*).
- Pronominal adverbs of location: *er, hier, daar, waar, waarin, ...*

[1] Fixing this would require correcting the tokenization and parse tree, which is outside the scope of these annotation guidelines.

Exclude non-referential, pleonastic, and/or expletive pronouns:

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

Indefinite pronouns require judgment to determine whether they refer to an identifiable person/object and whether they can be referred to again:

- The word *iets* often occurs in a negative context, such as *nauwelijks*, *zonder*, *alsof*, which indicates that there is no identifiable referent:
  - Van [die schaamte] is nauwelijks [iets] terug te vinden in [[zijn] romans].
  - [Esmée] nam [ze] in [de hand] , staarde er langdurig naar , legde [ze] zonder [iets] te zeggen terug op [de kist] .
  - *[Oscar] bleef in [de verte] turen alsof [hij] [iets] verwachtte : [een plots inzicht] , [een verklaring] ?*
- Conversely, in the following sentence there is a concrete referent: *"[Ik] zie [iets]," zegt [Jan]. [Een eiland] verschijnt aan de horizon.*

2.2 Proper nouns (named entities)

- One-word names: *[Jan]*, *[Amerika]*.
- Multiword names form a single mention: *[Jan de Vries]*, *[de Verenigde Staten]*.
- Geographical: *[Los Angeles, [California]]*
- Possessive: *[[Jans] moeder]*

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]
  Quantifiers are also determiners: [iedere buitenlandse televisiezender]
- 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.

- Disjunctions (*tram 18 of 22*). Include the whole disjunction if there is a single intended referent; otherwise only annotate the disjuncts as mentions separately:
  [We] zijn in [Praag] op [de hoek van [de Vyšehradska] en [de Trojicka]]. [Tram 18 of 22] staat stil bij [de Botanische Tuin].
  While the description is imprecise, the narrator has a specific tram in mind, so the whole is a mention; later, *de tram* is used to refer to the same tram again.
  Contrast with the following, where there are two separate options, which do not form a single mention:
  Vanaf [het station] kan je [tram 18] of [22] nemen.
- 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]]*
  - Adjectives: *[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 and other NP-modifying clauses: The relative pronoun (*die, dat, waar, waarover,* ...) indicates the end of the mention:
  - *[[De burgemeester]$_1$ [die]$_1$ de vergadering opende] was behoorlijk nors.*
  - *[Hij] en [ik], na [een aperitief]$_1$ [dat]$_1$ hij aanduidde als '[Rotkäppchen]$_1$'*
  - *... [een kroeg op [de Schönhauser Allee]$_1$] , [een buurt]$_1$ [waar]$_1$ volgens [hem] ondanks [de Wende] niets veranderd was .*

  The same holds for other clauses modifying an NP, e.g. *[NP] om te ...*:
  Dan overviel [mij] [de onweerstaanbare drang] om te vluchten , in [grote haast] , [de duivel] op [[mijn] hielen] .

What to exclude:

- Time-related NPs: *[gisteren] , [de langste dag van de zomer]*
- Actions, manners, verb phrases: *[het verzamelen van liquide middelen], [de wijze] [waarop] [die oude communisten] alles rechtpraten wat krom is*
- Adjectives, demonyms: *[de [Nederlandse] soldaten]*
- 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]. (there is an implied conversation, but the common noun *gesprek* is not a mention that can be referred to again)
- Material, substances, and other non-specific mass nouns:
  [het deksel van [blank hout]]
- Descriptions in a negative context (*niet*, *geen*, *nooit*, . . .) do not refer, and are therefore not mentions:
  Maar nee, geen [glimmende regenjassen en gleufhoeden] 's nachts aan [de deur van [[mijn] hotelkamer]] , [geen enkele toespeling op [mijn] geschrijf] van de kant van [het Presseamt] , [waar] [ik] [mij] bij ieder bezoek aan [de DDR] nederig meldde , nooit [gezeur met visa] , . . .

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]$_{2j}$.
- 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]
- Acronyms: [De Partij van de Arbeid]$_1$ ([PvdA]$_1$)
- Generic entities. Only add a link when there is a clear anaphoric relation:
  [Men]$_1$ verloor [elkaar]$_1$ makkelijk uit het oog na [de Wende] .
  If a later sentence mentions a generic *men*, annotate it as a different entity.
- 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 referents of 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.
  However, this only holds when the intended referent is strictly equivalent. The following cases are not coreferent:
  - *[westerse critici] hadden [het boek]$_1$ ([zevenhonderd pagina's]$_2$) onder hoongelach verguisd als '[hagiografie van een meeloper]$_1$'*
    (*zevenhonderd pagina's* refers only to a physical aspect of the book)
  - *[De gastheer]$_1$ begon met [een loflied op [Nederland]$_2$] , [[zijn]$_1$ aangenaamste buitenlandse post]$_3$, en ook [[zijn]$_1$ laatste]$_3$.* (*Nederland* and *buitenlandse post*, a country and a job, are not equivalent)
  - *[Hij]$_1$ werd voorgesteld als '[mein Mitarbeiter Herr . . .]$_1$' ([Naam]$_2$ niet verstaan.)* (person and name are not equivalent)
- Use–mention distinction:[2]
  [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/clausal coreference:
  [Mijn fiets was gestolen] . [Dat] vond ik jammer .
  [Heeft u ook een nieuwsbericht] , dan vernemen wij [dat] graag .
  In addition to not annotating a link, these are not mentions because they do not refer to objects or persons. In the following example, the clause is not a mention (see use–mention distinction above), *deze opmerking* is a mention (a mental object), but again there is no coreference link:
  ["Ik ben onschuldig,"] zei hij. Na [deze opmerking] bleef het stil.
- 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 .
  The entities of *het college* and *alle wethouders* are related but 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]*

[2] Cf. https://en.wikipedia.org/wiki/Use%E2%80%93mention_distinction
NB: 'mention' in this terminology is used in a different sense as in these guidelines.

4 Comparison with related annotation schemes

4.1 Differences with the Corea annotation scheme for Dutch

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 a person or object are annotated (including singletons), non-referential spans are not included as mentions.
- Relative pronouns are considered mentions and coreferent.
  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 Dutch 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.

4.3 The proposed annotation scheme of Rösiger et al. (2018)

Rösiger et al. (2018) propose an annotation scheme for German literary texts which provided the inspiration for these annotation guidelines.

Commonalities:

- Mentions are manually corrected.
- Coreference annotation of entity clusters instead of binary anaphora-antecedent links. No annotation of link type.
- NP mentions in idiomatic expressions are excluded.
- Bridging relations are excluded.

Differences:

- Singleton mentions are included.
- Non-nominal antecedents (VPs, clauses) are not annotated.
- Generic mentions/entities do not receive a special label.
- Group mentions/entities do not receive a special label.
  No relations between entities are annotated.
- Subtoken annotation is not allowed, to be compatible with the CoNLL 2012 format.
- Discontinuous mentions are not allowed, for the same reason.

References

Gosse Bouma, Walter Daelemans, Iris Hendrickx, Véronique Hoste, and A Mineur. The COREA-project: Manual for the annotation of coreference in Dutch texts. Technical report, University of Groningen, 2007. https://www.researchgate.net/publication/252395718.

Ina Rösiger, Sarah Schulz, and Nils Reiter. Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of LaTeCH-CLfL*, pages 129–138, 2018. http://aclweb.org/anthology/W18-4515.

Anneleen Schoen, Chantal van Son, Marieke van Erp, and Hennie van Vliet. Newsreader document-level annotation guidelines - Dutch. Technical report, VU University, 2014. http://www.newsreader-project.eu/files/2013/01/8-AnnotationGuidelinesDutch.pdf.