A Dutch coreference resolution system with quote attribution

van Cranenburgh, Andreas

A Dutch coreference resolution system with quote attribution

A.W.van.Cranenburgh@rug.nl, University of Groningen, CLIN 2019.

Abstract

- Coreference resolution is the task of identifying spans in text (*mentions*) that refer to the same *entity*
- We present a rule-based system for Dutch, based on the Stanford deterministic multi-sieve architecture (1)
- Handles book-length documents (literature!)
- Heuristic rules attribute speaker and addressee of direct speech

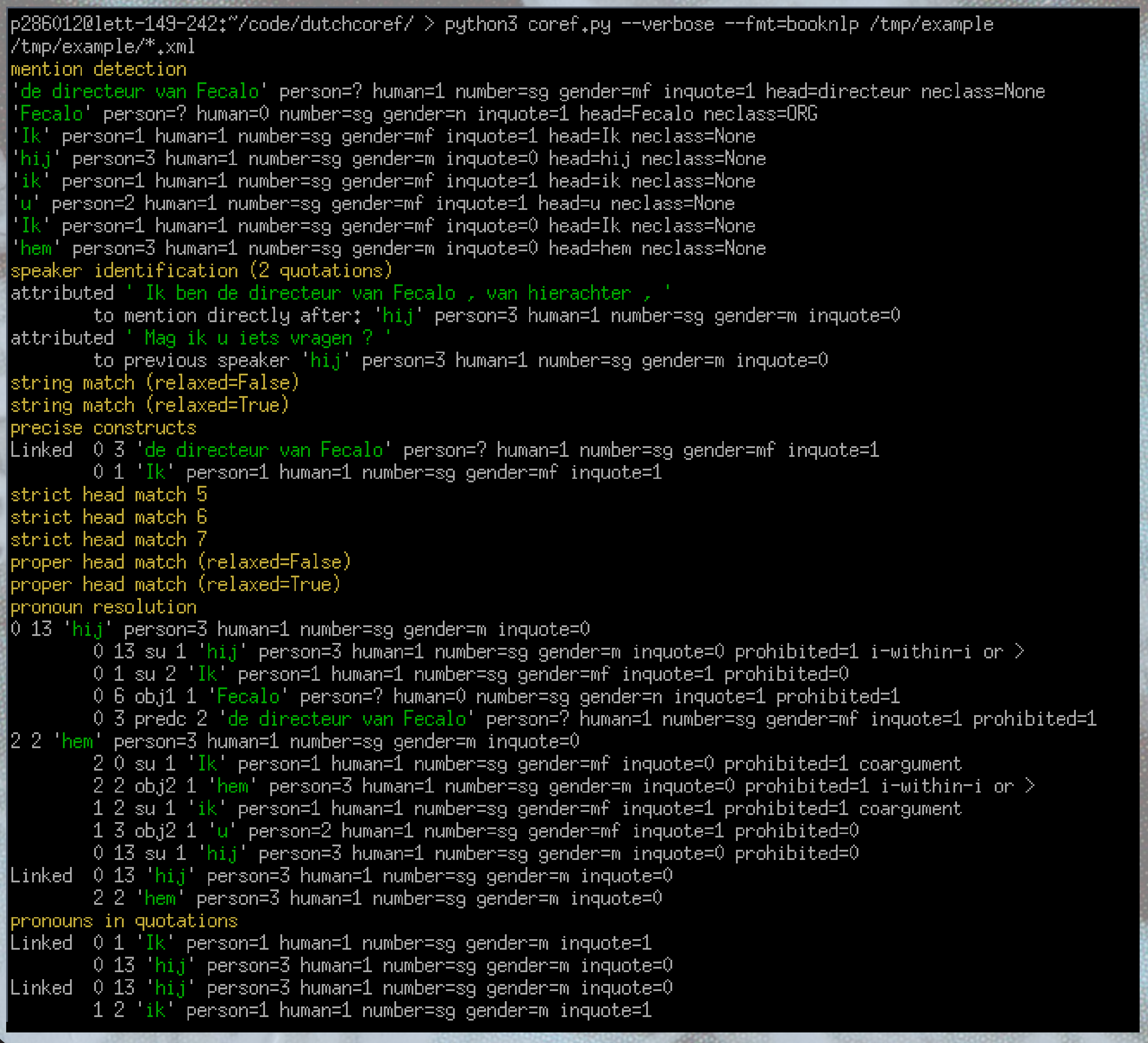
Input: Alpino parse trees (XML files); includes named entities
Output: tabular CoNLL file; columns:

- coreference clusters
- direct speech spans/speakers
- named entities
- universal dependencies

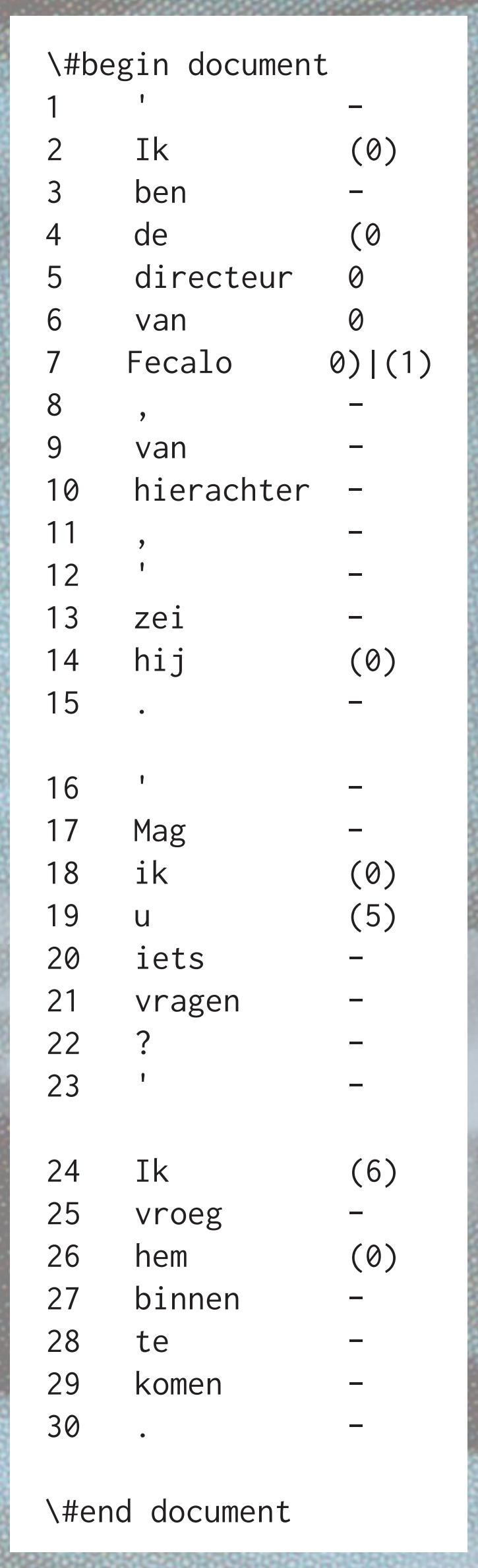
Code: `https://github.com/andreasvc/dutchcoref`

Example (Voskuil, De Buurman)

' Ik ben de directeur van Fecalo , van hierachter , ' zei hij .
' Mag ik u iets vragen ? '
Ik vroeg hem binnen te komen .



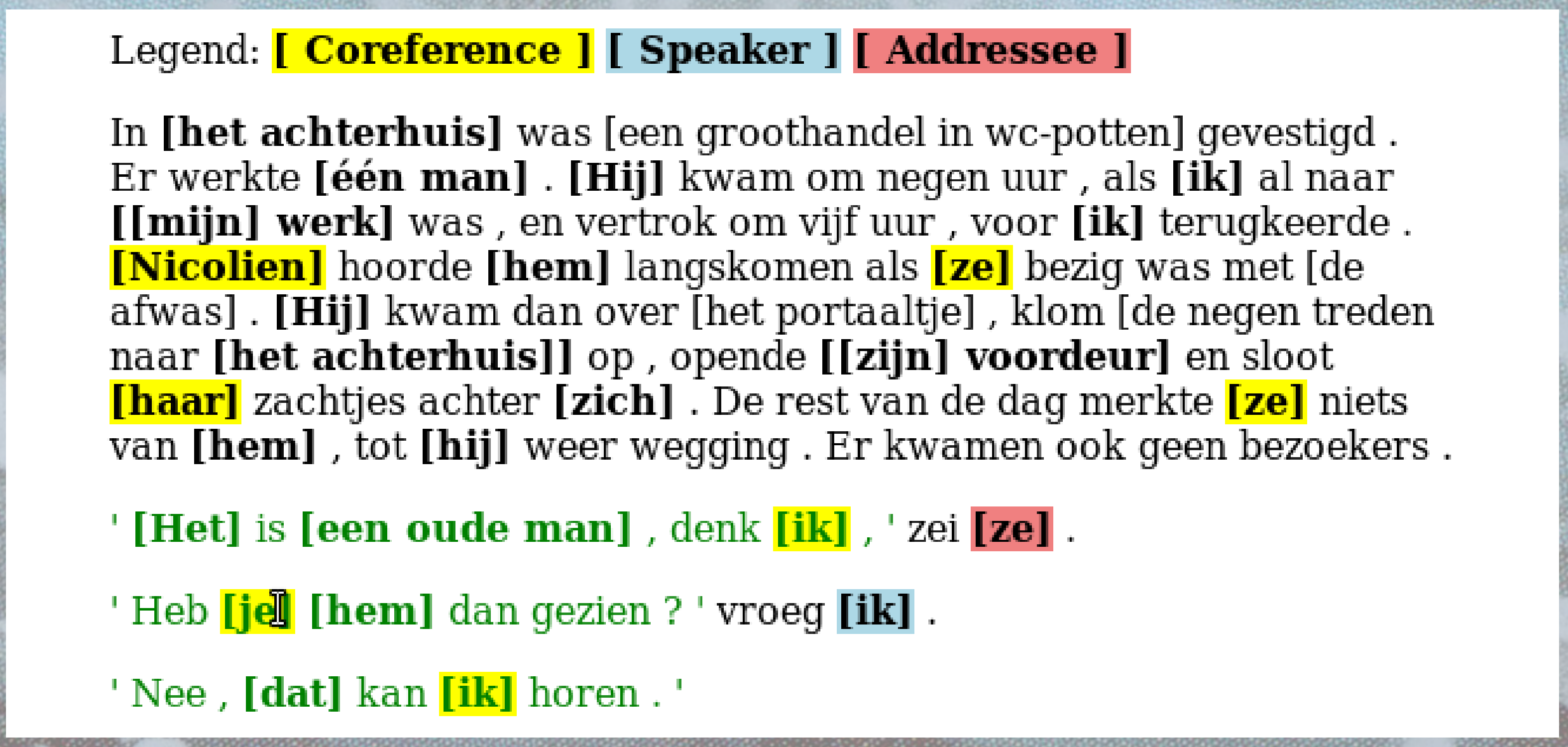
```
p286012@lett-149-242:~/code/dutchcoref/ > python3 coref.py --verbose --fmt=booknlp /tmp/example
/tmp/example/*.xml
mention detection
'de directeur van Fecalo' person=? human=1 number=sg gender=mf inquote=1 head=directeur neclass=None
'Fecalo' person=? human=0 number=sg gender=n inquote=1 head=Fecalo neclass=ORG
'Ik' person=1 human=1 number=sg gender=mf inquote=1 head=Ik neclass=None
'hij' person=3 human=1 number=sg gender=m inquote=0 head=hij neclass=None
'ik' person=1 human=1 number=sg gender=mf inquote=1 head=ik neclass=None
'u' person=2 human=1 number=sg gender=mf inquote=1 head=u neclass=None
'Ik' person=1 human=1 number=sg gender=mf inquote=0 head=Ik neclass=None
'hem' person=3 human=1 number=sg gender=m inquote=0 head=hem neclass=None
speaker identification (2 quotations)
attributed ' Ik ben de directeur van Fecalo , van hierachter , '
        to mention directly after: 'hij' person=3 human=1 number=sg gender=m inquote=0
attributed ' Mag ik u iets vragen ? '
        to previous speaker 'hij' person=3 human=1 number=sg gender=m inquote=0
string match (relaxed=False)
string match (relaxed=True)
precise constructs
Linked  0 3 'de directeur van Fecalo' person=? human=1 number=sg gender=mf inquote=1
        0 1 'Ik' person=1 human=1 number=sg gender=mf inquote=1
strict head match 5
strict head match 6
strict head match 7
proper head match (relaxed=False)
proper head match (relaxed=True)
pronoun resolution
0 13 'hij' person=3 human=1 number=sg gender=m inquote=0
        0 13 su 1 'hij' person=3 human=1 number=sg gender=m inquote=0 prohibited=1 i-within-i or >
        0 1 su 2 'Ik' person=1 human=1 number=sg gender=mf inquote=1 prohibited=0
        0 6 obj1 1 'Fecalo' person=? human=0 number=sg gender=n inquote=1 prohibited=1
        0 3 predc 2 'de directeur van Fecalo' person=? human=1 number=sg gender=mf inquote=1 prohibited=1
2 2 'hem' person=3 human=1 number=sg gender=m inquote=0
        2 0 su 1 'Ik' person=1 human=1 number=sg gender=mf inquote=0 prohibited=1 coargument
        2 2 obj2 1 'hem' person=3 human=1 number=sg gender=m inquote=0 prohibited=1 i-within-i or >
        1 2 su 1 'ik' person=1 human=1 number=sg gender=mf inquote=1 prohibited=1 coargument
        1 3 obj2 1 'u' person=2 human=1 number=sg gender=mf inquote=1 prohibited=0
        0 13 su 1 'hij' person=3 human=1 number=sg gender=m inquote=0 prohibited=0
Linked  0 13 'hij' person=3 human=1 number=sg gender=m inquote=0
        2 2 'hem' person=3 human=1 number=sg gender=m inquote=0
pronouns in quotations
Linked  0 1 'Ik' person=1 human=1 number=sg gender=m inquote=1
        0 13 'hij' person=3 human=1 number=sg gender=m inquote=0
Linked  0 13 'hij' person=3 human=1 number=sg gender=m inquote=0
        1 2 'ik' person=1 human=1 number=sg gender=m inquote=1
```



```
\#begin document
1    '          -
2    Ik         (0)
3    ben        -
4    de         (0
5    directeur  0
6    van        0
7   Fecalo     0)|(1)
8    ,          -
9    van        -
10   hierachter -
11   ,          -
12   '          -
13   zei        -
14   hij        (0)
15   .          -

16   '          -
17   Mag        -
18   ik         (0)
19   u          (5)
20   iets       -
21   vragen     -
22   ?          -
23   '          -

24   Ik         (6)
25   vroeg      -
26   hem        (0)
27   binnen     -
28   te         -
29   komen      -
30   .          -

\#end document
```

Dialogue attribution

Speakers are detected where explicitly mentioned, and this information is extrapolated assuming turn-taking of alternating interlocutors. Interactive HTML visualization:



Legend: **[Coreference]** **[Speaker]** **[Addressee]**

In **[het achterhuis]** was [een groothandel in wc-potten] gevestigd . Er werkte **[één man]** . **[Hij]** kwam om negen uur , als **[ik]** al naar **[[mijn] werk]** was , en vertrok om vijf uur , voor **[ik]** terugkeerde . **[Nicolien]** hoorde **[hem]** langskomen als **[ze]** bezig was met [de afwas] . **[Hij]** kwam dan over [het portaaltje] , klom [de negen treden naar **[het achterhuis]]** op , opende **[[zijn] voordeur]** en sloot **[haar]** zachtjes achter **[zich]** . De rest van de dag merkte **[ze]** niets van **[hem]** , tot **[hij]** weer wegging . Er kwamen ook geen bezoekers .

' **[Het]** is **[een oude man]** , denk **[ik]** , ' zei **[ze]** .

' Heb **[je]** **[hem]** dan gezien ? ' vroeg **[ik]** .

' Nee , **[dat]** kan **[ik]** horen . '

Lexical resources

Pronouns must agree in number, gender, and animacy with names and nouns they corefer with. Look up in external datasets:

- Meertens Voornamenbank (3); e.g., *Marie* ⇒ animate, female
- For nouns, Cornetto (2); e.g., *zoon* ⇒ animate, male; Manually disambiguated multiple senses; e.g., *apparaat* ⇒ inanimate, neuter
- Gender and animacy data extracted with heuristic patterns from web text; e.g., *Barack Obama* ⇒ animate, male

Evaluation: shared tasks

CLIN26 shared task dev. set	Mentions	BLANC
GroRef (4)	60.66	31.48
This Work	**62.01**	**33.21**

SemEval 2010 Dutch dev. set	Mentions	BLANC
Best Dutch SemEval 2010 system	100	65.3
This Work	100	**66.73**

With predicted mentions, performance not good due to different annotation conventions.

Evaluation: Literature

Annotated first 100 sentences of 10 Dutch novels by manually correcting our system output.

Novel	BLANC	mentions	entities
Barnes, AlsofVoorbijls	69.2	372	155
Carré, OnsSoortVerrader	45.0	552	250
Eco, BegraafplaatsVanPraag	65.3	871	465
Eggers, WatIsWat	78.4	411	126
Grunberg, HuidEnHaar	52.1	309	120
James, VijftigTintenGrijs	76.2	328	108
Koch, Diner	71.6	375	136
DeMoor, SchilderEnMeisje	40.6	347	192
Voskuil, Buurman	58.7	198	62
Yalom, RaadselSpinoza	71.7	474	185
Overall	64.4		

Speaker attribution: 45%; addressee: 33%

Comparison with similar work:

	MUC	B^3	BLANC
Krug et al. 2015 (6), German	85.5	56.0	-
This work, Dutch	71.5	65.8	64.4

Challenges, future work

1. Simplified annotation scheme:
   - Only one link type (no bound, bridge, predicative links)
   - Cut off mentions at commas, discontinuity
   - Avoid redundant/overlapping spans
     ((the man) (who) stole my bike)
     ((John) (the painter))
2. Evaluation metrics are problematic, hard to interpret.
3. Train classifiers for:
   - Better quote attribution
   - Mention and singleton detection
   - End-to-end deep learning system based on Sonar 1M word coref. dataset.

References

(1) Lee et al., 2013. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*, 39(4).
(2) Vossen et al., 2009. Cornetto Lexical Database.
(3) Meertens instituut KNAW, 2010. Nederlandse Voornamenbank (Dutch first name database).
(4) van der Goot et al., 2015. GroRef: Rule-Based Coreference Resolution for Dutch. CLIN26 shared task.
(5) Recasens et al., 2010. SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. Proc. of SemEval, pp. 1–8.
(6) Krug et al., 2015. Rule-based coreference resolution in German historic novels. In Proc. of CLFL.

Painting: oil on canvas, 15×10 cm, `www.bittremieux.nl`