

# A Dutch coreference resolution system with an evaluation on literary fiction

Andreas van Cranenburgh

University of Groningen, The Netherlands

November 7, 2019, Düsseldorf, CL colloquium talk

**WHAT DO WE WANT?**



**NATURAL LANGUAGE  
PROCESSING**



**WHEN DO WE WANT IT?**



imgflip.com

**WHEN DO WE WANT WHAT?**



# Plan for today

1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work

1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work

## Definition

Coreference resolution is the task of clustering mentions in text that refer to the same underlying real world entities.



## Definition

Coreference resolution is the task of clustering mentions in text that refer to the same underlying real world entities.

[illegible]

- ▶ Entity 1 = {Obama, he}
- ▶ Entity 2 = {I, my, she}

# Mentions

## Definition

**Mention** or referring expression: span of text that refers to a person or object in the real or mention world.

NB: contrast with **markable**, a *potentially* referring expression.



# Mentions

## Definition

**Mention** or referring expression: span of text that refers to a person or object in the real or mention world.

NB: contrast with **markable**, a *potentially* referring expression.

**Pronouns** I, he, my, his, that, [each other], himself, ...

**Names** [John], [John Smith], [Mr. Smith], ...

**Nominals** [the man], [the flowers on [the table]], ...

But not:

- Events, actions, times

## Coreference: mentions of the same entity

- ▶ [John]<sub>1</sub> sees [Mary]<sub>2</sub>. [He]<sub>1</sub> waves at [her]<sub>2</sub>.
- ▶ [Bond]<sub>1</sub>, [James Bond]<sub>1</sub>.
- ▶ [I]<sub>1</sub> took [[my]<sub>1</sub> bike]<sub>2</sub>.

# Winograd schemes

The [city councilmen]<sub>1</sub> refused [the demonstrators]<sub>2</sub> a permit because ...

1. ...[they]<sub>1</sub> feared violence.
2. ...[they]<sub>2</sub> advocated violence.

“AI-complete” problem

# History of coreference resolution

Various datasets, languages:

1996 MUC-6 shared task, English

2004 ACE shared task, English/Chinese/Arabic

2010 SemEval shared task, multilingual including Dutch

2011 CoNLL shared task, English

2012 CoNLL shared task, English/Chinese/Arabic

## State of the art: from rules to a neural arms race ...

OntoNotes (English), CoNLL scores:

CoNLL 2011	shared task, winner: Lee et al., rule-based	58.3%
CoNLL 2012	shared task, winner: Fernandes et al., perceptron	58.7%
EMNLP 2017	end-to-end coref. resolution, deep learning	67.2%
NAACL 2018	e2e + ELMO + c2f, deeper learning	73.0%
EMNLP 2019	e2e + BERT Large, even deeper learning	76.9%

# Evaluation metrics

Coreference evaluation is a **mess**!

Fatally flawed metrics:

**1996** MUC

**1998** B<sup>3</sup>

**2005** CEAFm, CEAF<sub>e</sub>

**2011** CoNLL score (= avg of MUC, B<sup>3</sup>, CEAF<sub>e</sub>)

**2011** BLANC

No known issues (yet!):

**2016** Link-based Entity-Aware metric (LEA)

Moosavi & Strube (ACL 2016) Which coreference evaluation metric do you trust?  
A proposal for a link-based entity aware metric

1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work

By the way ...

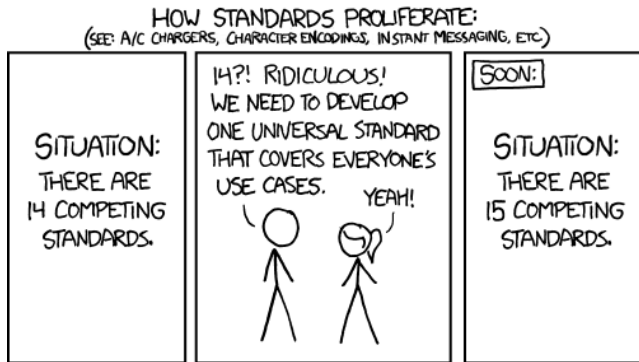
## #BenderRule:

The rest of this talk is about Dutch!

<https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/>



# I made a new annotation scheme ...



# Annotation scheme

Simplified annotation scheme:

- ▶ Annotate mentions: include singletons, exclude non-referring expressions.
- ▶ Avoid difficult mention boundaries: no discontinuity, relative clauses
- ▶ Only annotate entity clusters, not directed anaphor-antecedent relations

# Annotation workflow

1. Tokenize, parse with Alpino
2. Run coreference system
3. Manually correct output with CorefAnnotator
4. Optional: correction by second annotator

<http://www.let.rug.nl/vannoord/alp/Alpino/>  
<https://github.com/nilsreiter/CorefAnnotator/>

# Annotation workflow

1. Tokenize, parse with Alpino
2. Run coreference system
3. Manually correct output with CorefAnnotator
4. Optional: correction by second annotator

**Result:** tabular CoNLL 2012 file

```
#begin document (example); part 000
example 1      John      (0)
example 2      sees      -
example 3      Mary      (1)

#end document
```

## Annotated texts

	CLIN26 dev set	SemEval 2010 dev	Novels, dev set	Novels, test set
documents	30	23	10	11
tokens	4018	9164	19,051	88,092
sents per doc	7	21.4	100	491.5
avg sent len	19.3	18.4	19.0	16.3

## Annotated texts

	CLIN26 dev set	SemEval 2010 dev	Novels, dev set	Novels, test set
documents	30	23	10	11
tokens	4018	9164	19,051	88,092
sents per doc	7	21.4	100	491.5
avg sent len	19.3	18.4	19.0	16.3
entities	273	424	1798	8337
mentions	663	1010	4243	20,873
% pronouns	7.69	14.45	43.3	36.5
% nominal	52.34	54.35	46.2	52.2
% names	39.97	31.20	10.5	11.2

107k tokens of annotated literary text!

# Mention detection

1. Extract candidate constituents
2. Adjust spans
3. Filter with patterns
4. Detect features

- ▶ [John], [who], [the book on [the table]], ...
- ▶ [Mr Smith], [San Jose, California], ...
- ▶ Drop: yesterday, about two meters, ...
- ▶ gender, animacy, number

# Mention feature detection

**parse** features in the Alpino parse tree (HPSG-inspired)

**NER** part of Alpino; person/org/loc/misc

**wordnet** animacy & gender of head nouns (hand-corrected)

**web text** names extracted w/heuristic patterns from 30GB English text



# Mention feature detection

**parse** features in the Alpino parse tree (HPSG-inspired)

**NER** part of Alpino; person/org/loc/misc

**wordnet** animacy & gender of head nouns (hand-corrected)

**web text** names extracted w/heuristic patterns from 30GB English text

	Pronouns	Nominals	Names
Number	parse	parse	web text
Gender	parse	wordnet	web text
Animacy	parse	wordnet	NER, web text

# Sieves: link mentions with deterministic rules

**Quote attribution** find speaker & addressee of direct speech

**String match** [The boy]<sub>1</sub> ... [The boy]<sub>1</sub> ...

**Precise constructs** [the boy]<sub>1</sub> [who]<sub>1</sub> ...

**Head match** [The clever boy]<sub>1</sub> ... [the boy]<sub>1</sub>

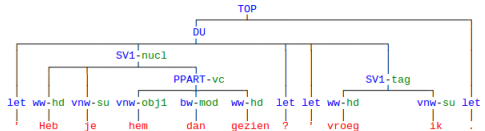
**Proper head noun match** [Bond]<sub>2</sub>, [James Bond]<sub>2</sub>

**Pronoun resolution** [He]<sub>1</sub> ... [his]<sub>1</sub> ...

# Demo

Move mouse over bracketed text to highlight coreference. Move mouse over direct speech to highlight speaker and addressee. Click on a sentence to toggle the display of its parse tree.

Legend: [ Singleton ] [ **Coreference** ] [ **Speaker** ] [ **Addressee** ] ' Direct speech '



[XML source](#)

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 portaalje] , 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

8 3 human=1 gender=m number=sg person=3 inquote=1 necclass=None  
head=hem

Er werd gebeld  
slecht zittend , wat te raam , grijs kostuum .

<https://andreascvc.github.io/voskuil.html>

1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work

## Evaluation: shared tasks

CLIN26 shared task	Mentions	BLANC			
GroRef, Boeing test set	59.34	30.96			
This Work, Boeing test set	<b>59.49</b>	<b>31.48</b>			
GroRef, GM test set	<b>60.40</b>	<b>31.31</b>			
This Work, GM test set	59.26	31.07			
GroRef, Stock test set	53.70	25.40			
This Work, Stock test set	<b>54.68</b>	<b>26.09</b>			
SemEval 2010, Dutch, test set	Mentions	BLANC	MUC	B <sup>3</sup>	CEAFm
SemEval 2010: Sucre	42.3	<b>46.9</b>	29.7	11.7	15.9
SemEval 2010: UBIU	34.7	32.3	8.3	17.0	17.0
This Work	<b>64.27</b>	41.48	<b>51.95</b>	<b>45.85</b>	<b>51.20</b>

## Evaluation: novels

	mention F1	recall	precision	LEA F1
SemEval 2010 (test set)	64.27	36.00	39.96	37.88
CLIN26 shared task (Boeing test set)	59.49	29.83	33.95	31.76
Literary texts (dev set)	87.05	57.13	61.71	59.33
Literary texts (test set)	87.10	49.27	57.45	53.05

# Discussion

- ▶ High variance among novels; matter of style?
- ▶ Better performance on novels than news! Surprising?
  - ▶ More dialogue and pronouns in novels (some long chains)
  - ▶ Novels are longer documents (including our annotated fragments)
  - ▶ Not all errors are created equal ...

1. Background
2. Annotating Dutch novels
3. The coreference system
4. Evaluation
5. Future work



# Improve components

## Components:

- ▶ Mention detection/spans
- ▶ Pleonastic pronoun detection
- ▶ Gender/animacy
- ▶ Quote attribution
- ▶ Pronoun resolution

# Improve components

## Components:

- ▶ Mention detection/spans
- ▶ Pleonastic pronoun detection
- ▶ Gender/animacy
- ▶ Quote attribution
- ▶ Pronoun resolution

## Procedure:

1. Acquire/annotate more data
2. Train supervised classifier
3. ???
4. Profit!

# Neural coreference

Train  
End-to-end coreference system  
with BERT  
for Dutch ...

Kenton Lee et al. (EMNLP 2017) End-to-end neural  
coreference resolution

**RULE-BASED**



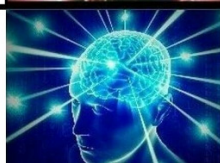
**STATISTICAL**



**NEURAL**



**BERT**



# THE END

**Code:** <https://github.com/andreascv/dutchcoref>

**Paper:** Coming soon™

**Thanks** to my BSc thesis students for helping with annotation!

