

# An End-to-end Environment for Research Question-Driven Entity Extraction and Network Analysis

Andre Blessing, Nora Echelmeyer, Markus John, and Nils Reiter

## Eschenbach's *Parzival*

- Arthurian grail novel
- Written 1200 - 1210 CE
- Versed Middle High German
- 16 books, over 20k lines

...  
"dû nennest ritter: waz ist daz?  
hâstu niht gotlîcher kraft,  
sô sage mir, wer gît ritterschaft?"  
"daz tuot **der kûnec Artûs**.  
**junchêrre**, komt ir in **des hûs**,  
der bringet iuch an ritters namen,  
daz irs iuch niemer durfet schamen.  
ir mugt wol sîn von ritters art."  
...

## Entities & Entity References

- NLP-like workflow: Annotation guidelines, parallel annotation, adjudication
  - training of prediction tools independent of research question
- Annotation Concept
  - Named entities, e.g., 'Parzivâl'
  - Appellative NPs, e.g., 'the knight'
- Annotated Corpus

Book	Lines	Tokens	PER	LOC
III	1,898	12,015	610	120
IV	1,338	8,035	464	122
V	1,682	10,441	472	140
VI	1,740	10,918	594	144
VII	1,800	11,358	687	134
Mean	1,691.6	10,553.4	565.4	132
SD	213.2	1,522.5	95.7	10.7

- Automatic Detection
  - CRF-based system
  - Features: Surface, PoS (Echelmeyer et al., 2017), Case Lookup, Unicode character pattern, Gazetter

		Person		Location	
		Prec	Rec	Prec	Rec
strict	BL <sub>NER</sub>	27.3	1.2	27.6	2.4
	BL <sub>Case</sub>	36.2	19	0	0
	ERT	71.2	56.8	71.8	48
loose	BL <sub>NER</sub>	72.9	3.6	41.9	3.9
	BL <sub>Case</sub>	74.8	38.5	0	0
	ERT	91.6	76.1	85.3	57.9

## Entity Grounding

- Disambiguation of entity references wrt to given cast

Character	#ER	#Proper	Ratio
Parzivâl	427	111	25.8
Gâwân	185	118	63.8
Artûs	128	88	68.8
Jeschûte	103	30	29.1
Clâmidê	74	47	63.5
Herzeloyde	69	9	13

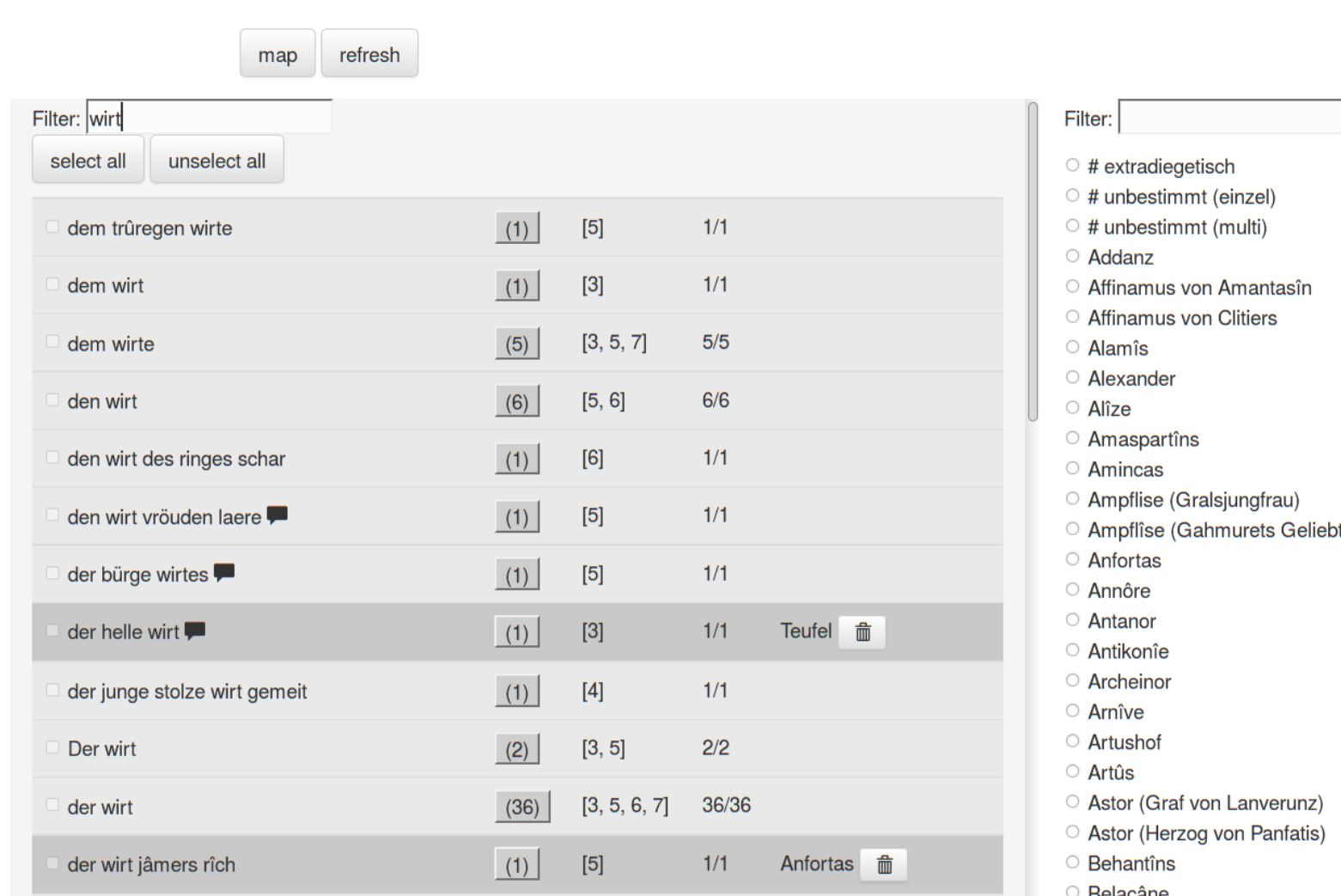


Fig. 1: Entity Grounding Tool

## Segmentation

- Closely tied to H research question
  - No generic abstraction layer
  - ⇒ Interactive experimentation with different options (John et al., 2016)
- Segmentation based on
  - Linguistics: Sentences
  - Structure: Strophes (30 lines)
  - Content: Episodes in the plot

## Network Creation

- Network
  - Vertices: Entities
  - Edges: If the two entities appear in one segment, *except* direct speech, narrator comments, embedded entities

## Network Analysis

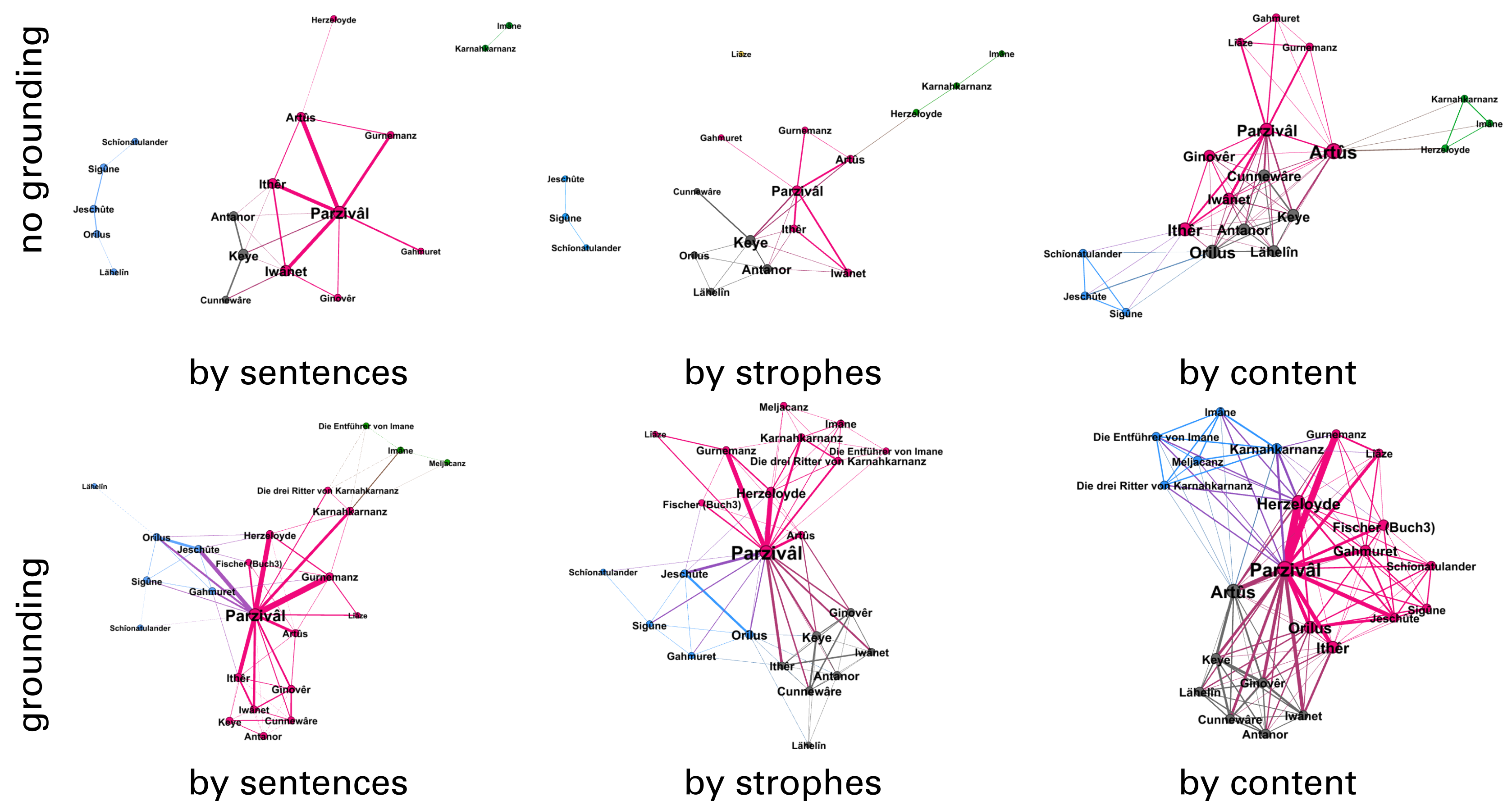


Fig. 2: Comparison of Resulting Networks (Book III)

## Insights

- Shorter segments make grounding more important
  - Otherwise relations are missing
- Exact boundaries of automatic entity reference detection not that important

## References

- Nora Echelmeyer, Nils Reiter, and Sarah Schulz. 2017. Ein PoS-Tagger für "das" Mittelhochdeutsche. In *Book of Abstracts of DHD 2017*. Bern, Switzerland, pages 141–147.
- Markus John, Steffen Lohmann, Steffen Koch, Michael Wörner, and Thomas Ertl. 2016. Visual analytics for narrative text - visualizing characters and their relationships as extracted from novels. In *Proceedings of the IVAPP*. pages 27–38.