# Identifying similarities in text analysis: Hierarchical clustering (linkage) versus network clustering (community detection)

Jeremi K. Ochab (Jagiellonian University, Kraków)
Joanna Byszuk (Institute of Polish Language PAS, Kraków) @jbyszuk
Steffen Pielström (Universität Würzburg, Würzburg) @StPielstroem
Maciej Eder (Institute of Polish Language PAS, Kraków) @MaciejEder

DH2019 Conference in Utrecht

# Problem

# Why is this important?

- Clusters and networks are commonplace as visualizations of stylometry results
- Clustering techniques have so far not been evaluated systematically
- Stylometrists have to rely on habit rather than fact-based recommendations

# Experiment setup

# Data

## 25 expected authorial clusters

- English, French and German literature corpora
- previously used in various studies on text distance measures (e.g. Jannidis et al. 2015)

## binary problems

- 17th century French drama (Schöch) labeled as comedies or tragedies
- Latin verse and prose from the so-called Golden Age
- Latin historiography – texts from Golden Age (late first century BCE), and the “Silver Age”

# Methods – 3 clustering quality measures

- Adjusted for baseline value in case of random clustering, but not selection bias:
  - ARI (Adjusted Rand Index; Hubert and Arabie 1985),
  - AMI (Adjusted Mutual Information)
- Not adjusted:
  - NMI (Normalized Mutual Information)

# Clustering setup

I. number of MFWs: 100 – 1000, iterated by 100,

II. distance measure: Classic and Cosine Deltas,

III. linkage method:

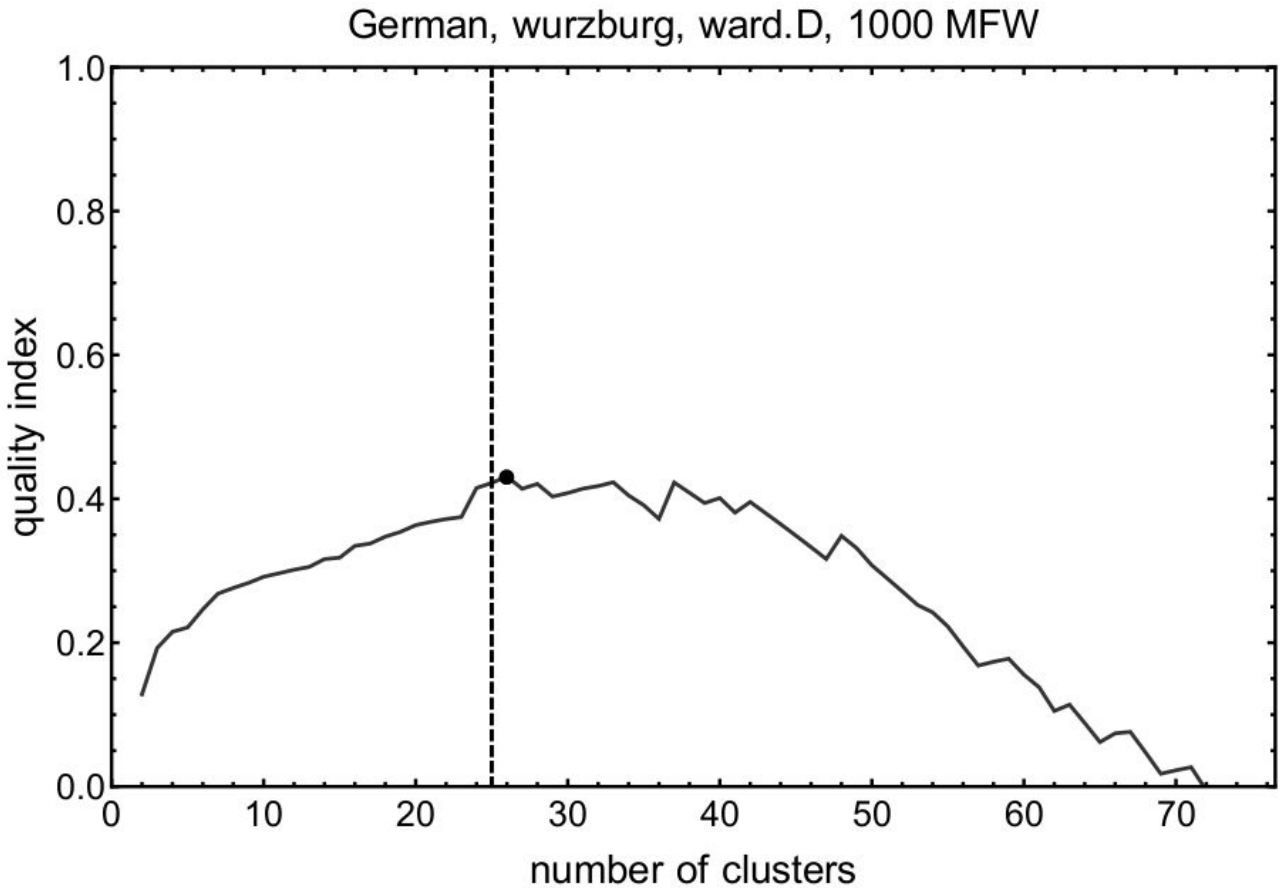
- Ward in two implementations - “ward.D”, “ward.D2”,
- Single link – "single",
- Complete-link – "complete",
- Average-link – "average",
- McQuitty's – "mcquitty,
- k-median – "median",
- k-means – "centroid".

# Methods of community detection in networks

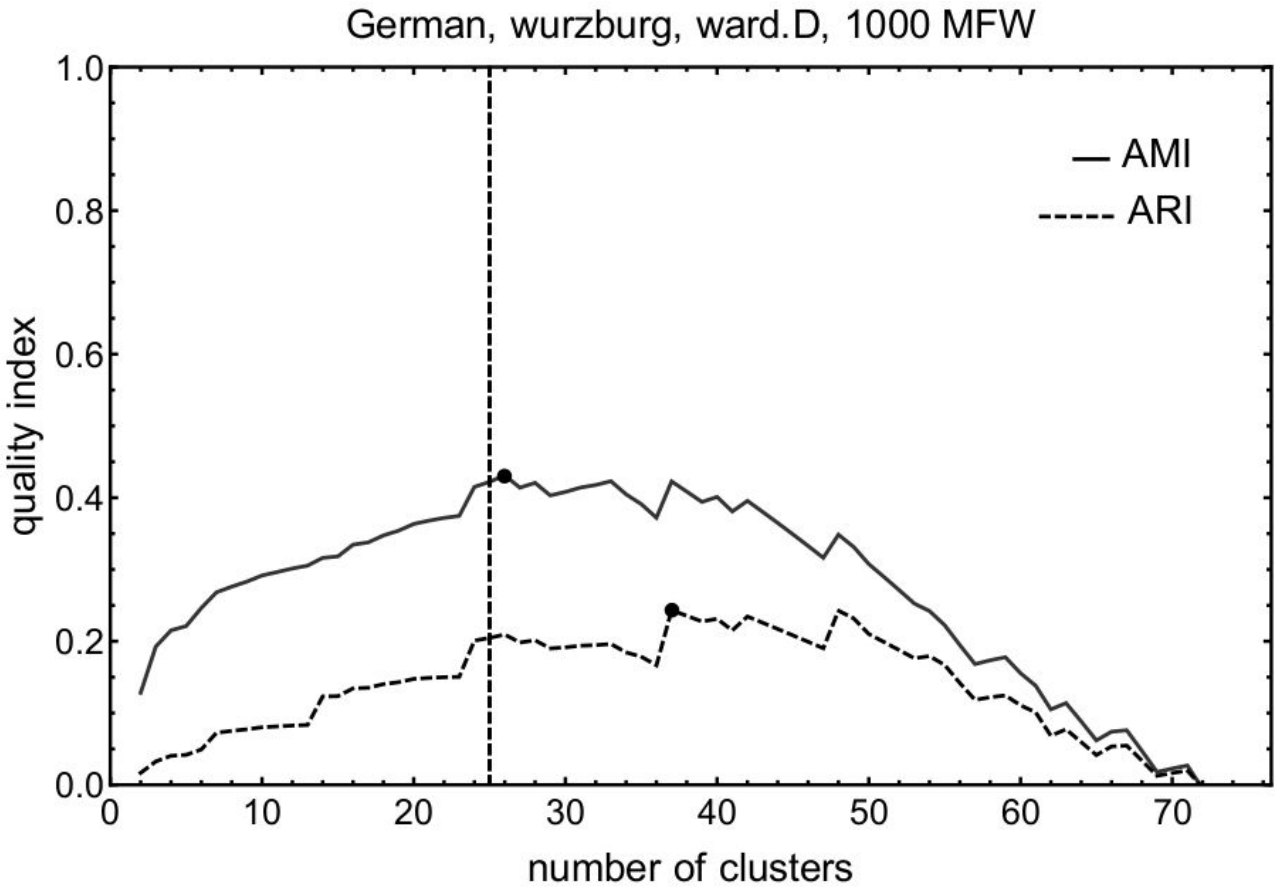
- Networks based on Bootstrap Consensus Tree, 100-1000 MFW, Delta and Cosine Delta
- "A not so small collection of clustering methods" (Lancichinetti and Fortunato 2012):
  - OSLOM,
  - Infomap,
  - label propagation method,
  - modularity optimization by simulated annealing,
  - Louvain method

# Results

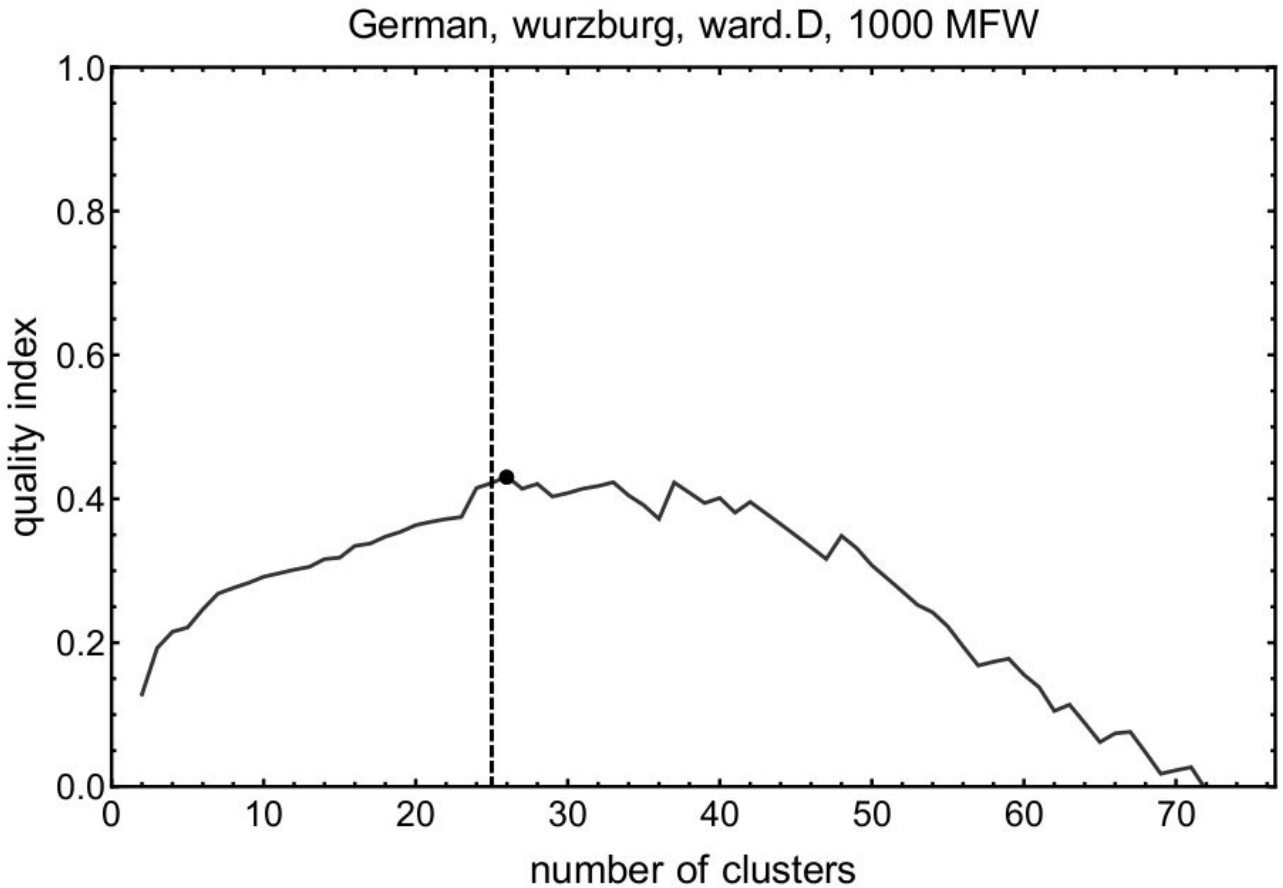
# Which evaluation measure?



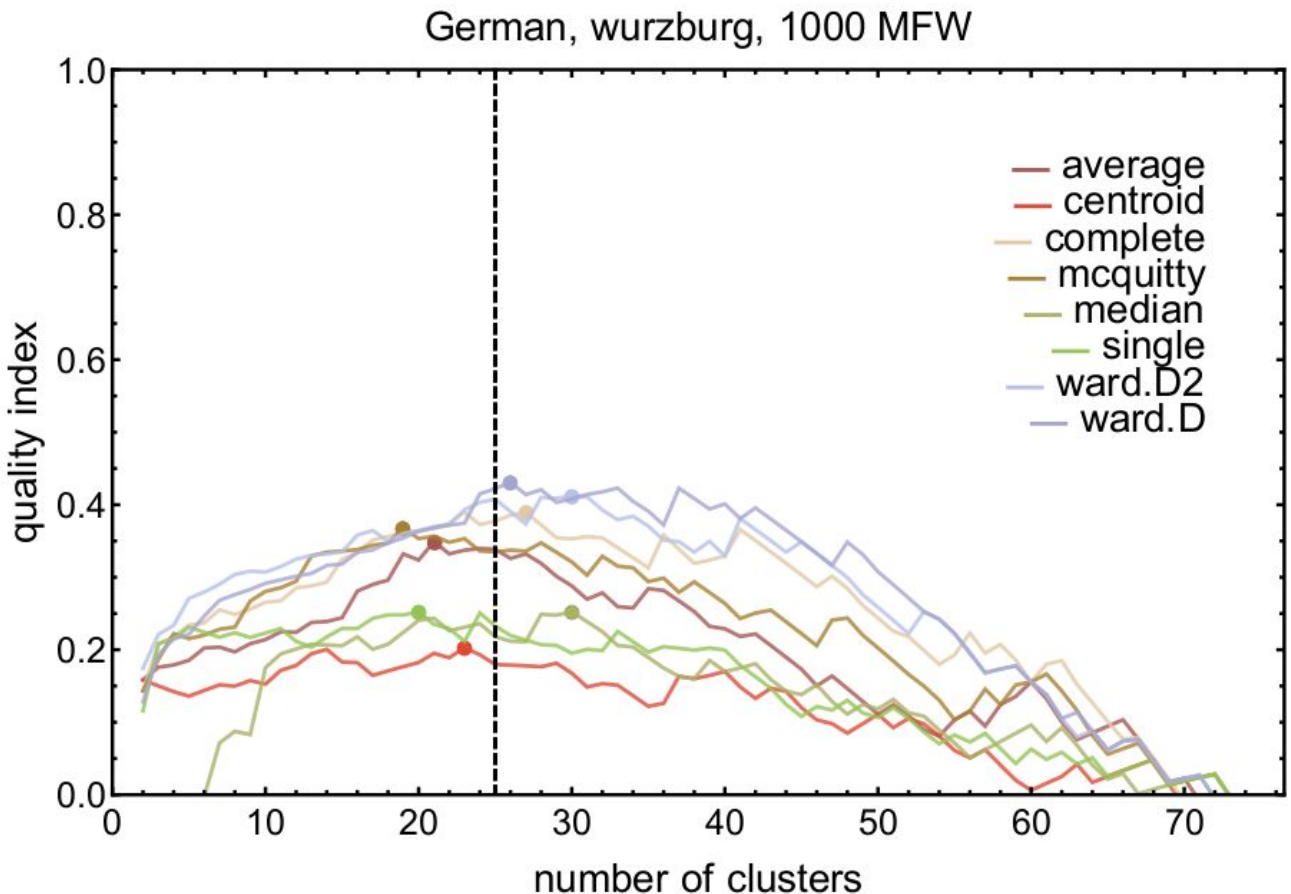
# Which evaluation measure?



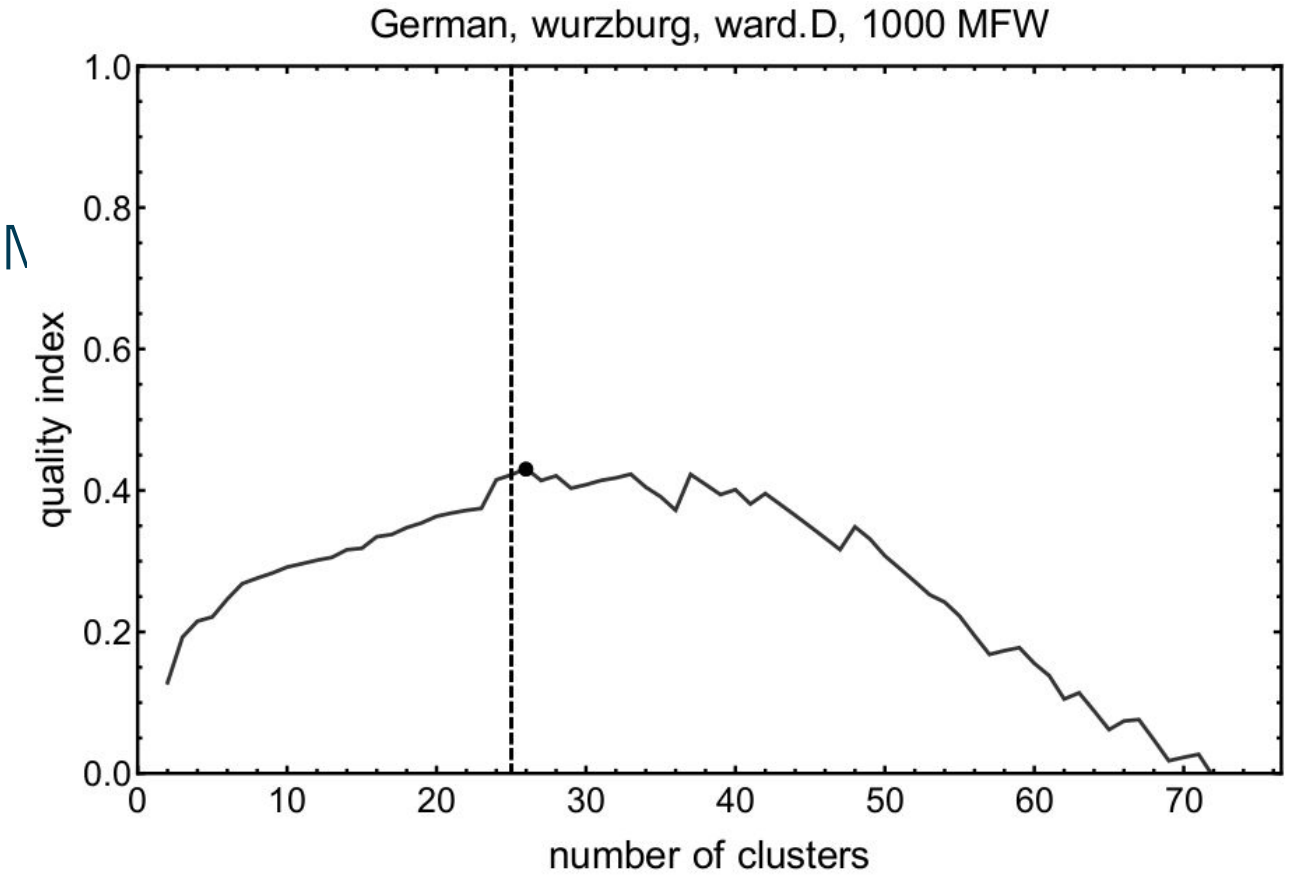
# Which evaluation measure?



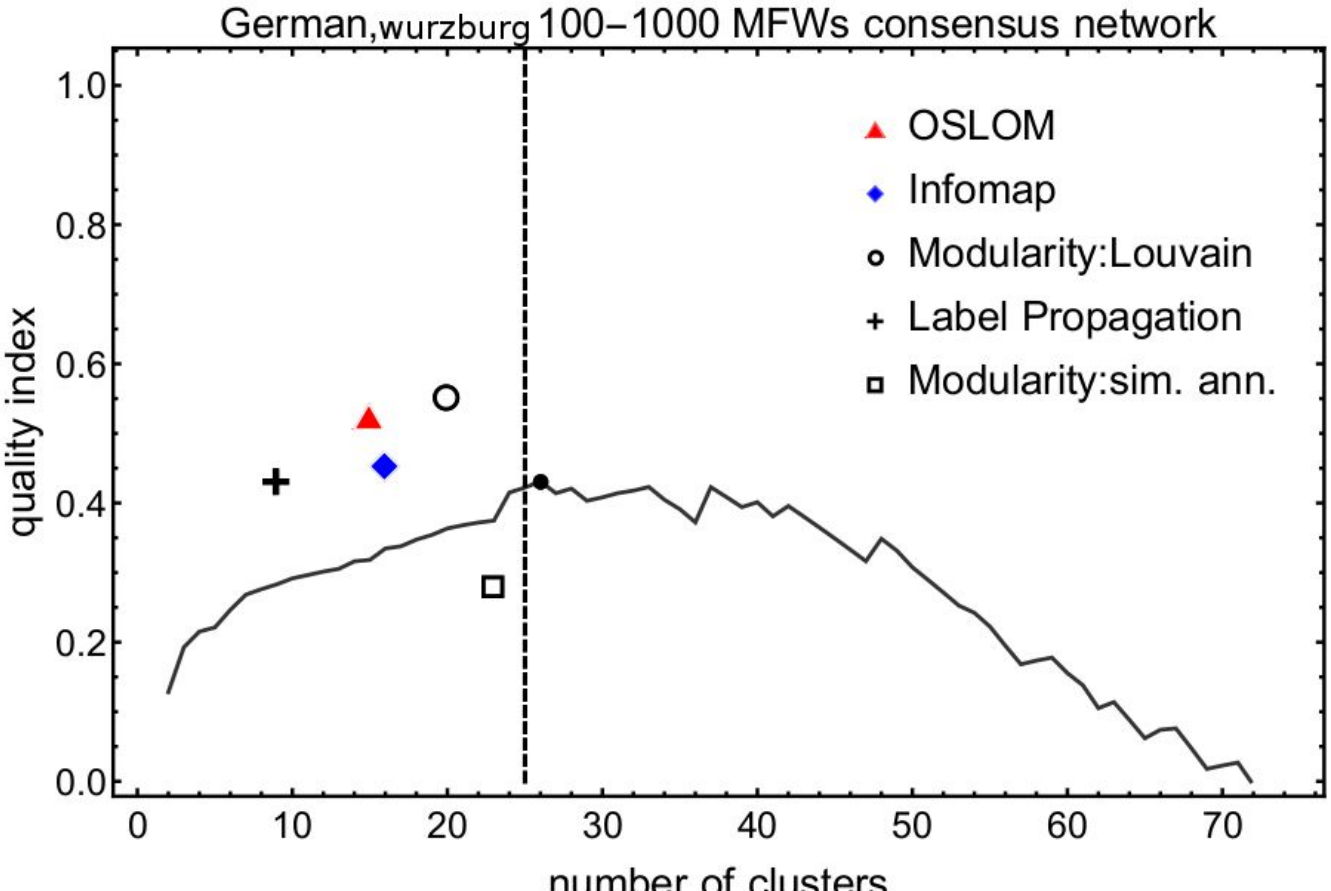
# Which linkage method?



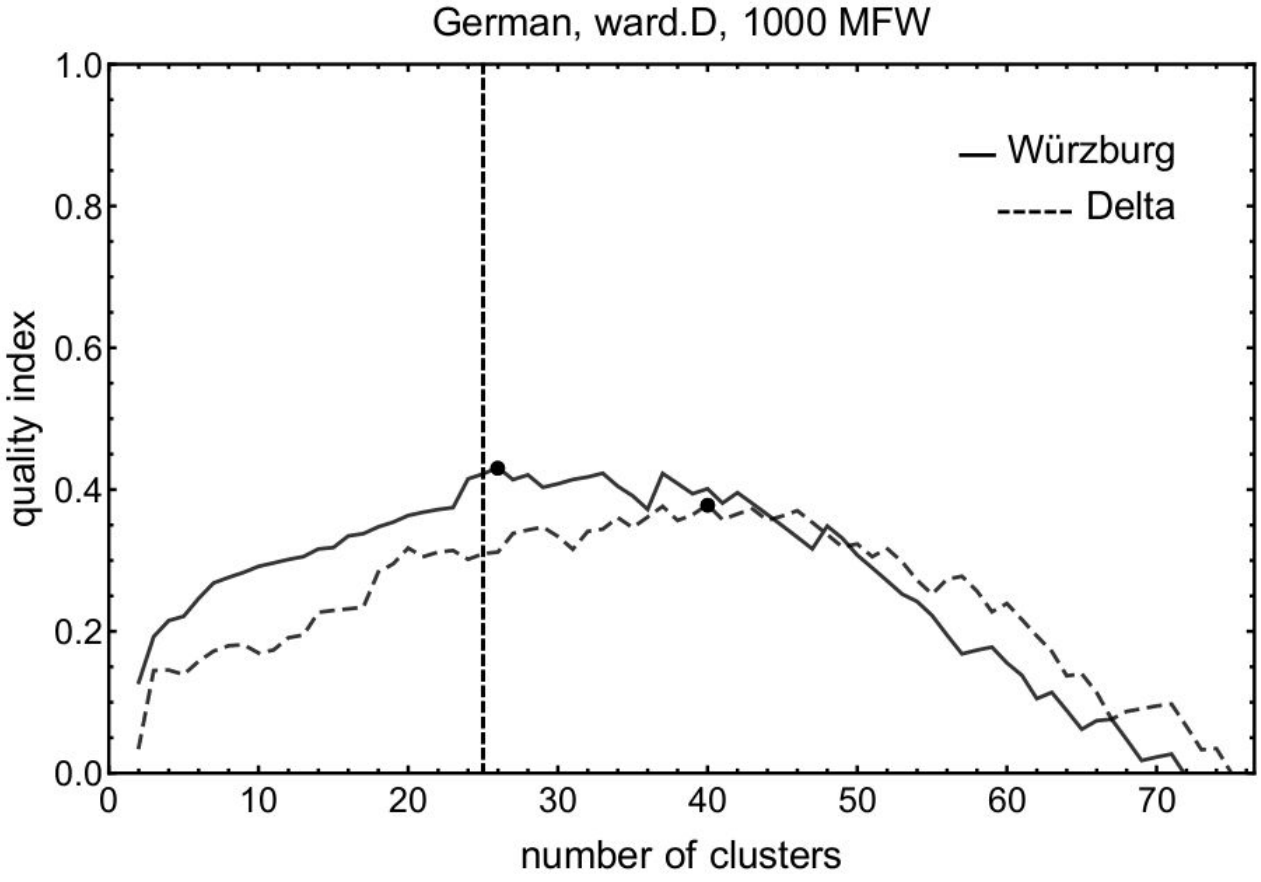
# Which linkage method?



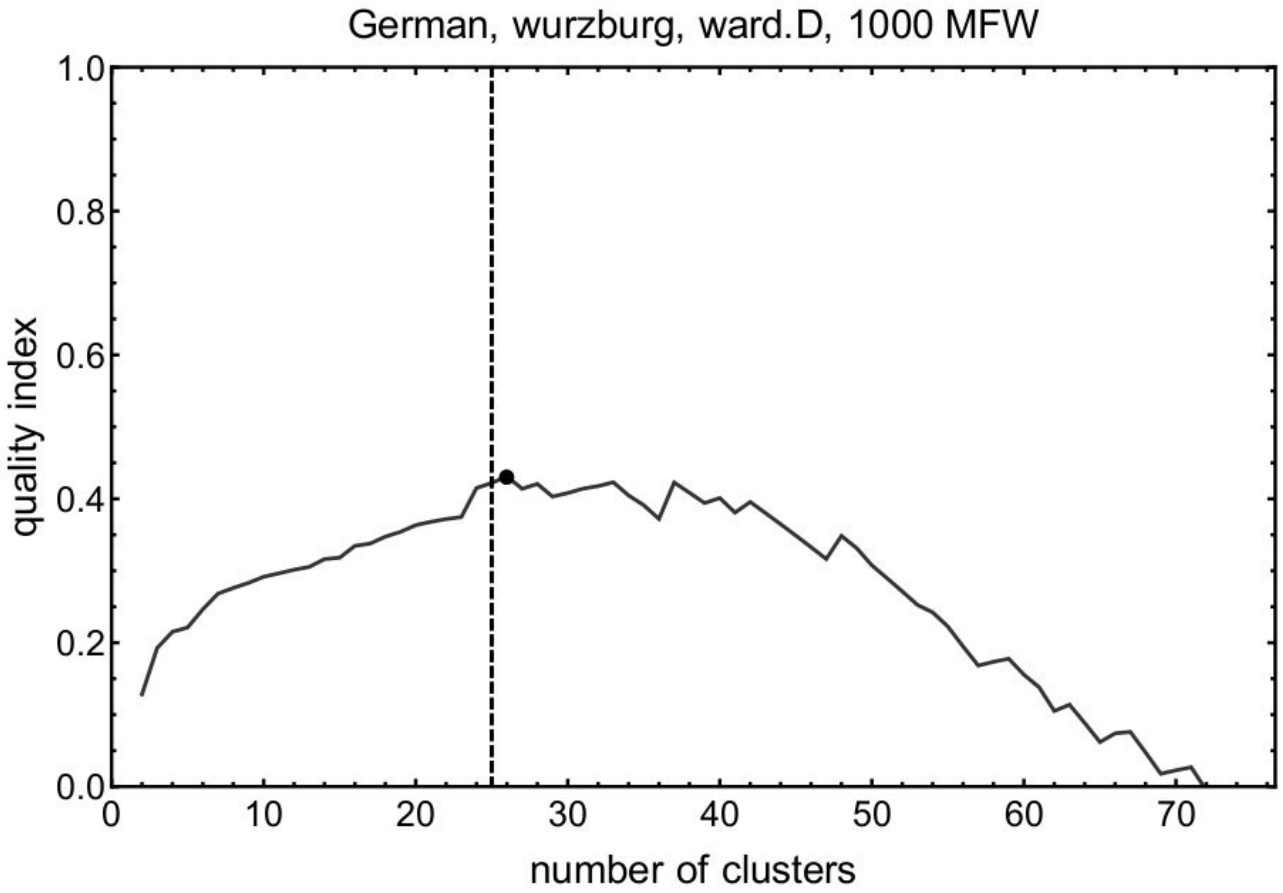
# Which community detection method?



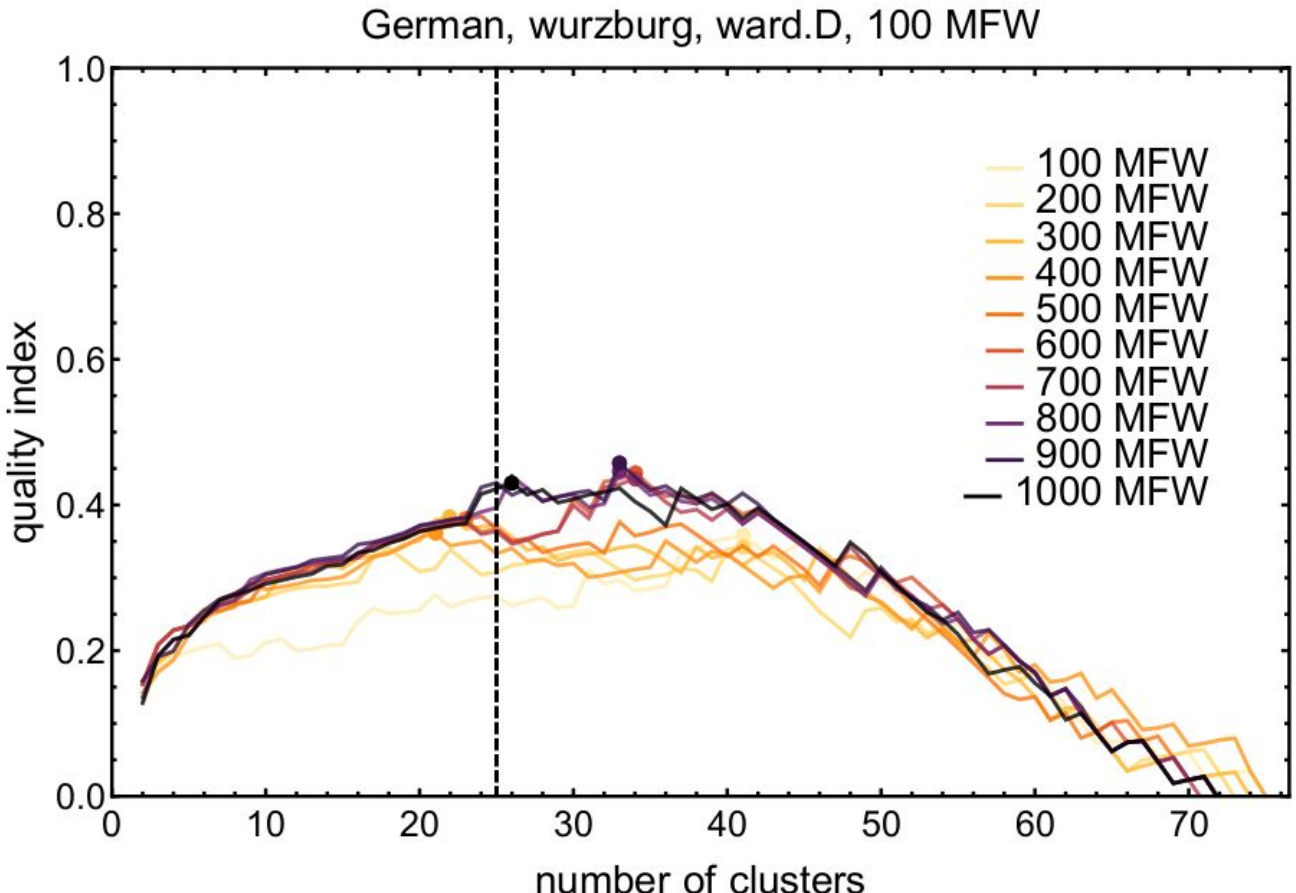
# Which distance measure?



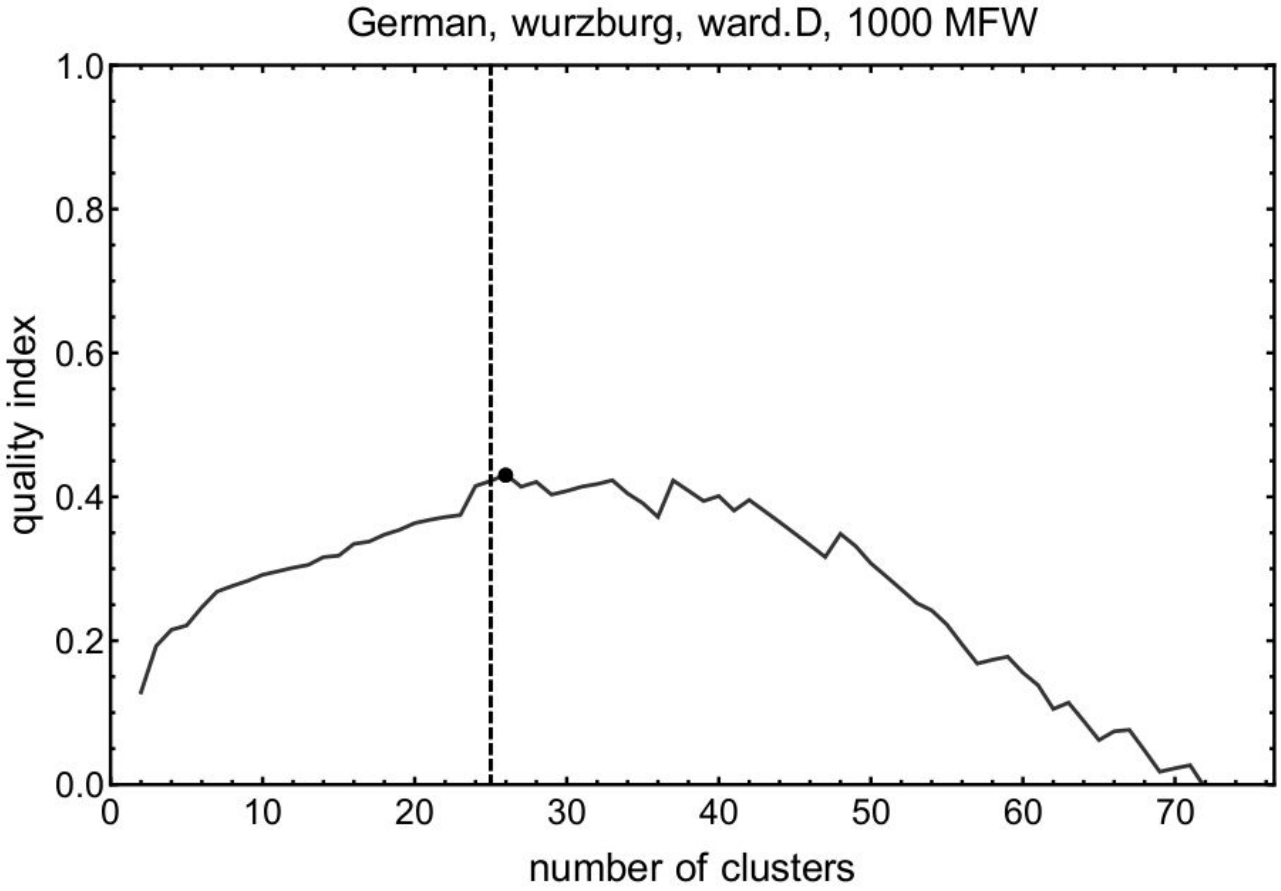
# Which distance measure ?



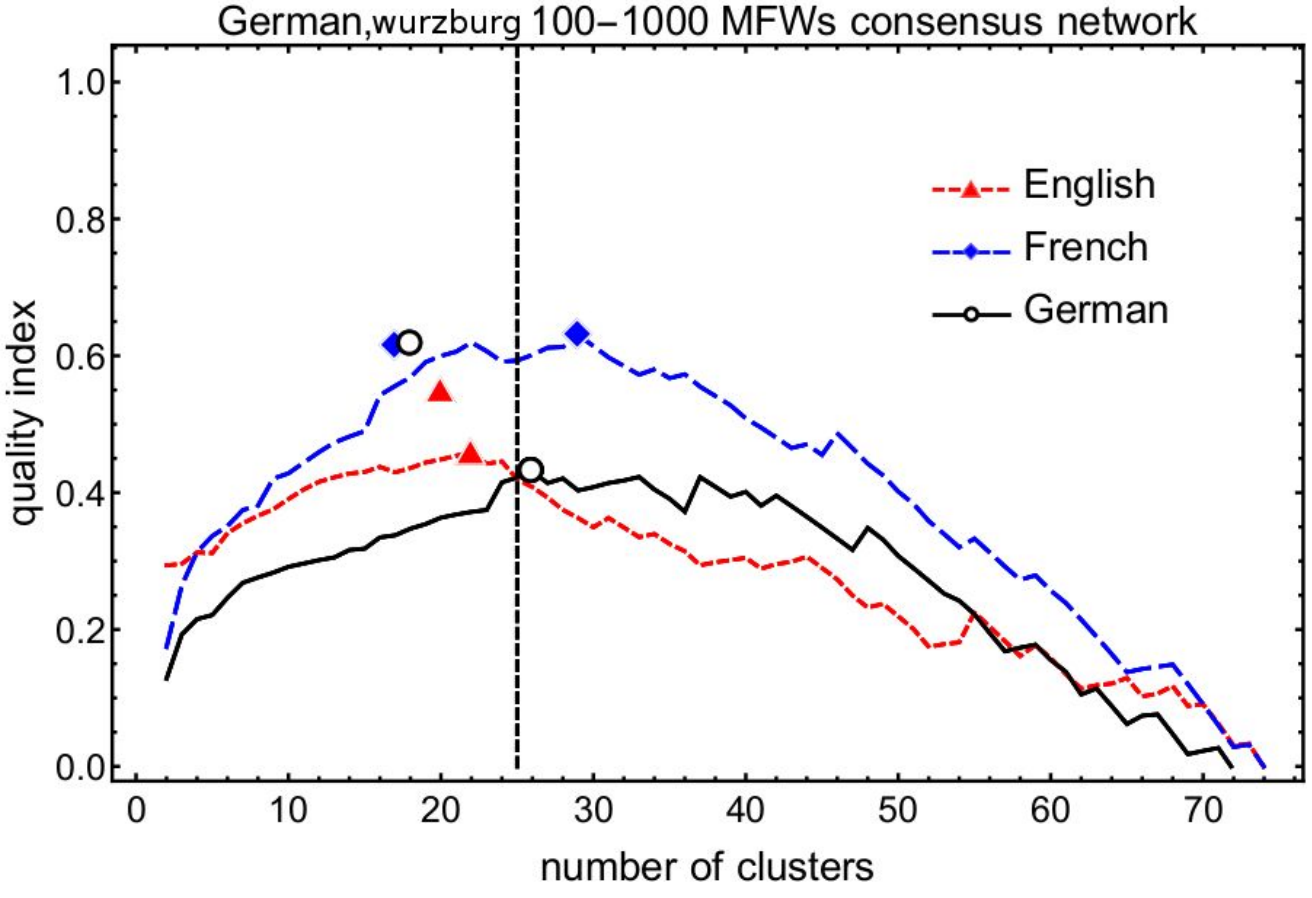
# How many MFW?


German, wurzburg, ward.D, 100 MFW
quality index
number of clusters
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
100 MFW
200 MFW
300 MFW
400 MFW
500 MFW
600 MFW
700 MFW
800 MFW
900 MFW
1000 MFW

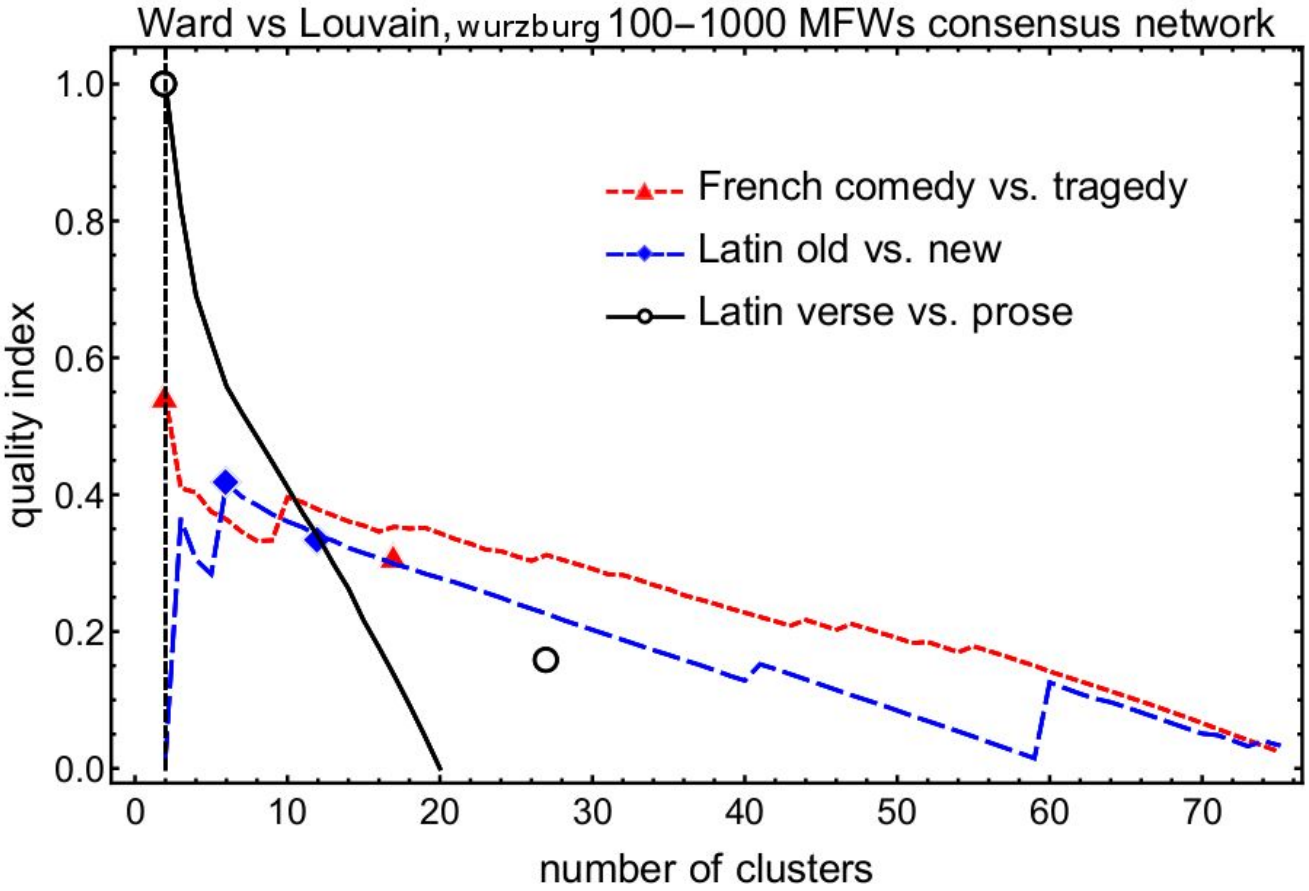
# How many MFW?


German, wurzburg, ward.D, 1000 MFW
quality index
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
number of clusters

# Stability across languages



# Binary problems



# Concluding remarks

# Recommendations

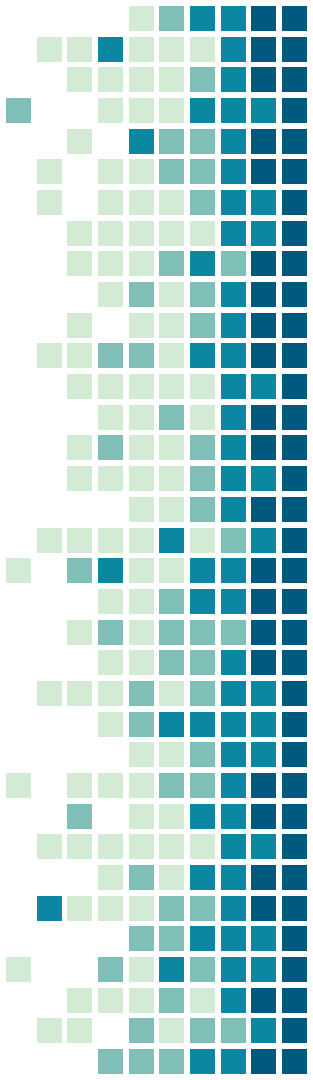
## Best measures:

- Ward (clustering)
- Cosine Delta aka Wurzburg (distance)

## Networks:

- Louvain method
- Results depend on the number of classes

Hierarchical clustering approximates number of clusters from above, network methods from below.



# Acknowledgements

We are grateful for the financial support we received:

JB was partially funded for the research by Poland's National Science Centre (grant number 2017/26/HS2/01019),

SP contributed to this research as part of a Short Term Scientific Mission financed by the EU COST Action "Distant Reading" (CA16204),

ME was partially funded by the National Science Centre (grant number 2014/12/W/ST5/00592).

Distant Reading

National Science Centre Poland

Julius-Maximilians-Universität Würzburg

Jagiellonian University in Kraków

PAN IJP Institute of Polish Language Polish Academy of Sciences

10 Computational 01
01 Stylistics 0101000
11 Group 011010110

# Thank you!

Documentation and resources:
https://github.com/JoannaBy/hierarchical-vs-network-clustering