6th Workshop of the
Cross-Language
Evaluation Forum (CLEF)
Vienna 23 Sept. 2005

Web Retrieval Experiments with one Multilingual Index at the University of Hildesheim

**Niels Jensen, René Hackl,
Thomas Mandl, Robert Strötgen**

Information Science

Universität Hildesheim

mandl@uni-hildesheim.de

Overview

- Challenges
- Approach
- Experiments in Hildesheim
- Post Submission Runs

Language Identification List

- 15 % of all docs: unknown
- other 85 %: 2.3 languages on average
- Intellectual analysis of some 700 pages from CZ domain (annotated list)
  - Set of pages not identified as Czech
  - 85 % inaccurate + 4 % wrong (Hofman Miquel 2005)
-

  -> development of new language identification tool (evaluation ongoing)
  (Artemenko et al. @ LWA 2005)

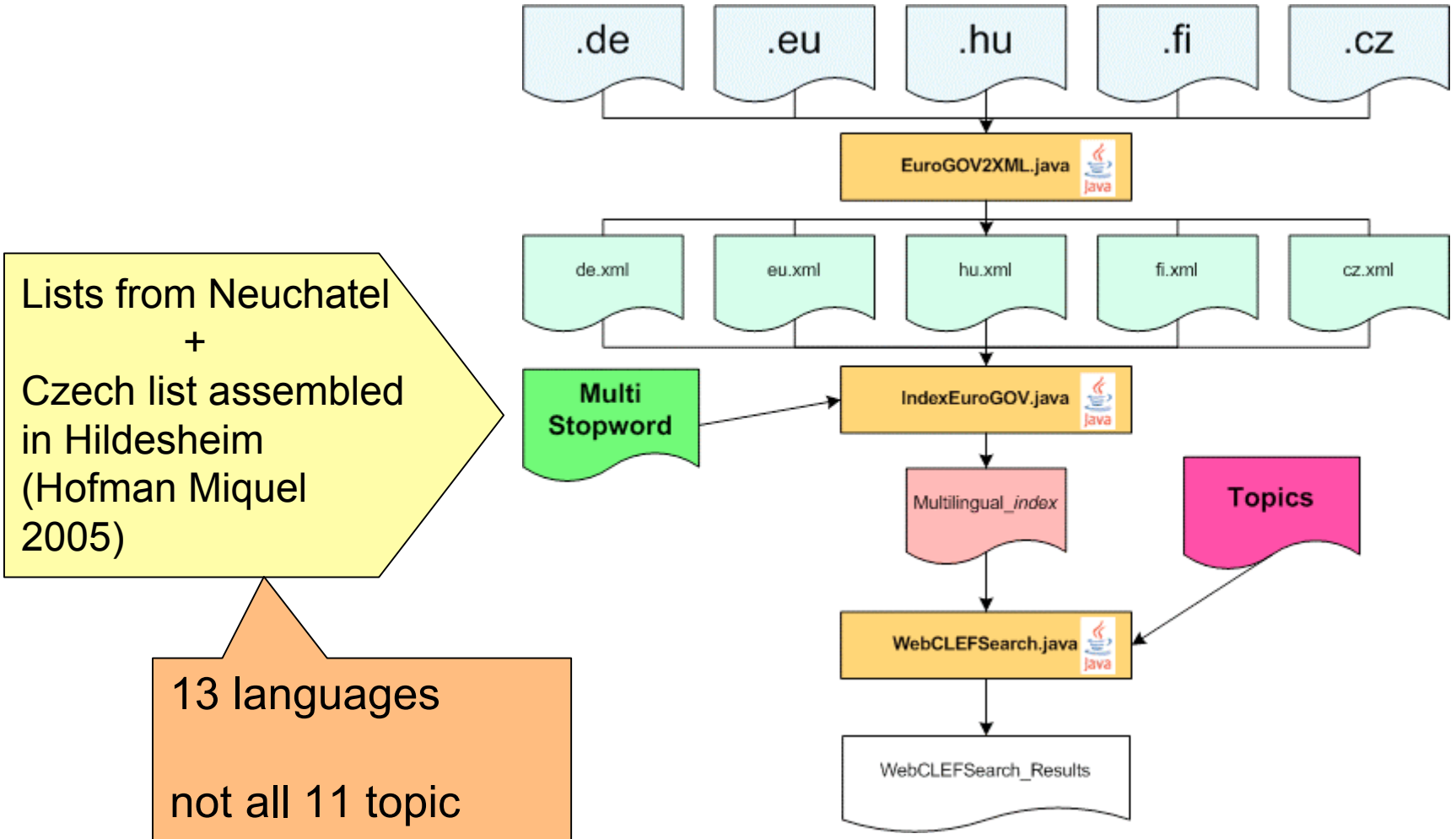
Problems

- Content of some 20% of the documents was only partially indexed (unresolved XML parsing errors)
- Time constraints (Server only available three weeks prior to deadline)
- -> no meta data used
  no stemming
  no BRF

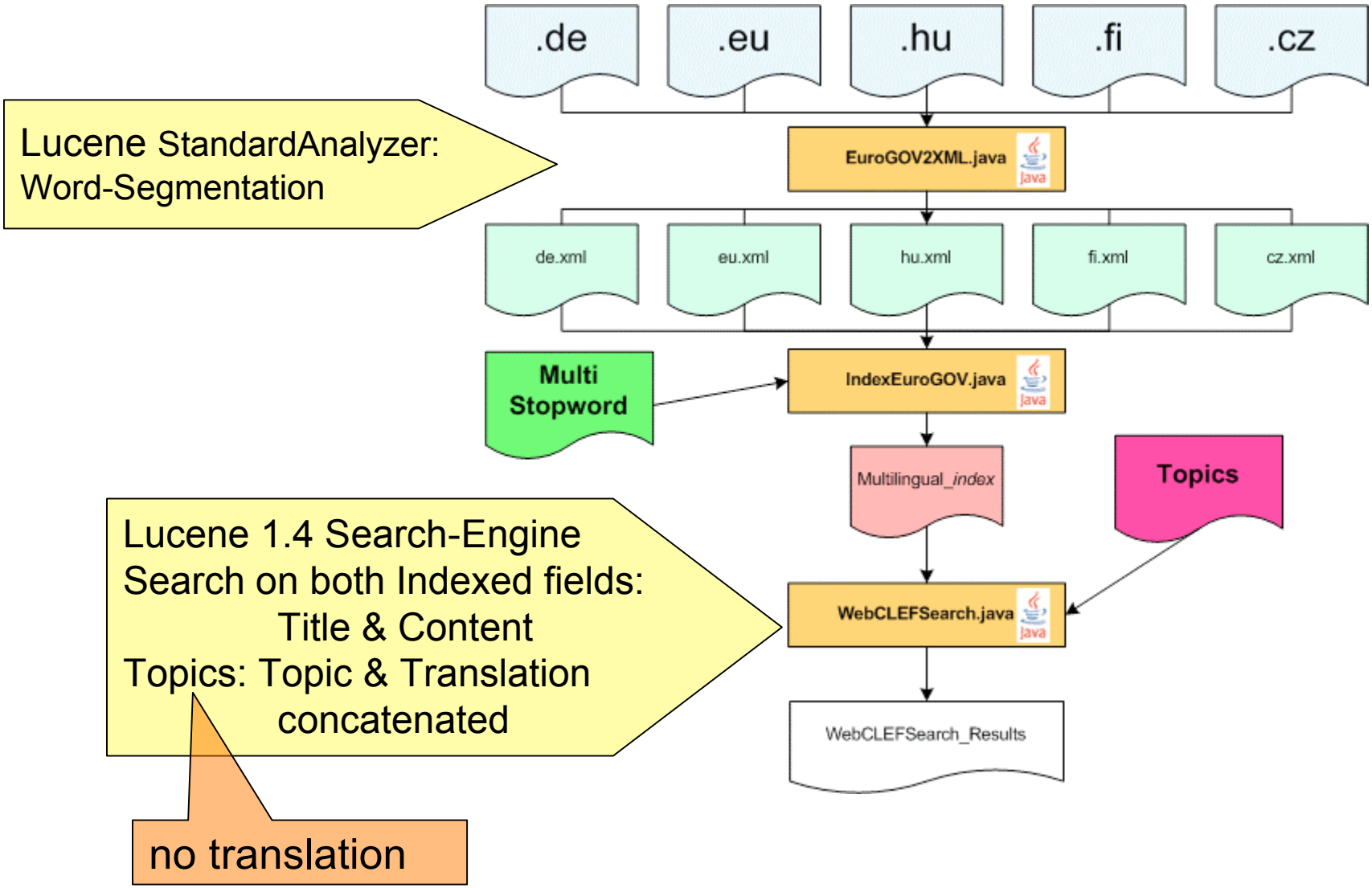
Indexing and Retrieval Approaches

- Indexed Fields:
  - Title and Content
  - for some runs content cutoff at 100 chars
- Multilingual stopword list
- One index for all languages
  - Words: no stemming
  - Tri-, Four- and Five-Grams
- -> no fusion problem, no language identification problem
- Boosting: weighting topic to topic translation

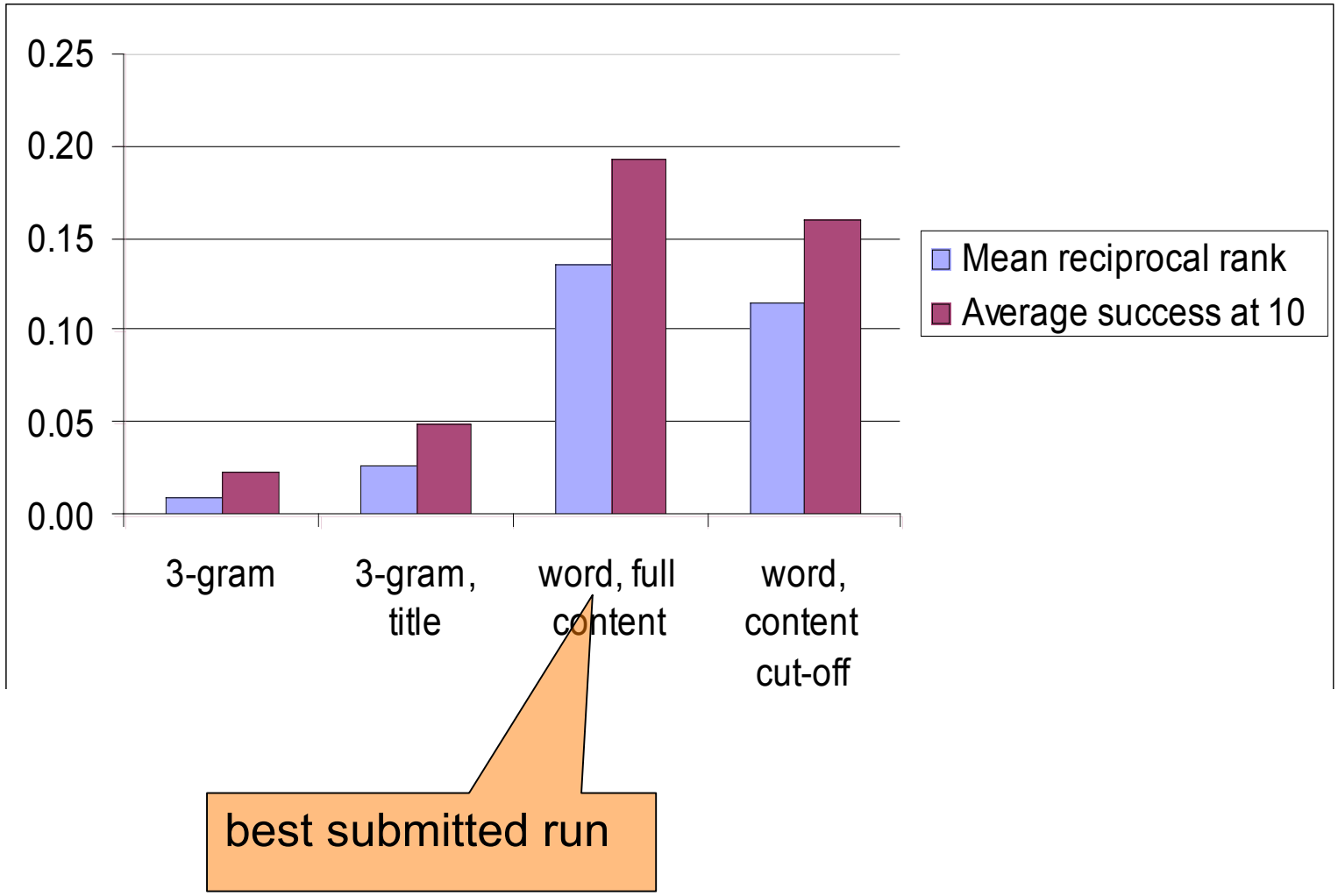
WebCLEFSearch Prozess



WebCLEFSearch Prozess



Submitted Multilingual Results



Multilingual n-gram Runs

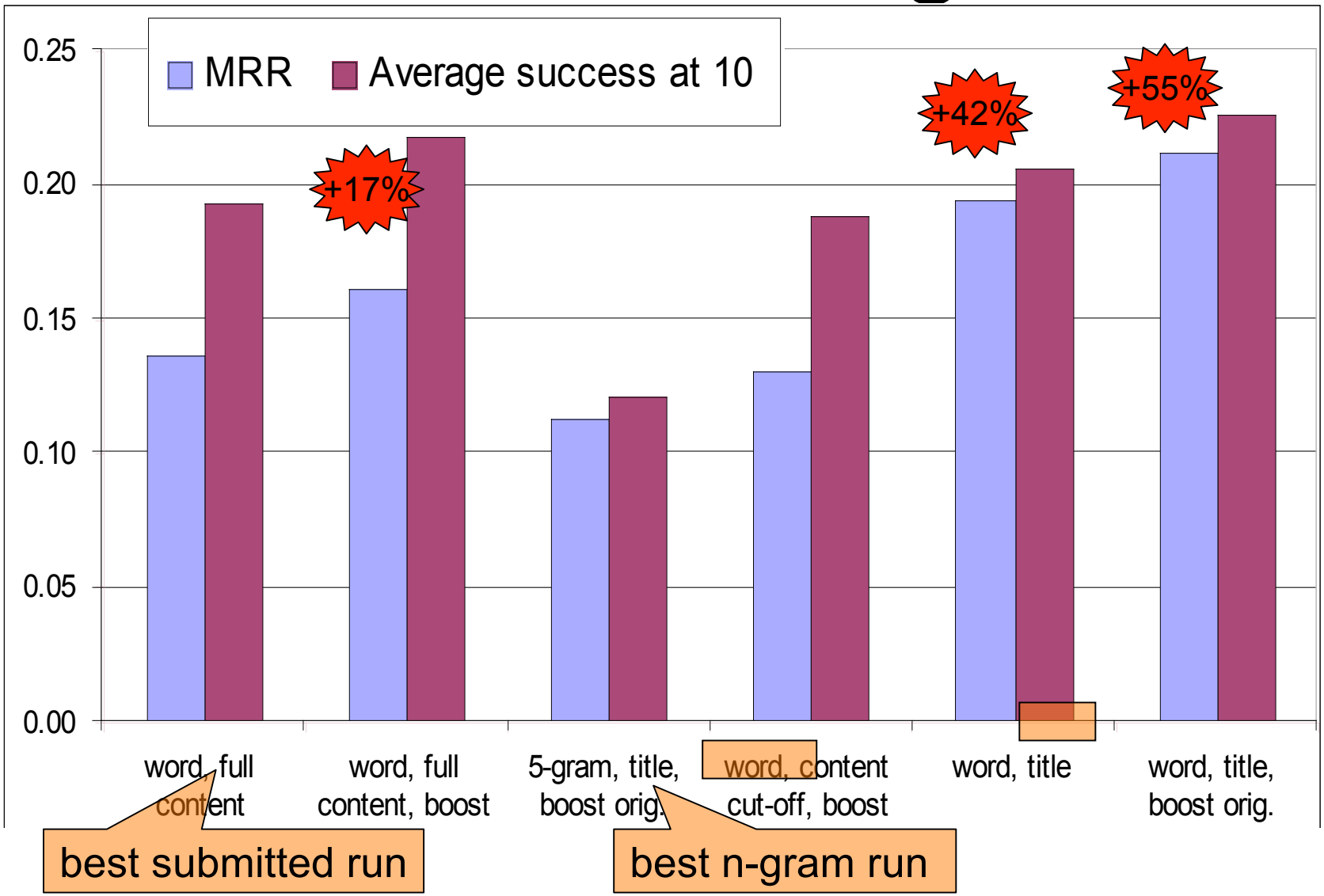
	title	boost orig. topic	boost tranls. topic	content, cut-off	boost orig. topic	boost tranls. topic	content	boost orig. topic	boost tranls. topic
3-gram									
MRR	0.027	0.038	0.014	0.099	0.108	0.099	0.010	0.017	0.006
Average success at 10	0.049	0.062	0.024	0.106	0.114	0.106	0.024	0.033	0.015
4-gram									
MRR	0.084	0.102	0.048	0.050	0.053	0.036			
Average success at 10	0.092	0.112	0.053	0.056	0.057	0.041			
5-gram									
MRR	0.095	0.113	0.057						
Average success									

missing runs ongoing

- n-gram always worse than word index
- Boosting original topic always helps

Highlights of Post Submission Multilingual Runs

no Meta-Data used



A simple approach worked, why not even simplify more?

Conclusion

Surprising! Titles are not always meaningful

- Best run with **title only**
  - however, cutoff for content not helpful
- Boosting original topic always helps
  - Translation seems harmful (see other participants)

Web IR different from ad-hoc?

- Evaluation issue for multi-lingual task
  - maybe calculate MRR for each language with at least one known relevant item

Conclusion

- A great corpus with many topics! Let‘s continue!
- Ample room for improve-ment at multilingual ?

Plans @ Hildesheim

- Do lots of other things next year, but run the same setup again as a benchmark
- We will try to provide an alternative language identification list