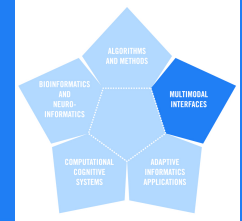


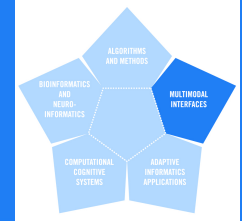
Evaluation by IR experiments - Competition 2

Mikko Kurimo and Ville Turunen



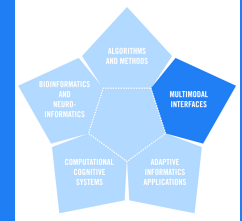
Introduction

- real world application for morpheme analysis: Information Retrieval
- analysis is needed to handle morphology
 - inflection
 - compounding
- CLEF collections for Finnish, German and English



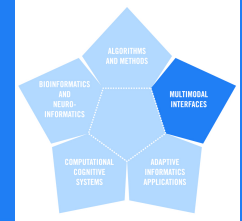
Data sets

- Finnish (CLEF 2004)
 - 55K documents from articles in Aamulehti 1994-95
 - 50 test queries
 - 23K binary relevance assessments
- English (CLEF 2005)
 - 107K documents from articles in Los Angeles Times 1994 and Glasgow Herald 1995
 - 50 test queries
 - 20K binary relevance assessments



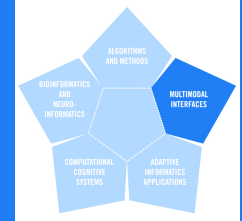
Data sets

- German (CLEF 2003)
 - 300K documents from short articles in Frankfurter Rundschau 1994, Der Spiegel 1994-95 and SDA German 1994-95
 - 60 test queries
 - 23K binary relevance assessments



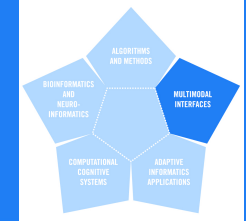
Submissions

- **Bernhard 1, Bernhard 2**
- **Bordag 5, Bordag 5a**
- **McNamee 3, McNamee 4, McNamee 5**
- **Zeman**
- **Monson Morfessor, Monson ParaMor, Monson ParaMor-Morfessor**
- **Pitler**



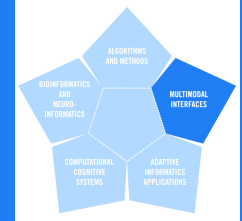
Reference methods

- **Morfessor Categories-Map**
- **Morfessor Baseline**
- **dummy**: no segmentation
- **grammatical**: gold standard segmentations
 - **all**: all alternatives included
 - **first**: only first alternative
- **Porter**: LEMUR's default stemmer
- **Tepper**: hybrid method based on Morfessor MAP



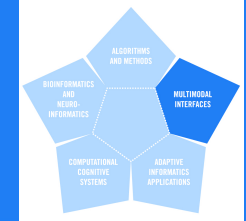
Evaluation 1/2

- words in the documents and queries were replaced by the submitted segmentations
- new words
 - the collections contained words that were not in the original word list
 - additional segmentations were requested
 - if segmentation was not available, words were indexed as such



Evaluation 2/2

- LEMUR-toolkit (<http://www.lemurproject.org/>)
- TFIDF and Okapi BM25 retrieval methods
- default parameter settings
- Okapi seems to handle common morphemes very poorly
 - affixes etc.
 - => stoplist of most common terms
- for TFIDF stoplist was not used
- best results with Okapi

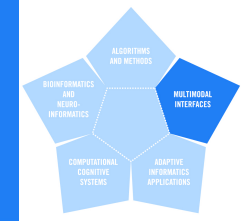


Performance measures

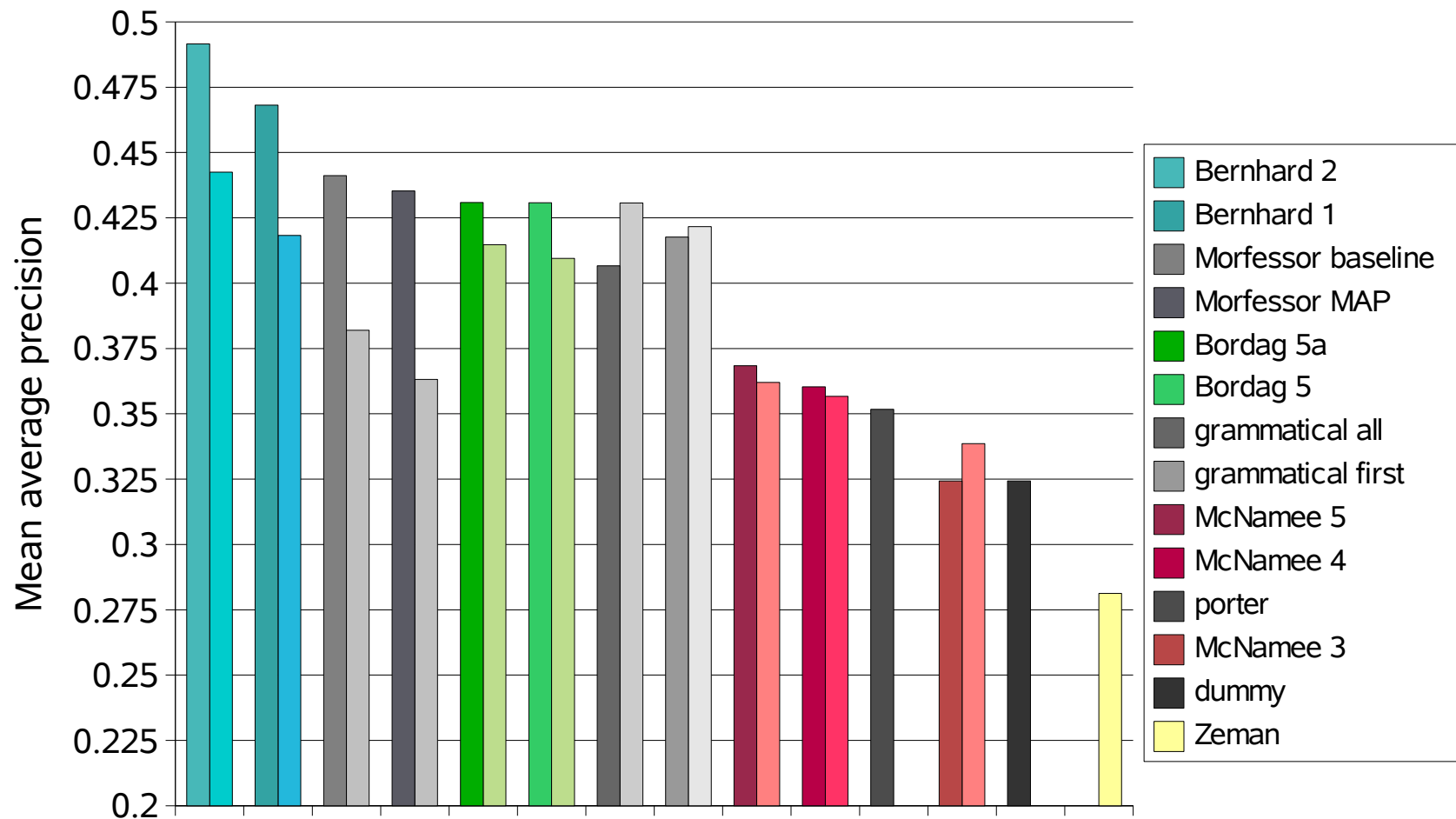
$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

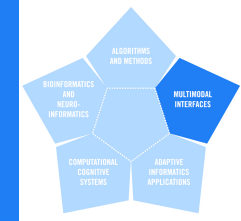
$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

- average precision: $AP = \frac{\sum_{r=1}^N (P(r) \times rel(r))}{|\{relevant\ documents\}|}$
- mean over all queries

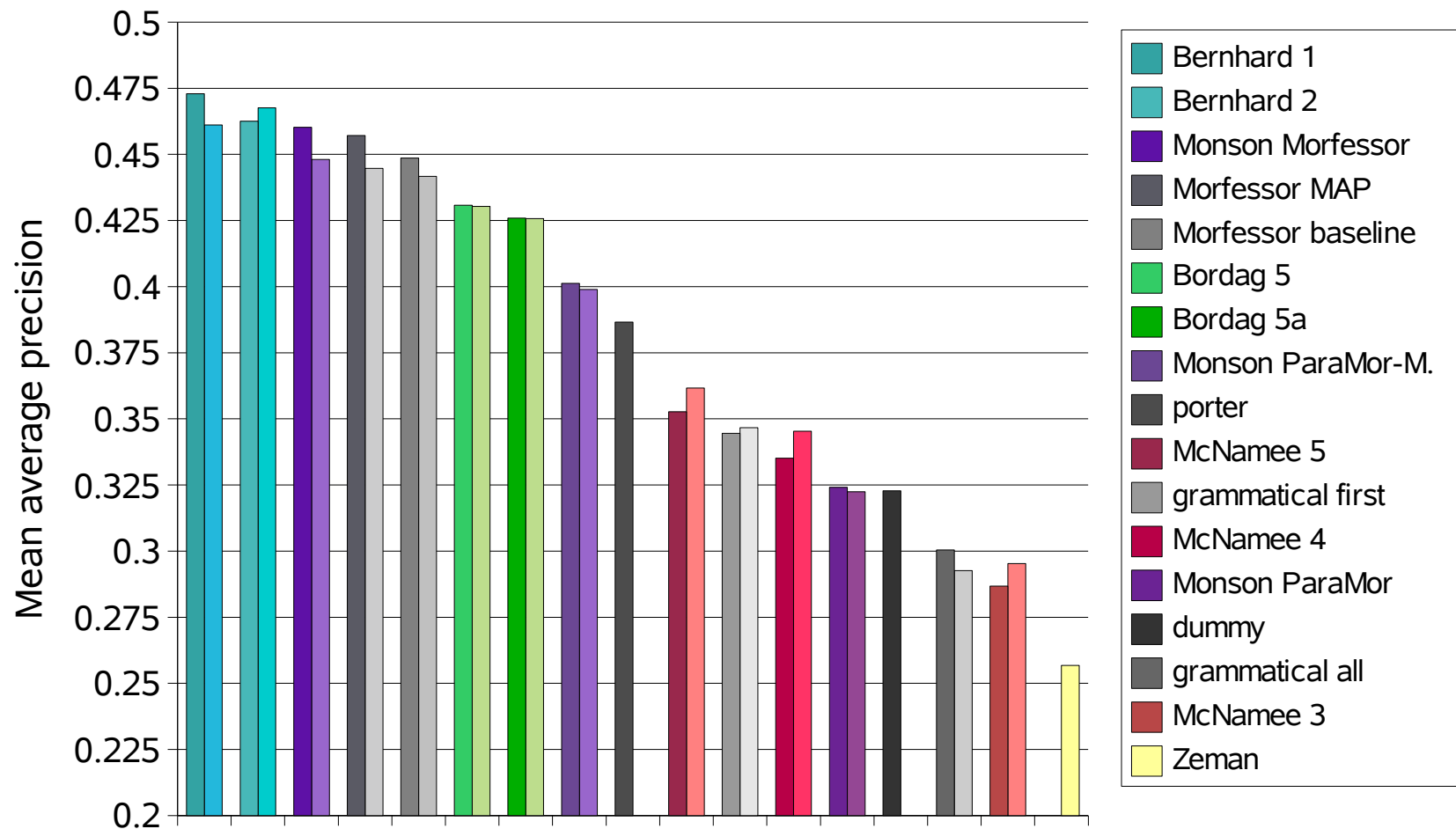


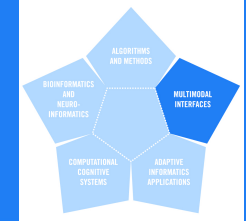
Results: Finnish



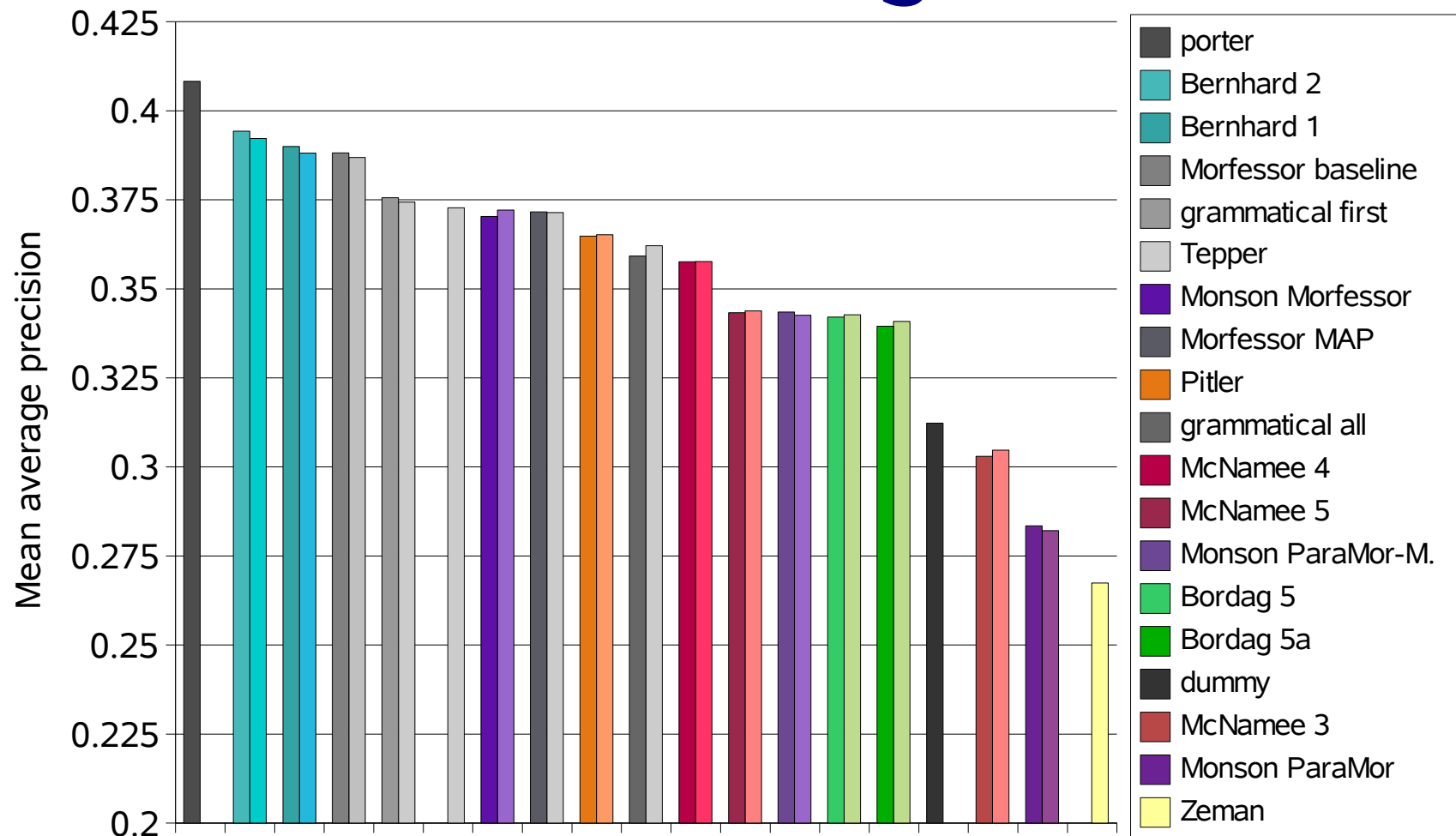


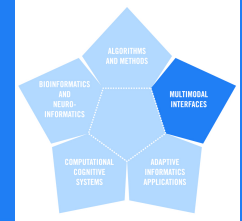
Results: German





Results: English





Conclusions

- new words important for Finnish, less so for German and English
- Porter stemming unbeaten for English (so far)
- unsupervised morpheme analysis methods work very well for IR!