09:10 Mikko Kurimo: "*Morpho Challenge Workshop 2008*"

09:20 **Mikko Kurimo**: "Evaluation by a Comparison to a Linguistic Gold Standard – Competition 1"

09:40 **Mikko Kurimo**:"Evaluation by IR experiments – Competition 2"

10:00 **Christian Monson**: "ParaMor and Morpho Challenge 2008"

10:30 Break

10:50 **Paul McNamee**: "Retrieval Experiments at Morpho Challenge 2008"

11:10 **Daniel Zeman**: "Using Unsupervised Paradigm Acquisition for Prefixes"

11:30 **Oskar Kohonen**: "Allomorfessor: Towards Unsupervised Morpheme Analysis"

11:50 **Sarah A. Goodman:** "Morphological Induction Through Linguistic Productivity"

12:10 Discussion

13:00 Conclusion

ADAPTIVE INFORMATICS
RESEARCH CENTRE

PASCAL
Pattern Analysis, Statistical Modelling and
Computational Learning

CLEF

Unsupervised Morpheme Analysis

Morpho Challenge Workshop 2008

Mikko Kurimo, Matti Varjokallio
and Ville Turunen

Helsinki University of Technology, Finland

Opening

Welcome to the Morpho Challenge 2008 workshop:

- challenge participants
- workshop speakers
- other CLEF researchers
- everybody who is interested in the topic!

Motivation

- To design statistical machine learning algorithms that discover which morphemes words consist of
- Follow-up to Morpho Challenge 2005 and 2007
- Find morphemes that are useful as vocabulary units for statistical language modeling in: *Speech recognition, Machine translation, Information retrieval*

Discussion topics for the end

- New ways to evaluate morphemes ?
- Use context for more accurate gold standard and evaluation, also in IR ?
- New test languages: Hungarian, Estonian, Russian, Korean, Japanese, Chinese ?
- New application evaluations: MT,..?
- New organizing partners ?
- Next Morpho Challenge 2009 / 2010 ?
- Journal special issue ?
- Next Morpho Challenge workshop ?

Thanks

Thanks to all who made Morpho Challenge 2008 possible:

- PASCAL network, CLEF, Leipzig corpora collection
- Gold standard providers: Nizar Habash, Ebru Arisoy, Stefan Bordag and Mathias Creutz
- Morpho Challenge organizing committee, program committee and evaluation team
- Morpho Challenge participants
- CLEF 2008 workshop organizers

09:10 **Mikko Kurimo**: "*Morpho Challenge Workshop 2008*"

09:20 Mikko Kurimo: "Evaluation by a Comparison to a Linguistic Gold Standard – Competition 1"

09:40 **Mikko Kurimo**:"Evaluation by IR experiments – Competition 2"

10:00 **Christian Monson**: "ParaMor and Morpho Challenge 2008"

10:30 Break

10:50 **Paul McNamee**: "Retrieval Experiments at Morpho Challenge 2008"

11:10 **Daniel Zeman**: "Using Unsupervised Paradigm Acquisition for Prefixes"

11:30 **Oskar Kohonen**: "Allomorfessor: Towards Unsupervised Morpheme Analysis"

11:50 **Sarah A. Goodman:** "Morphological Induction Through Linguistic Productivity"

12:10 Discussion

13:00 Conclusion

Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard – Competition 1

Mikko Kurimo and Matti Varjokallio

Contents

- Objectives
- Call for participation, Rules, Datasets
- Evaluation
- Participants
- Results
- Conclusion

Scientific objectives

- To learn of the phenomena underlying **word construction** in natural languages
- To discover approaches suitable for a wide **range of languages**
- To advance **machine learning** methodology

Call for participation

- Part of the EU Network of Excellence **PASCAL**’s Challenge Program
- Organized in collaboration with **CLEF**
- Participation is open to all and **free** of charge
- Word sets are provided for: *Finnish, English, German, Turkish and **Arabic***
- **Implement an unsupervised algorithm** that discovers morpheme analysis of words in each language!

Rules

- Morpheme analysis are submitted to the organizers for two different evaluations:
- **Competition 1**: Comparison to a linguistic morpheme "gold standard“
- **Competition 2**: Information retrieval experiments, where the indexing is based on morphemes instead of entire words.

Datasets

- Word lists downloadable at our home page
- Each word in the list is preceded by its frequency
- **Finnish**: 3M sentences, 2.2M word types
- **Turkish**: 1M sentences, 620K word types
- **German**: 3M sentences, 1.3M word types
- **English**: 3M sentences, 380K word types
- **Arabic**: no context, 140K* word types
- Small gold standard sample available in each language

Examples of gold standard analyses

- **English**: baby-sitters: baby_N sit_V er_s +PL
- **Finnish**: linuxiin: linux_N +ILL
- **Turkish**: kontrole: kontrol +DAT
- **German**:zurueckzubehalten:
  zurueck_B zu be halt_V +INF
- **Arabic**: Algbn: gabon_POS:N Al+ +SG

Evaluation method

- **Problem**: The unsupervised morphemes may have **arbitrary names**, not the same as the ”real” linguistic morphemes, nor just subword strings
- **Solution**: Compare to the linguistic gold standard analysis by **matching the morpheme-sharing word pairs**
- Compute matches from a large random sample of word pairs where both words in the pair have a common morpheme

Evaluation measures

- *F-measure* = 1/(1/*Precision* + 1/*Recall*)
- *Precision* is the proportion of suggested word pairs that also have a morpheme in common according to the gold standard
- *Recall* is the proportion of word pairs *sampled from the gold standard* that also have a morpheme in common according to the suggested algorithm

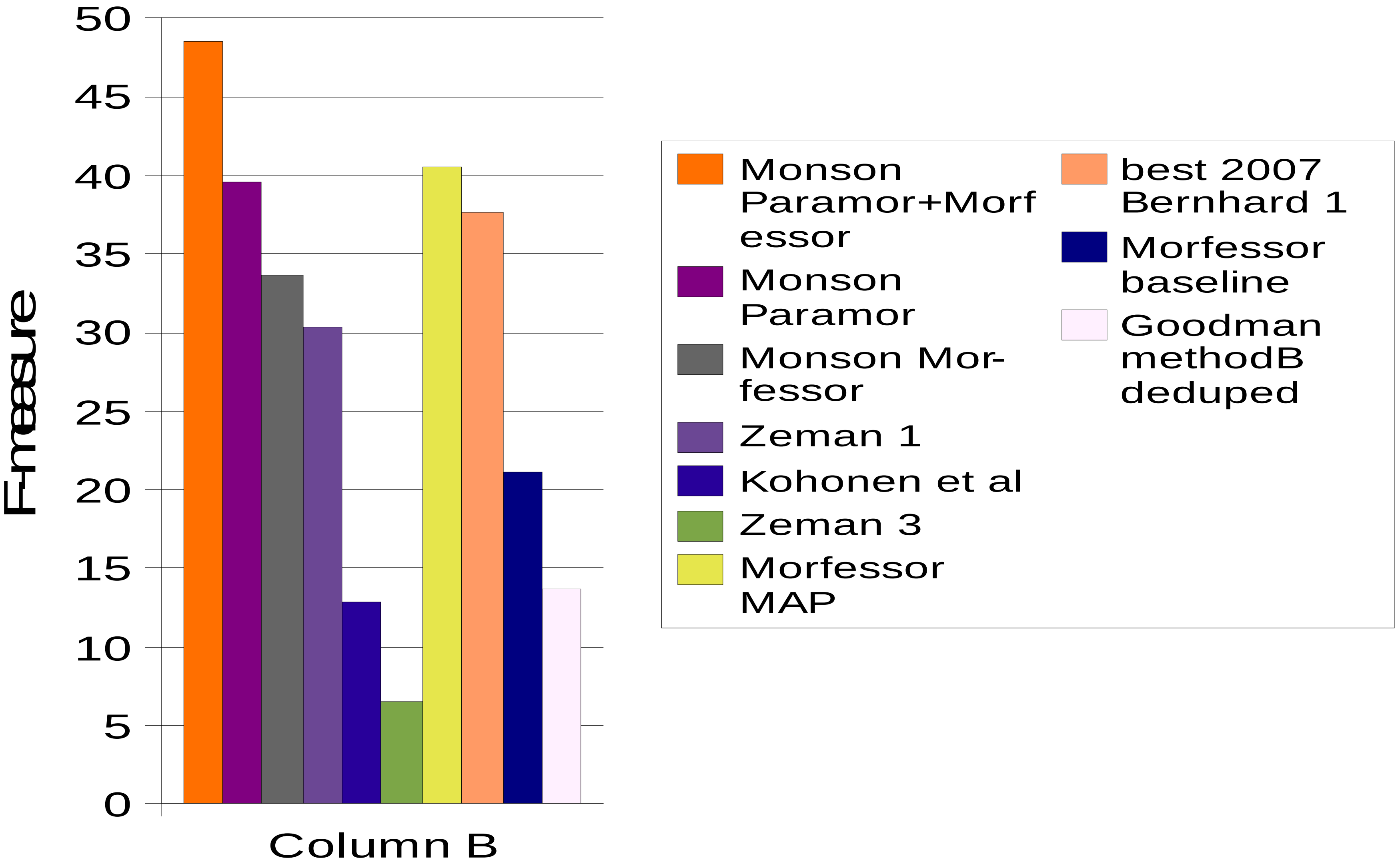
Participants

- (Burcu Can, Univ. York, UK – no submission)
- Sarah A. Goodman, Univ. Maryland, USA
  - late submission
- Oskar Kohonen et al., Helsinki Univ. Tech, FI
- Paul McNamee , JHU, USA
  - only in Competition 2 (IR evaluation)
- Daniel Zeman, Karlova Univ., CZ
- Christian Monson et al., CMU, USA

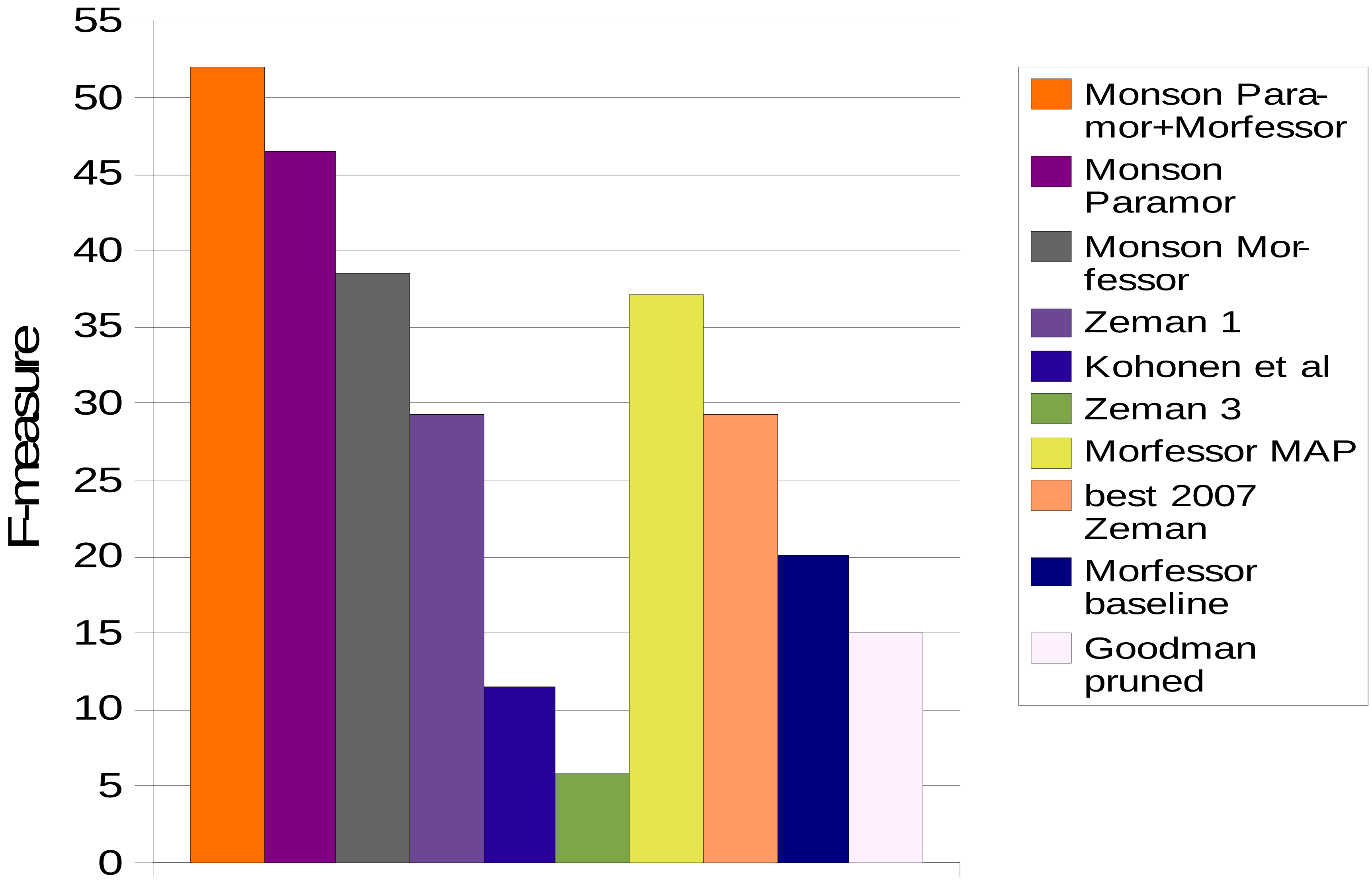
Example morphemes for “baby-sitters”

Gold Standard:	baby_N sit_V er_s +PL
Morfessor:	baby- sitters
Kohonen:	baby- sitters
Monson paramor:	bab +y, sitt +er +s
Monson Morfessor:	+baby-/PRE sitter/STM +s/SUF
Zeman1:	baby-sitter s, baby-sitt ers
Zeman3:	baby-sitt ers, baby-sitter s

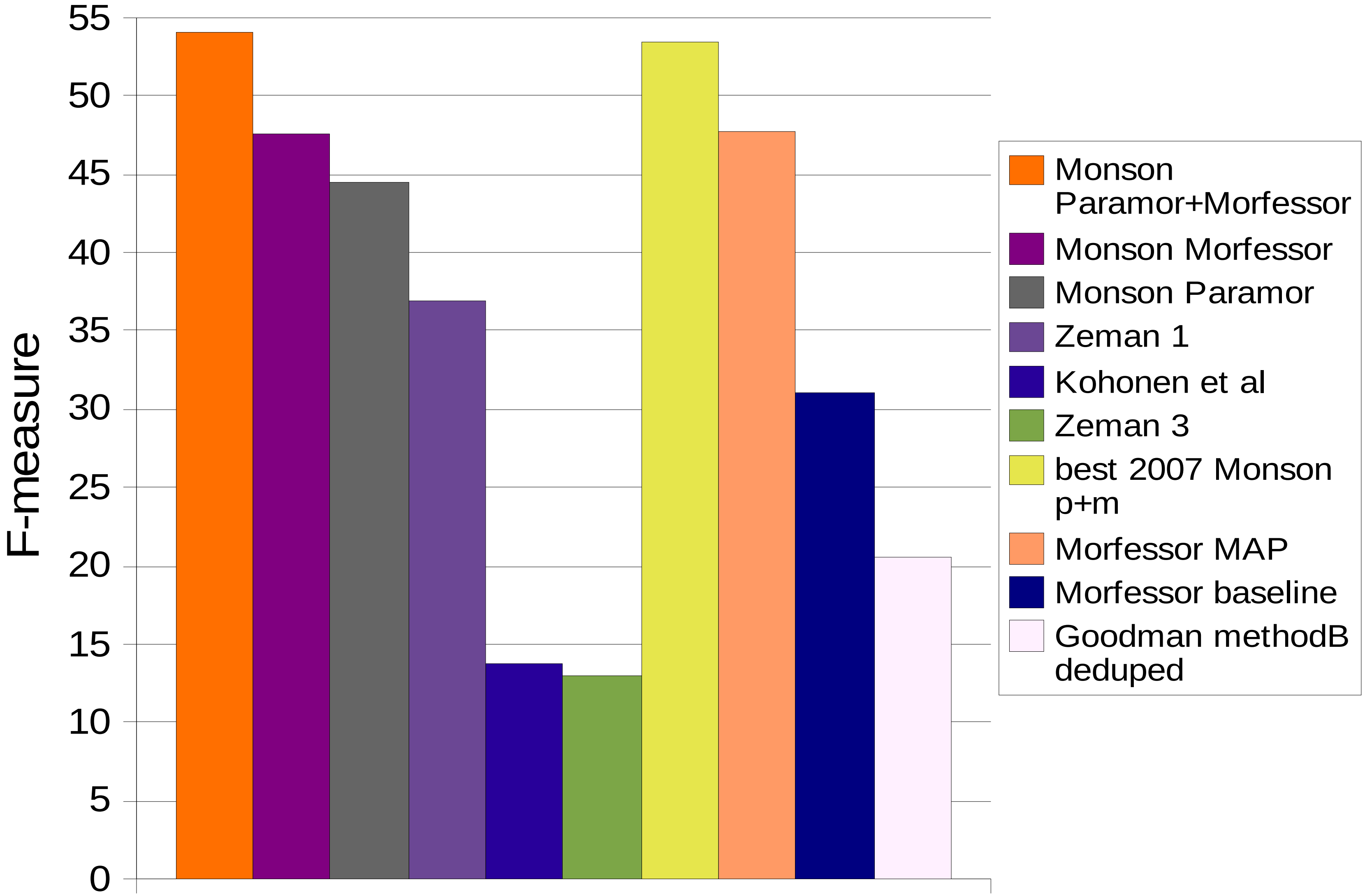
Results: Finnish, 2.2M word types



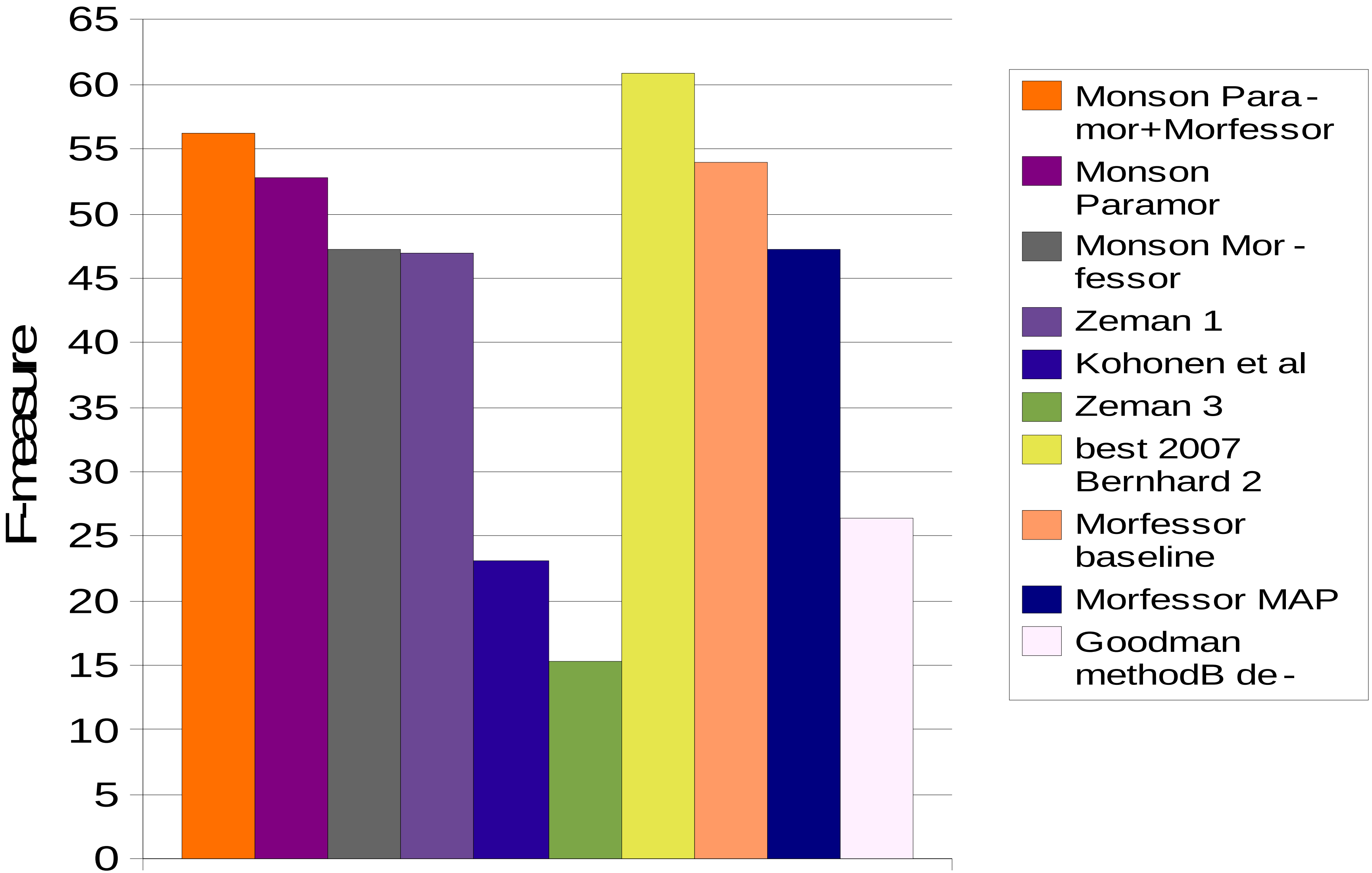
Results: Turkish, 620K word types



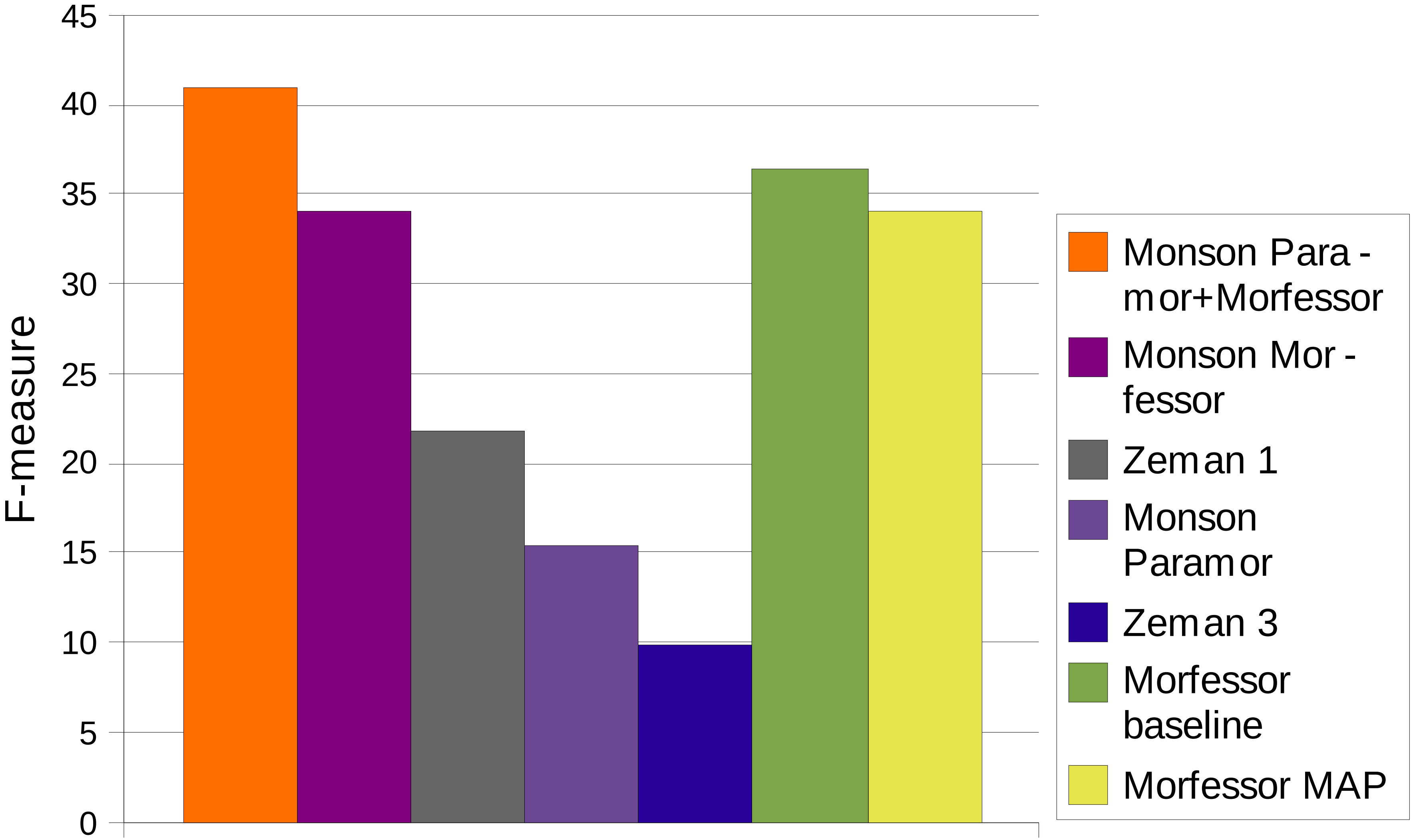
Results: German, 1.3M word types



Results: English, 380K word types



Results: Arabic, 140K word types



About 2008 results

- One algorithm best in all tasks
- Monson ParaMor better than Morfessor in TUR but worse in ARA
- The ”simple” Morfessor Baseline still hard to beat in ENG and ARA
- Large improvements over 2007 in FIN and TUR
- Highest F in ENG and lowest in ARA, but the best algorithms survived >30% in all tasks
- Features of the gold standard affect the results

Conclusion

- 10 different unsupervised algorithms
- 6 participating research groups
- Evaluations for 5 languages
- Good results in all languages
- Full report and papers in the CLEF proceedings
- Details, presentations, links, info at: *http://www.cis.hut.fi/morphochallenge2008/*

09:10 **Mikko Kurimo**: "*Morpho Challenge Workshop 2008*"

09:20 **Mikko Kurimo**: "Evaluation by a Comparison to a Linguistic Gold Standard – Competition 1"

⇨**09:40 Mikko Kurimo:"Evaluation by IR experiments – Competition 2"**

10:00 **Christian Monson**: "ParaMor and Morpho Challenge 2008"

10:30 Break

10:50 **Paul McNamee**: "Retrieval Experiments at Morpho Challenge 2008"

11:10 **Daniel Zeman**: "Using Unsupervised Paradigm Acquisition for Prefixes"

11:30 **Oskar Kohonen**: "Allomorfessor: Towards Unsupervised Morpheme Analysis"

11:50 **Sarah A. Goodman:** "Morphological Induction Through Linguistic Productivity"

12:10 Discussion

13:00 Conclusion

Unsupervised Morpheme Analysis Evaluation by IR experiments – Competition 2

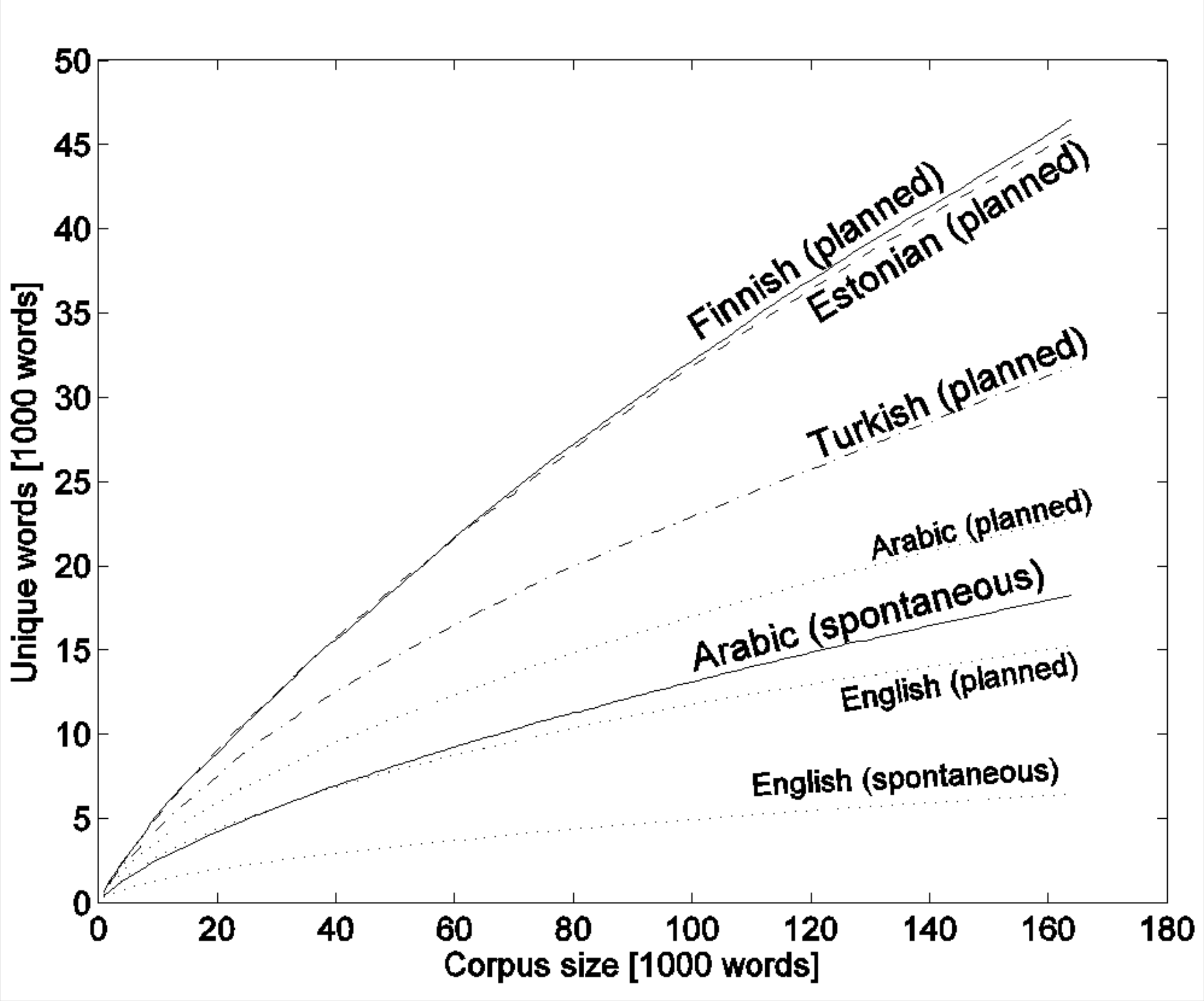
Mikko Kurimo and Ville Turunen

Motivation

- Real world application for morpheme analysis: Information Retrieval (IR)
- Analysis is needed to handle the inflection, compounding and agglutination of words
- IR tasks for Finnish, English and German used as in CLEF 2007

The vocabulary problem

- Speech recognition, information retrieval and machine translation require a **large vocabulary**
- **Agglutinative and highly-inflected** languages suffer from a severe **vocabulary explosion**
- More efficient representation units needed



IR data sets (as in CLEF 2007)

- **Finnish (CLEF 2004)**
  - **55K documents from articles in Aamulehti 1994-95**
  - **50 test queries, 23 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**
- **German (CLEF 2003)**
  - **300K documents from short articles in Frankfurter Rundschau 1994, Der Spiegel1994-95 and SDA German 1994-95**
  - **60 test queries, 23K binary relevance assessments**

IR evaluation

- words in the documents and queries were replaced by the suggested segmentations
- OOV words un-replaced
- all morphemes used for indexing
- stoplist for the most common ones (over a fixed frequency threshold)
- LEMUR-toolkit http://www.lemurproject.org/
- Okapi BM25 retrieval method (default)

Evaluation measure

- *Precision* is the proportion of retrieved documents that are relevant
- *Recall* is the proportion of relevant documents that are retrieved
- Compute the *average of precisions* after truncating the list of retrieved documents after each relevant document in turn
- Take the **mean of the average precision** over all queries

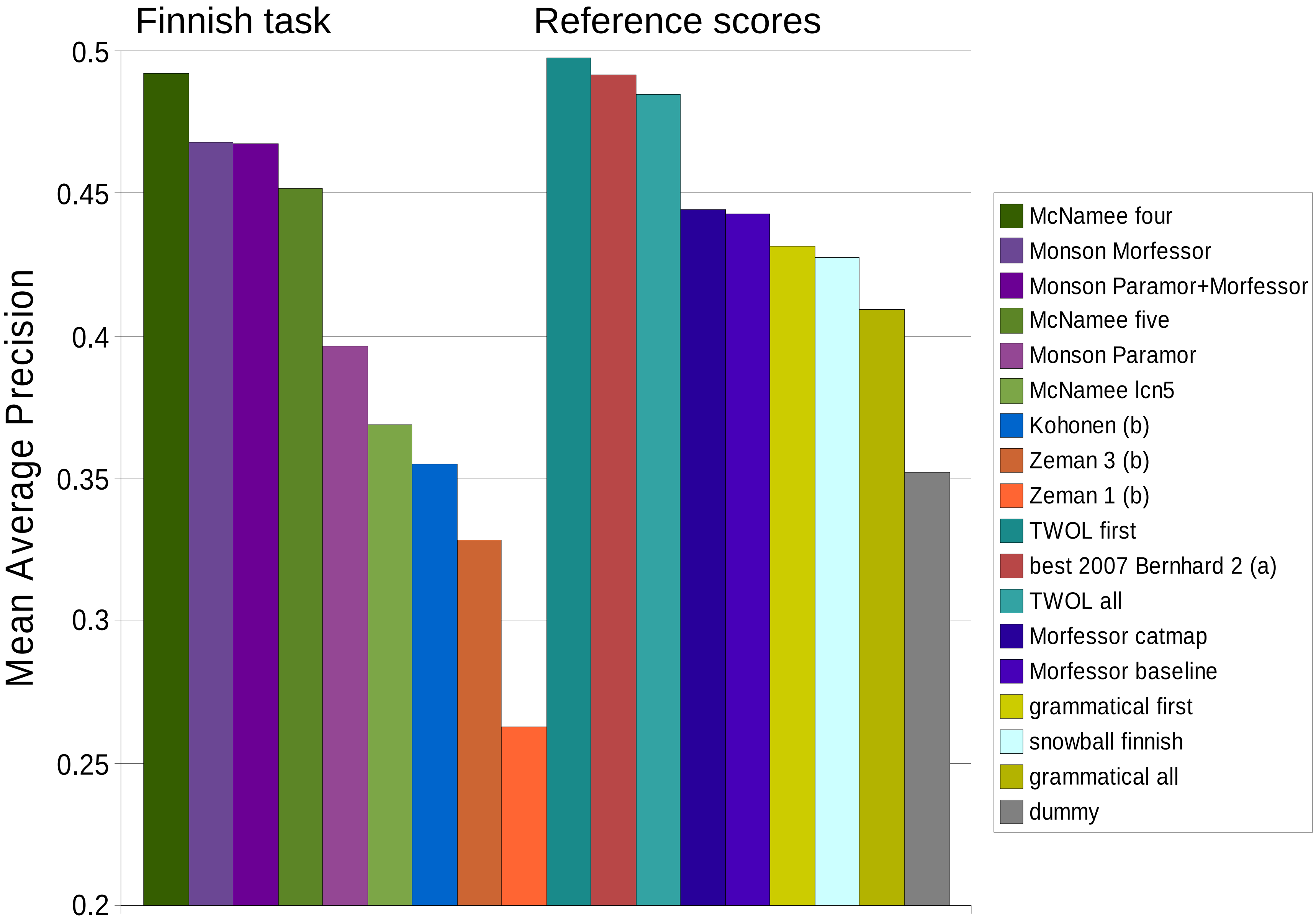
Submitted analysis

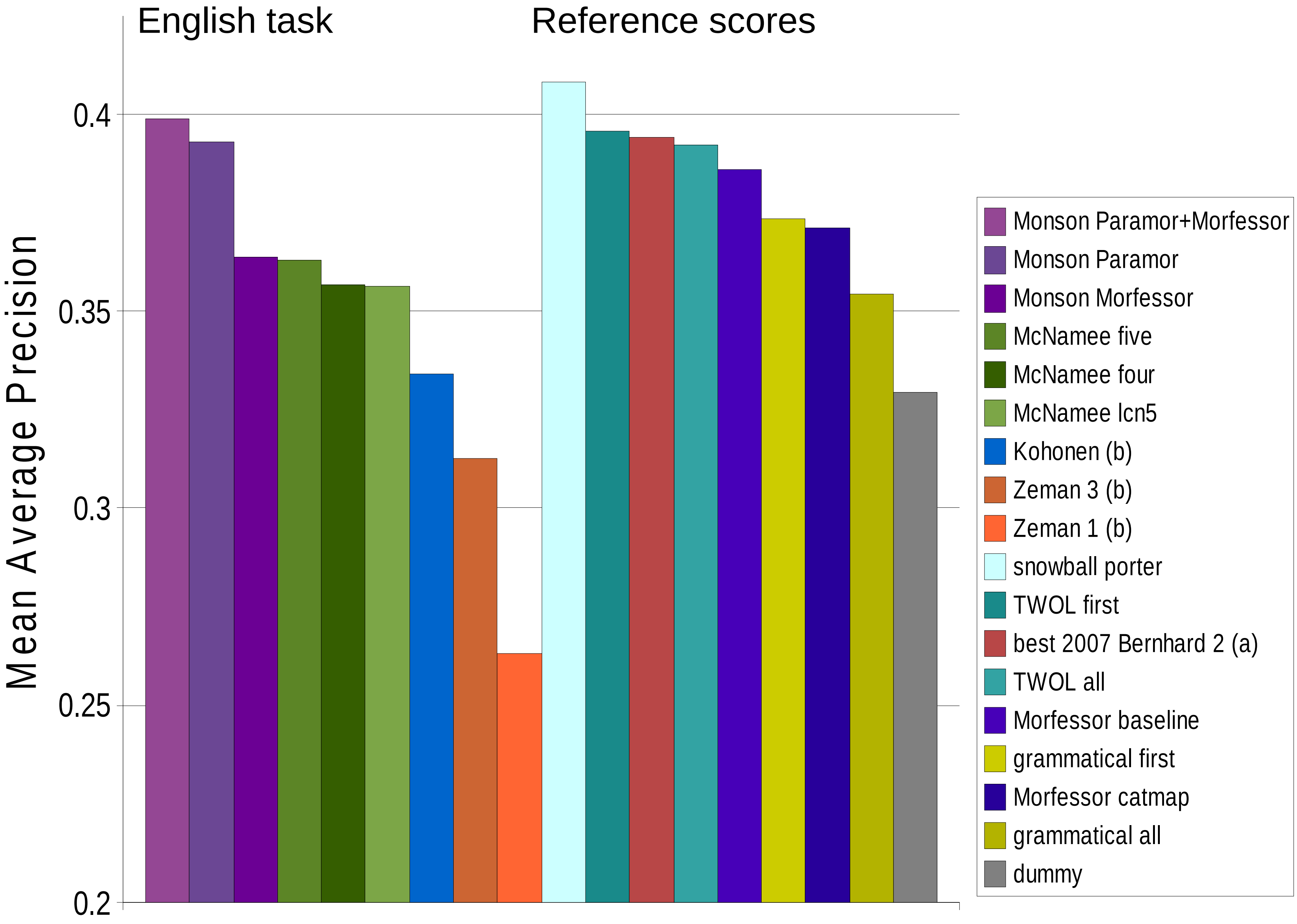
- Oskar Kohonen et al., Helsinki Univ. Tech, FI, (b)
- Paul McNamee , JHU, USA
- Daniel Zeman, Karlova Univ., CZ (b)
- Christian Monson et al., CMU, USA

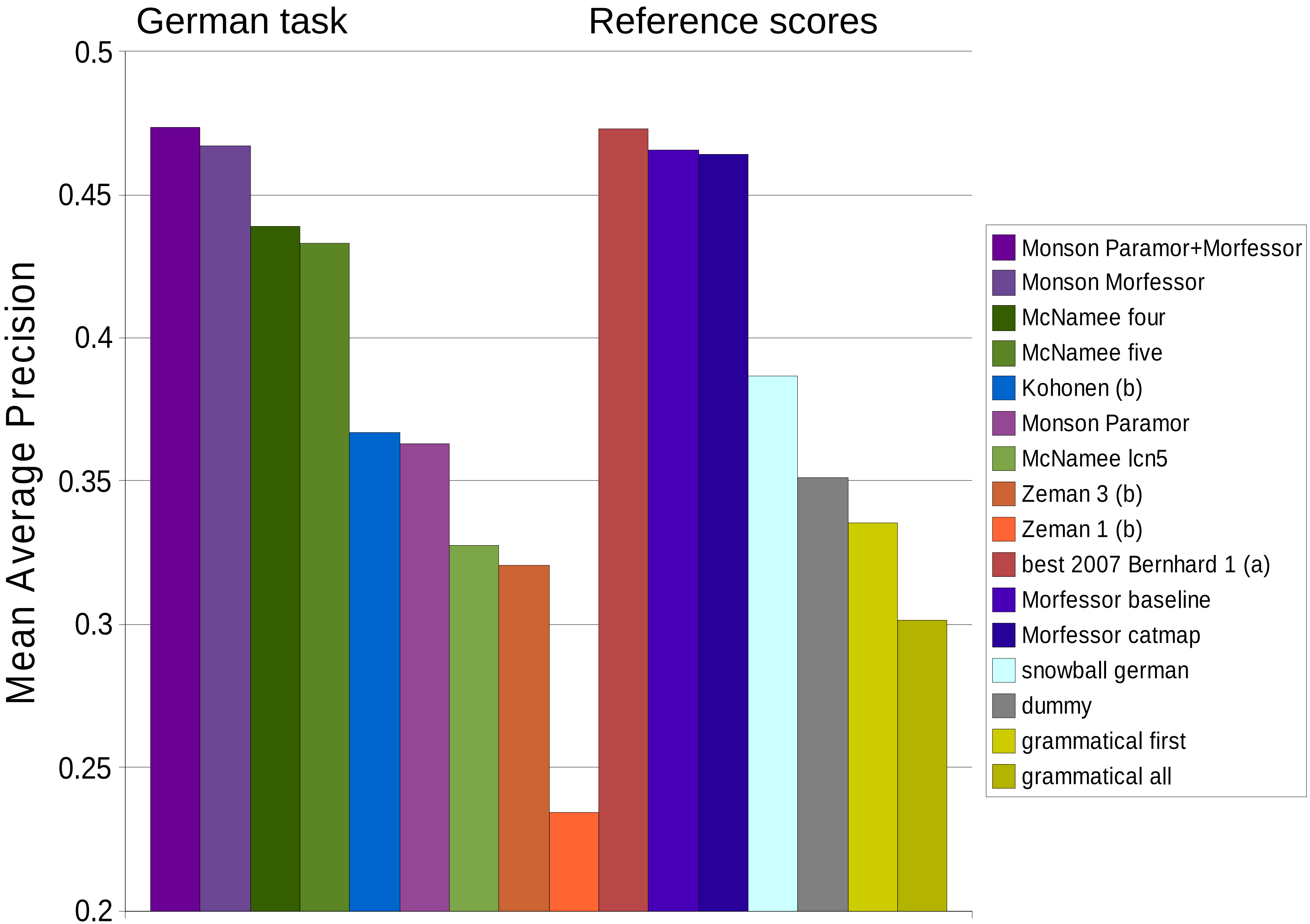
(b) Only analysis of Competition 1 words provided. OOVs unsplit.

Reference methods

- **Morfessor Baseline:** our public code since 2002
- **Morfessor Categories-MAP:** improved, public 2006
- **dummy**: no segmentation, all words unsplit
- **grammatical**: full gold standard segmentation (reference of competition 1)
  - all: all alternative segmentations included
  - first: only the first alternative chosen
- **TWOL**: word normalization by a commercial rule-based morphological analyzer (all & first)
- **Snowball**: Language specific stemming


Finnish task
Reference scores
Mean Average Precision
0.5
0.45
0.4
0.35
0.3
0.25
0.2
McNamee four
Monson Morfessor
Monson Paramor+Morfessor
McNamee five
Monson Paramor
McNamee lcn5
Kohonen (b)
Zeman 3 (b)
Zeman 1 (b)
TWOL first
best 2007 Bernhard 2 (a)
TWOL all
Morfessor catmap
Morfessor baseline
grammatical first
snowball finnish
grammatical all
dummy


English task
Reference scores
Mean Average Precision
0.4
0.35
0.3
0.25
0.2
Monson Paramor+Morfessor
Monson Paramor
Monson Morfessor
McNamee five
McNamee four
McNamee lcn5
Kohonen (b)
Zeman 3 (b)
Zeman 1 (b)
snowball porter
TWOL first
best 2007 Bernhard 2 (a)
TWOL all
Morfessor baseline
grammatical first
Morfessor catmap
grammatical all
dummy


German task
Reference scores
Mean Average Precision
0.5
0.45
0.4
0.35
0.3
0.25
0.2
Monson Paramor+Morfessor
Monson Morfessor
McNamee four
McNamee five
Kohonen (b)
Monson Paramor
McNamee lcn5
Zeman 3 (b)
Zeman 1 (b)
best 2007 Bernhard 1 (a)
Morfessor baseline
Morfessor catmap
snowball german
dummy
grammatical first
grammatical all

About 2008 results

- Bernhard 2007 only very narrowly beaten
- McNamee4 best in FIN, Monson P+M best in ENG,GER
- Monson ParaMor better than Morfessor in ENG, but worse in FIN,GER
- Highest MAP in FIN and lowest in ENG, but the best algorithms survived well in all tasks
- TWOL good, grammatical not, Snowball only good in ENG

Conclusions

- IR evaluations for 3 languages (out of 5)
- Good results in all languages
- Winner not as clear as in Competition 1
- Full report and papers in the CLEF proceedings
- Details, presentations, links, info at: *http://www.cis.hut.fi/morphochallenge2008/*

09:10 **Mikko Kurimo**: "*Morpho Challenge Workshop 2008*"

09:20 **Mikko Kurimo**: "Evaluation by a Comparison to a Linguistic Gold Standard – Competition 1"

09:40 **Mikko Kurimo**:"Evaluation by IR experiments – Competition 2"

10:00 Christian Monson: "ParaMor and Morpho Challenge 2008"

10:30 Break

10:50 **Paul McNamee**: "Retrieval Experiments at Morpho Challenge 2008"

11:10 **Daniel Zeman**: "Using Unsupervised Paradigm Acquisition for Prefixes"

11:30 **Oskar Kohonen**: "Allomorfessor: Towards Unsupervised Morpheme Analysis"

11:50 **Sarah A. Goodman:** "Morphological Induction Through Linguistic Productivity"

12:10 Discussion

13:00 Conclusion