PASCAL
Pattern Analysis, Statistical Modelling and Computational Learning

CLEF at *Morpho Challenge 2008* - Unsupervised Morpheme Analysis

Mikko Kurimo, Matti Varjokallio and Ville Turunen

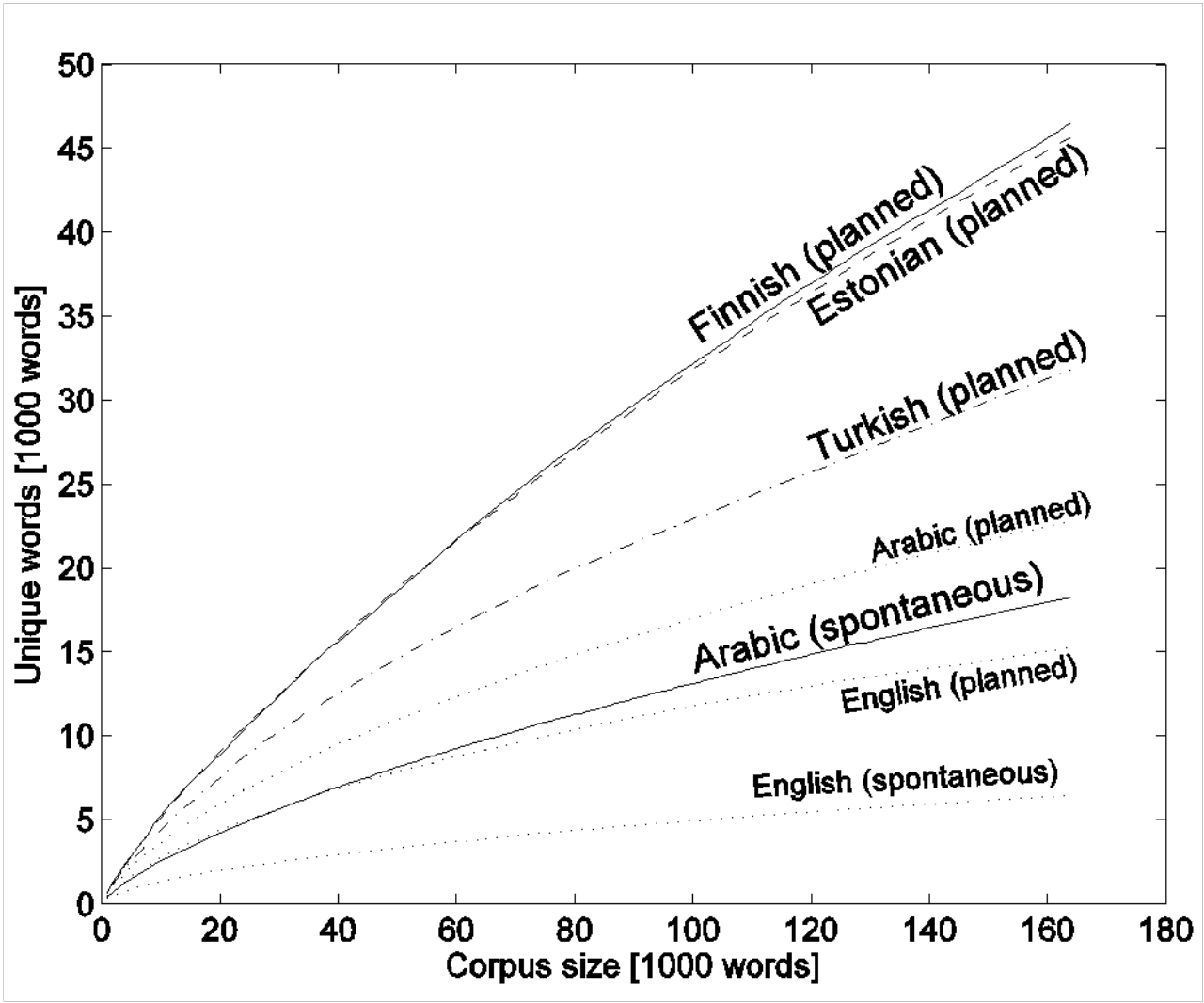
Helsinki University of Technology, Finland

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*
- Discover approaches suitable for a **wide range of languages and tasks**

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



Motivation for being at CLEF

- 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

Morpho Challenge 2008

- 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.

Training data

- Downloadable texts and word frequency lists
- **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

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

1. Linguistic evaluation

- **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

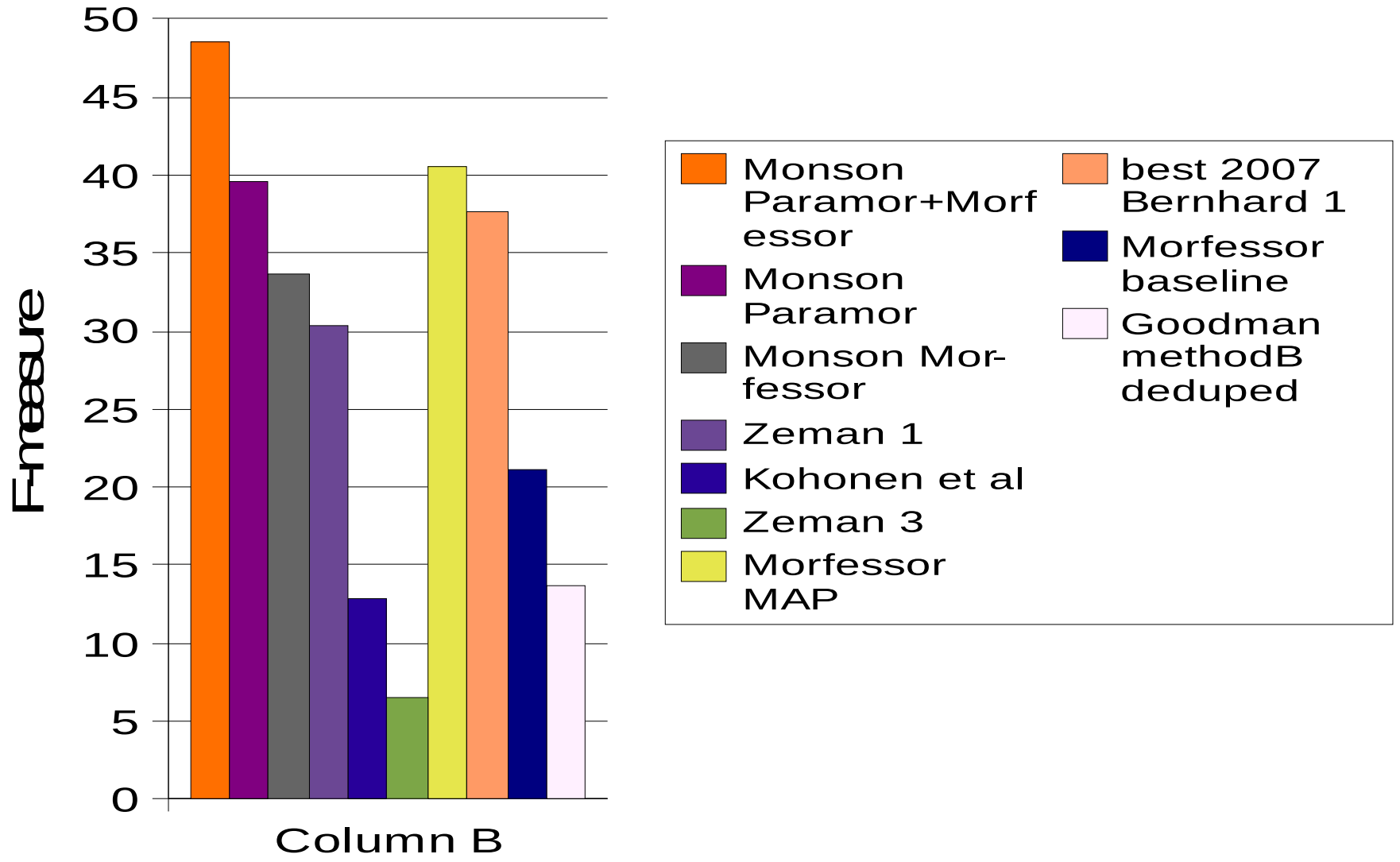
Evaluation measures

- *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

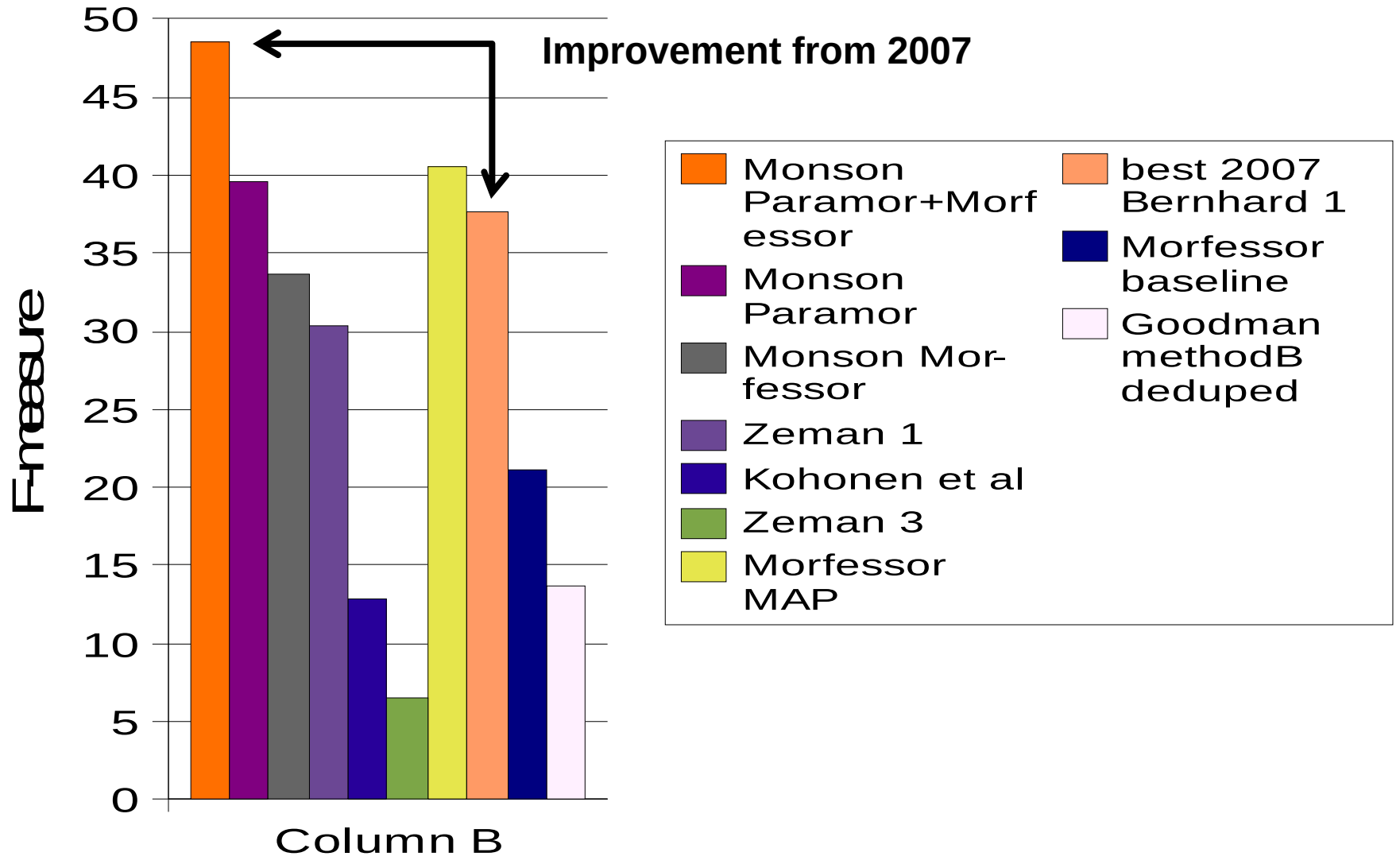
Participants

- (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

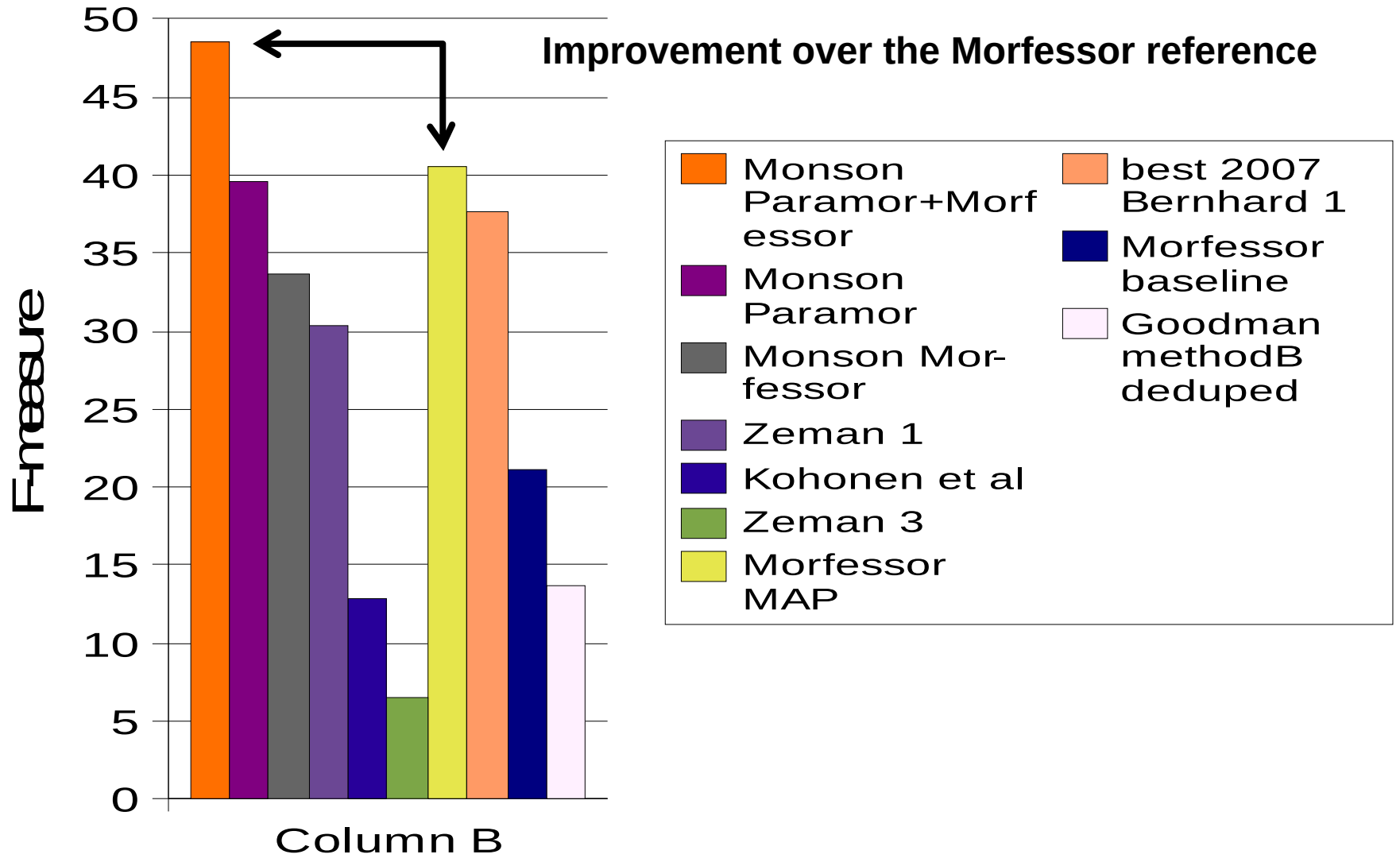
Results: Finnish, 2.2M word types



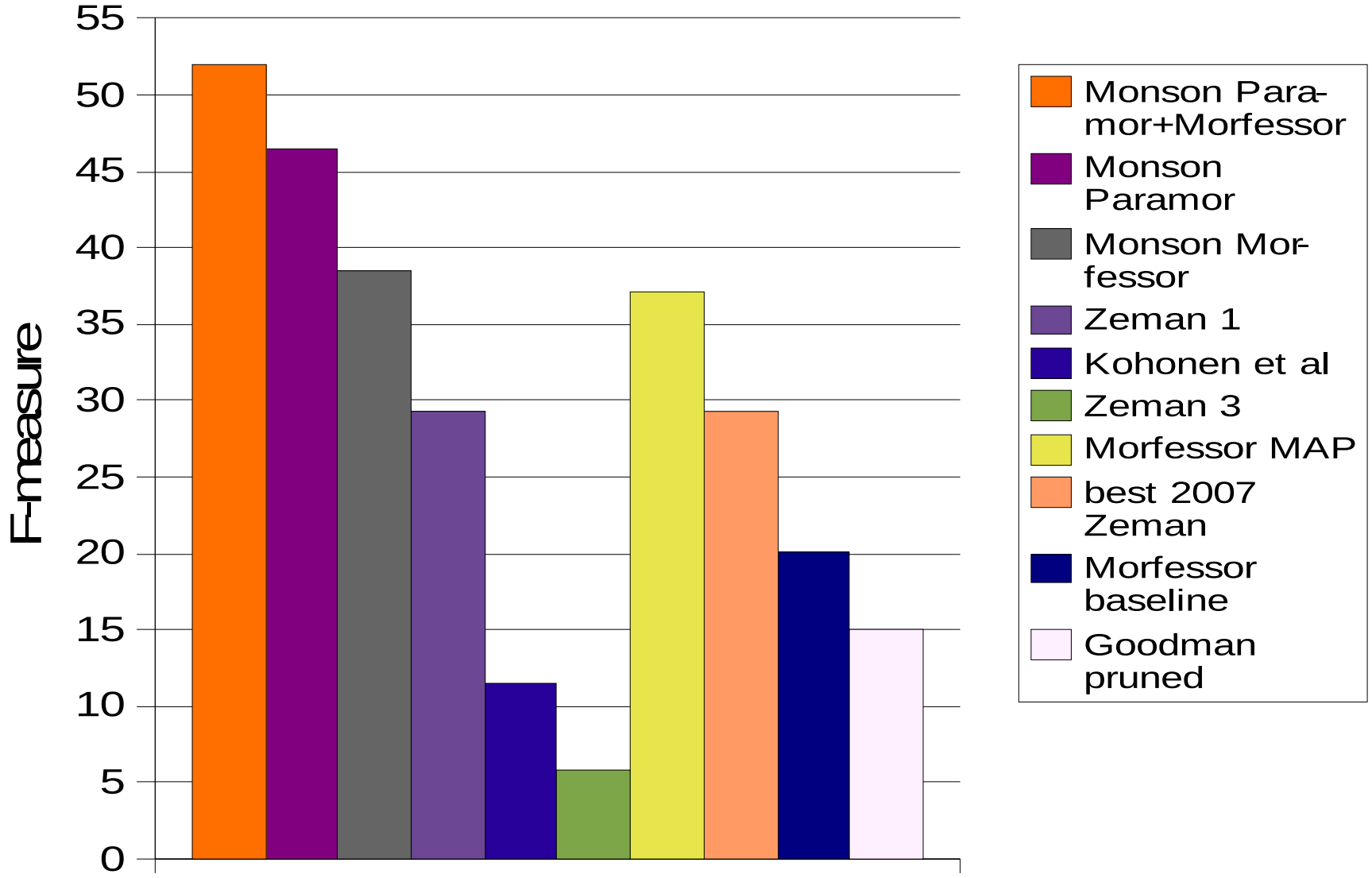
Results: Finnish, 2.2M word types



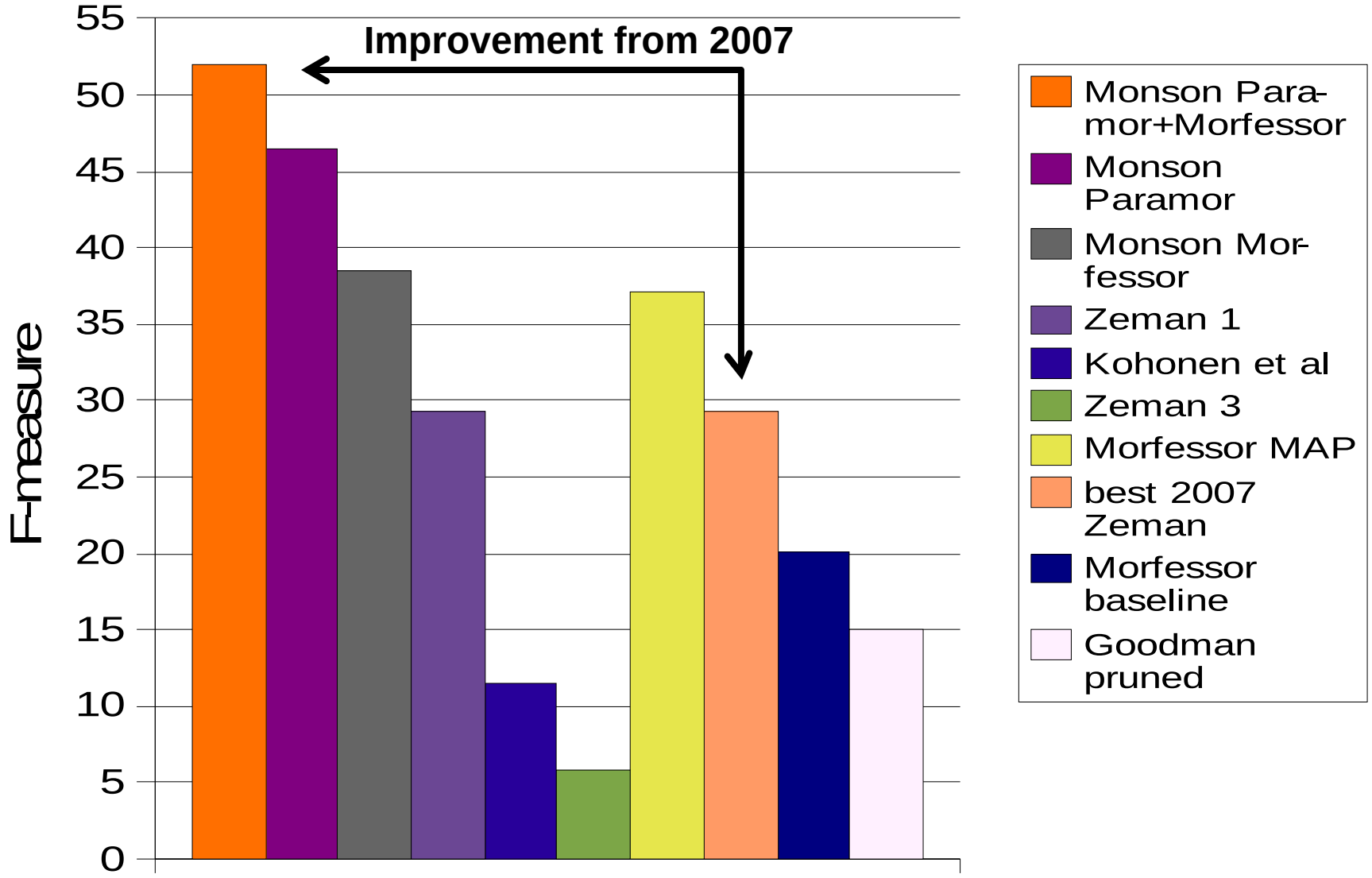
Results: Finnish, 2.2M word types



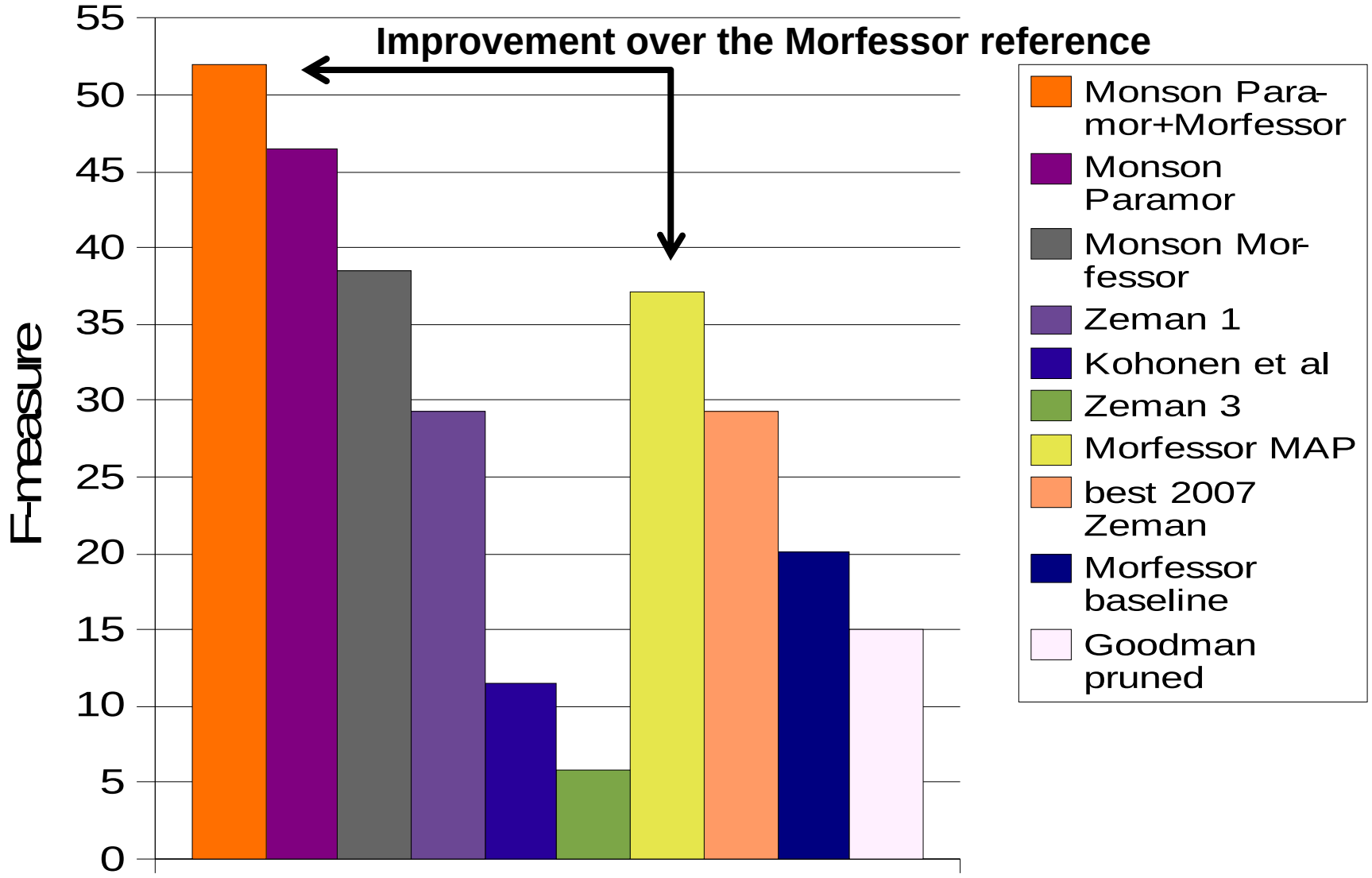
Results: Turkish, 620K word types



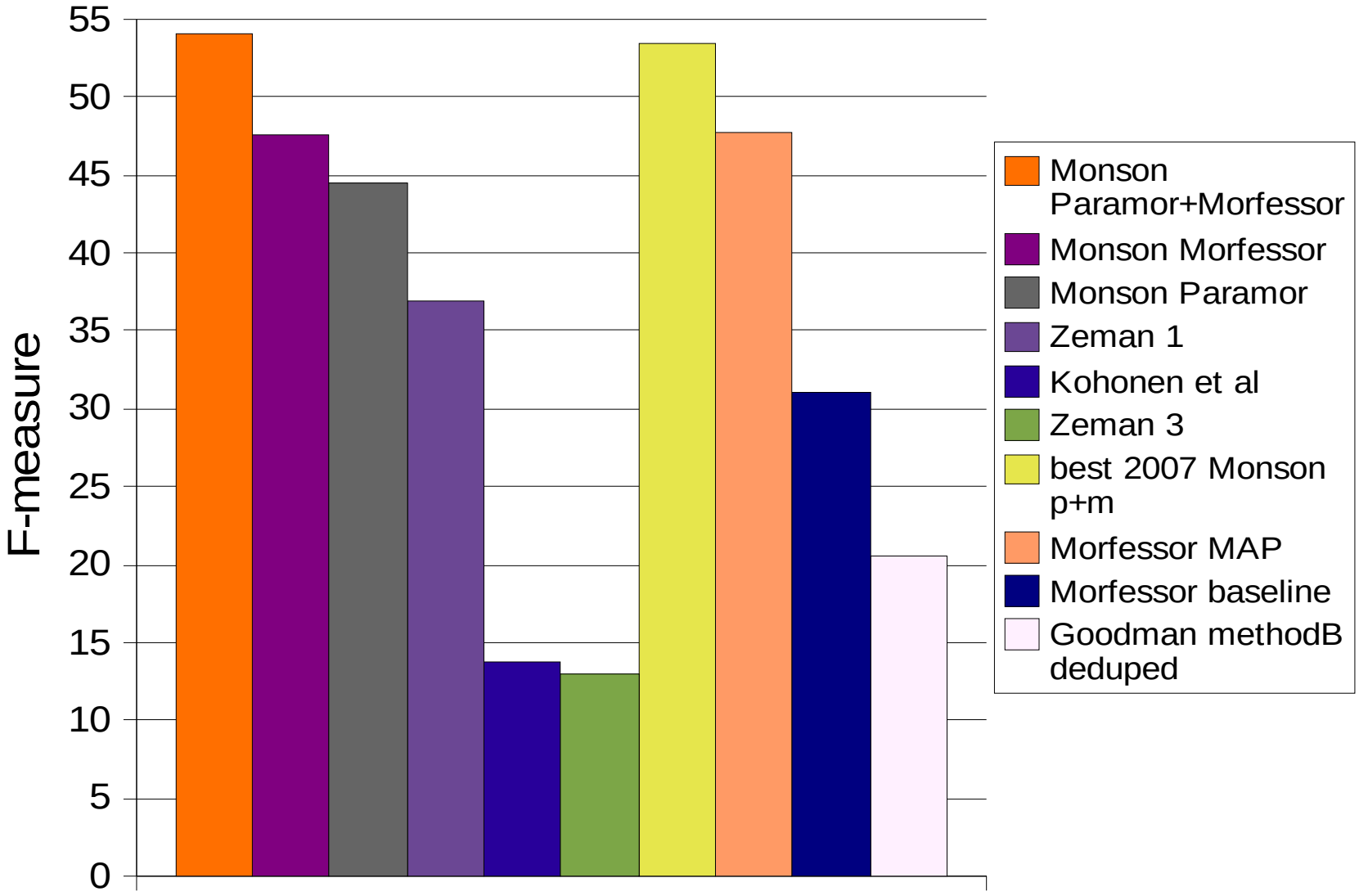
Results: Turkish, 620K word types



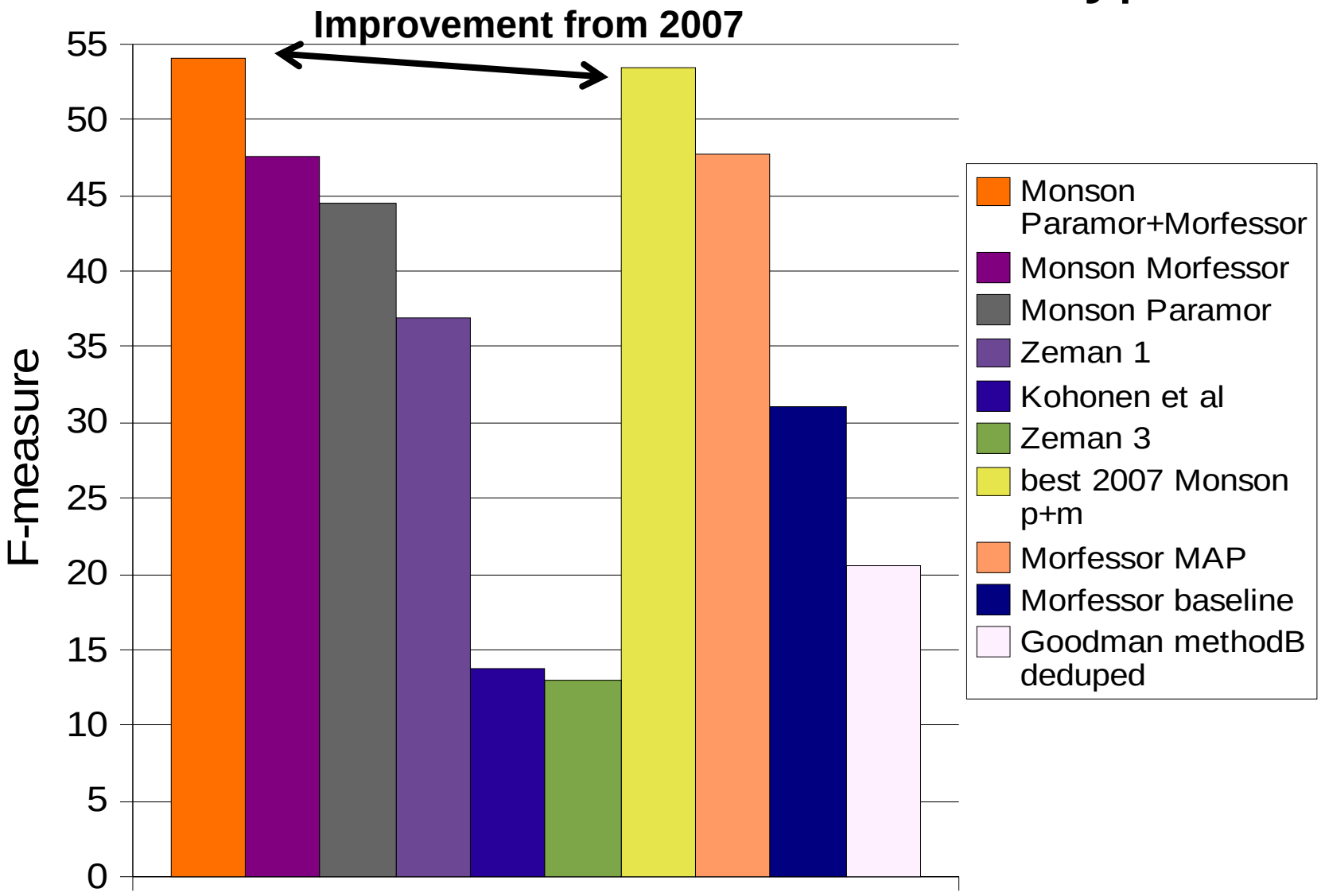
Results: Turkish, 620K word types



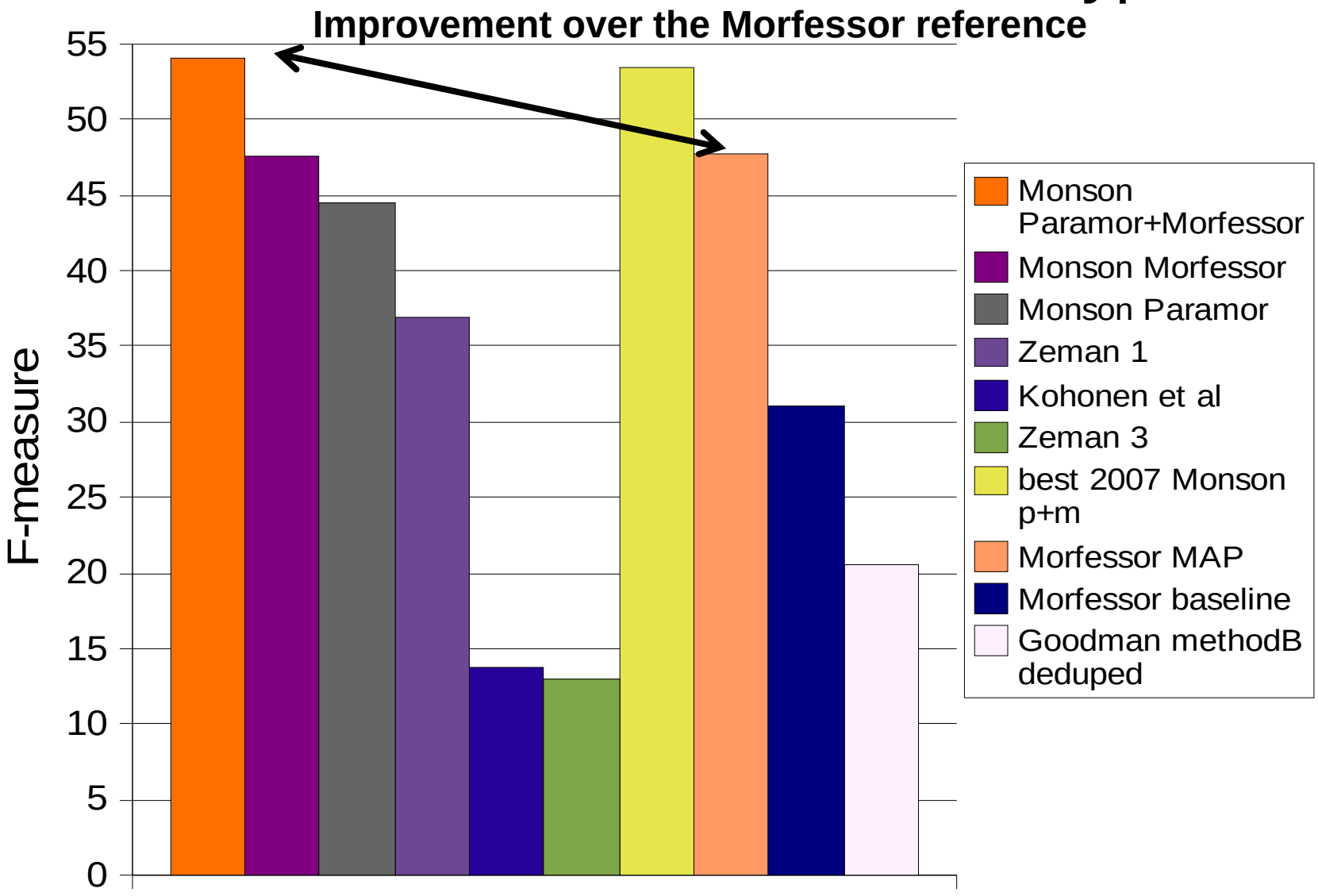
Results: German, 1.3M word types



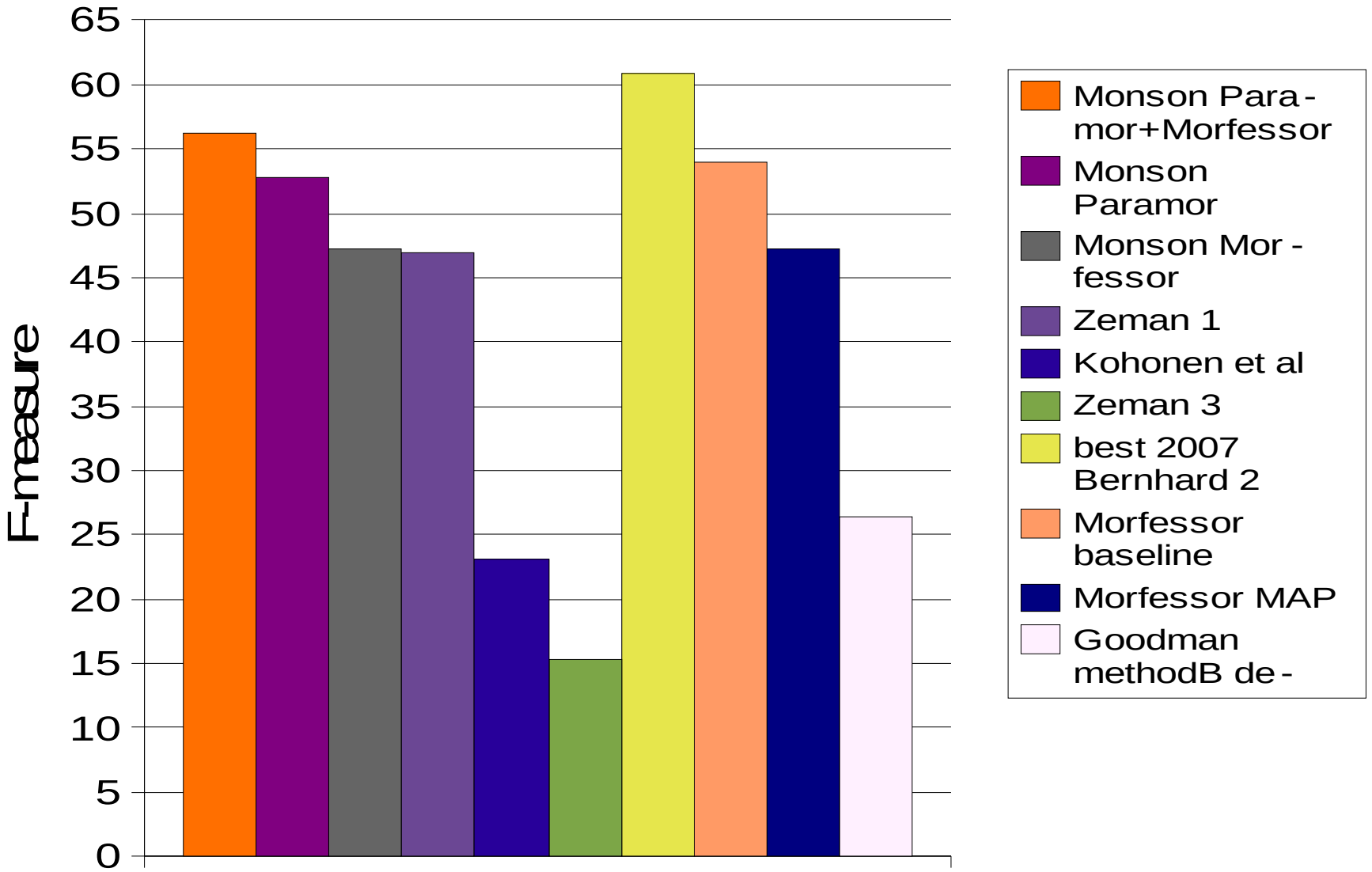
Results: German, 1.3M word types



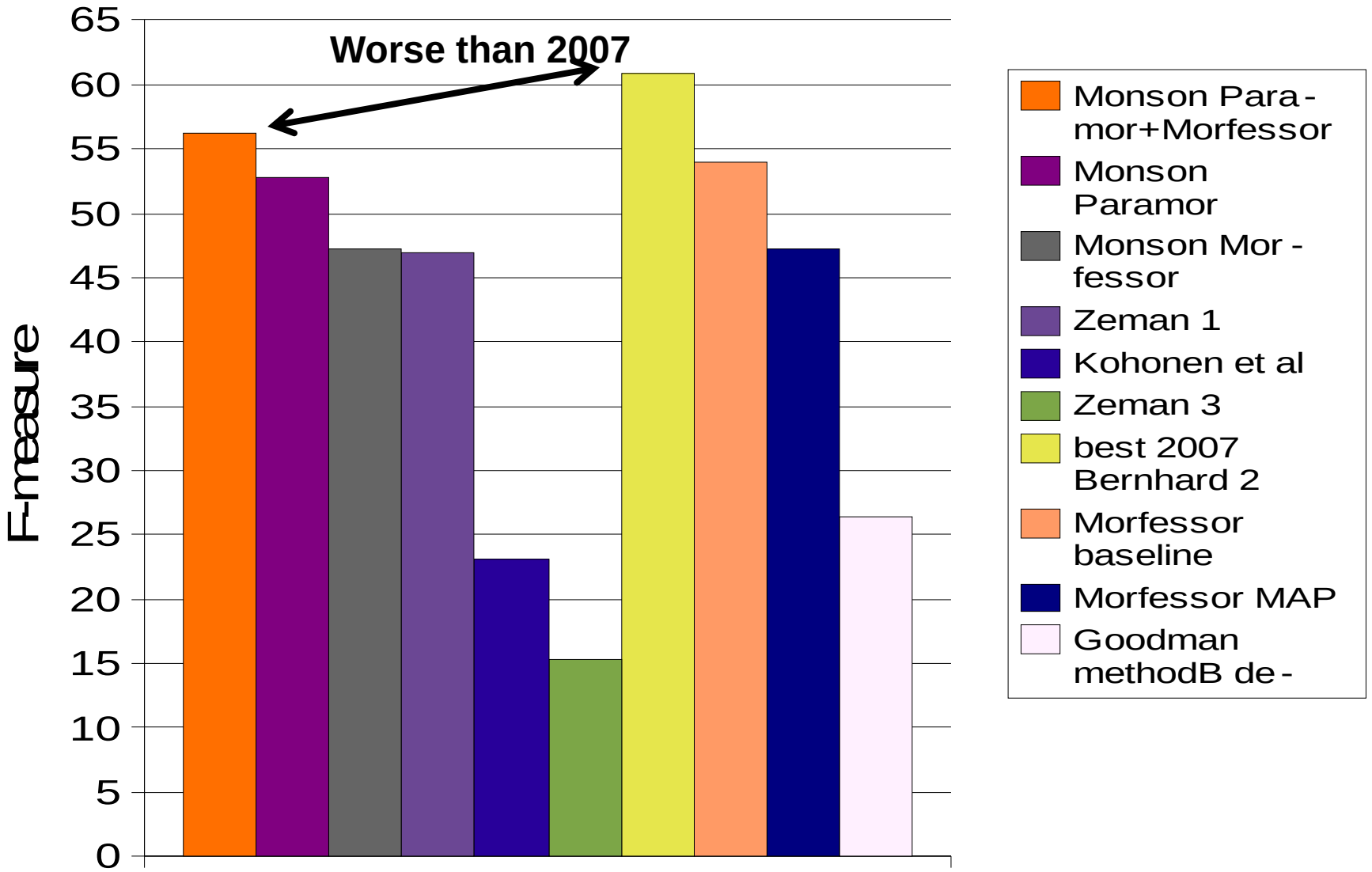
Results: German, 1.3M word types



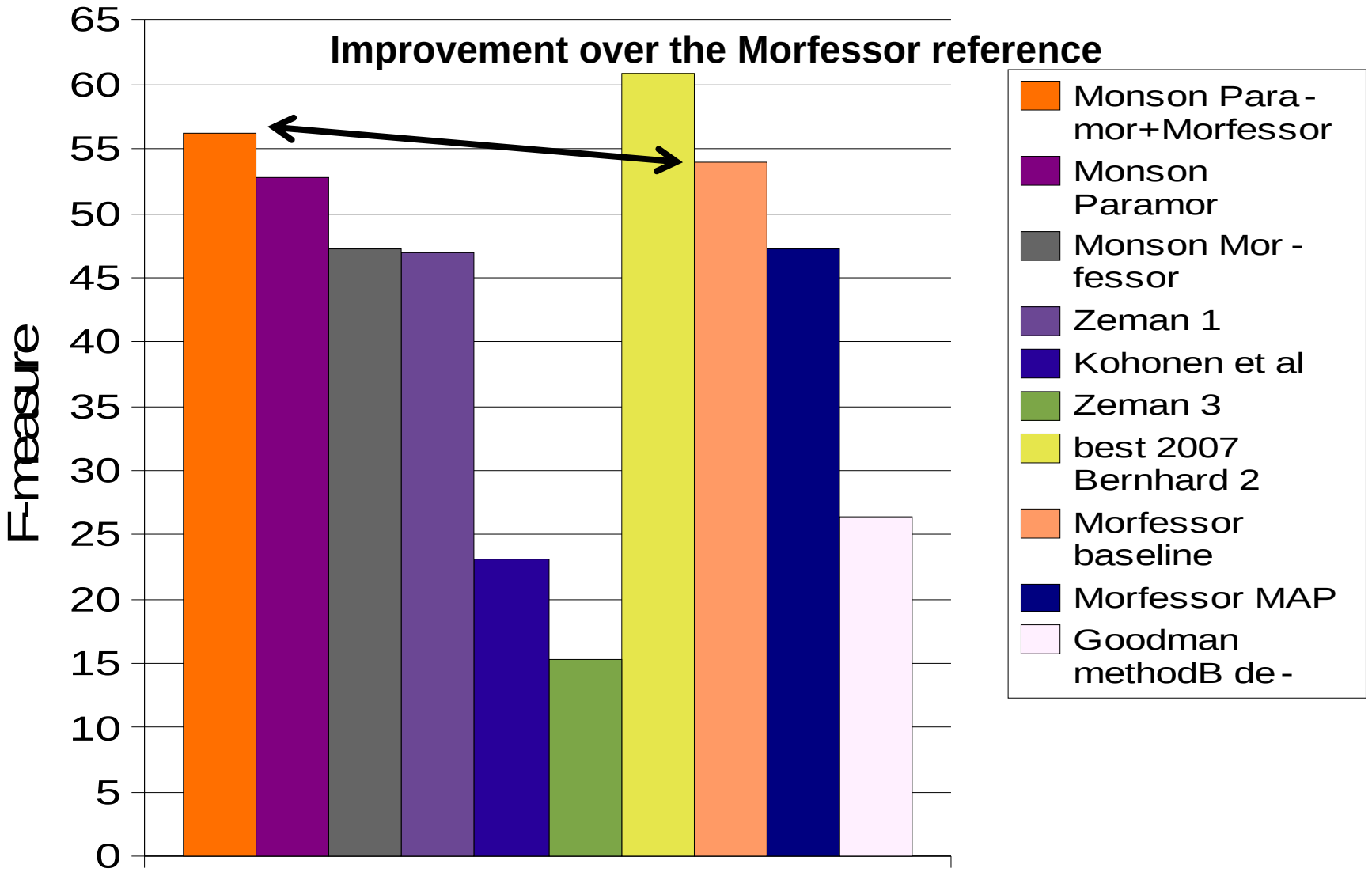
Results: English, 380K word types



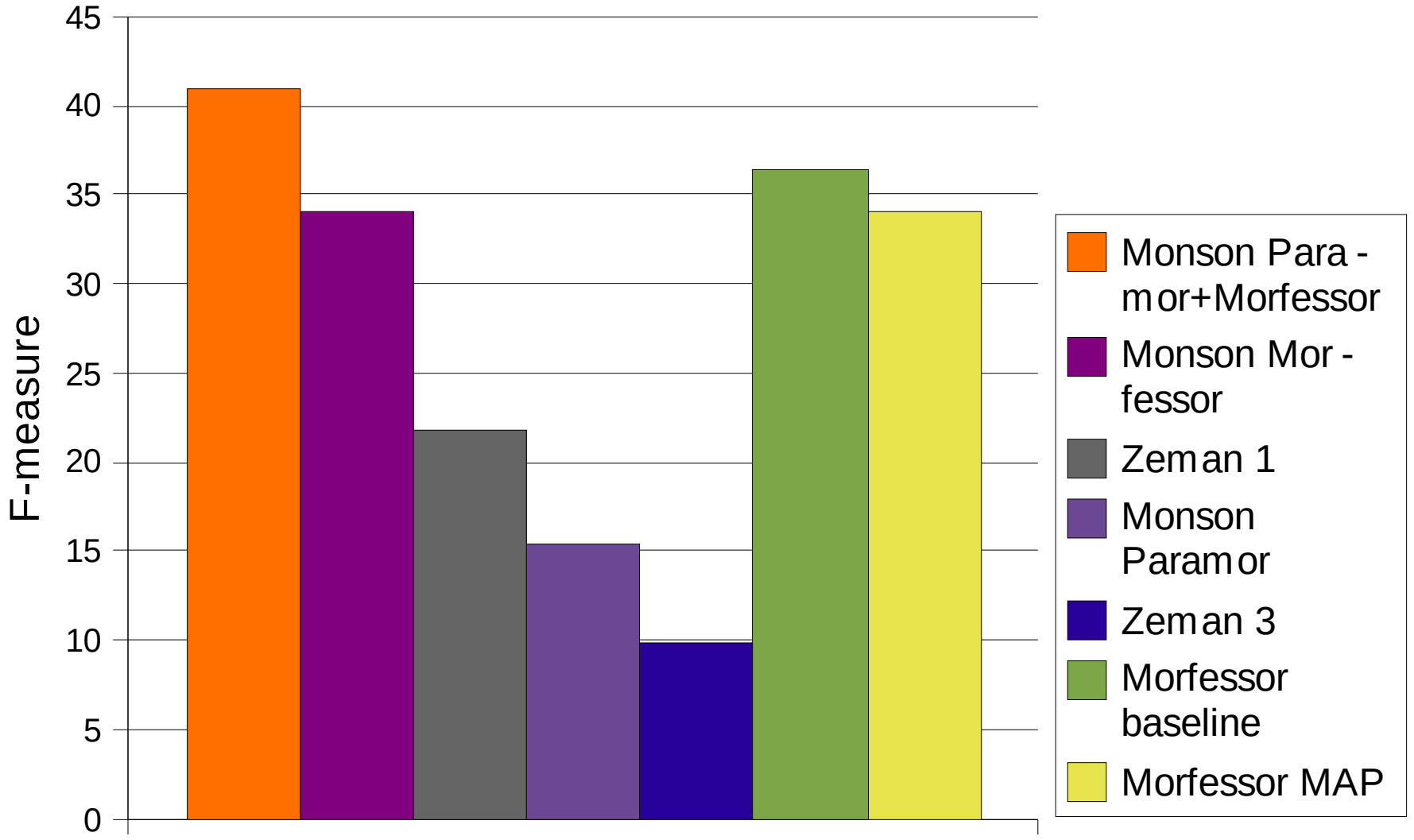
Results: English, 380K word types



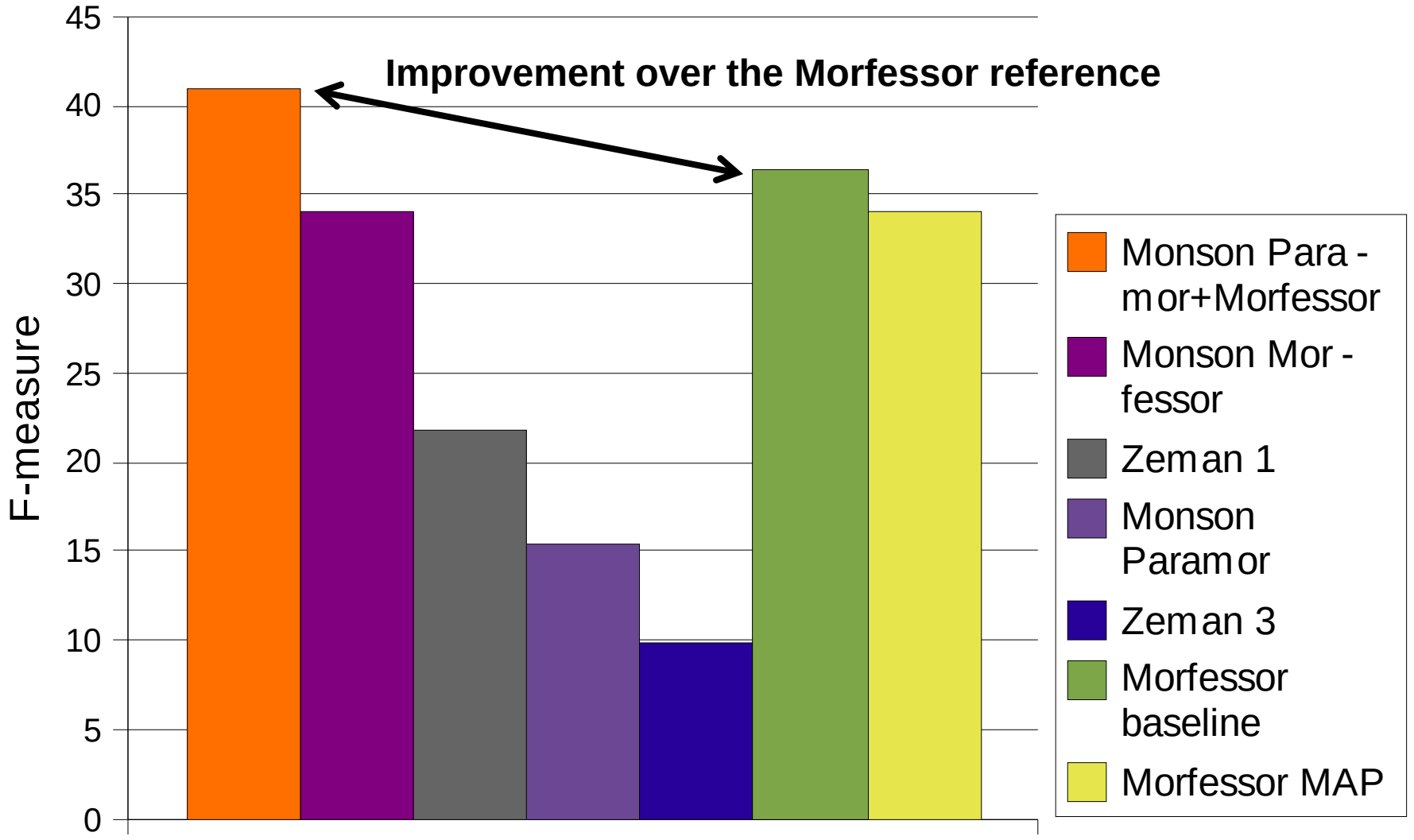
Results: English, 380K word types



Results: Arabic, 140K word types



Results: Arabic, 140K word types



About 2008 results

- Large **improvements over 2007** in FIN and TUR
- Highest F-score in ENG and lowest in ARA, but the best algorithms survived >30% in all tasks
- Features of the gold standard affect the level of F-scores in each language
- Best algorithm in all tasks: **Monson ParaMor + Morfessor**, combined analysis
- The ”simple” Morfessor Baseline still hard to beat in ENG and ARA

2. IR evaluation

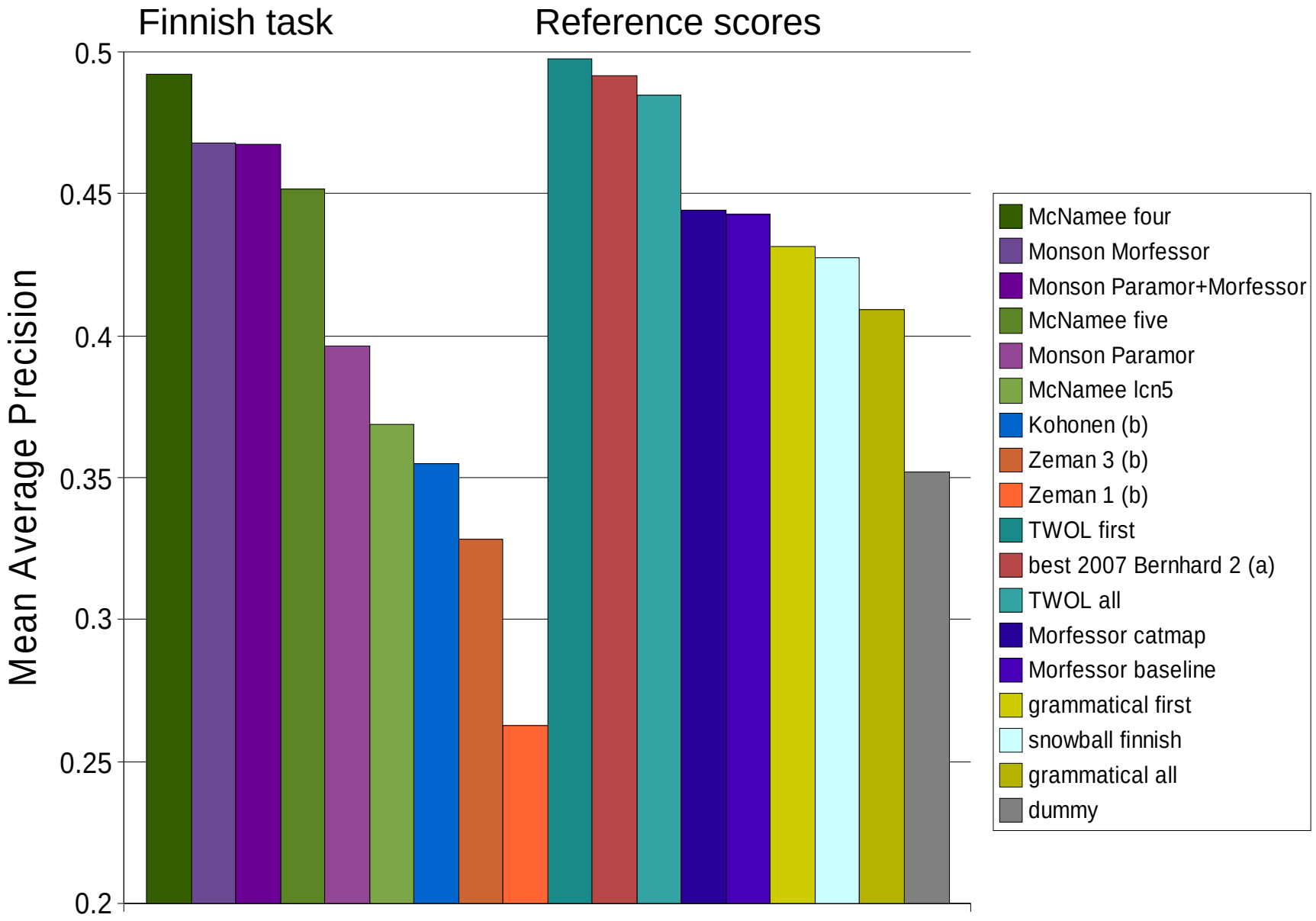
- words in the documents and queries were replaced by the suggested segmentations
- OOV words un-replaced
- all morphemes used for indexing
- stoplist for the most common ones (over a fixed frequency threshold)
- LEMUR-toolkit http://www.lemurproject.org/
- Okapi BM25 retrieval method (default)

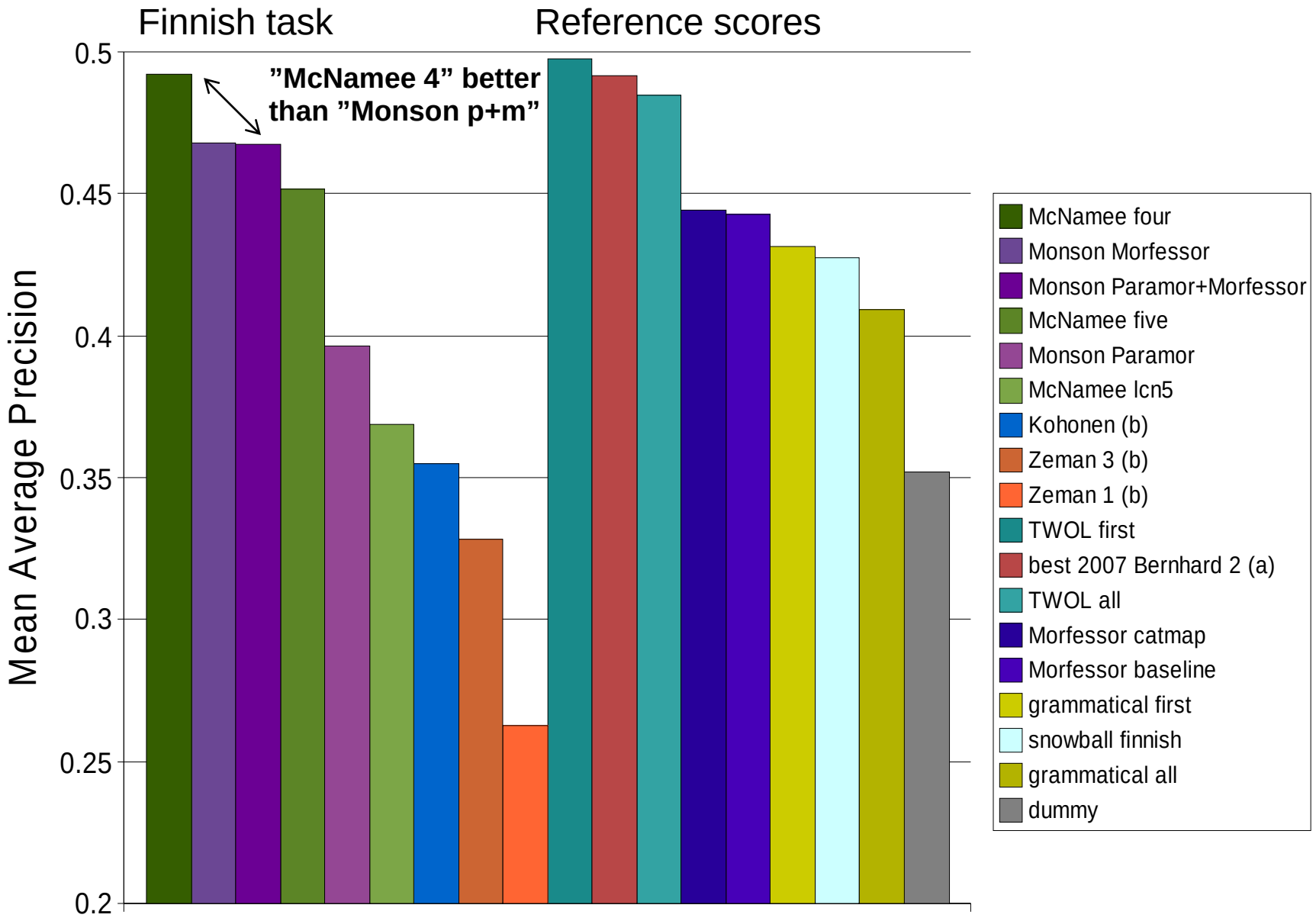
IR data sets (as in CLEF 2007)

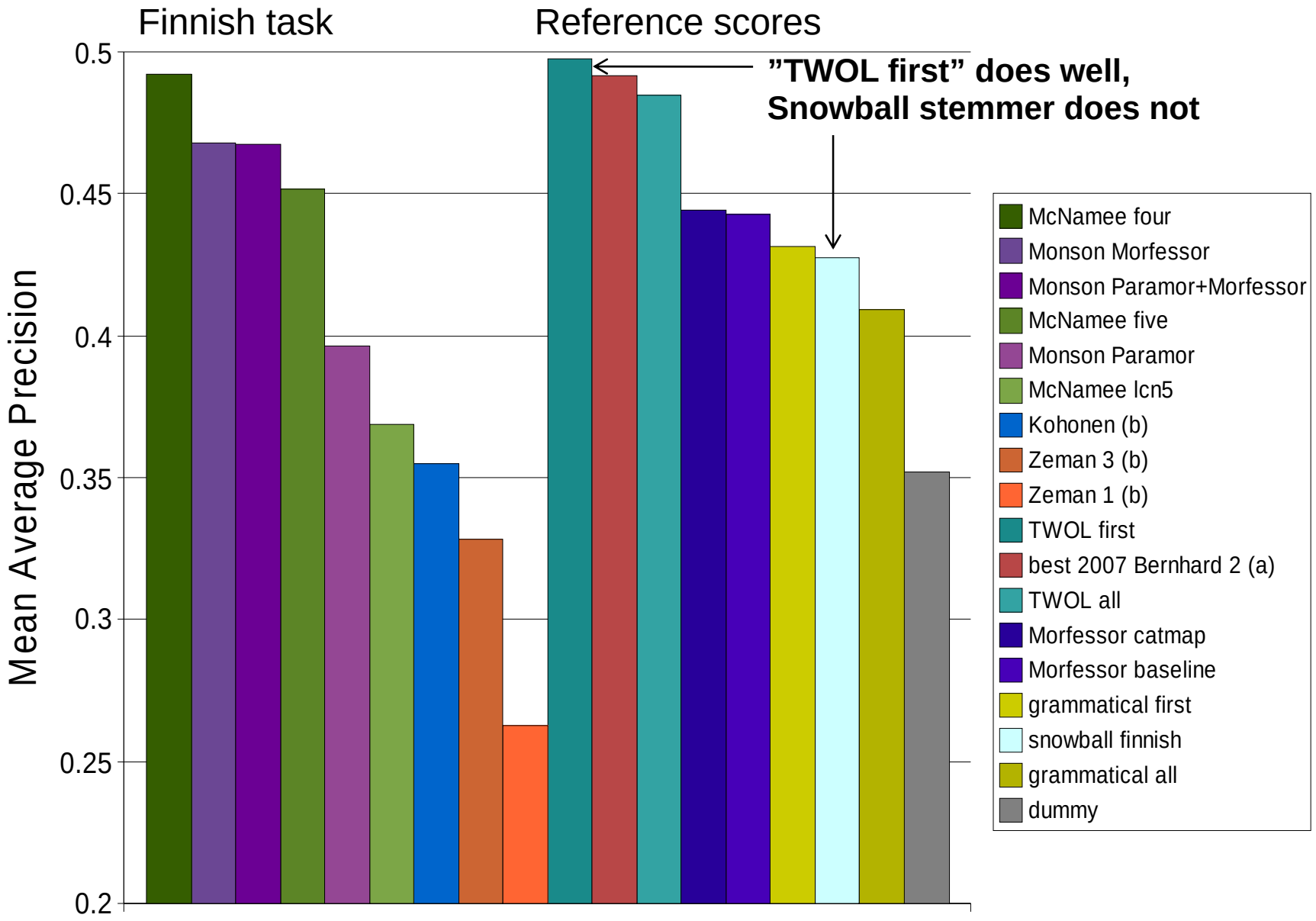
- **Finnish (CLEF 2004)**
  - **55K documents from articles in Aamulehti 1994-95**
  - **50 test queries, 23 binary relevance assessments**
- **English (CLEF 2005)**
  - **107K documents from articles in Los Angeles Times 1994 and Glasgow Herald 1995**
  - **50 test queries, 20K binary relevance assessments**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

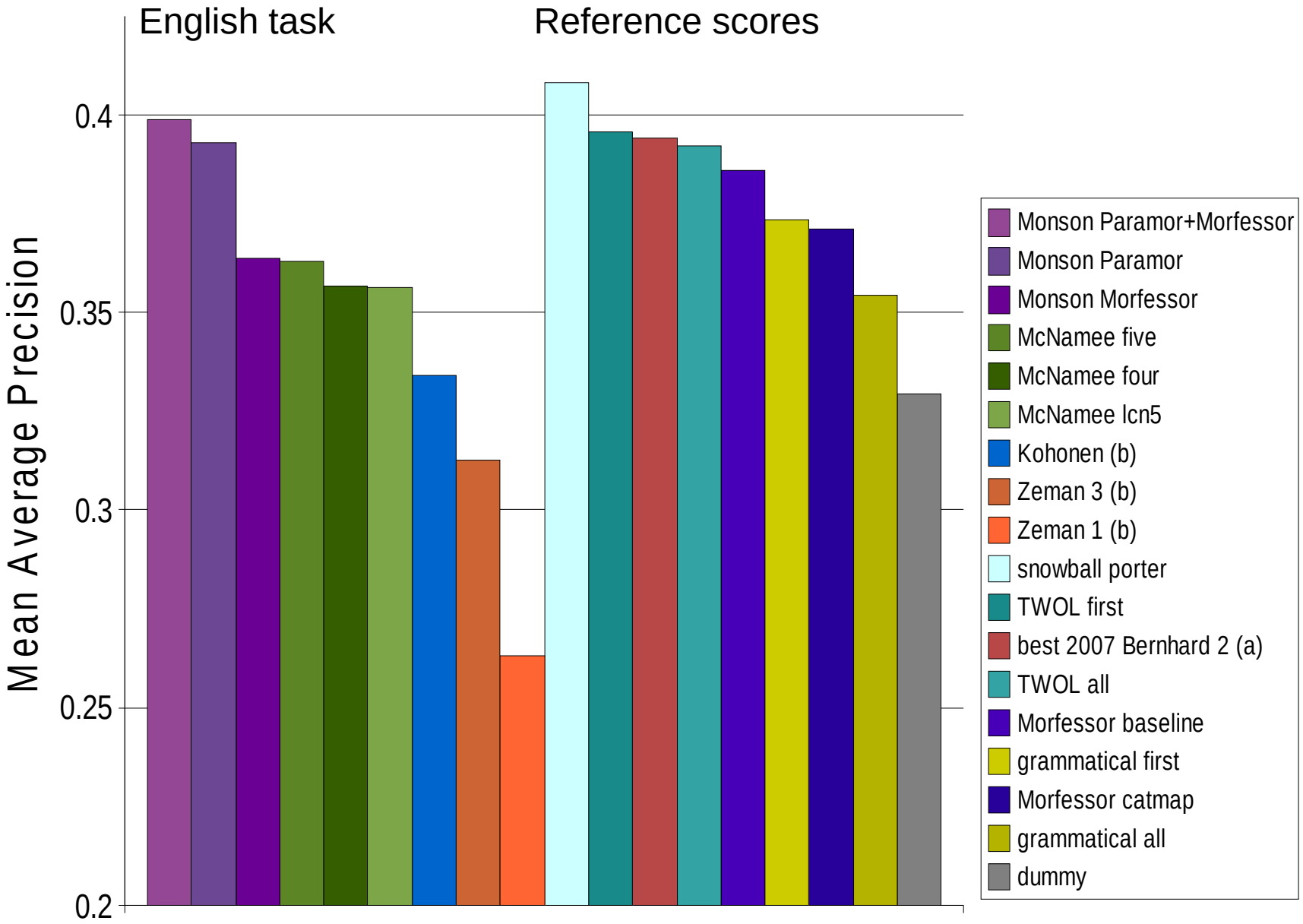
Reference methods

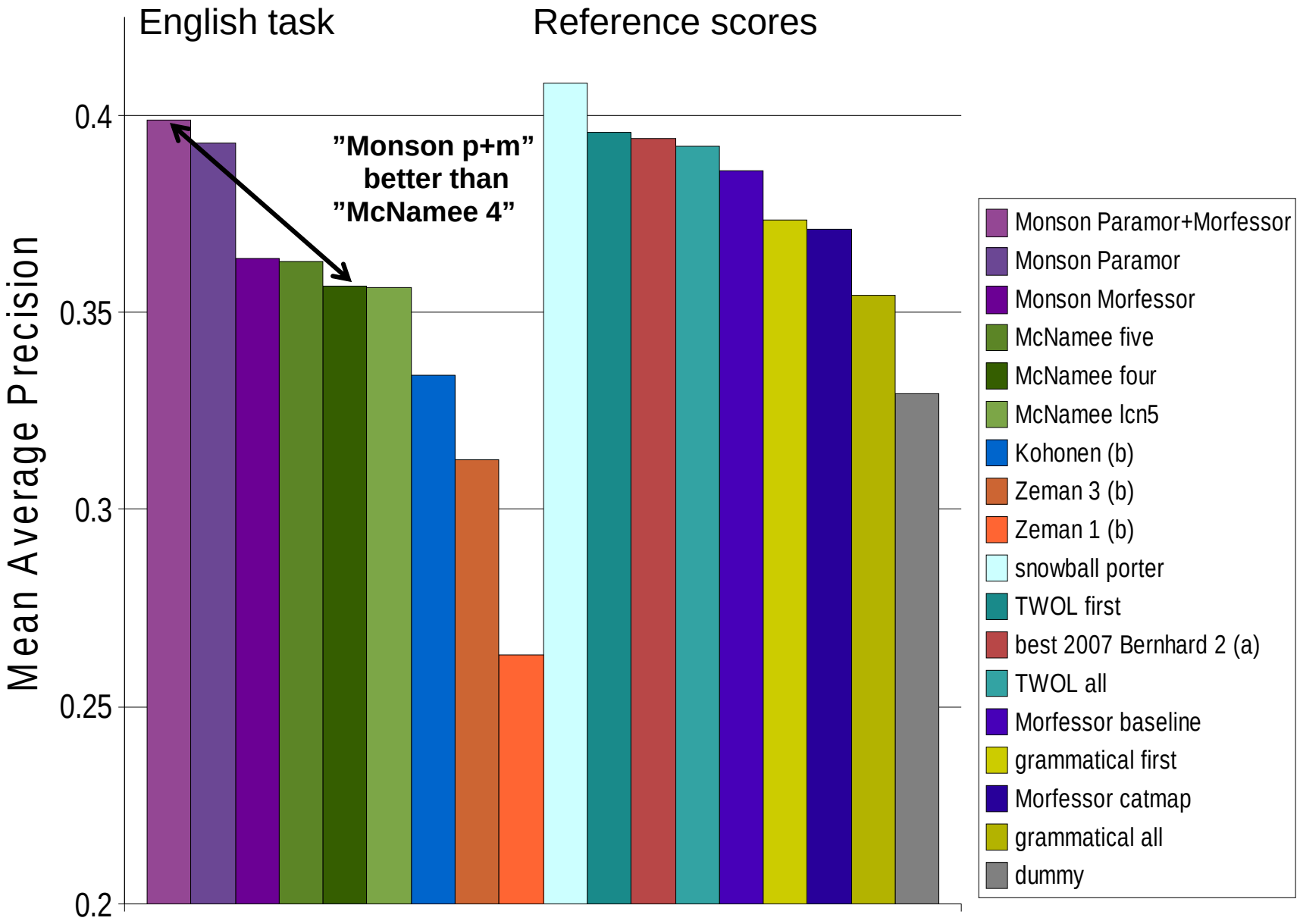
- **Morfessor Baseline:** our public code since 2002
- **Morfessor Categories-MAP:** improved, public 2006
- **dummy**: no segmentation, all words unsplit
- **grammatical**: full gold standard segmentation (reference of competition 1)
  - all: all alternative segmentations included
  - first: only the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

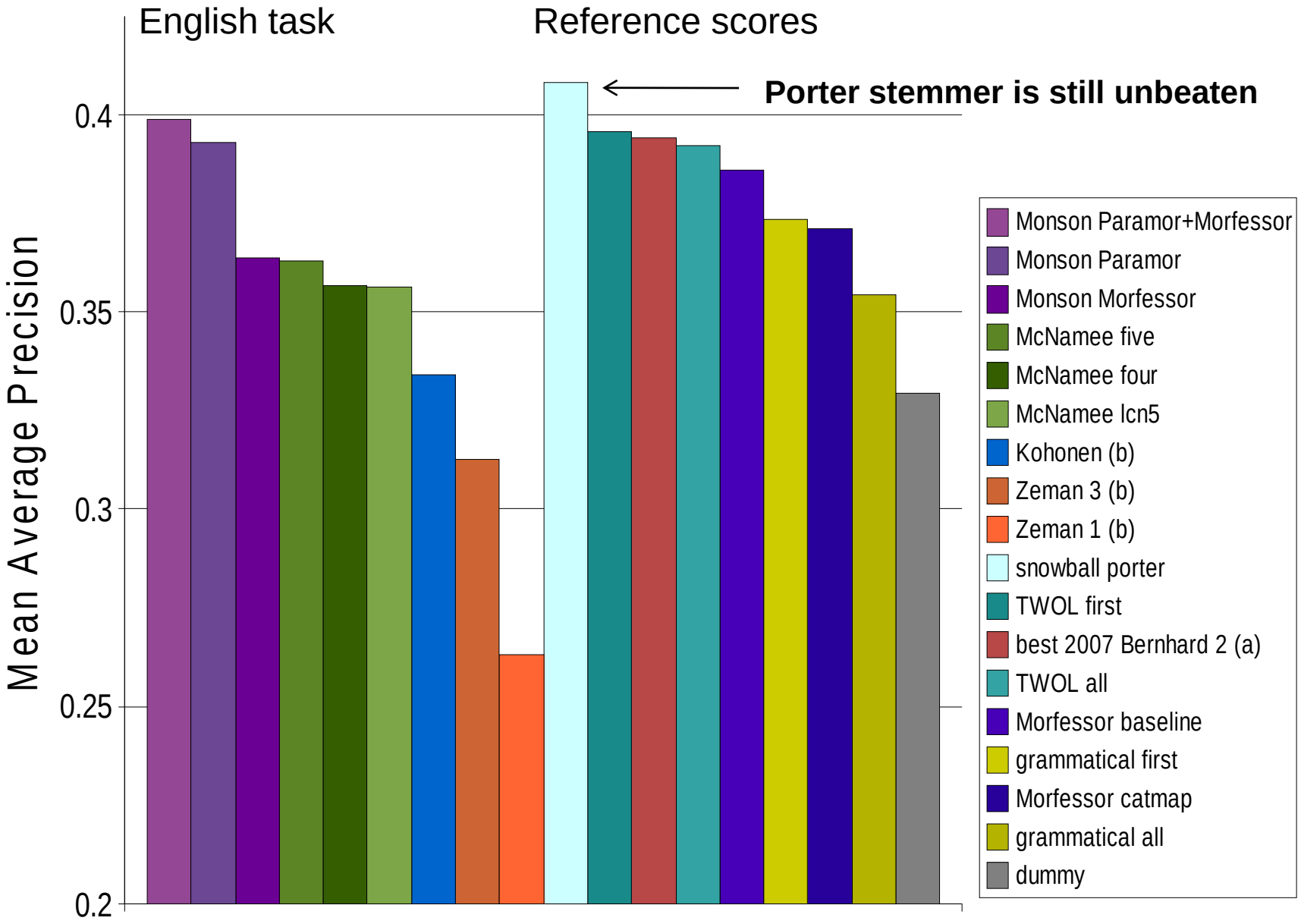

Finnish task
Reference scores
Mean Average Precision
0.5
0.45
0.4
0.35
0.3
0.25
0.2
McNamee four
Monson Morfessor
Monson Paramor+Morfessor
McNamee five
Monson Paramor
McNamee lcn5
Kohonen (b)
Zeman 3 (b)
Zeman 1 (b)
TWOL first
best 2007 Bernhard 2 (a)
TWOL all
Morfessor catmap
Morfessor baseline
grammatical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

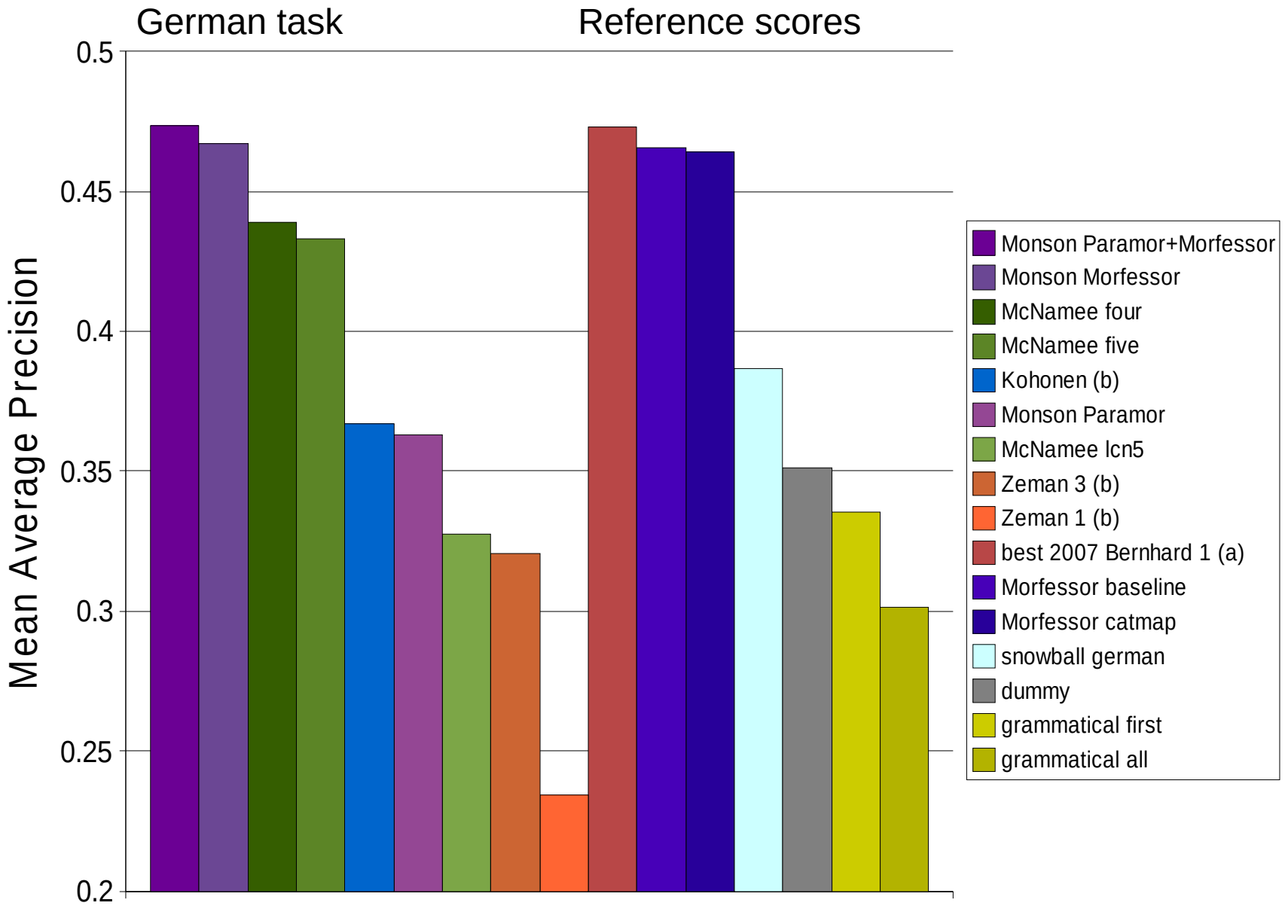

Finnish task
Reference scores
”McNamee 4” better than ”Monson p+m”
Mean Average Precision
0.5
0.45
0.4
0.35
0.3
0.25
0.2
McNamee four
Monson Morfessor
Monson Paramor+Morfessor
McNamee five
Monson Paramor
McNamee lcn5
Kohonen (b)
Zeman 3 (b)
Zeman 1 (b)
TWOL first
best 2007 Bernhard 2 (a)
TWOL all
Morfessor catmap
Morfessor baseline
grammatical first
snowball finnish
grammatical all
dummy

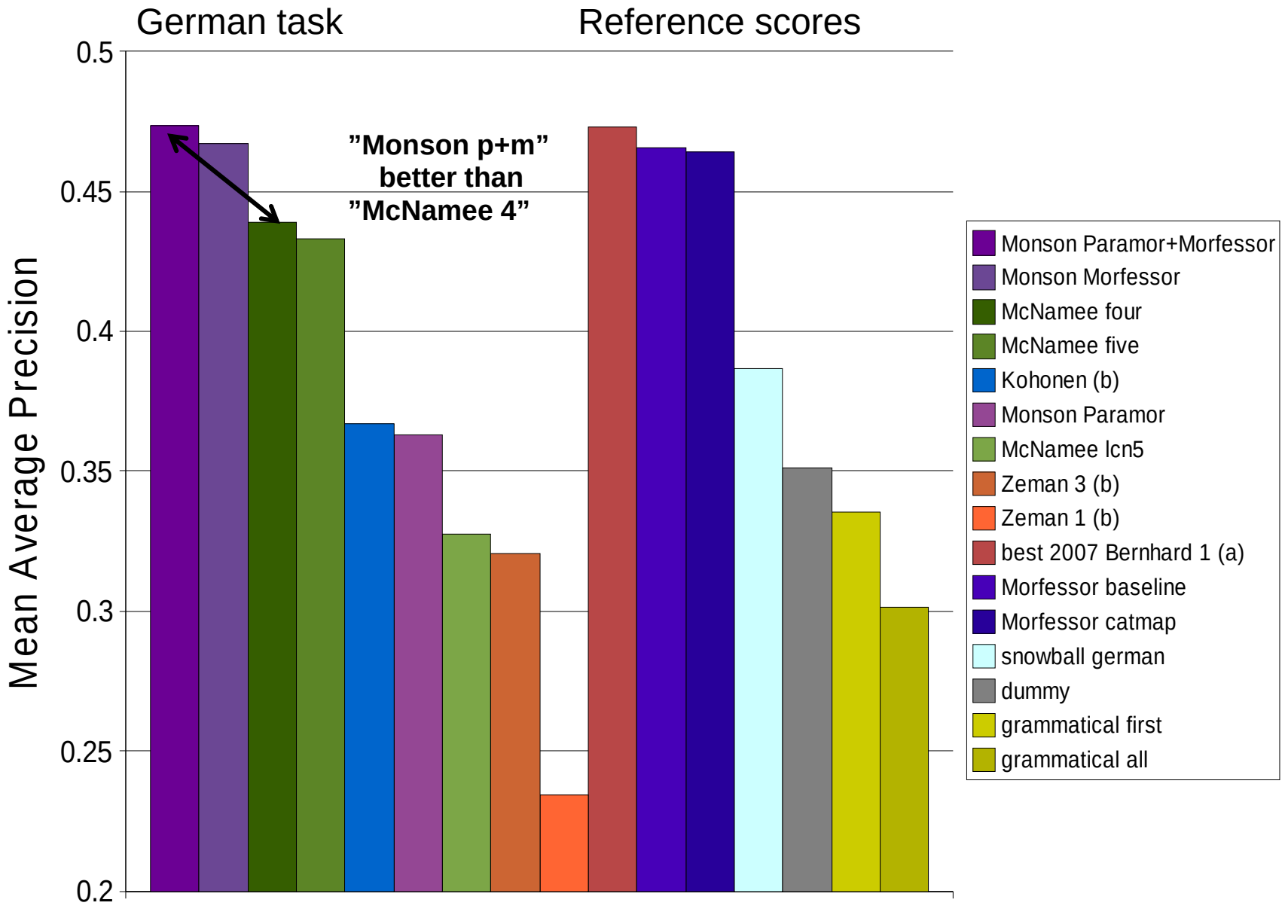

Finnish task
Reference scores
"TWOL first" does well,
Snowball stemmer does not
Mean Average Precision
0.5
0.45
0.4
0.35
0.3
0.25
0.2
McNamee four
Monson Morfessor
Monson Paramor+Morfessor
McNamee five
Monson Paramor
McNamee lcn5
Kohonen (b)
Zeman 3 (b)
Zeman 1 (b)
TWOL first
best 2007 Bernhard 2 (a)
TWOL all
Morfessor catmap
Morfessor baseline
grammatical first
snowball finnish
grammatical all
dummy

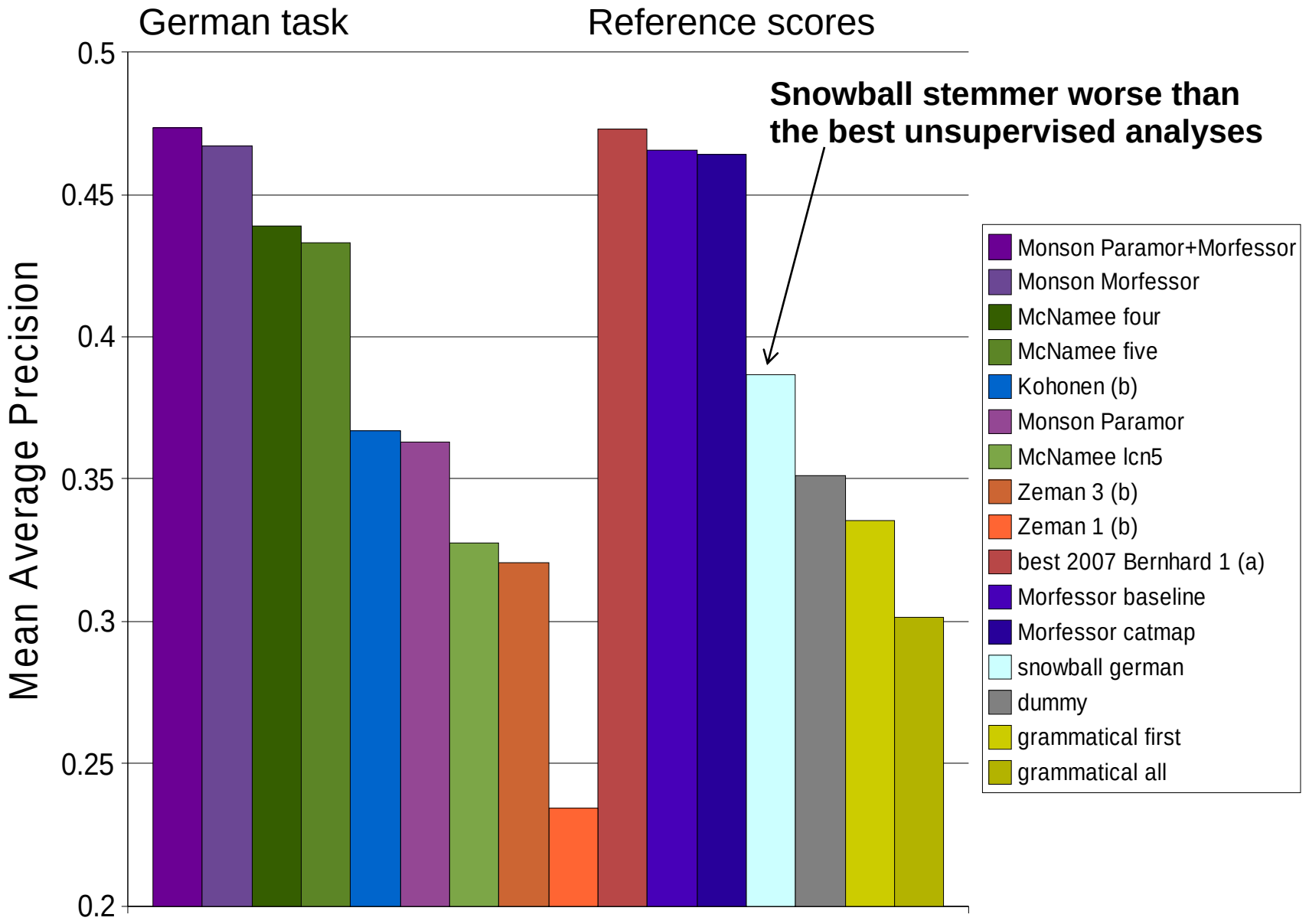

English task
Reference scores
Mean Average Precision
0.4
0.35
0.3
0.25
0.2
Monson Paramor+Morfessor
Monson Paramor
Monson Morfessor
McNamee five
McNamee four
McNamee lcn5
Kohonen (b)
Zeman 3 (b)
Zeman 1 (b)
snowball porter
TWOL first
best 2007 Bernhard 2 (a)
TWOL all
Morfessor baseline
grammatical first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


English task
Reference scores
Mean Average Precision
0.4
0.35
0.3
0.25
0.2
”Monson p+m” better than ”McNamee 4”
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


English task
Reference scores
Porter stemmer is still unbeaten
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


German task
Reference scores
Mean Average Precision
0.5
0.45
0.4
0.35
0.3
0.25
0.2
”Monson p+m”
better than
”McNamee 4”
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


German task
Reference scores
Snowball stemmer worse than the best unsupervised analyses
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/*

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

ParaMor: State-of-the-Art Unsupervised Morphology Induction System

Poster by Christian Monson et al.

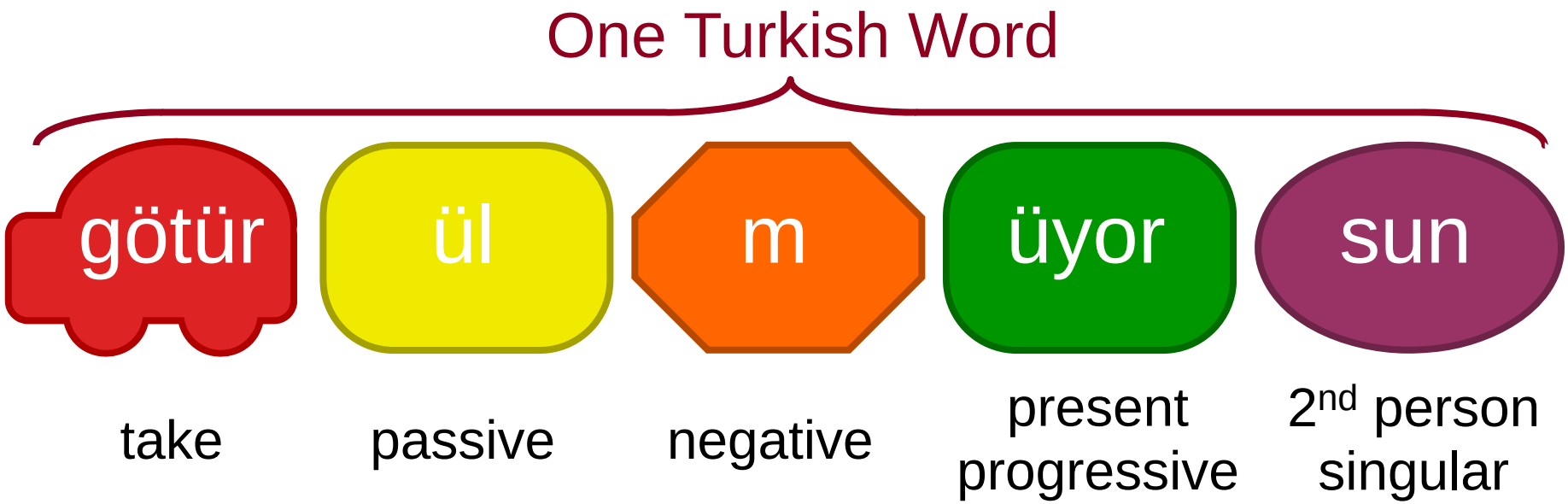
ParaMor

- **I**dentifies paradigms
  - **T**he organizing **structure** of inflectional morphology
- **S**egments words
  - **A**s discovered paradigms suggest

Combined with Morfessor

- **A**mong the best in Morpho Challenge
- **C**onsistent across languages

Turkish Morphology – Beads on a String



“You are not being taken”