Multilingual Information Access: What and How Well?

Nicola Ferro

@frrncl

University of Padua, Italy

ecir PADUA 2016

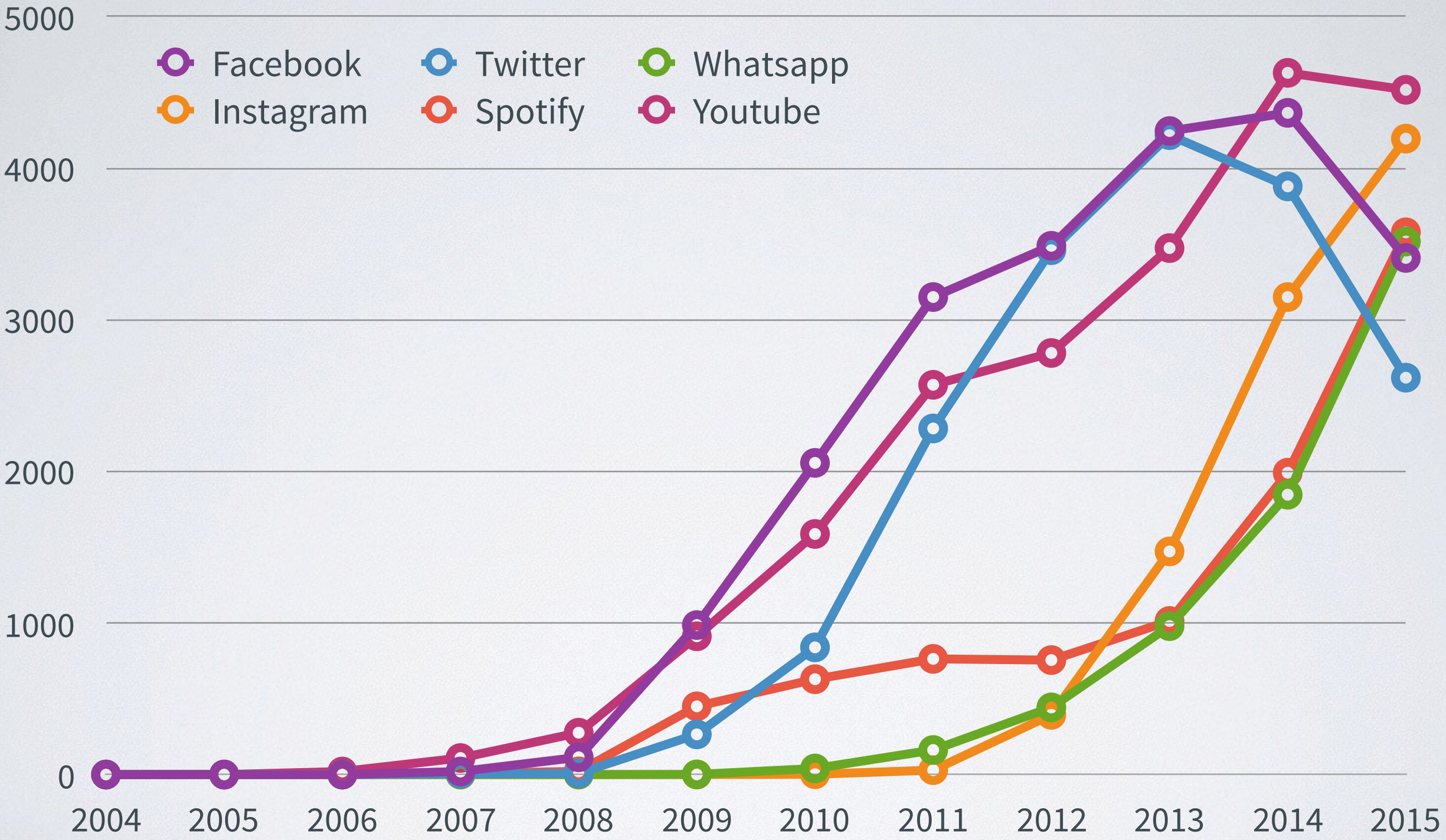
38th European Conference on Information Retrieval

Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine 2016)

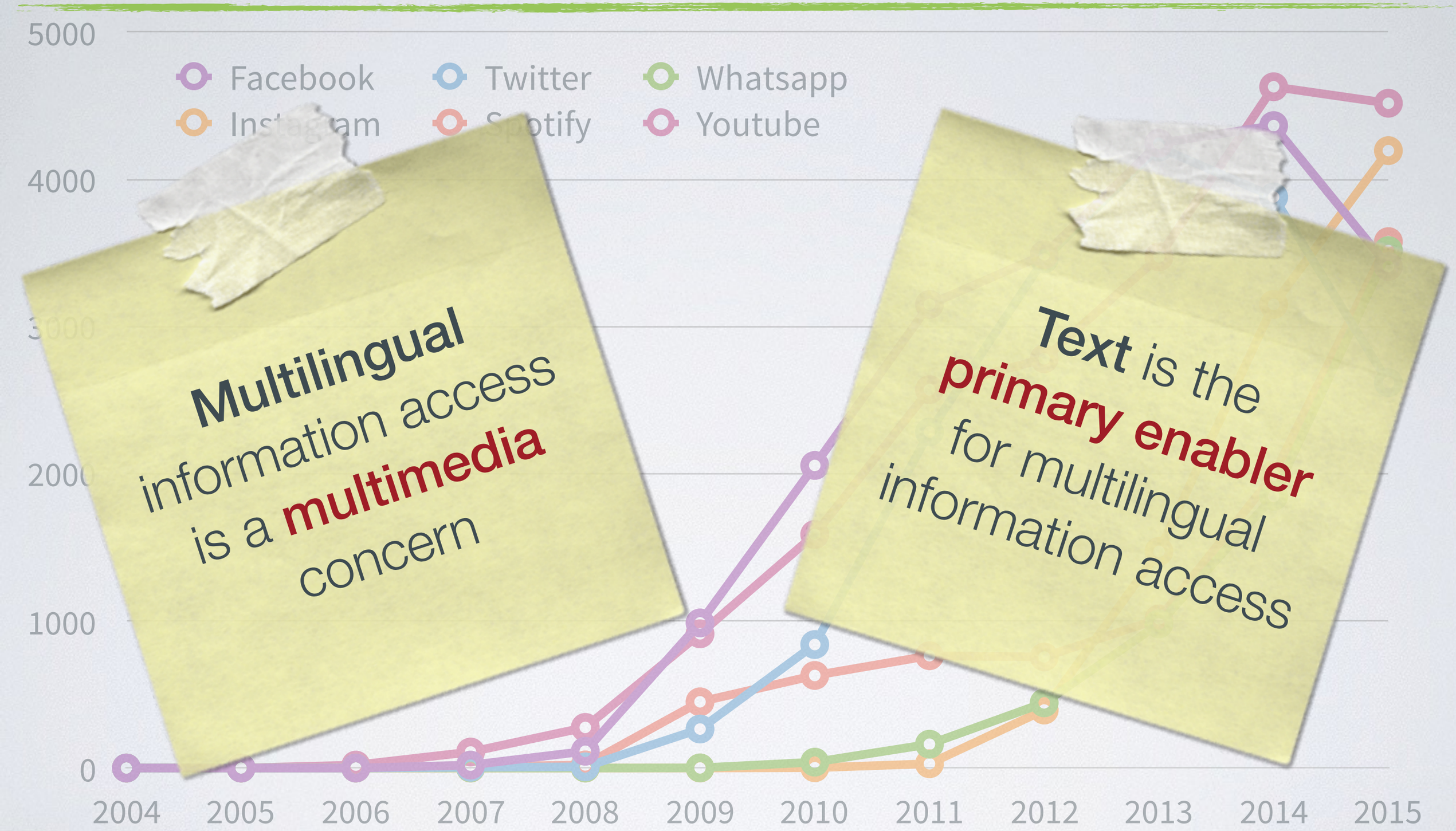
20 March 2016, Padua, Italy

What

Some search trends: translate + ...



Some search trends: translate + ...

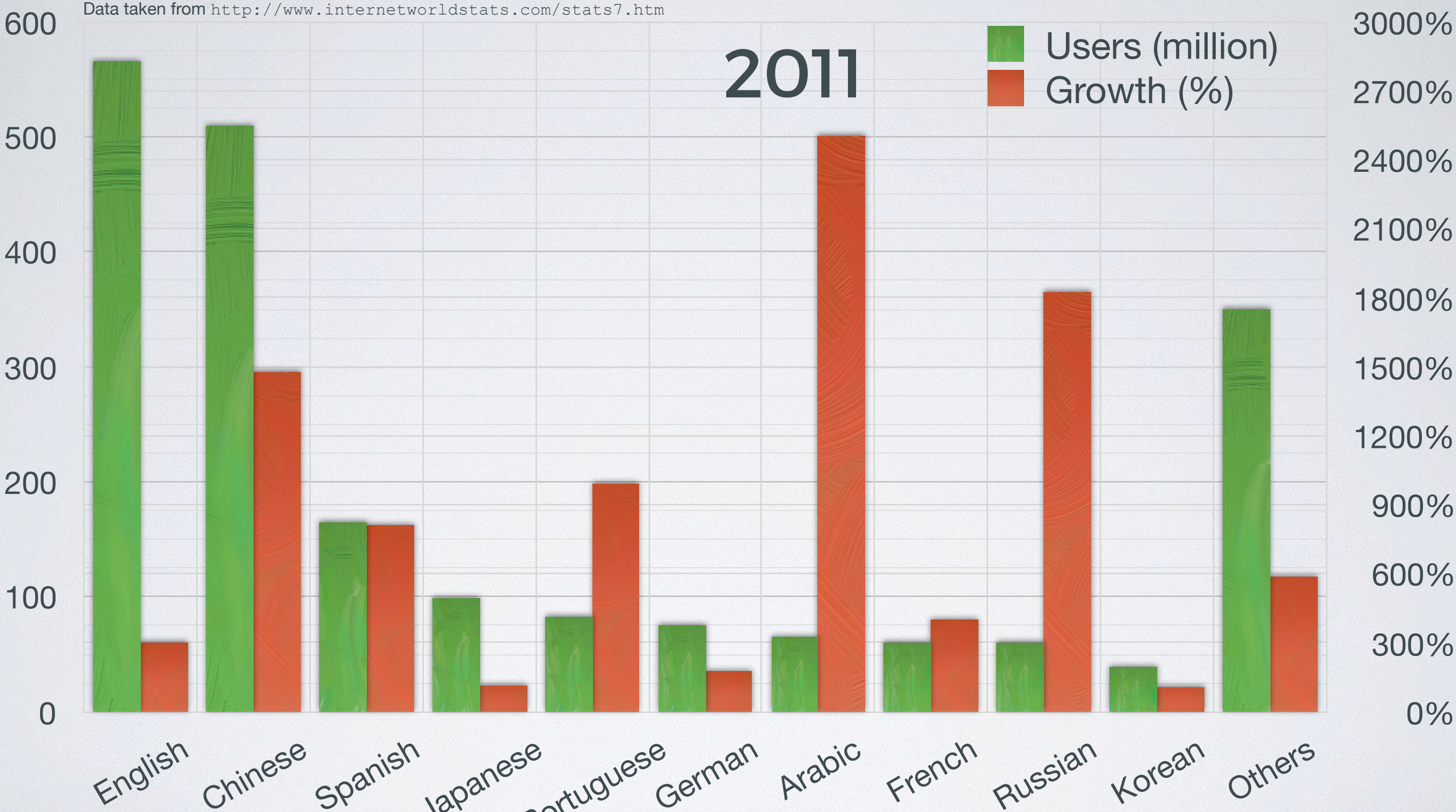


Multilingual information access is a **multimedia** concern

Text is the **primary enabler** for multilingual information access

Top Ten Languages in the Web

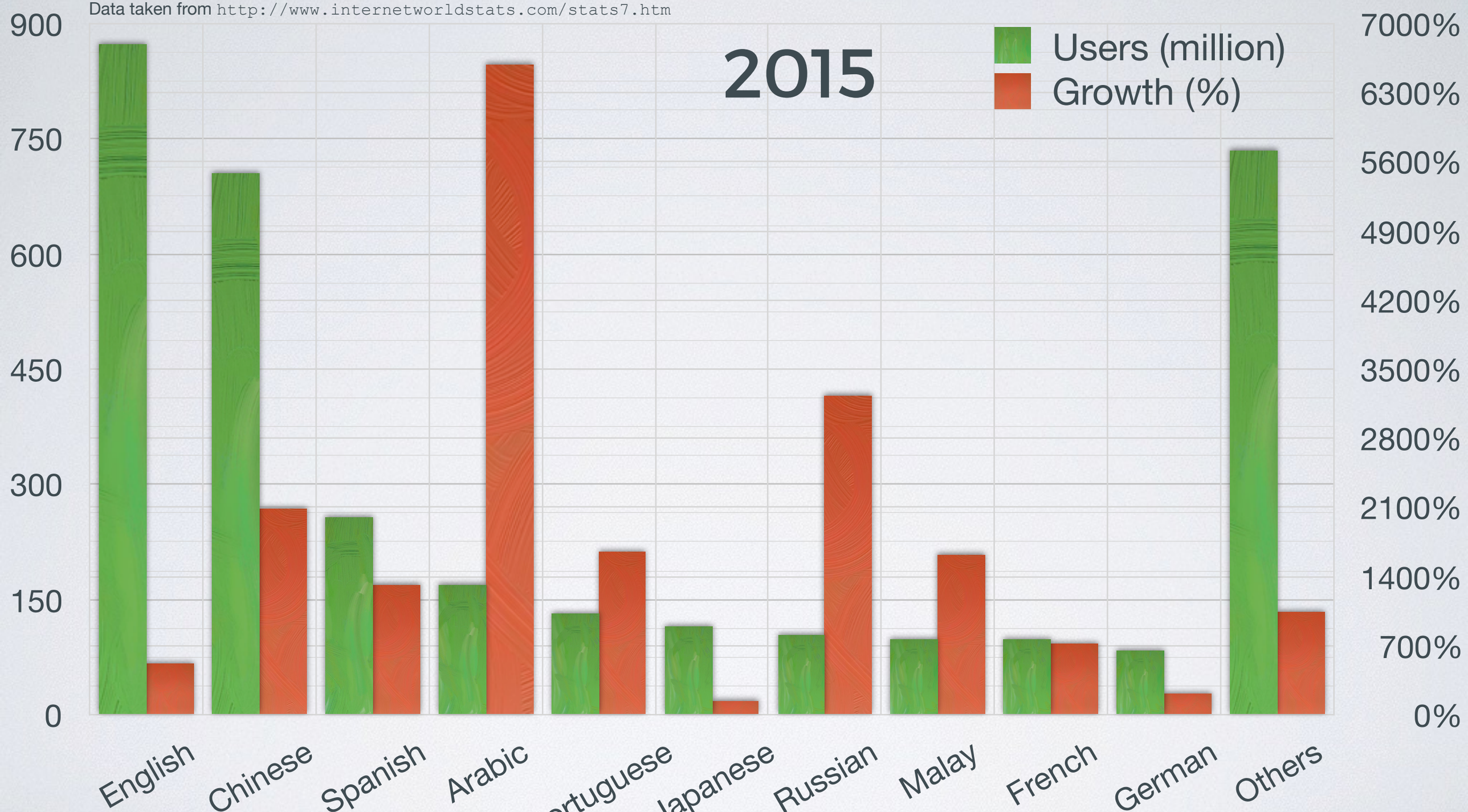
Data taken from http://www.internetworldstats.com/stats7.htm



@frrncl
#ecir2016

Top Ten Languages in the Web

Data taken from http://www.internetworldstats.com/stats7.htm



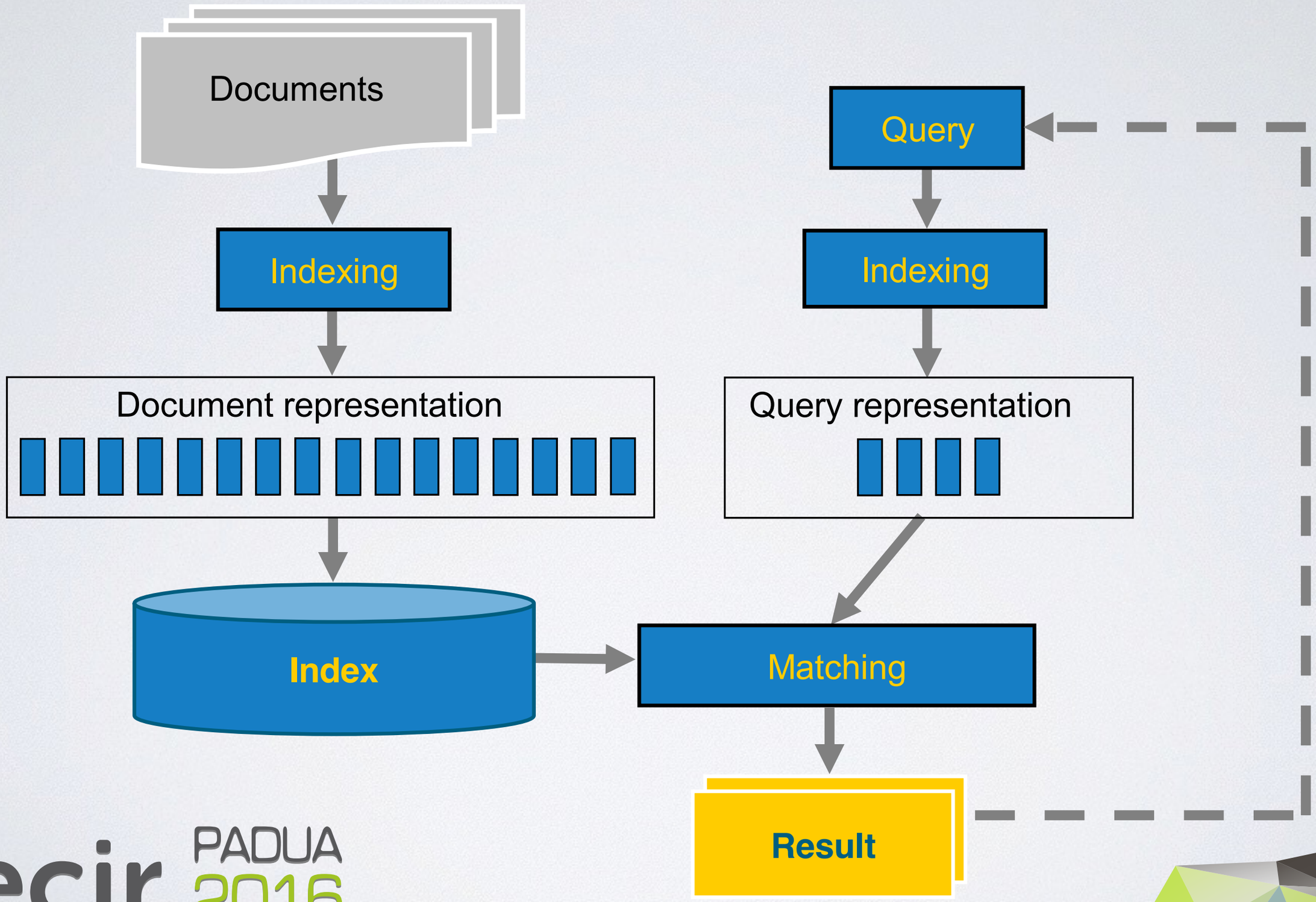
Multilingual Information Access

> **Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified.**
>
> [D. Oard & D. Hull, AAAI Symposium on Cross-Language IR, Spring 1997, Stanford]

- Monolingual retrieval in non-English languages
- Bilingual retrieval A → B
- Multilingual retrieval A → A, B, ...
- Multilingual retrieval AB → A, AB, AC, B, BC, ABC, ...

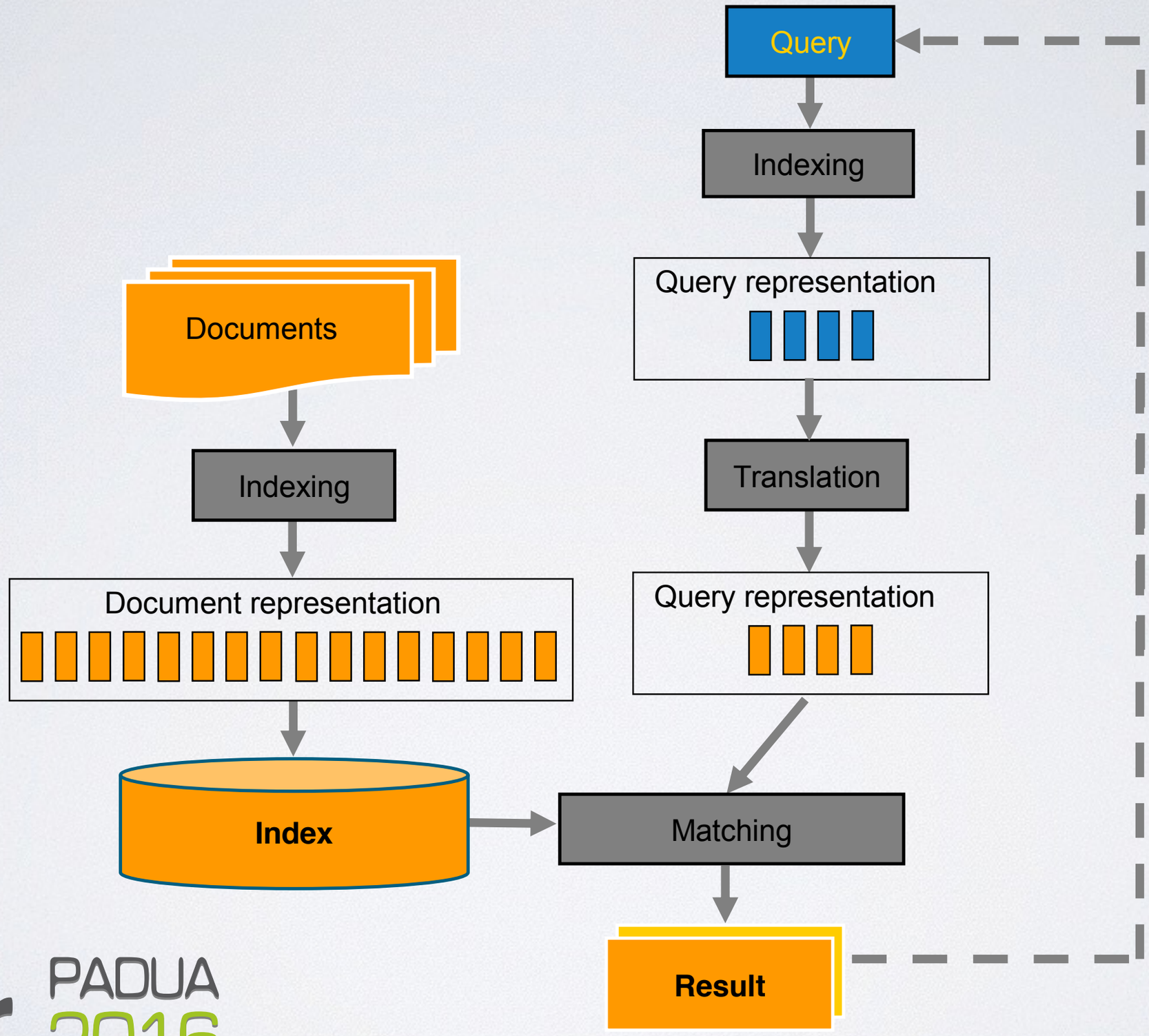
Typical IR Flow

[Figure taken from M. Braschler, *Multilingual Information Retrieval and Cross-Language Information Retrieval*, TrebleCLEF Summer School 2009, Italy]



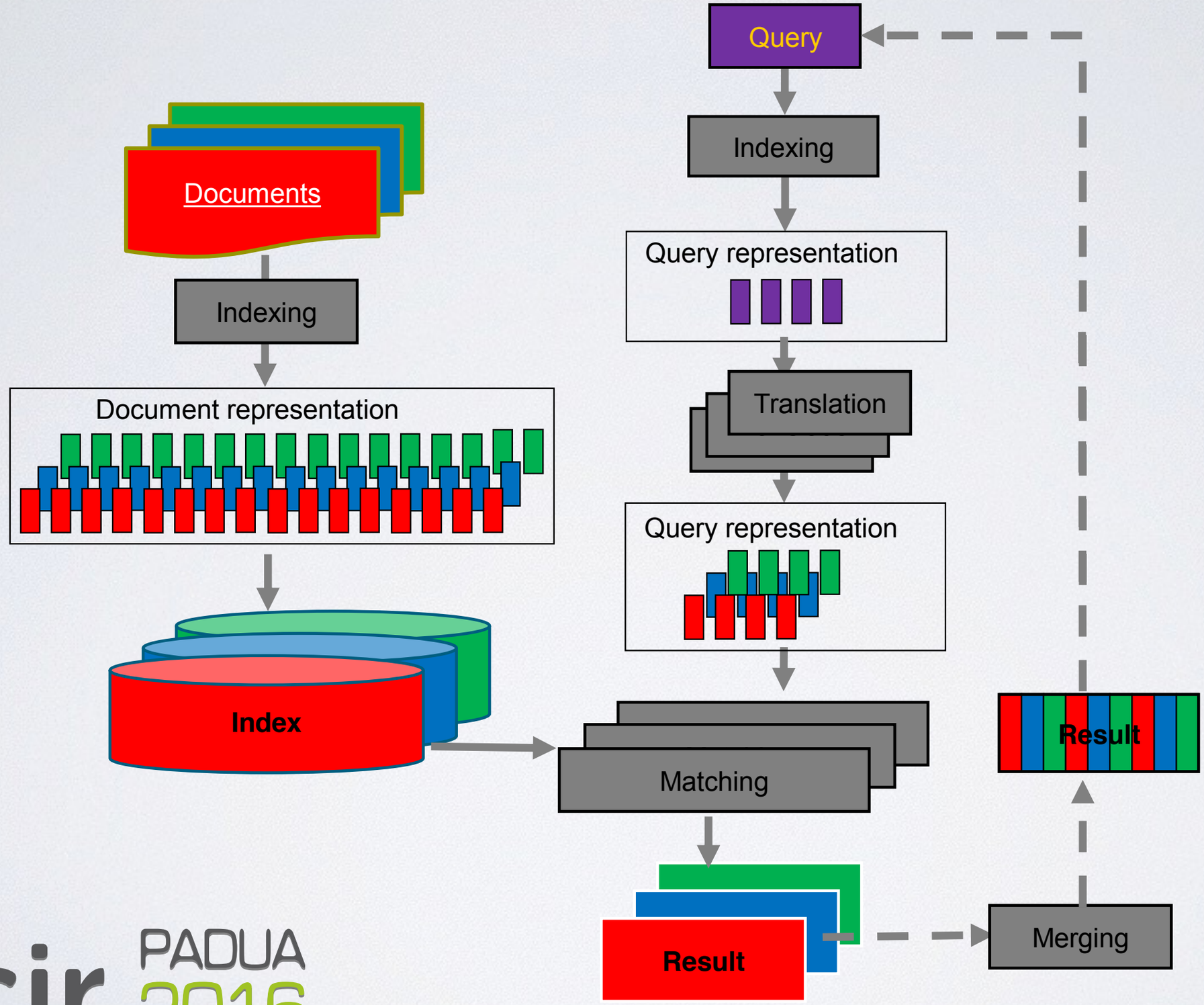
Possible CLIR Flow: Query Translation

[Figure taken from M. Braschler, *Multilingual Information Retrieval and Cross-Language Information Retrieval*, TrebleCLEF Summer School 2009, Italy]



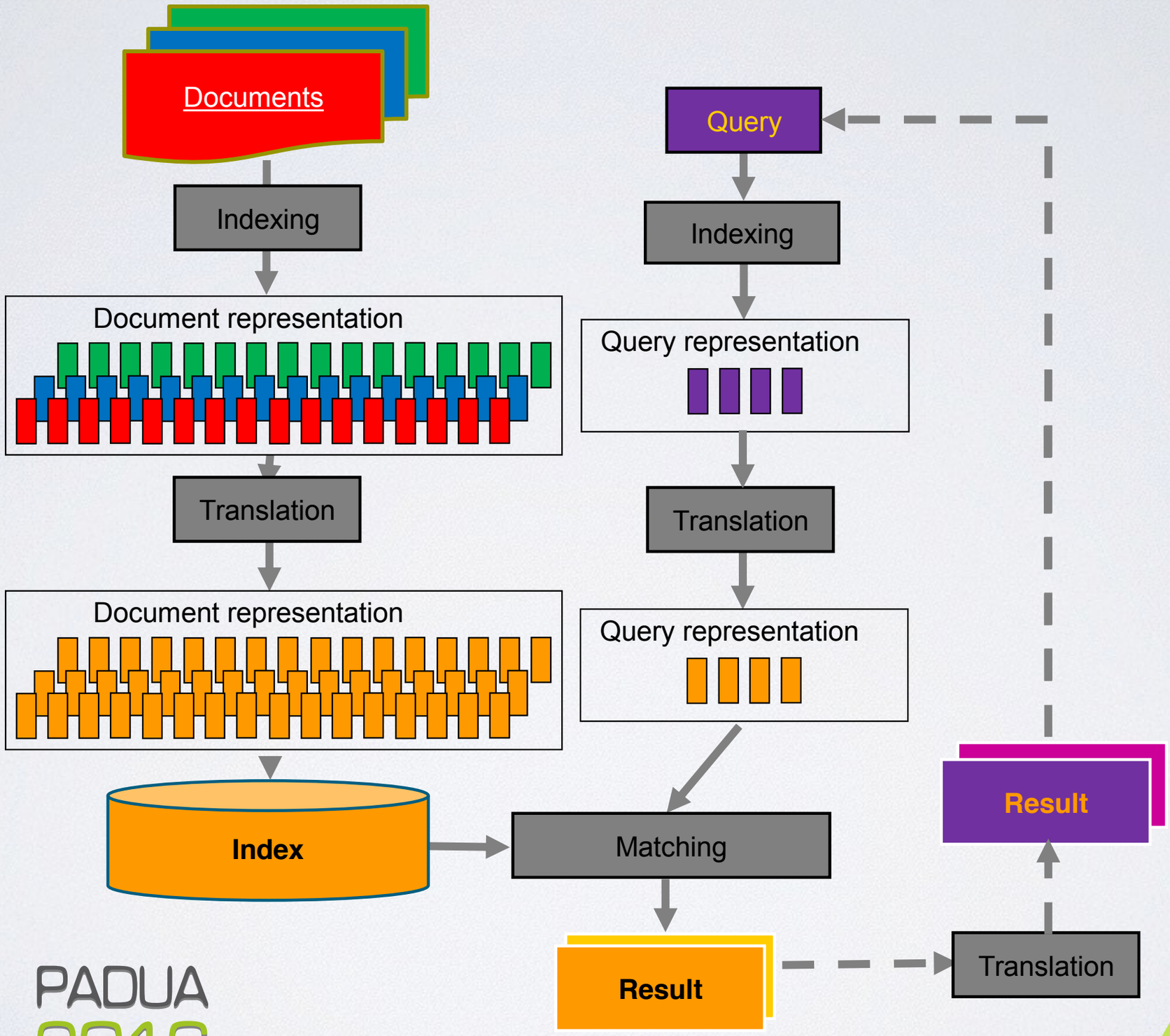
Possible MLIR Flow: Query Translation

[Figure taken from M. Braschler, *Multilingual Information Retrieval and Cross-Language Information Retrieval*, TrebleCLEF Summer School 2009, Italy]



Possible MLIR Flow: Query and Document Translation

[Figure taken from M. Braschler, *Multilingual Information Retrieval and Cross-Language Information Retrieval*, TrebleCLEF Summer School 2009, Italy]

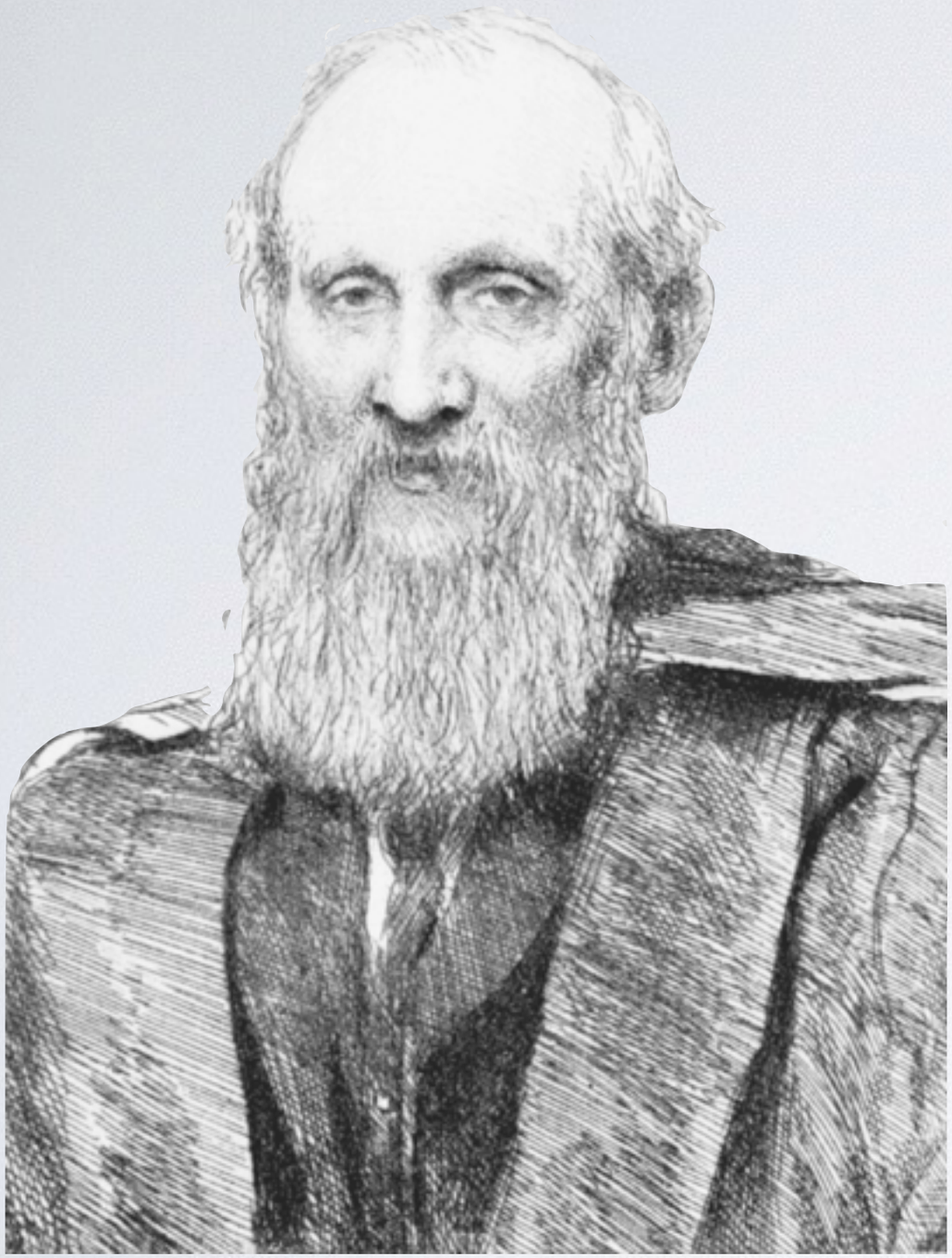


MLIA: Issues to Consider

- Document representation
- Language identification
  - cross-scripting
- Segmentation
  - compound words
  - N-grams
  - discourse
- Stop lists
  - varying length
- Normalization
  - diacritics
  - spellings
  - ...
- Stemming
  - rule-based
  - statistical / N-grams
- Enrichment
  - named entity recognition
  - thesaurus expansion
  - authorship
  - ...
- Translation
  - ambiguity
  - out-of-vocabulary terms
  - bag-of-words
  - ...

How Well

Why Evaluation?



"To measure is to know"

"If you cannot measure it, you cannot improve it"

Lord William Thompson,
first Baron Kelvin (1824-1907)

How Does Experimental Evaluation Work?

- Cranfield Paradigm
  - Dates back to mid 1960s
- Makes use of **experimental collections**
  - documents
  - topics
  - relevance judgments
- Ensures **comparability** and **repeatability** of the experiments

Large-scale Evaluation Campaigns

- Evaluation activities are conducted in large international fora
  - **TREC** (Text REtrieval Conference), USA, since 1992
  - **NTCIR** (NII Test Collection for IR Systems), Japan, since 1999
  - **CLEF** (Conference and Labs of the Evaluation Forum), Europe, since 2000
  - **FIRE** (Forum for Information Retrieval Evaluation), India, since 2008
- Share a common methodology, the **Cranfield Paradigm**

Adhoc-ish CLEF over the years

- We are interested in carrying out a longitudinal study on the Adhoc-ish CLEF tasks in order to gain an understanding on how performances behaved over the time

- It is difficult (or even impossible) to conduct this kind of studies because experimental collections are not directly comparable

Standardization

- It directly adjusts topic scores by the observed mean score and standard deviation for that topic in a sample of the systems.

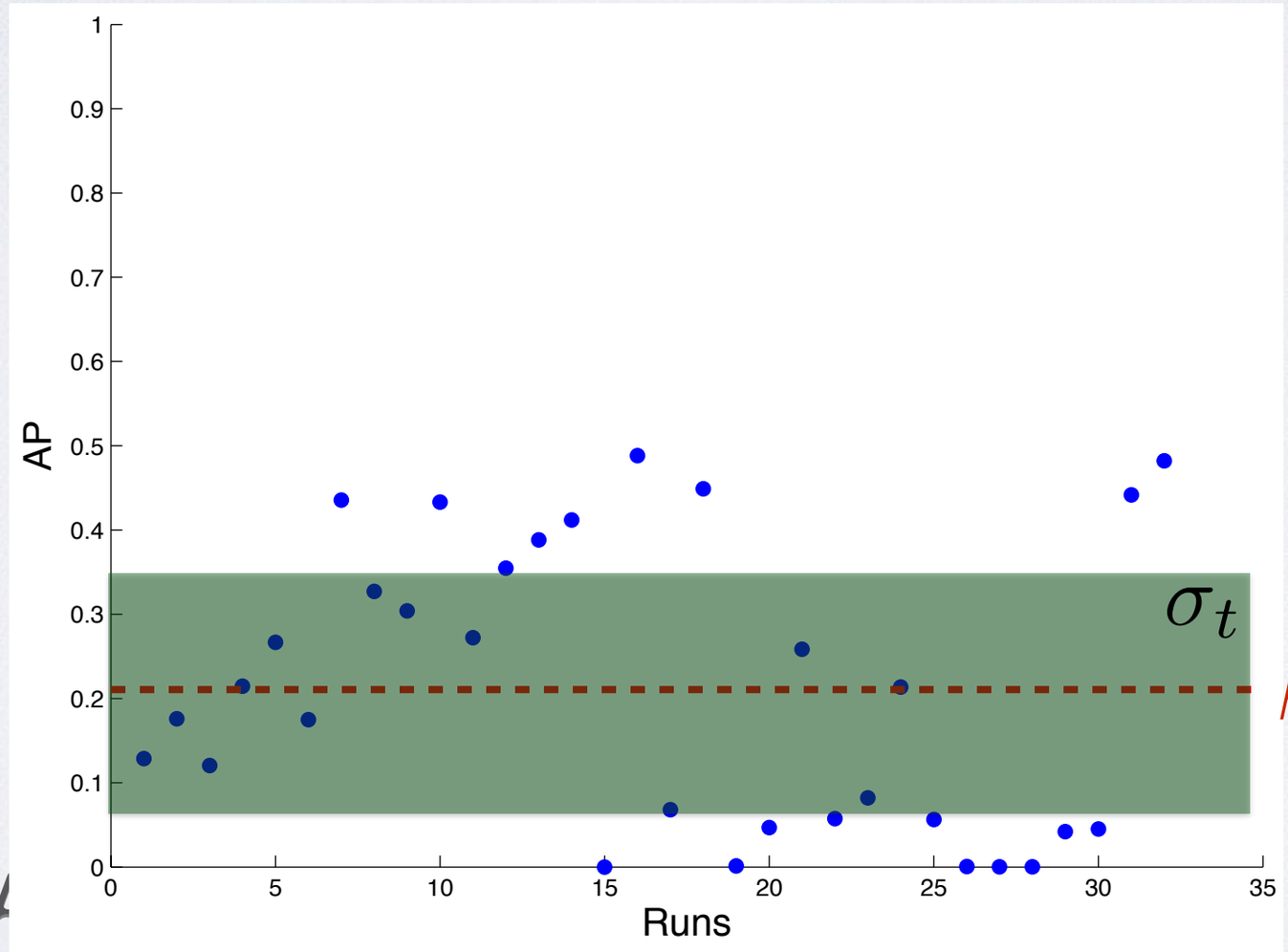
- The difficulty of a query is estimated from the scores achieved by systems, and parameters derived from these estimates are then used to normalize the systems scores

– W. Webber, A. Moffat, and J. Zobel. 2008
Score standardization for inter-collection comparison of retrieval systems.
In SIGIR 2008. ACM Press, 51-58.

How standardization works

- For every run in a collection, we have a measure m for each topic t with mean μt and standard deviation σt

These are AP values of all the runs for topic 301 of CLEF Ad-Hoc bilingual English 2006

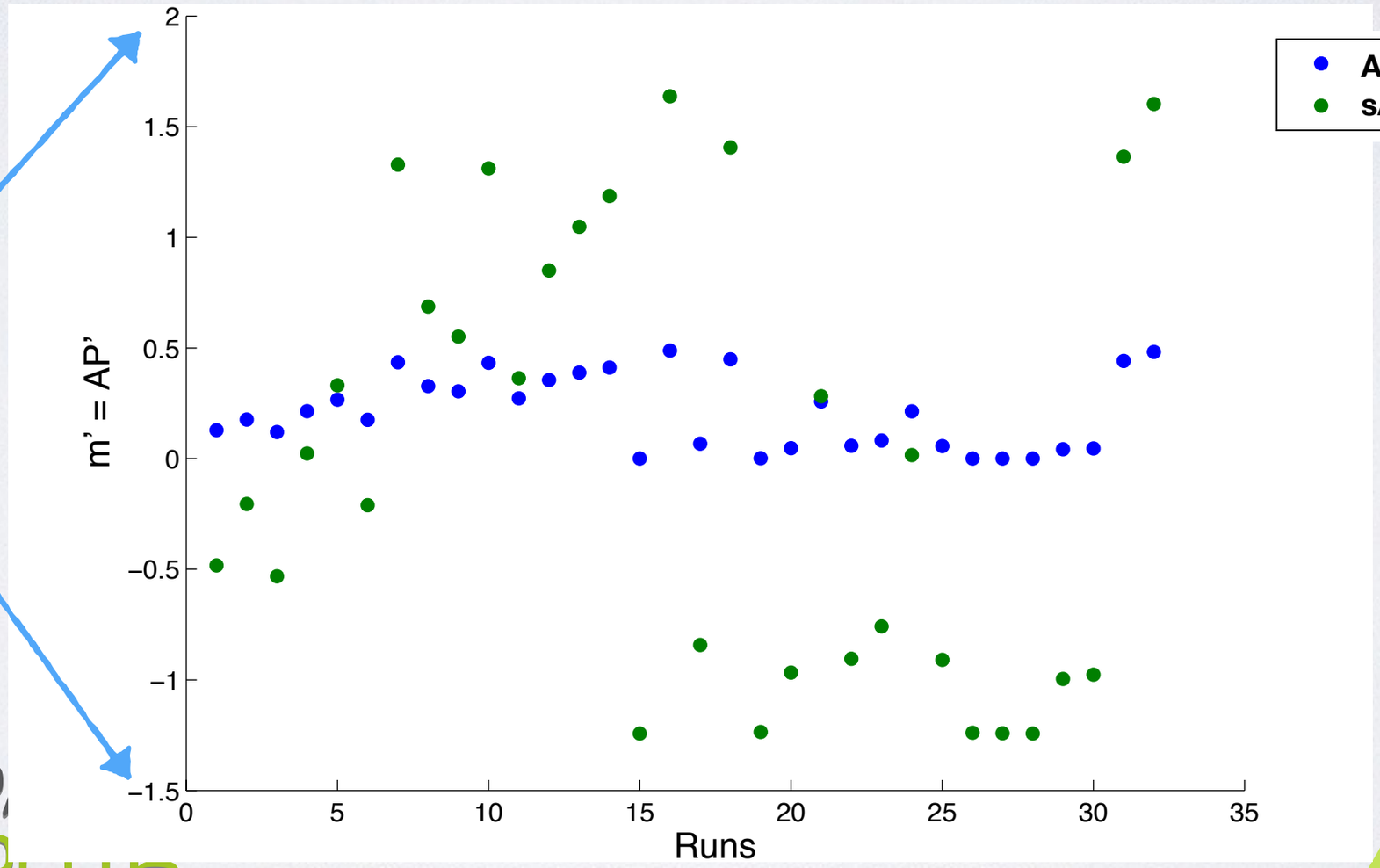


How standardization works

- For each topic in a collection we can calculate the *z-scores* of a measure m as

$$m' = \frac{m - \mu_t}{\sigma_t}$$

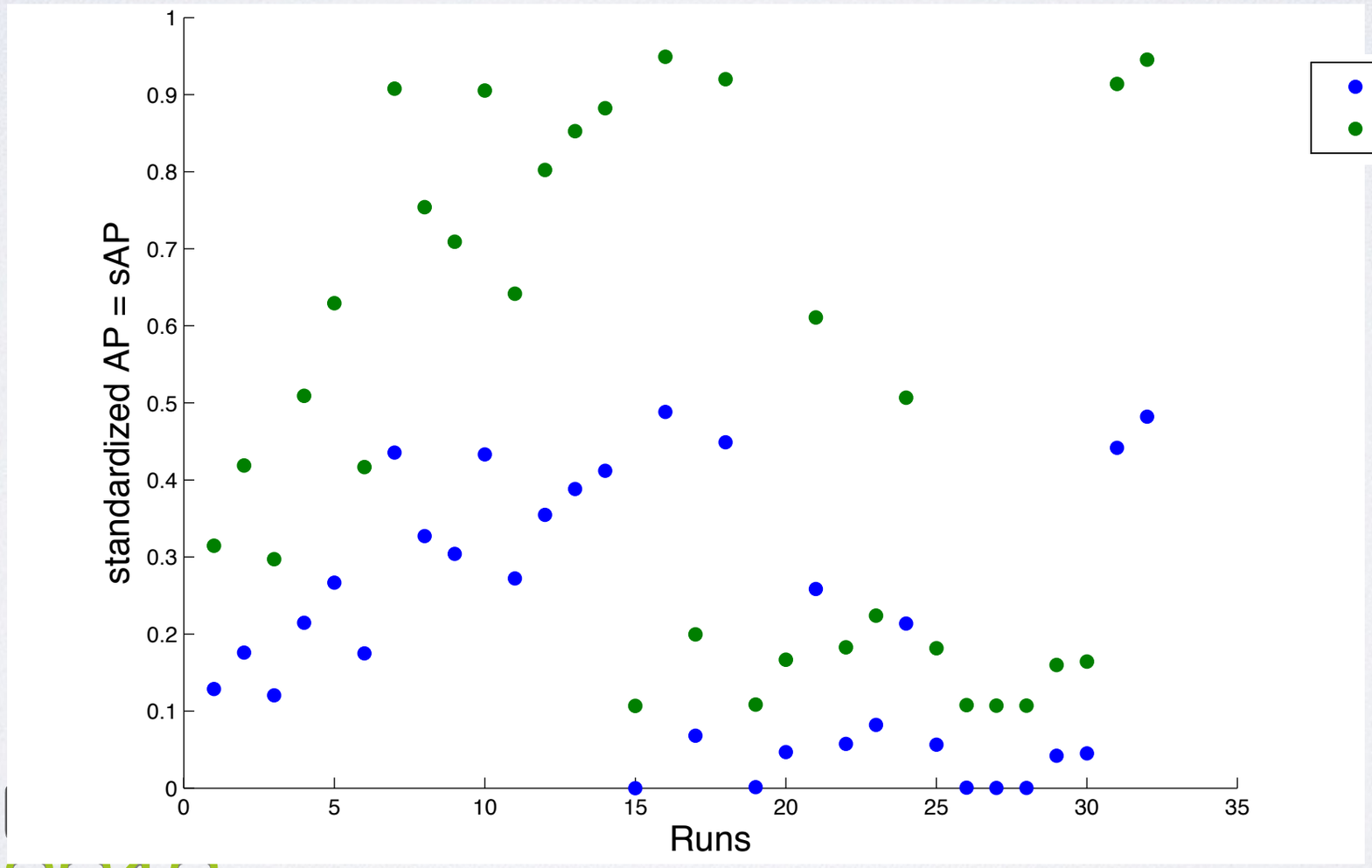
Not in the range [0, 1]



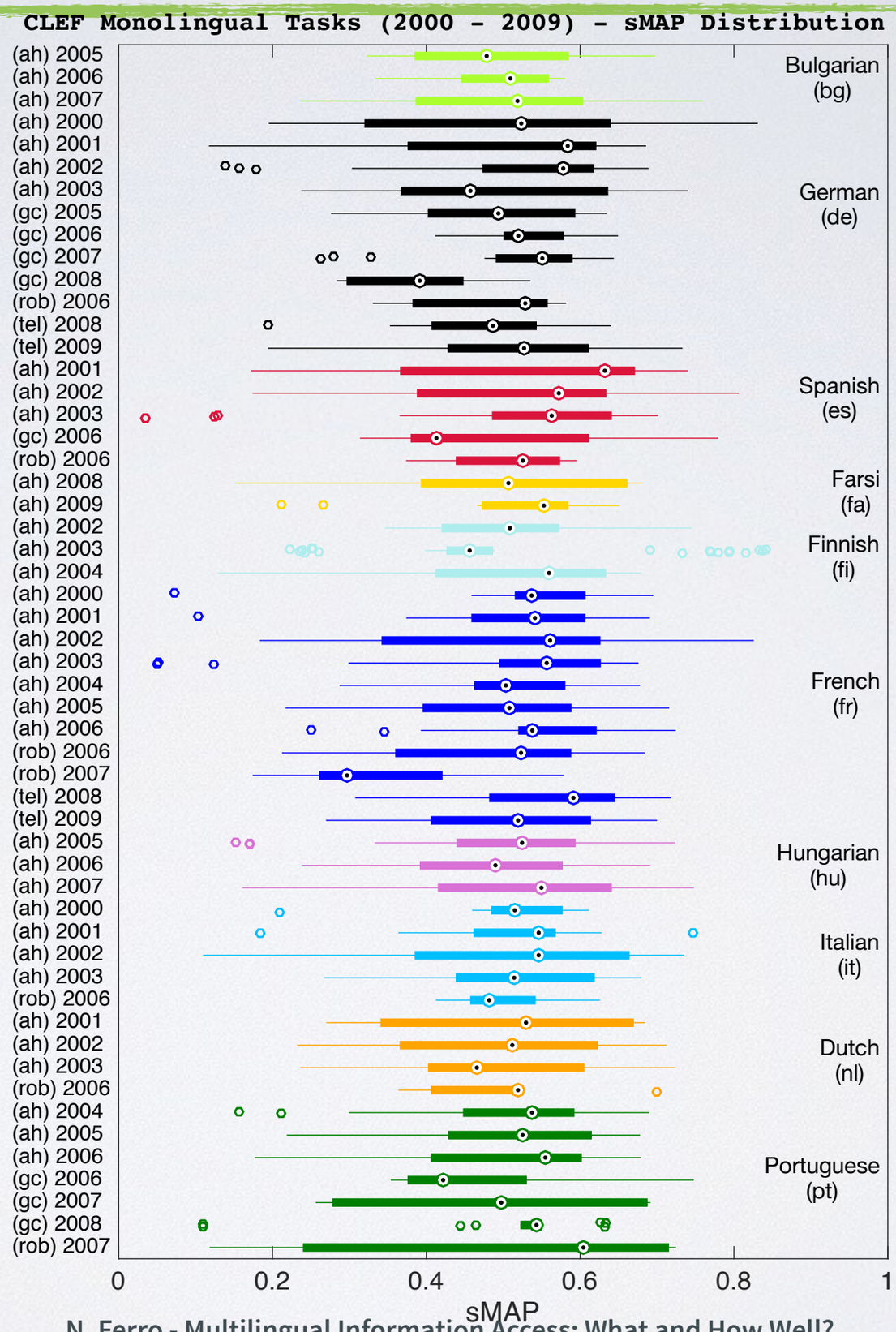
How standardization works

- Normalization in the [0,1] range by using the cumulative density function:

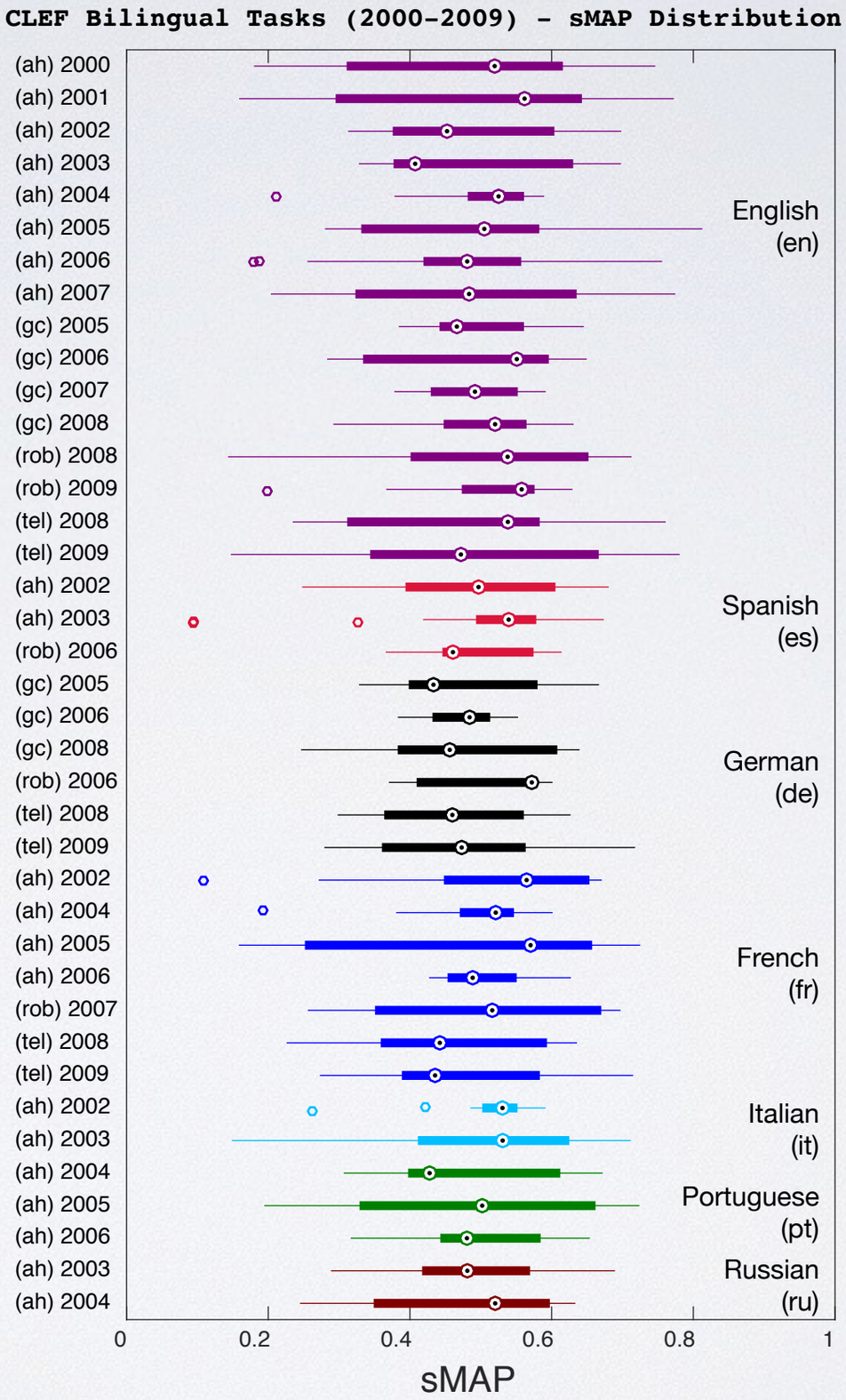
$$F_X(m') = \int_{-\infty}^{m'} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$



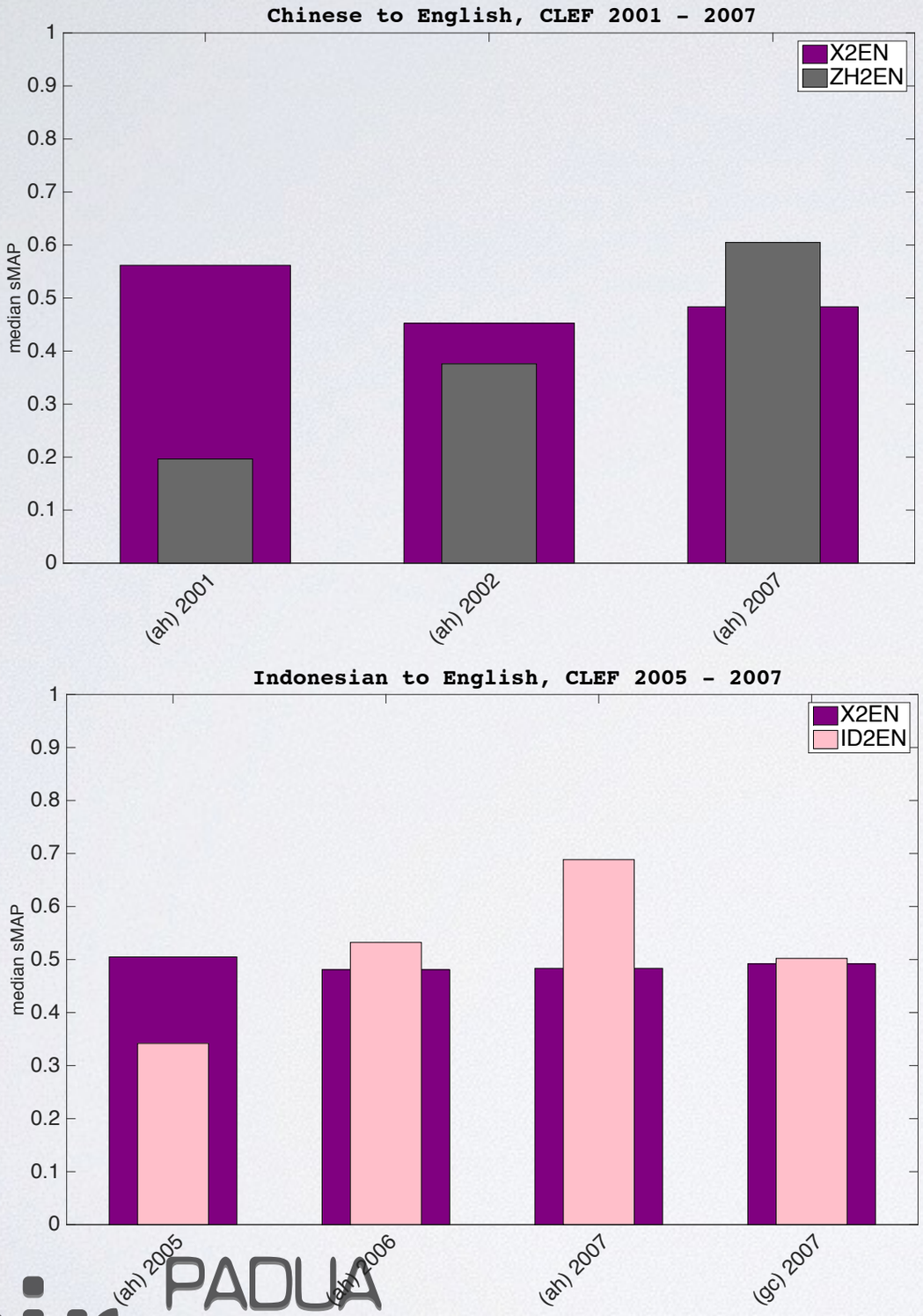
How do monolingual systems behave over the years?

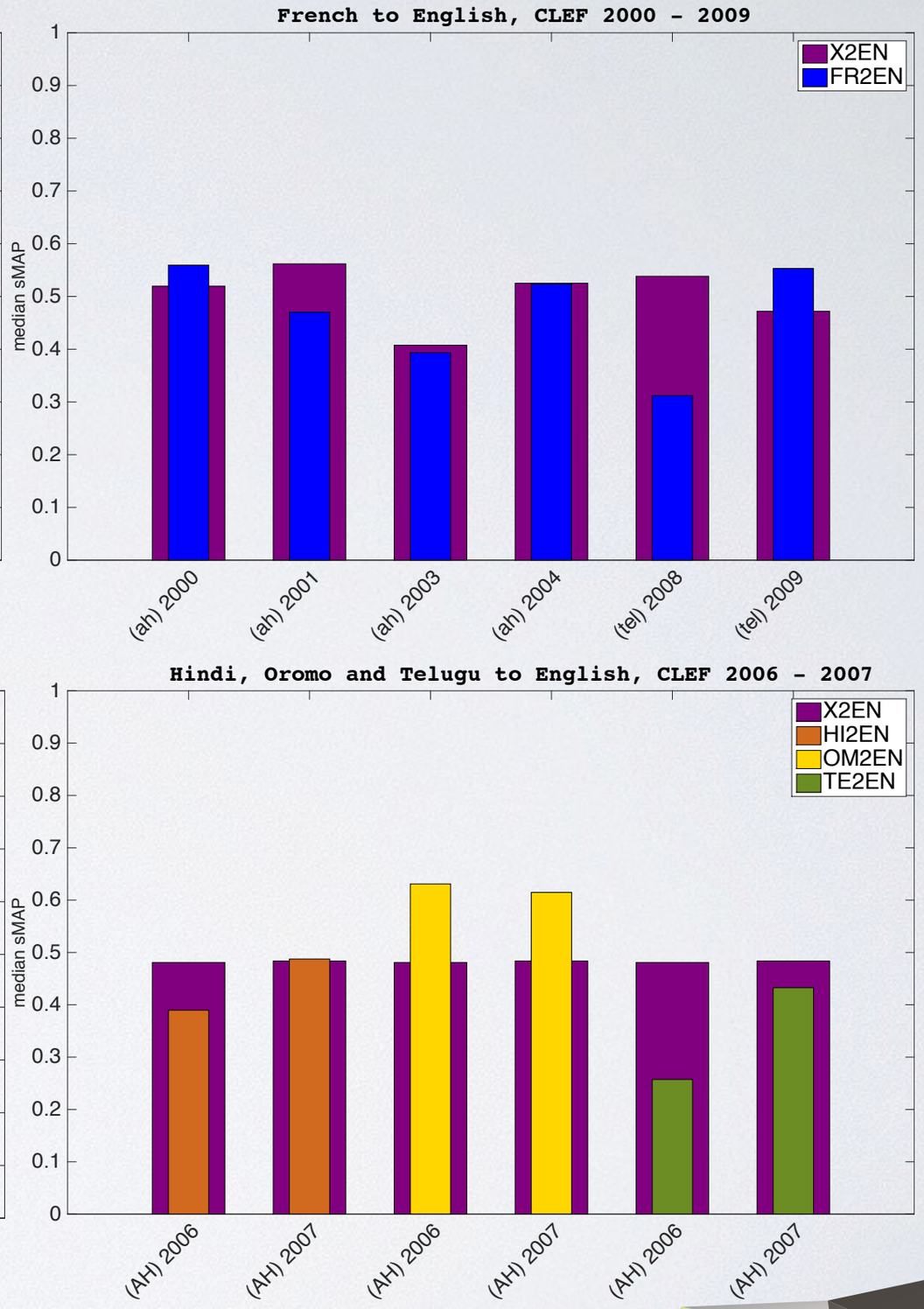


How do bilingual systems behave over the years?

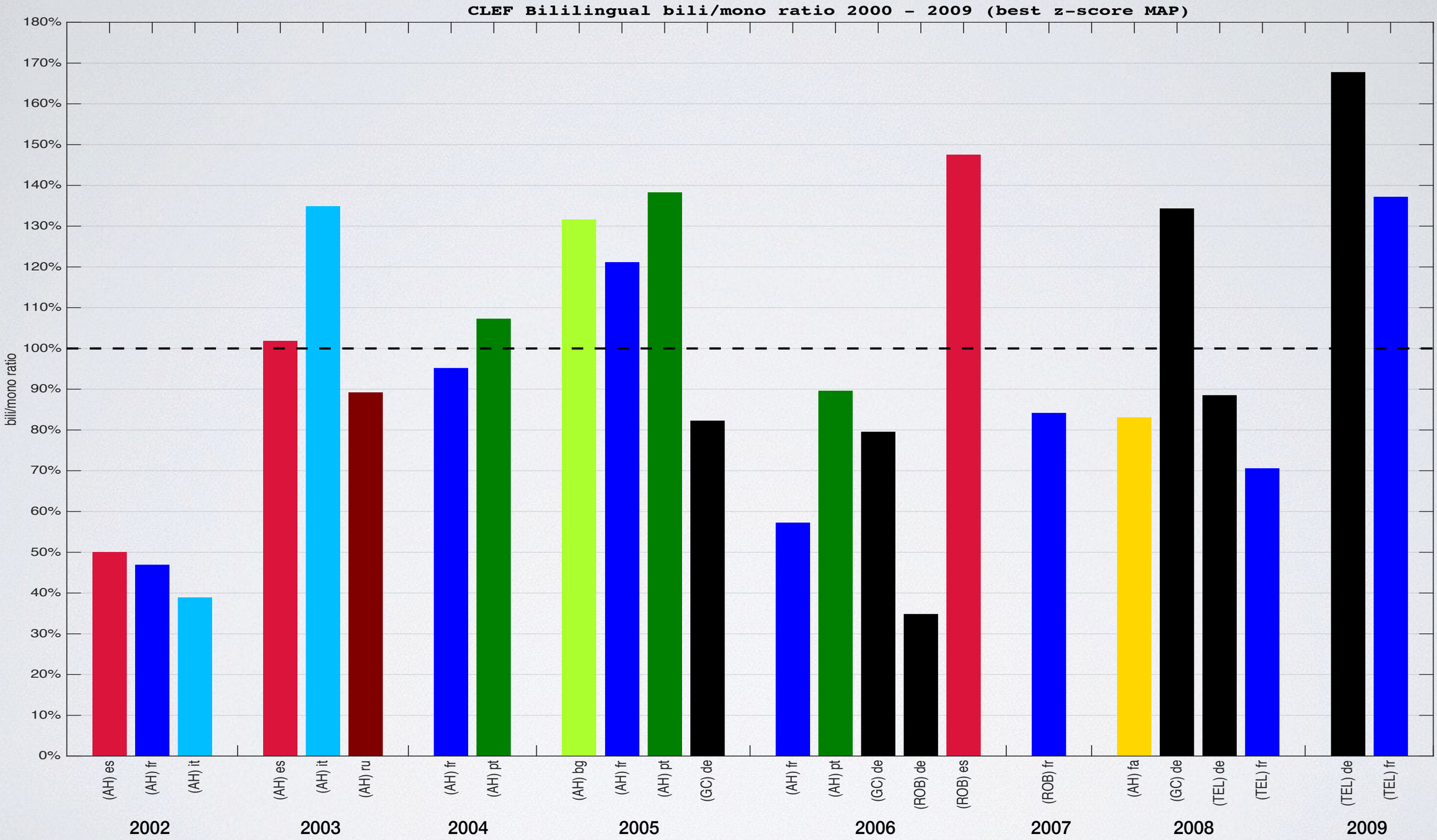


Bilingual Breakdown by Source Language

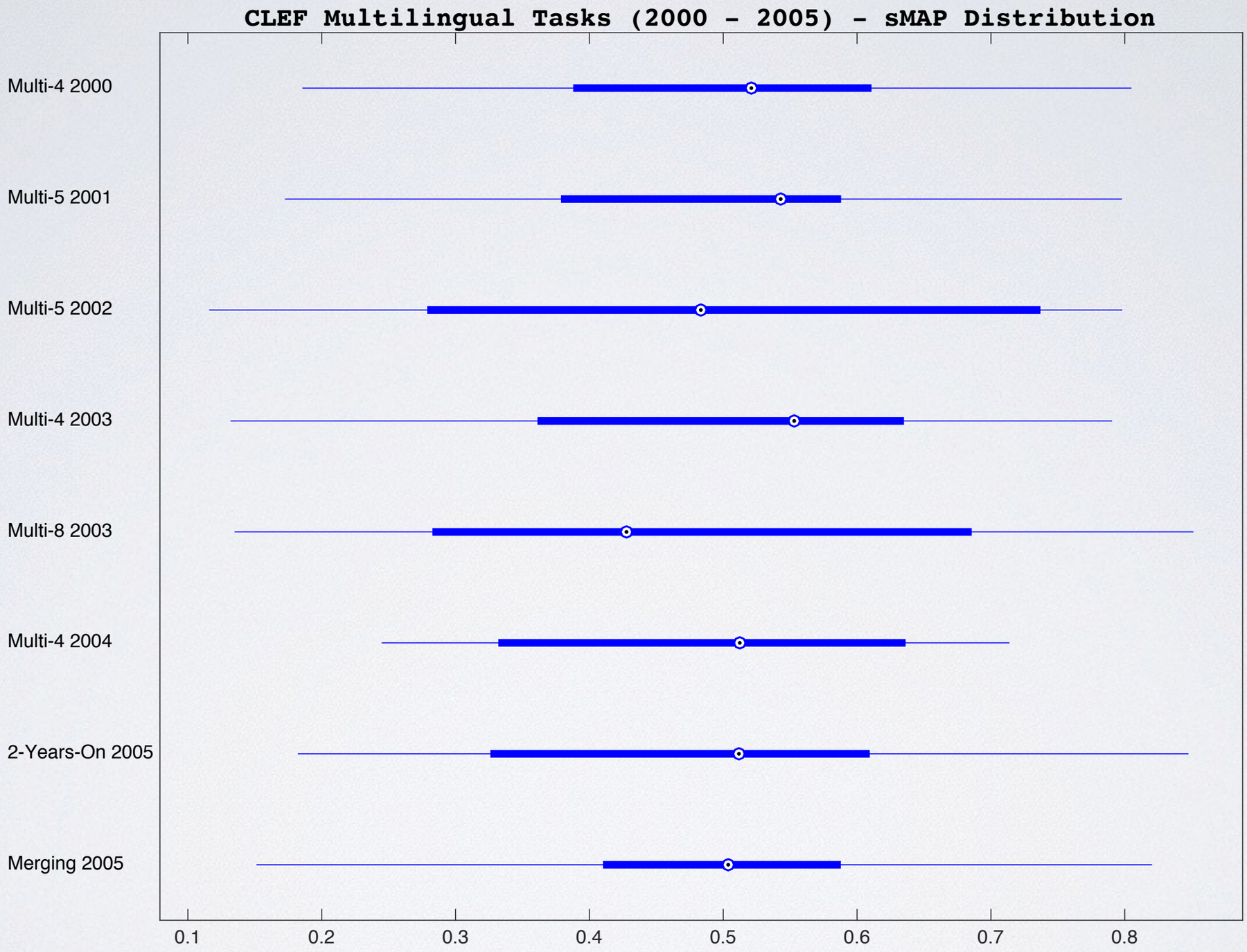




Bilingual to Monolingual Comparison



How do multilingual systems behave over the years?



Conclusions

- Multilingual information access is still an open challenge
  - user tasks are rapidly changing and evolving and new unprecedented needs emerge
- Multimodality and multimediality are more and more integral parts and key concerns for multilingual information access
  - the problem is no more only crossing the language barriers
- Evaluation has shown positive trends over the years
  - it is hard to keep consistent evaluation tasks able to represent the real challenges


ANY
QUESTIONS
?