

CLEF 2000 – 2014: Lessons Learnt from Ad Hoc Retrieval

Nicola Ferro and Gianmaria Silvello

Information Management Systems Research Group
Department of Information Engineering
University of Padua

`{ferro, silvello}@dei.unipd.it`

Outline

- Motivations
- Standardization
- Analysis of CLEF Ad-Hoc Campaigns
- Wrapping up

Motivations

Why Evaluation?



“To measure is to know.”

“If you cannot measure it, you cannot improve it.”

Lord William Thomson, 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
- Experimental collections and the results of the evaluation are **shared** and **re-usable**
- Ensures **comparability** and **reproducibility** of the experiments

Experimental Evaluation: Economic Impact

- The **TREC 2010 Economic Impact** study estimated in about **30 M\$** the overall **investment in TREC** by NIST
 - probably much much more if we had a means to estimate also the investment by participants in TREC
- They are the **pillars** for all the subsequent **scientific research** and **technology development**
 - TREC estimated the **return on investment** in the range of **3\$-5\$** for each invested dollar



The European contest: CLEF

The **CLEF Initiative** is a self-organized body whose main mission is to promote research, innovation, and development of information access systems with an emphasis on **multilingual and multimodal information with various levels of structure.**

- **multilingual** and multimodal system testing, tuning and evaluation;
- investigation of the use of unstructured, semi-structured, **highly-structured**, and semantically enriched data in information access;
- an open forum for **research groups** and **industries** to discuss and share competencies;
- discussion of results, comparison of approaches, exchange of ideas, and **transfer of knowledge.**

Our Goal

Our goal is to carry out a longitudinal study the **evolution of CLEF** in the last 15 years in order to understand its **impact** on monolingual, bilingual, and multilingual search

Motivations

- There have been very few systematic longitudinal studies about the impact of evaluation campaigns on the overall effectiveness of IR systems
 - e.g. SMART system tested on TREC 1 to 8

- C. Buckley, 2005.
The SMART project at TREC,
TREC - Experiment and Evaluation in Information Retrieval. MIT Press

- D. Harman, 2011,
Information Retrieval Evaluation. Synthesis Lectures on Information Concepts,
Retrieval, and Services, Morgan & Claypool Publishers

Motivations

- It is not easy (or possible) to conduct that kind of study for CLEF, because we would need to:
 - Use different versions of one or more systems
 - Test them on many collections for a great number of tasks
- Today's systems increasingly rely on-line linguistic resources which continuously change over time, thus preventing comparable longitudinal studies even when using the same systems.

Motivations

- **RQ1**. Do performances of monolingual systems increase over the years? Are more recent systems better than older ones?
- **RQ2**. Do performances of bilingual systems increase over the years and what is the impact of source languages?
- **RQ3**. Do monolingual systems have better performances than bilingual and multilingual systems?

Standardization

Standardization

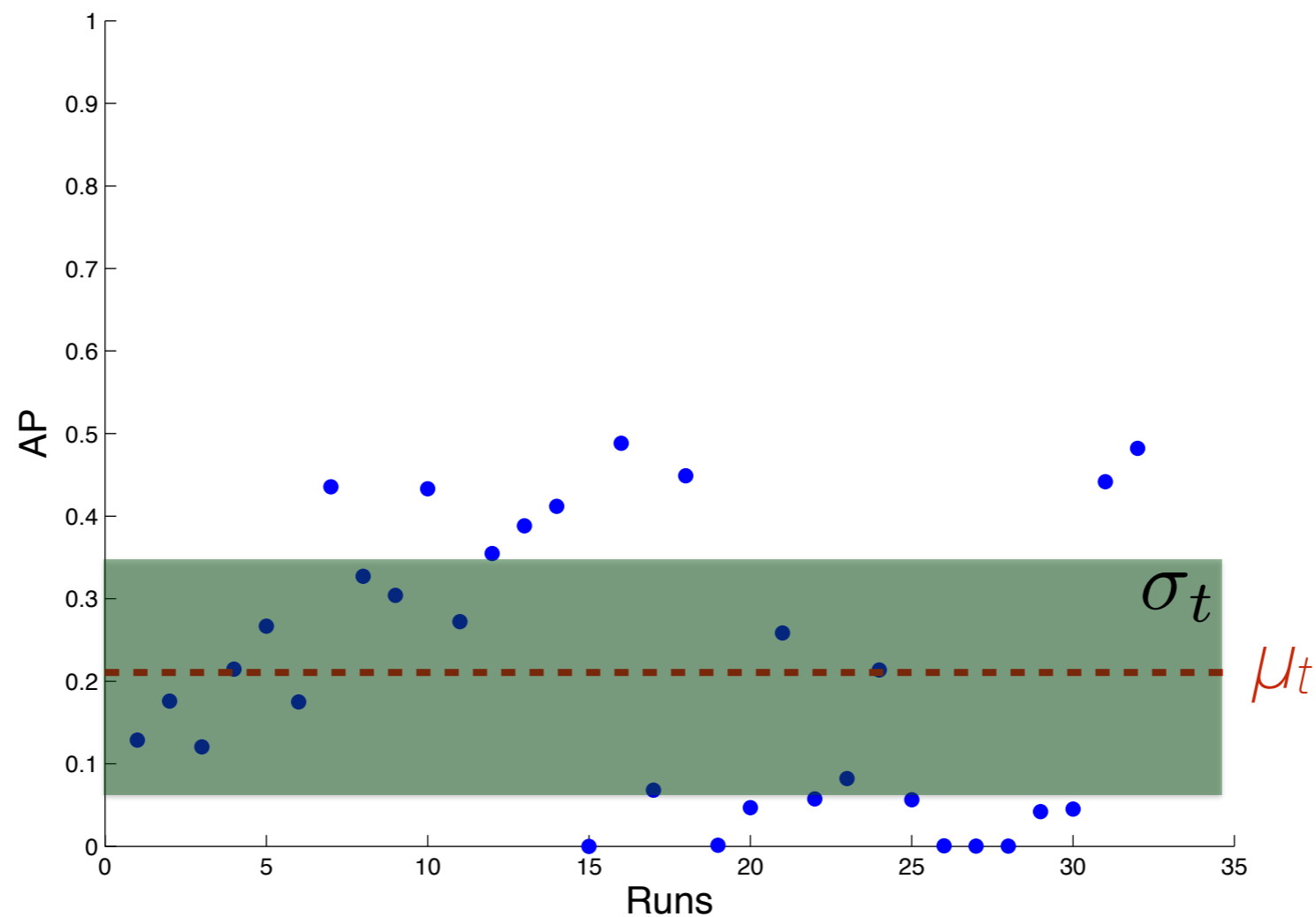
Inter-collection comparison between systems by limiting the effect of collections and by making system scores interpretable in themselves

– 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 (r) in a collection, we have a measure m for each topic t with mean μ_t and standard deviation σ_t

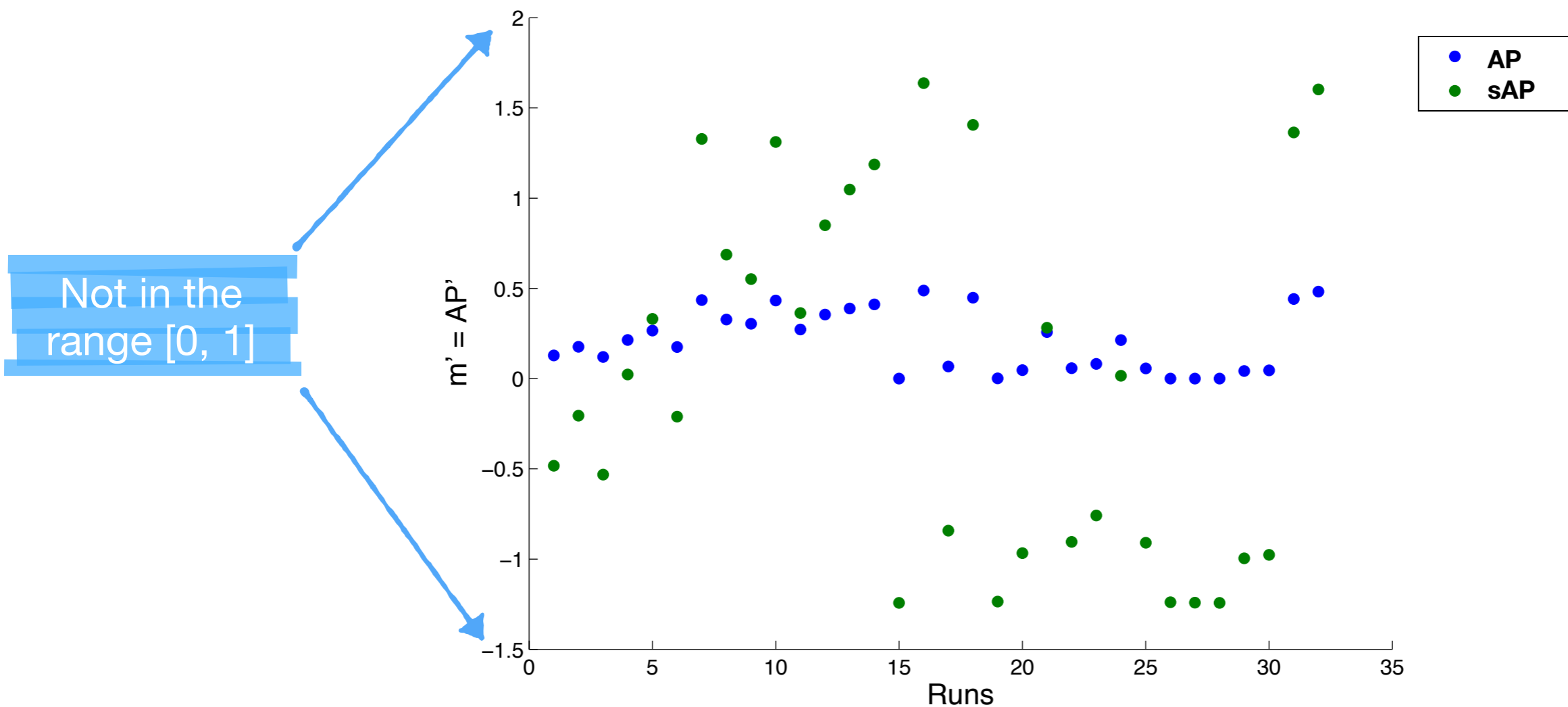
These are AP values of all the runs for topic 301 of CLEF Ad-Hoc bilingual English 2006



How standardization works

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

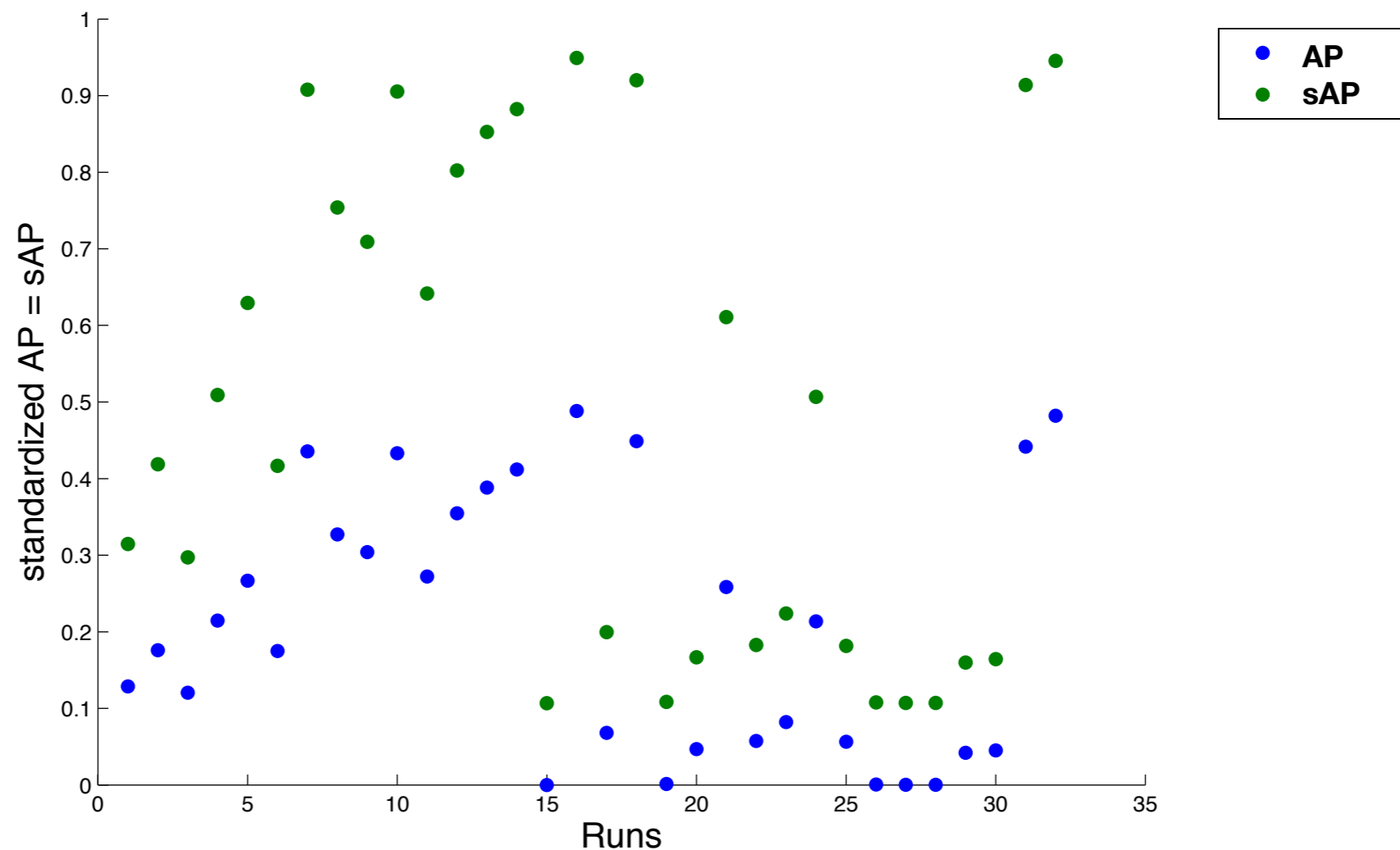
$$m' = \frac{m - \mu_t}{\sigma_t}$$



How standardization works

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

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

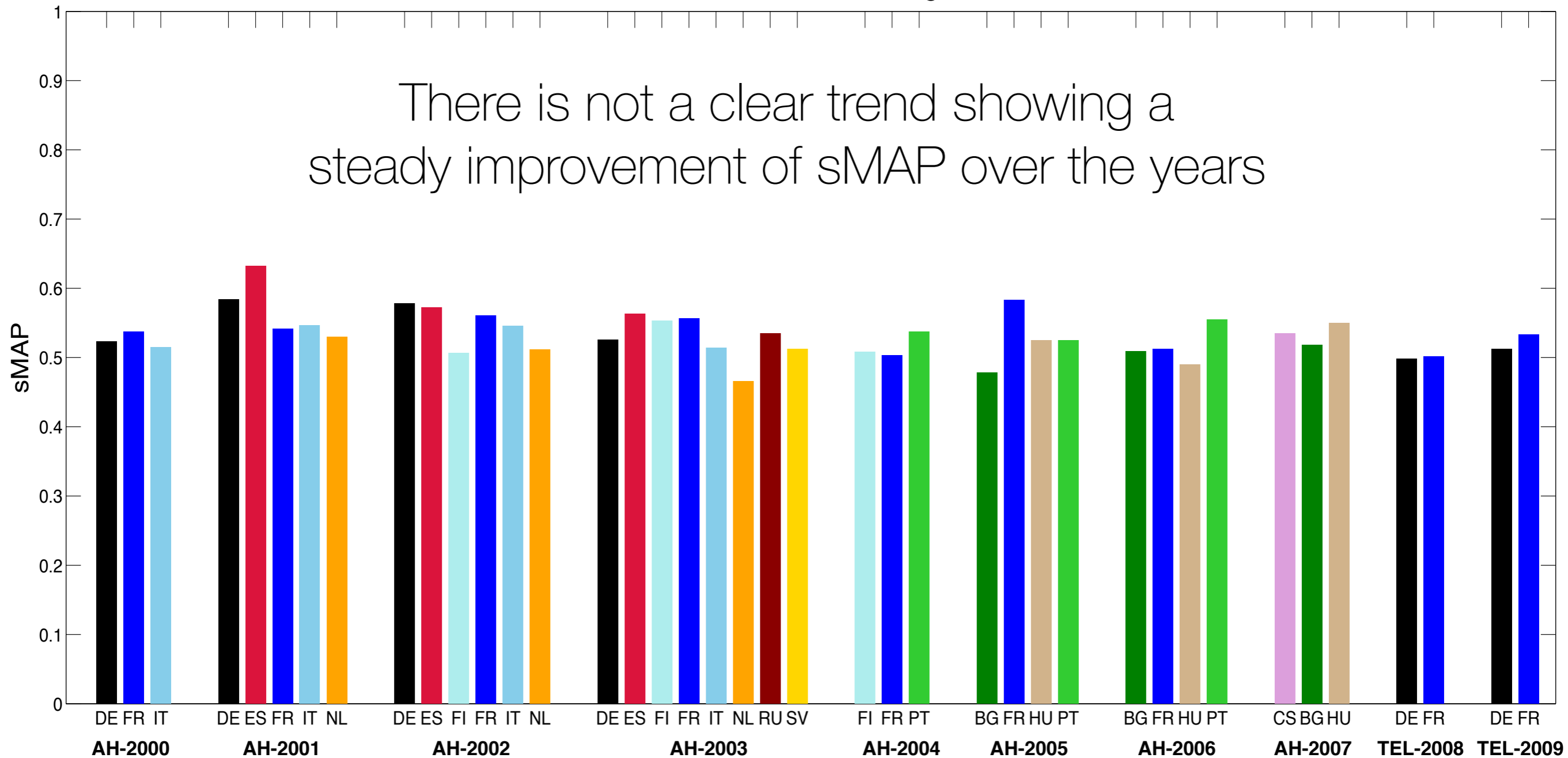


Analysis of CLEF Ad-Hoc Campaigns

RQ 1

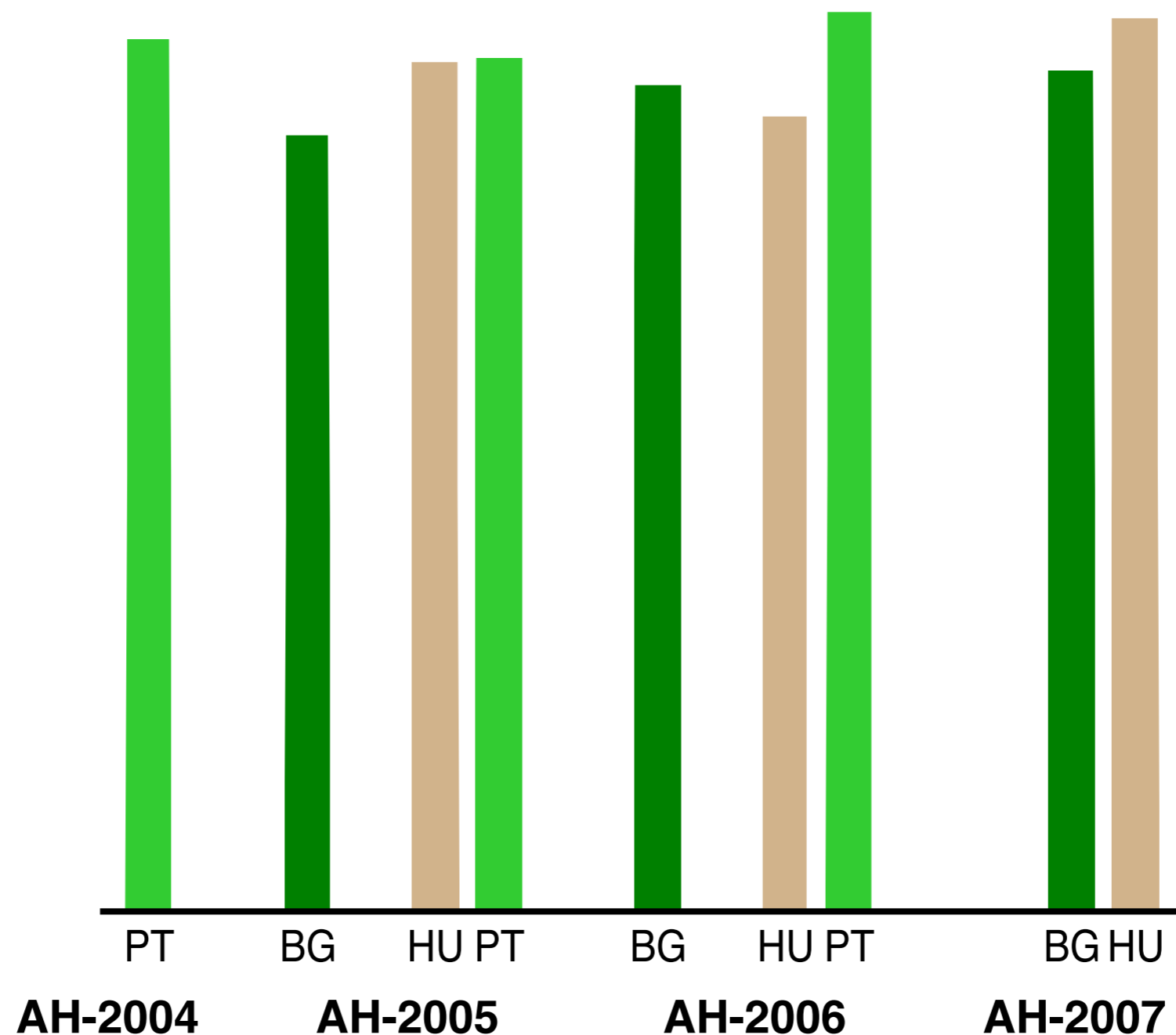
Do performances of monolingual systems increase over the years?
Are more recent systems better than older ones?

CLEF 2000 – 2009, Monolingual Tasks



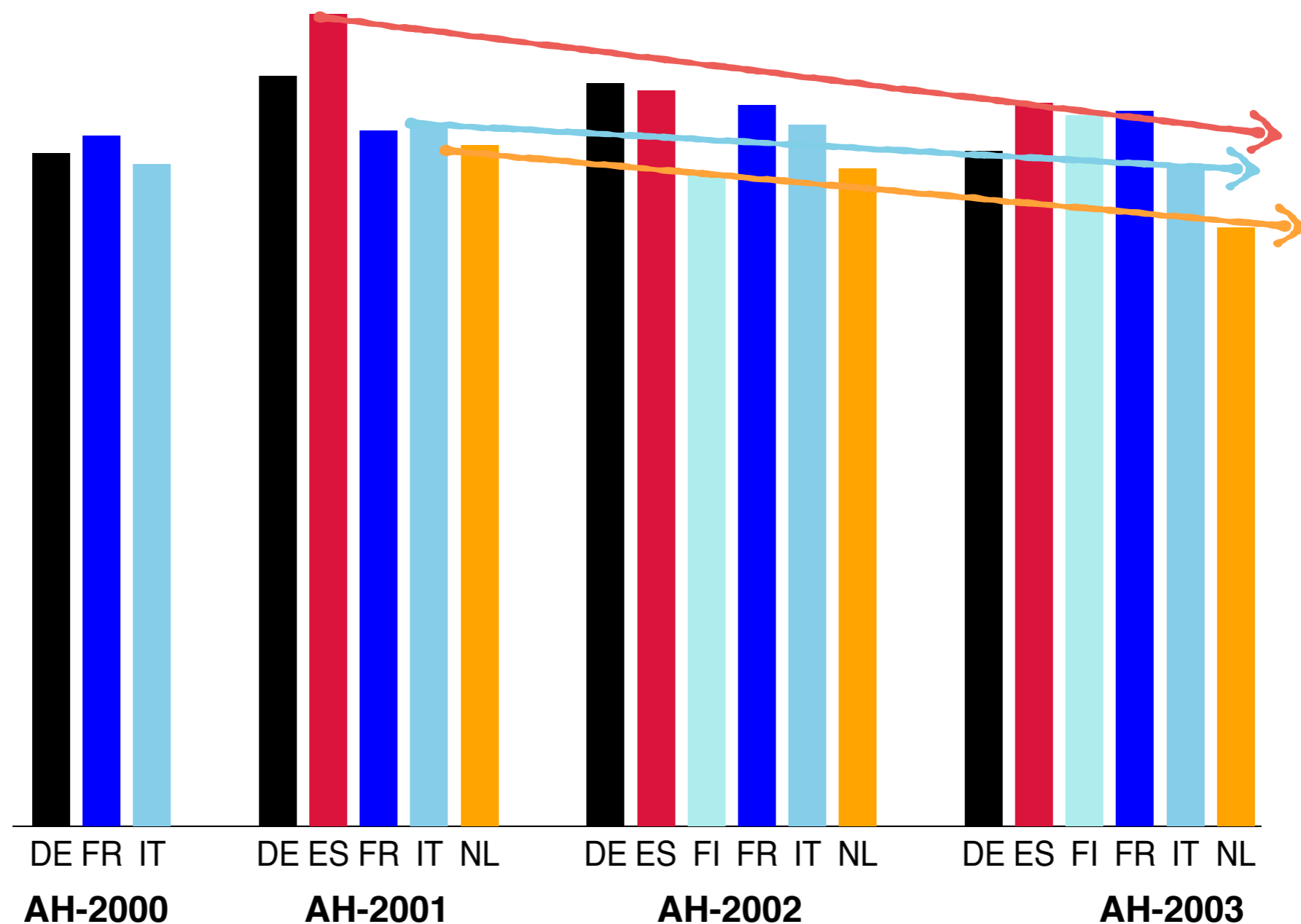
RQ 1: Do monolingual systems improve over the years?

The more evident improvement over the years is shown by the languages introduced in 2004 and 2005



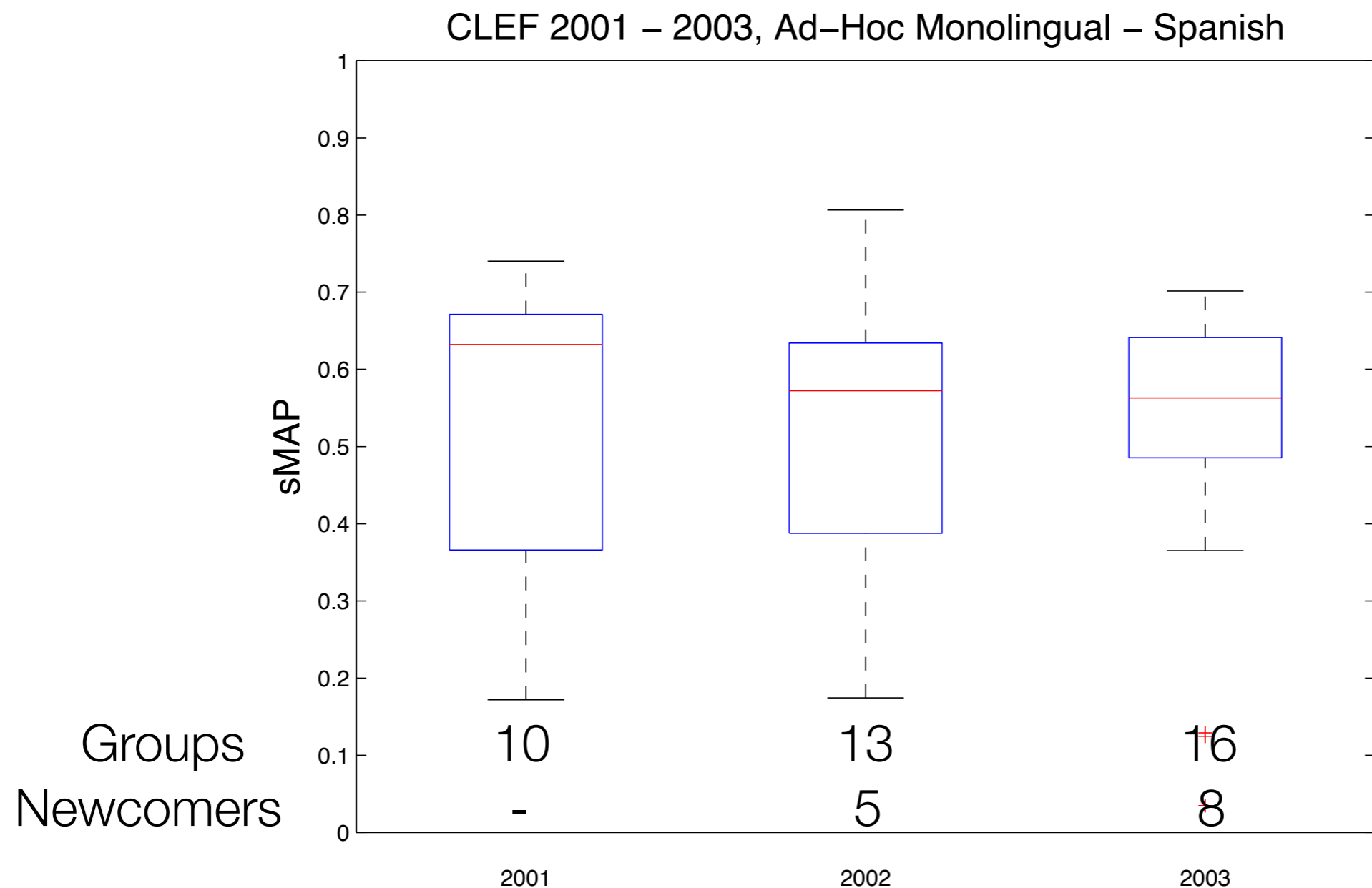
RQ 1: Do monolingual systems improve over the years?

The median sMAP of the monolingual tasks shows several examples of languages for which performances decrease over the years



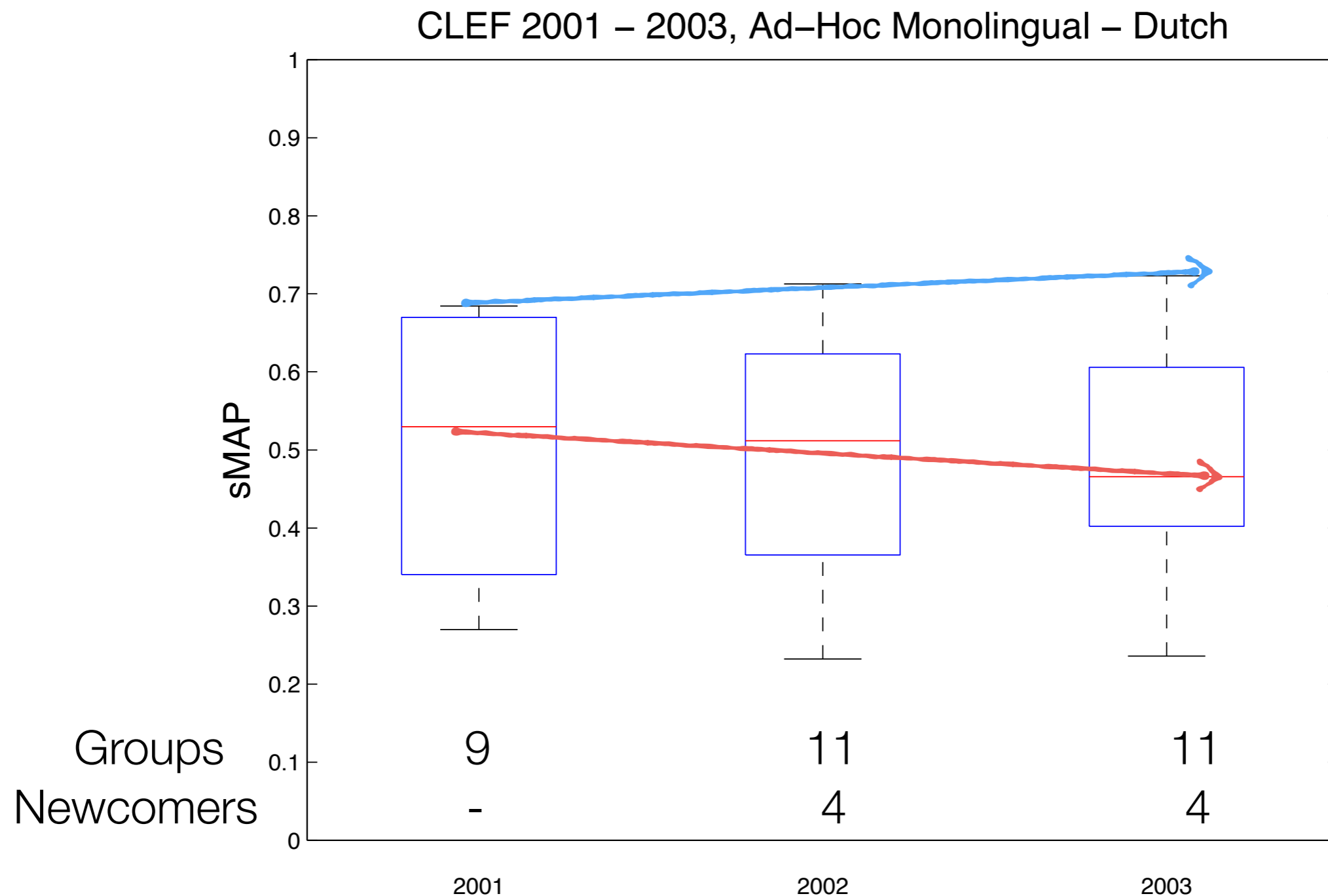
RQ 1: Do monolingual systems improve over the years?

The median sMAP is often influenced by the number of groups participating and by the number of newcomers with the positive effect of growing new local IR research communities



RQ 1: Do monolingual systems improve over the years?

... but looking at the best sMAP we find out that in several cases it grows through the years



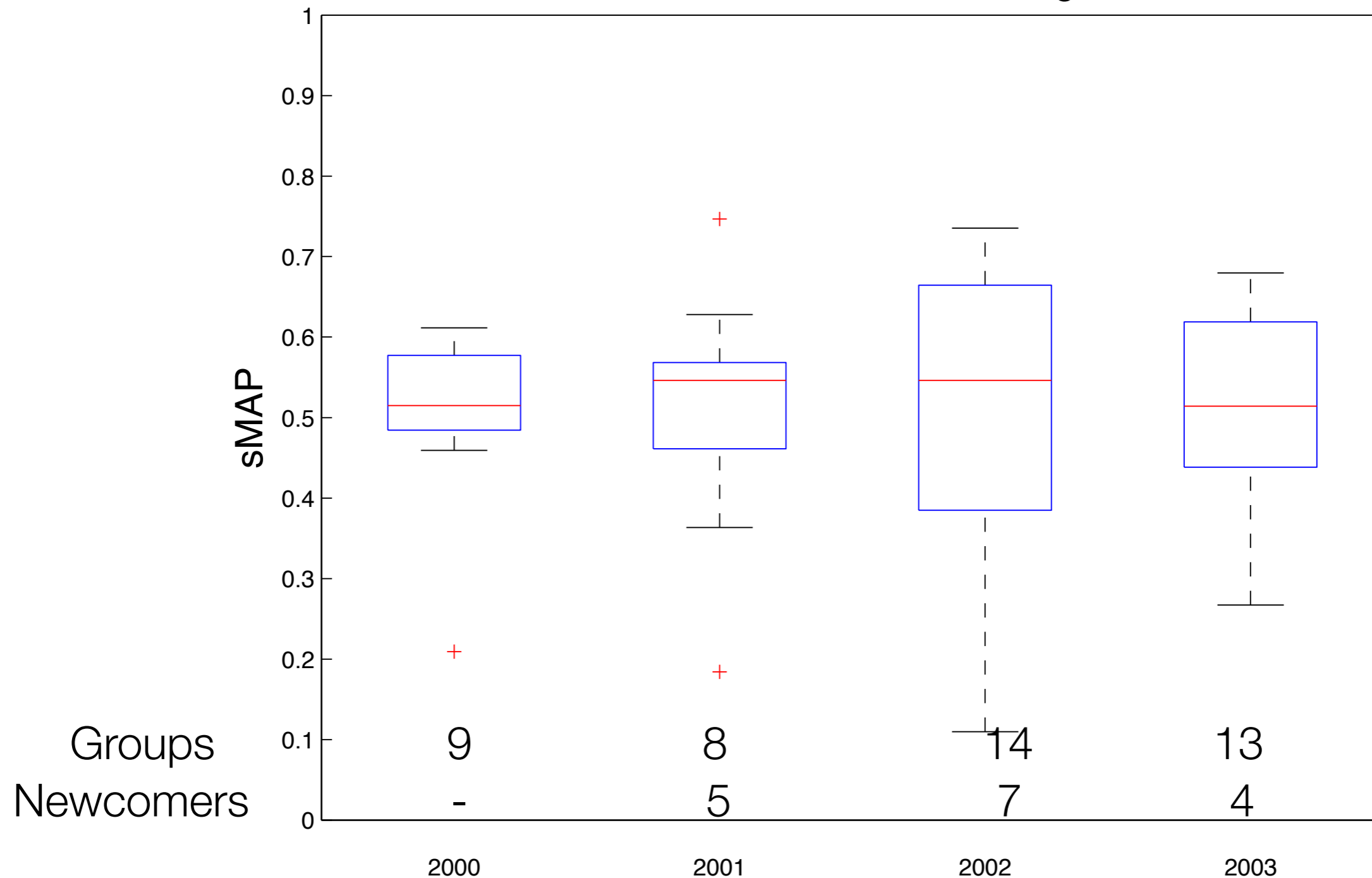
RQ 1: Do monolingual systems improve over the years?

But... Are general trends enough to explain phenomena?

RQ 1: Do monolingual systems improve over the years?

Let's take a look to the Italian case

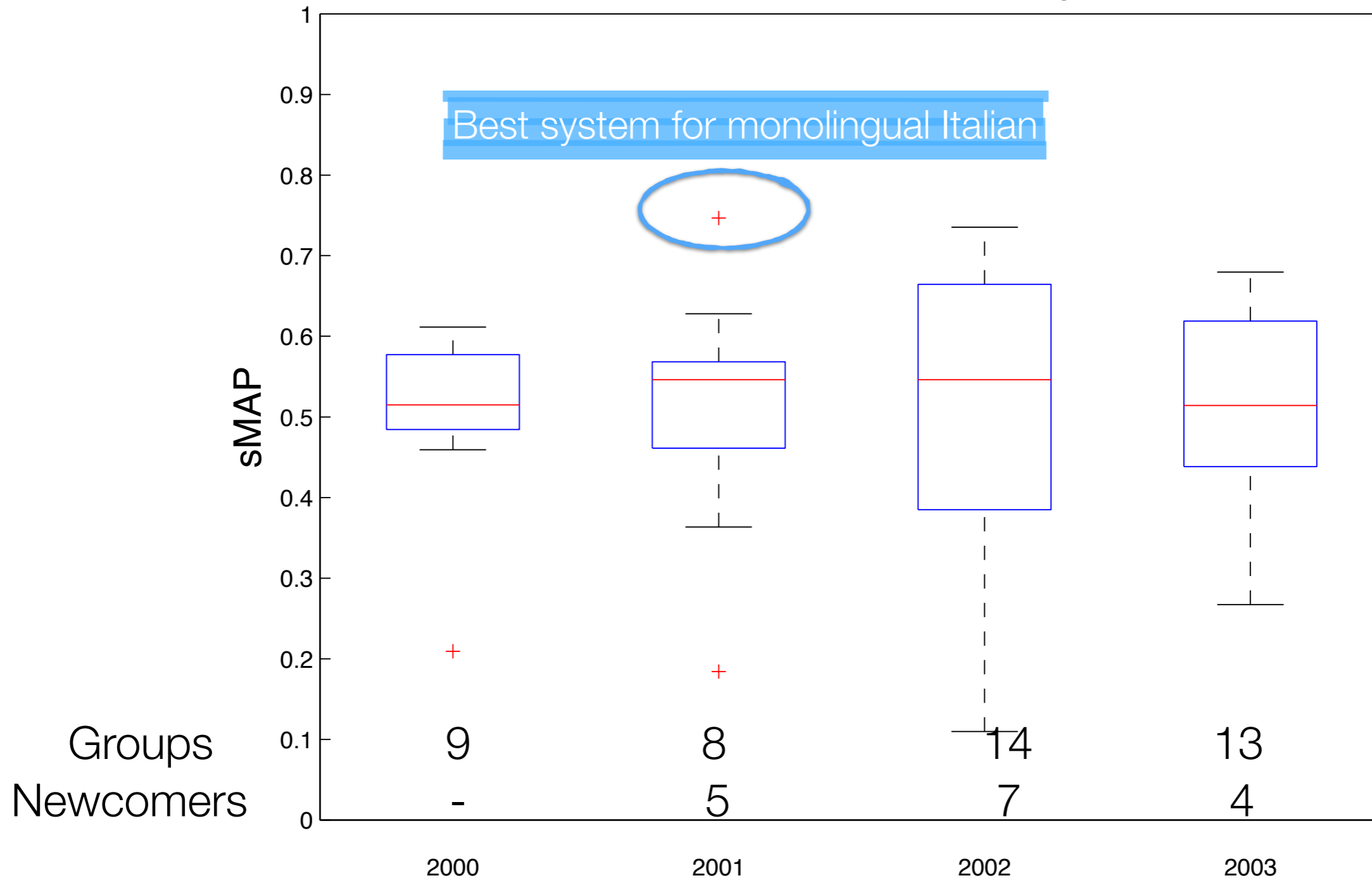
CLEF 2000 – 2003, Ad-Hoc Monolingual – Italian



RQ 1: Do monolingual systems improve over the years?

Let's take a look to the Italian case

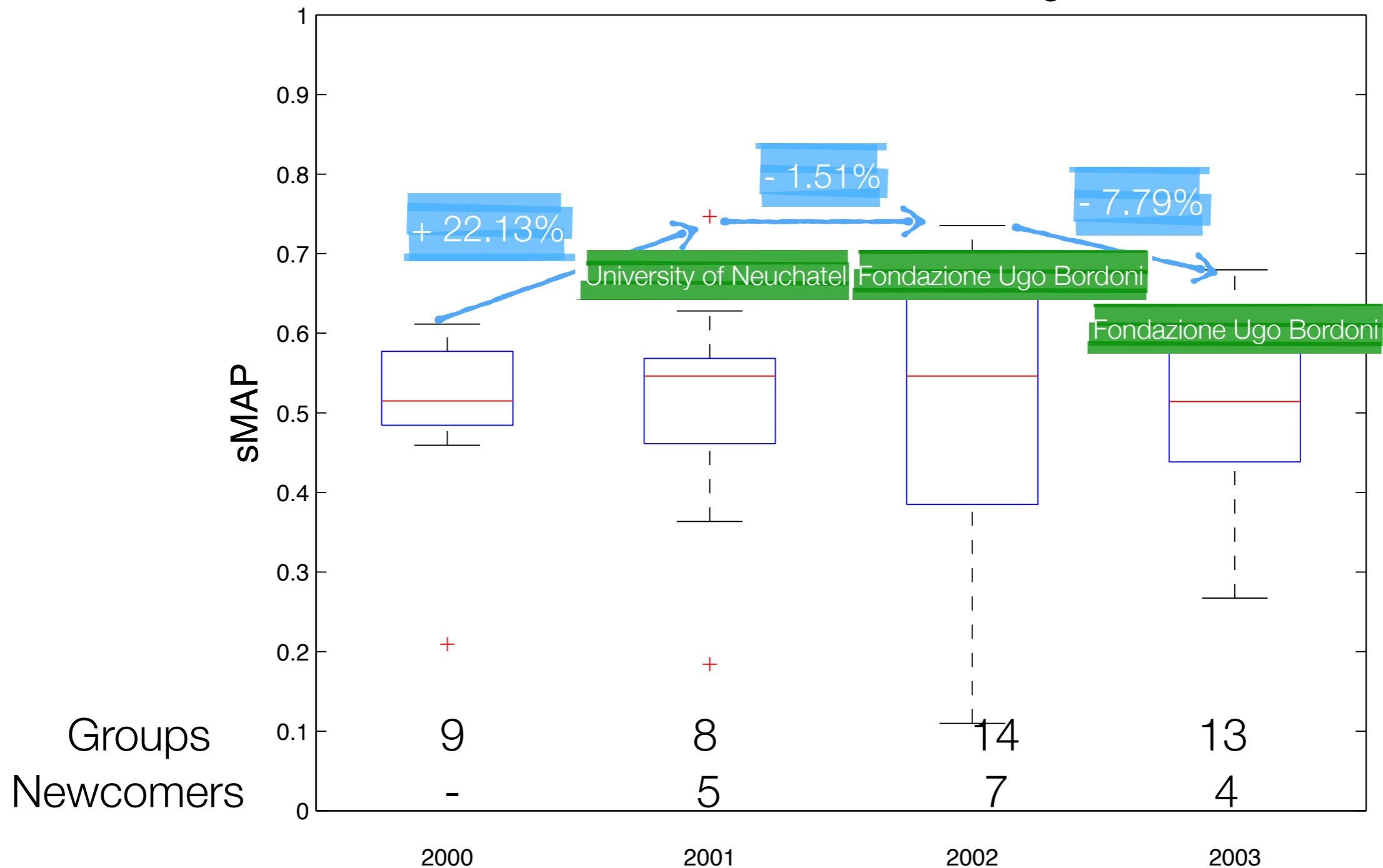
CLEF 2000 – 2003, Ad-Hoc Monolingual – Italian



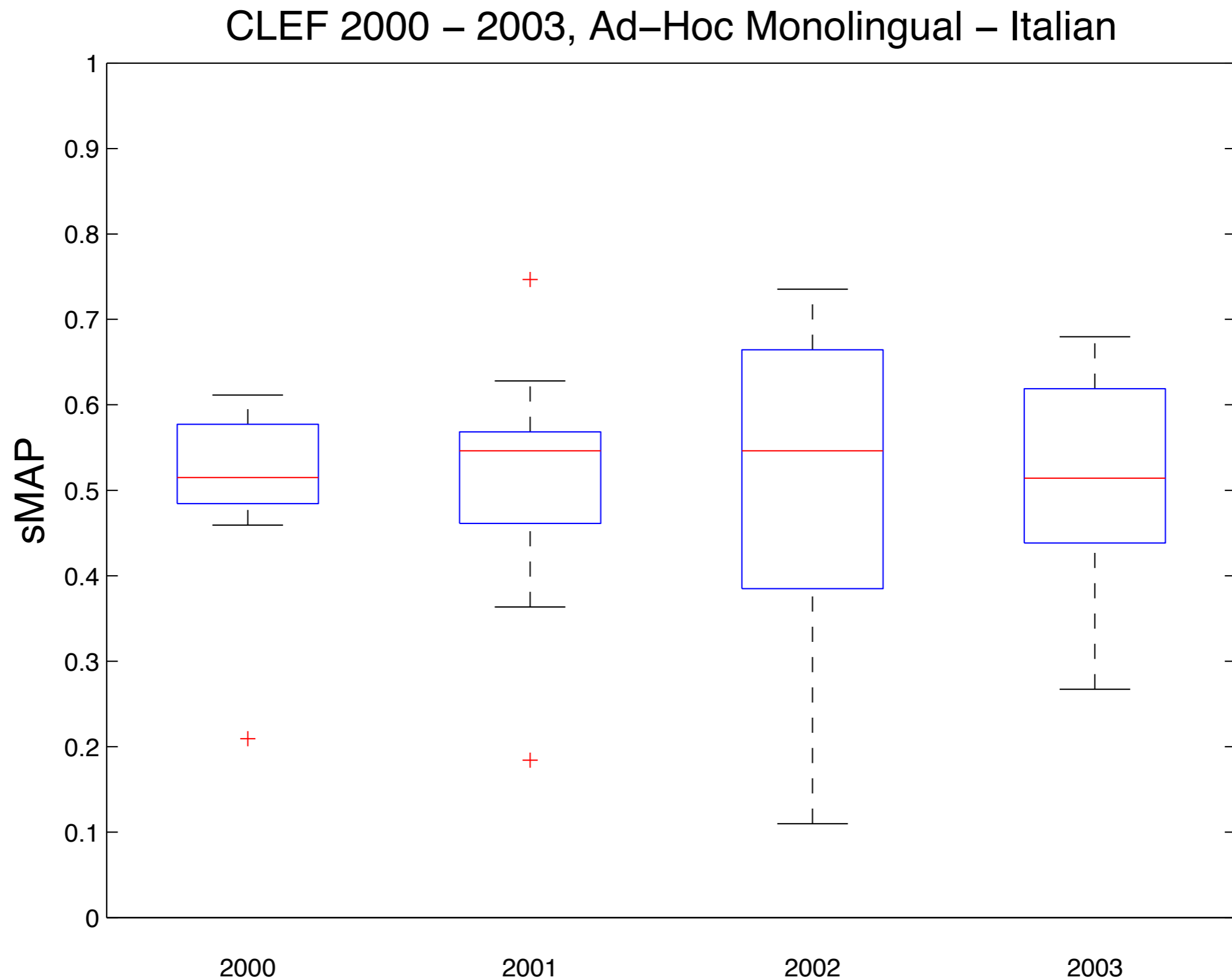
RQ 1: Do monolingual systems improve over the years?

Let's take a look to the Italian case

CLEF 2000 – 2003, Ad-Hoc Monolingual – Italian

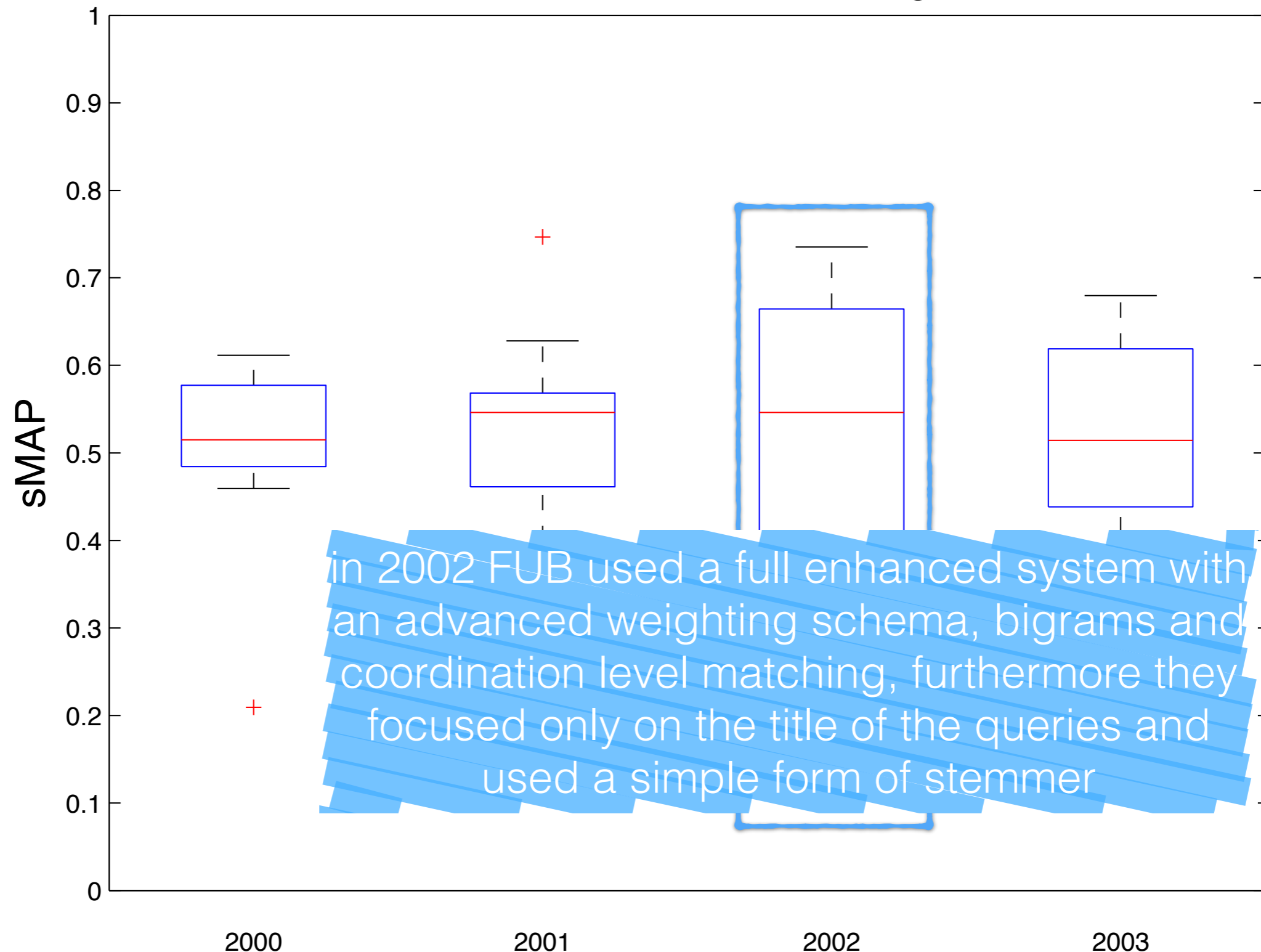


RQ 1: Do monolingual systems improve over the years?



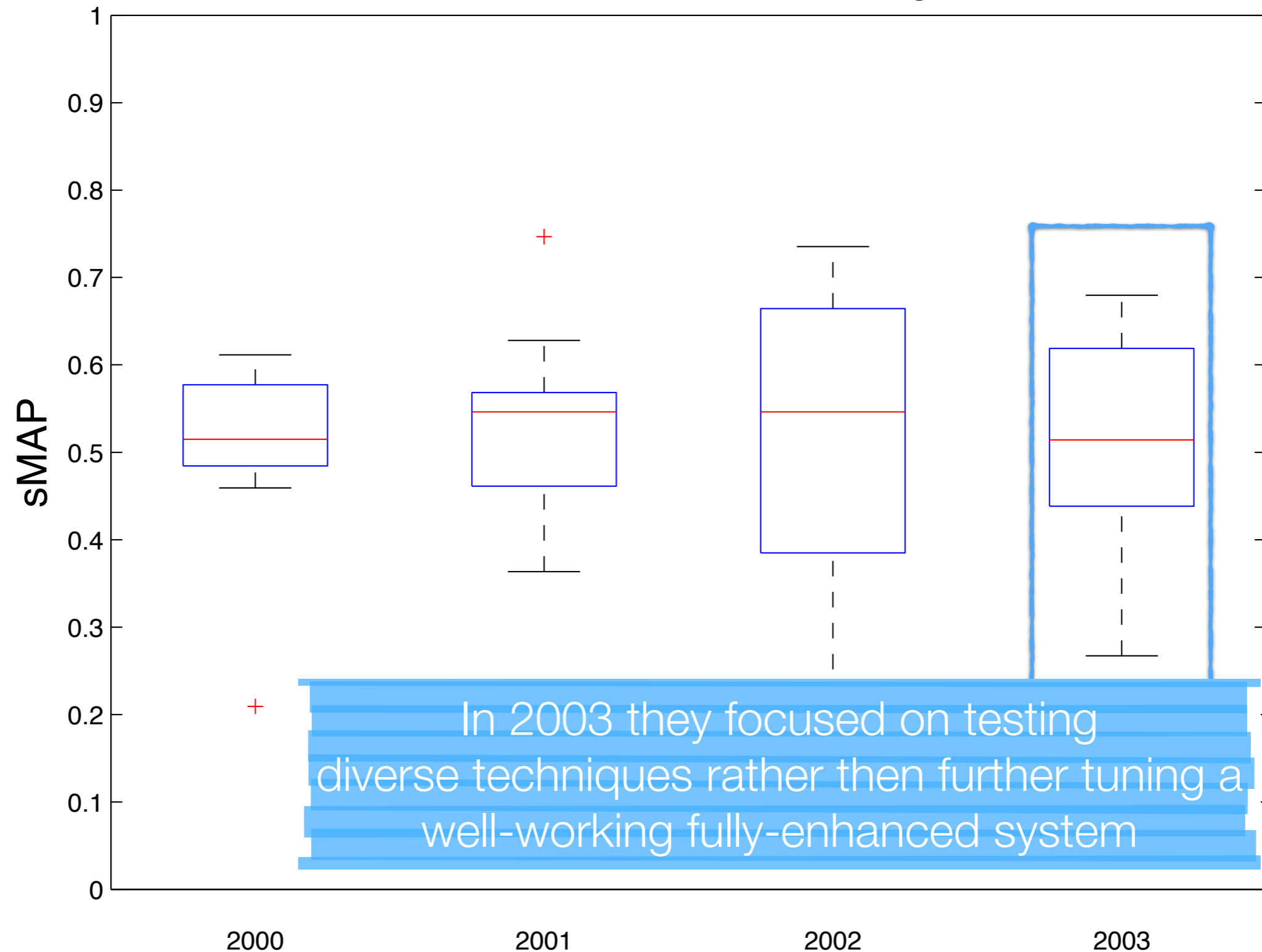
RQ 1: Do monolingual systems improve over the years?

CLEF 2000 – 2003, Ad-Hoc Monolingual – Italian



RQ 1: Do monolingual systems improve over the years?

CLEF 2000 – 2003, Ad-Hoc Monolingual – Italian



RQ 1: Outcomes

Performances of best groups tend to steadily increase over time

- Usually research groups that participated in previous years might have been *more interested in testing new techniques and retrieval settings* rather than tuning already well performing systems
- This hypothesis is corroborated by the best performances analysis, where *in the first years of a task research groups dedicated much effort to tuning and enhancing good systems* already tested in previous campaigns

RQ 2

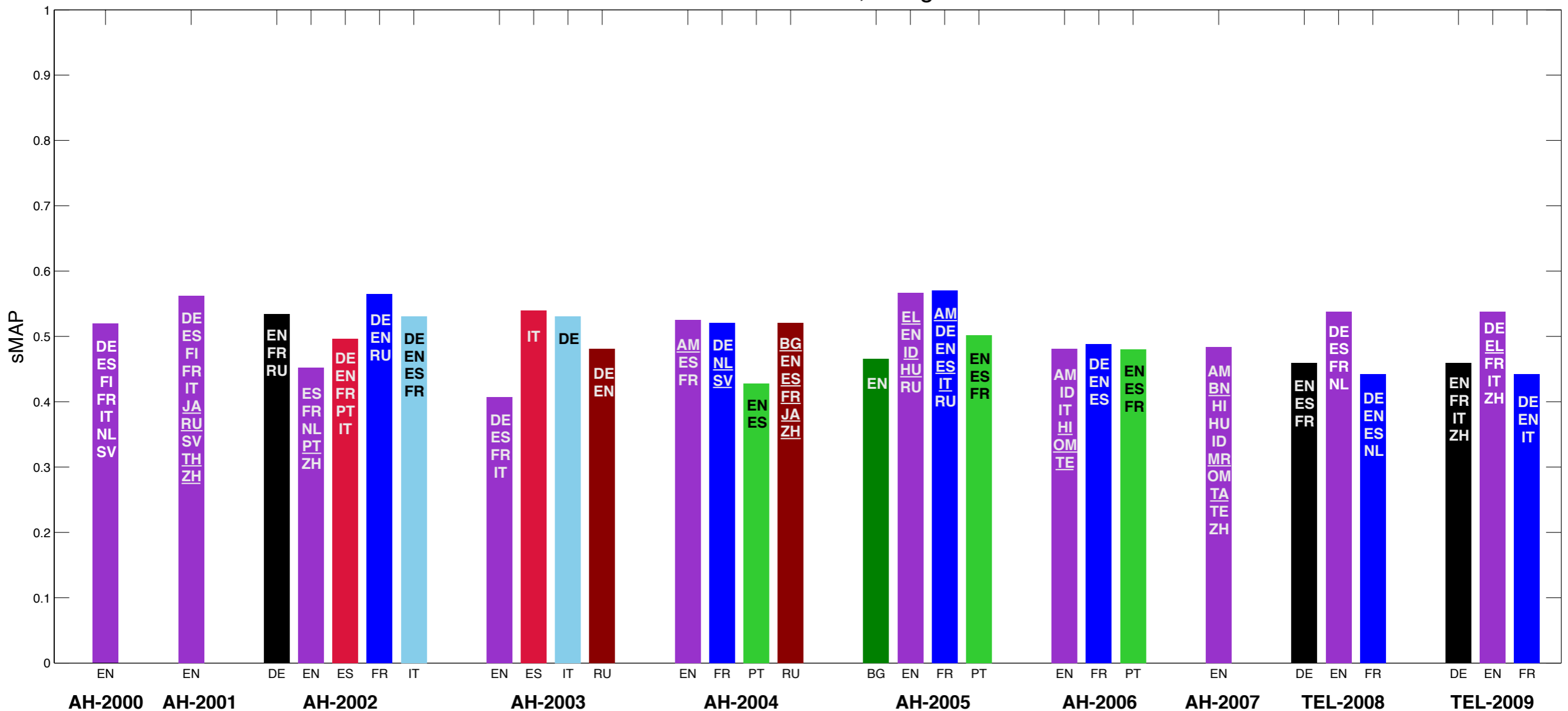
Do performances of bilingual systems increase over the years?

- For bilingual tasks we have to consider:
 - The target language: the language of the corpus
 - The source languages: the languages of the topics
- It is not always possible to identify a steady improvement of performances for a given target language over the years

RQ 2

Do performances of bilingual systems increase over the years?

CLEF 2000 – 2009, Bilingual



RQ 2: Do bilingual systems improve over the years?

Task	Year	Groups(new)	Runs	Best sMAP	Median sMAP
AH Bili DE	2002	6(-)	13	.6674 (-)	.5340 (-)
TEL Bili DE	2008	6(4)	17	.6268 (-6,08%)	.4599 (-13.88%)
	2009	6(3)	26	.7179 (14.53%)	.4731 (+2.87%)
AH Bili EN	2000	10(-)	26	.7463 (-)	.5196 (-)
	2001	19(15)	55	.7725 (+3.51%)	.5618 (+8.12%)
	2002	5(3)	16	.6983 (-9.60%)	.4524 (-19.47%)
	2003	3(3)	15	.6980 (-0.04%)	.4074 (-9.95%)
	2004	4(4)	11	.5895 (-15.54%)	.5251 (+28.89%)
	2005	8(8)	31	.7845 (+33.08%)	.5667 (+7.92%)
	2006	5(4)	32	.7559 (-3.64%)	.4808 (-15.16%)
TEL Bili EN	2007	10(9)	67	.7746 (+2.47%)	.4835 (0.56%)
	2008	8(7)	24	.7611 (-1,74%)	.5382 (+11.31%)
TEL Bili EN	2009	10(7)	43	.7808 (2.59%)	.4719 (-12.32%)
	2002	7(-)	16	.6805 (-)	.4969 (-)
AH Bili ES	2003	9(7)	15	.6737 (-1.01%)	.5394 (+8.55%)
AH Bili FR	2002	7(-)	14	.6708 (-)	.5647 (-)
	2004	7(5)	24	.6015 (-10.33%)	.5211 (-7.72%)
	2005	9(8)	31	.7250 (+20.53%)	.5703 (+9.44%)
	2006	4(3)	12	.6273 (-13.47%)	.4886 (-14.33%)
TEL Bili FR	2008	5(5)	15	.6358 (+1,35%)	.4422 (-9.50%)
	2009	6(4)	23	.7151 (+12.47%)	.4355 (-1.52%)
AH Bili IT	2002	6(-)	13	.5916 (-)	.5306 (-)
	2003	8(5)	21	.7119 (+20.34%)	.5309 (+0.05%)
AH Bili PT	2004	4(-)	15	.6721 (-)	.4278 (-)
	2005	8(5)	24	.7239 (+7.71%)	.5020 (+17.34%)
	2006	6(4)	22	.6539 (-9.67%)	.4804 (-4.30%)
AH Bili RU	2003	2(-)	9	.6894 (-)	.4810 (-)
	2004	8(7)	26	.6336 (-8.09%)	.5203 (+8.17%)

RQ 2: Do bilingual systems improve over the years?

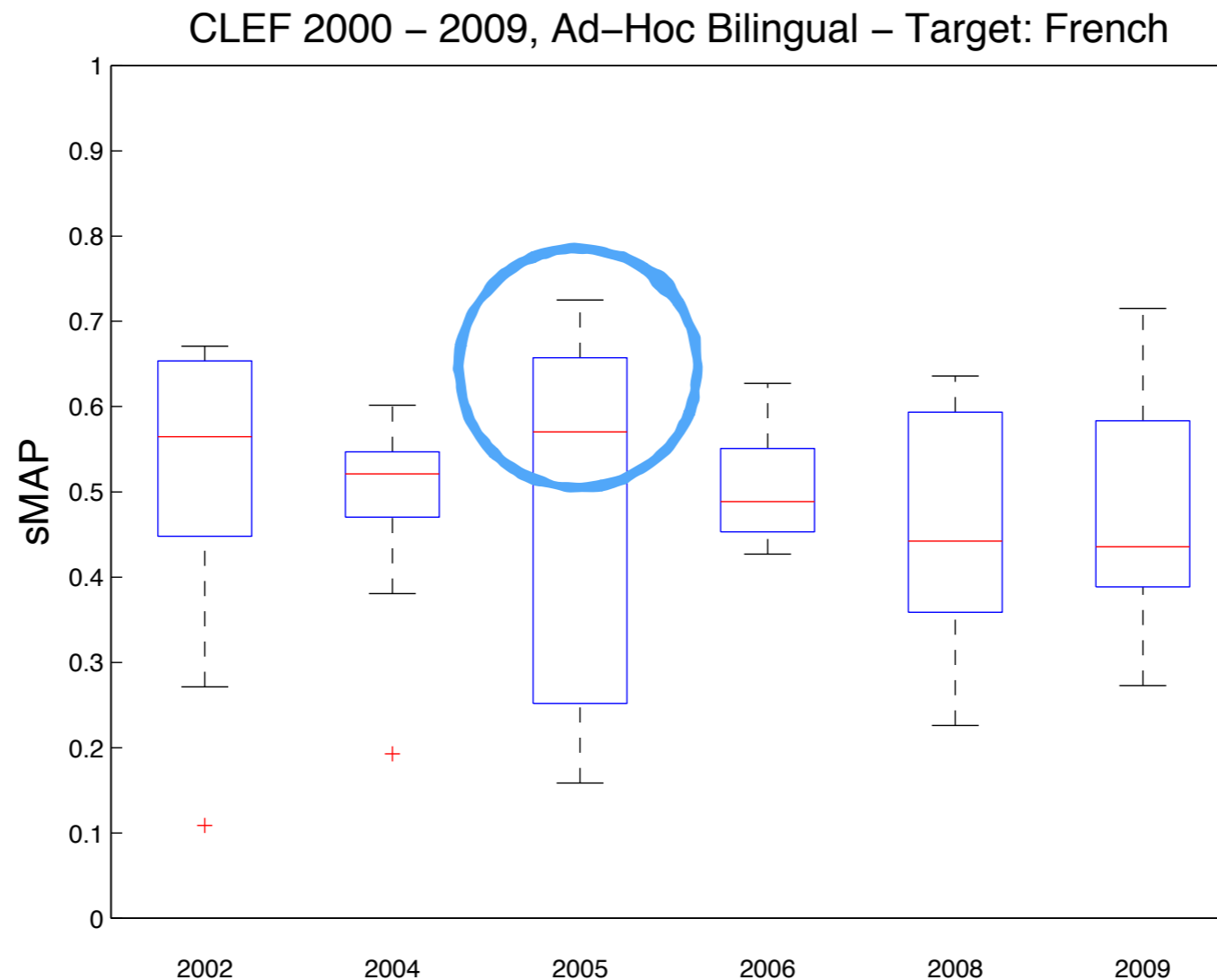
- Unlike for the monolingual tasks, that the **higher median sMAP** as well as the **best sMAP** are **achieved in the last years of each task.**
- This is an indicator of the **improvement of language resources** – e.g. dictionaries, external resources like Wikipedia and the use of semantic rather than syntactic resources
 - **CLEF Continued Impact!**

Task	Best sMAP	Median sMAP
	.6674 (-)	.5340 (-)
	.6268 (-6,08%)	.4599 (-13.88%)
	.7179 (14.53%)	.4731 (+2.87%)
	.7463 (-)	.5196 (-)
	.7725 (+3.51%)	.5618 (+8.12%)
	.6983 (-9.60%)	.4524 (-19.47%)
	.6980 (-0.04%)	.4074 (-9.95%)
	.5895 (-15.54%)	.5251 (+28.89%)
	.7845 (+33.08%)	.5667 (+7.92%)
	.7559 (-3.64%)	.4808 (-15.16%)
	.7746 (+2.47%)	.4835 (0.56%)
	.7611 (-1,74%)	.5382 (+11.31%)
	.7808 (2.59%)	.4719 (-12.32%)
	.6805 (-)	.4969 (-)
	.6737 (-1.01%)	.5394 (+8.55%)
	.6708 (-)	.5647 (-)
	.6015 (-10.33%)	.5211 (-7.72%)
	.7250 (+20.53%)	.5703 (+9.44%)
	.6273 (-13.47%)	.4886 (-14.33%)
	.6358 (+1,35%)	.4422 (-9.50%)
	.7151 (+12.47%)	.4355 (-1.52%)
	.5916 (-)	.5306 (-)
	.7119 (+20.34%)	.5309 (+0.05%)
	.6721 (-)	.4278 (-)
	.7239 (+7.71%)	.5020 (+17.34%)
	.6539 (-9.67%)	.4804 (-4.30%)
	.6894 (-)	.4810 (-)
	.6336 (-8.09%)	.5203 (+8.17%)

RQ 2: Do bilingual systems improve over the years?

Effect of the improvement of language resources

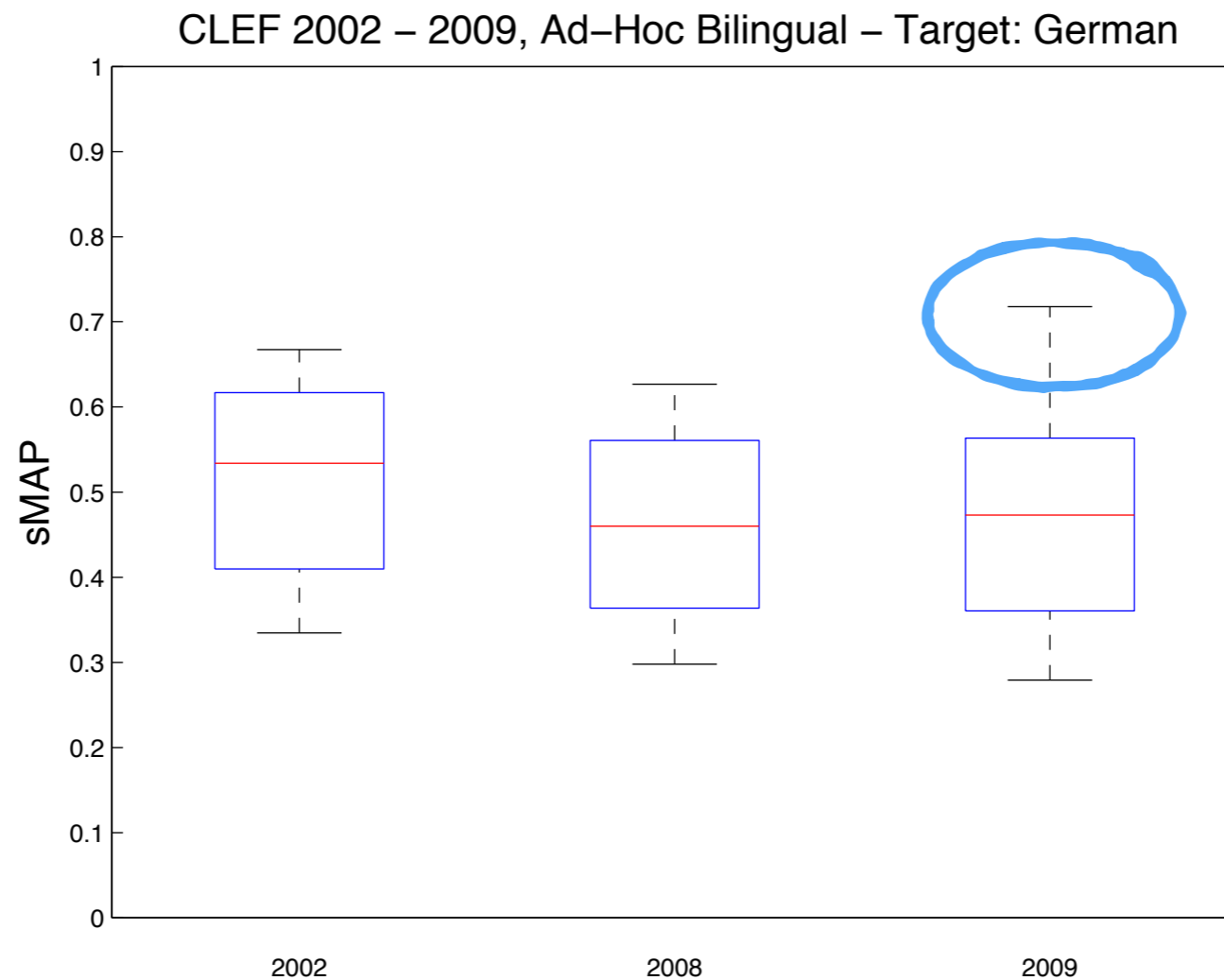
- The best bilingual system for the “X2FR” task exploited seven different machine translation systems, three bilingual dictionaries and ten freely available translation tools.



RQ 2: Do bilingual systems improve over the years?

Effect of the improvement of language resources

- The best bilingual system in the TEL “X2DE” task exploited three out-the-box retrieval systems (i.e. Lucene, Lemur and Terrier) and the high quality of the Google translation service contributed substantially to achieving the final result



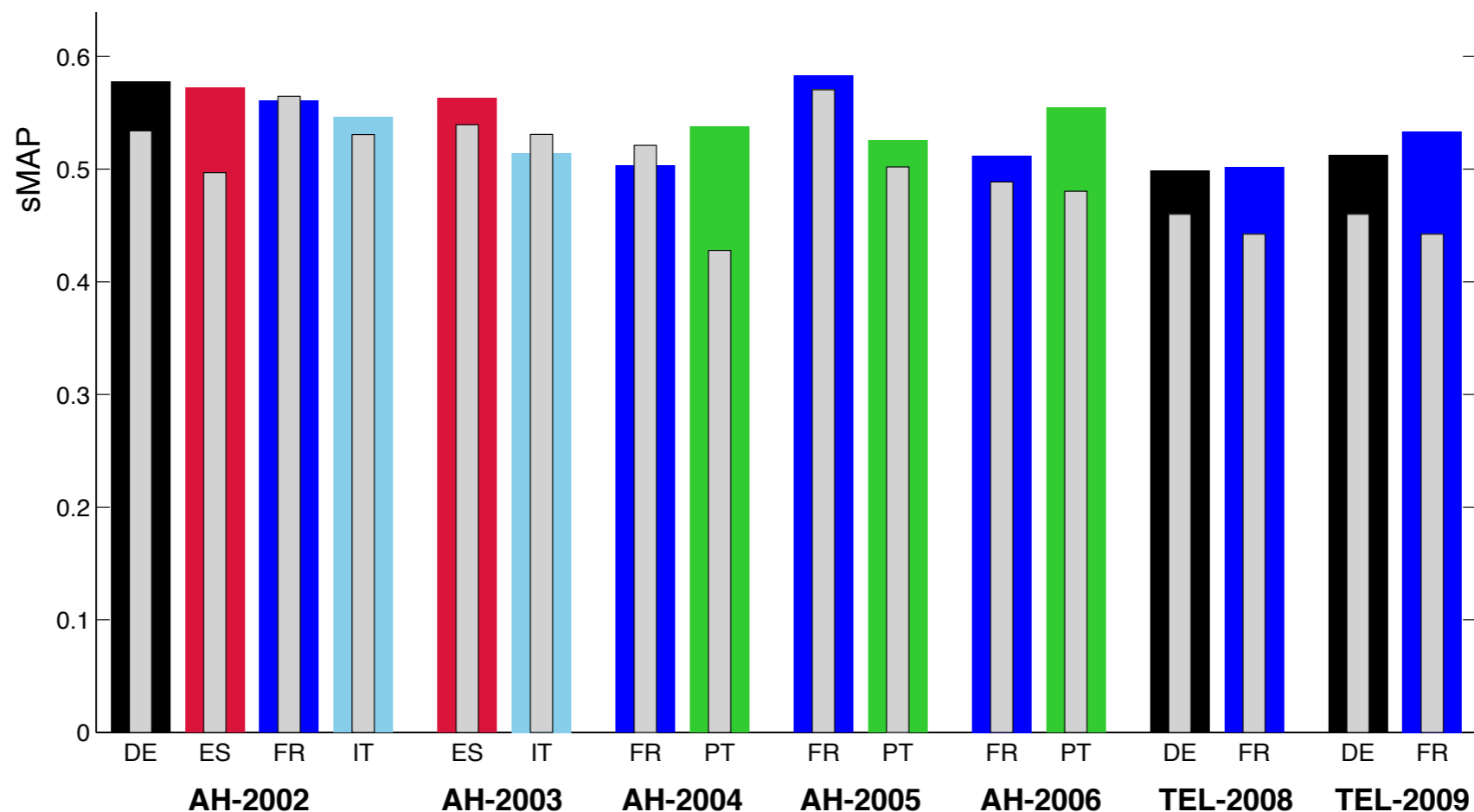
RQ3

Do monolingual systems have better performances than bilingual and multilingual systems?

CLEF 2002 – 2009, Mono/Bili Median MAP comparison



In most cases the median sMAP of the monolingual tasks overcome the median sMAP of the corresponding bilingual task



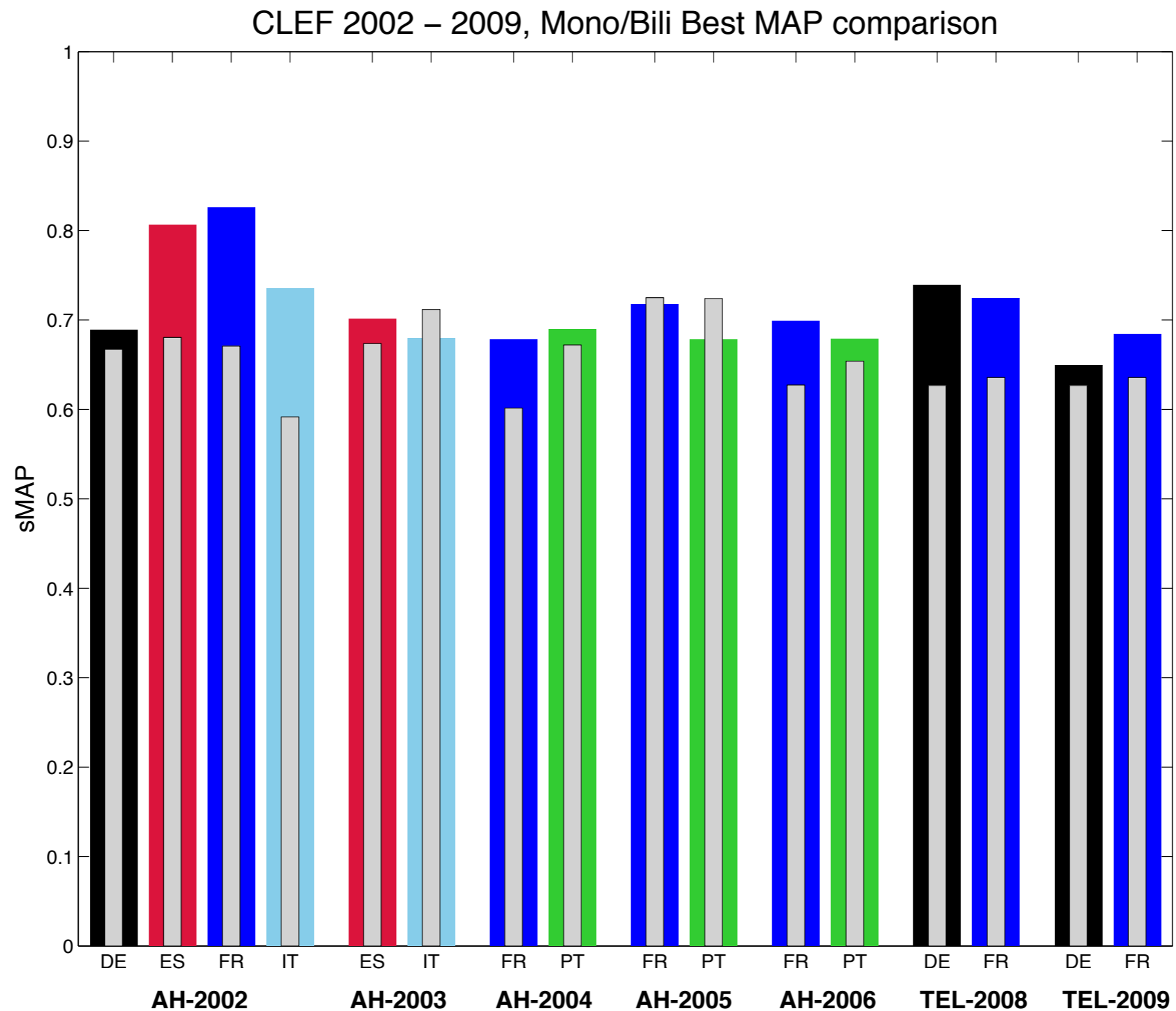
RQ3

Do monolingual systems have better performances than bilingual and multilingual systems?

Things are different when we consider the best performance ratios than the median ones...

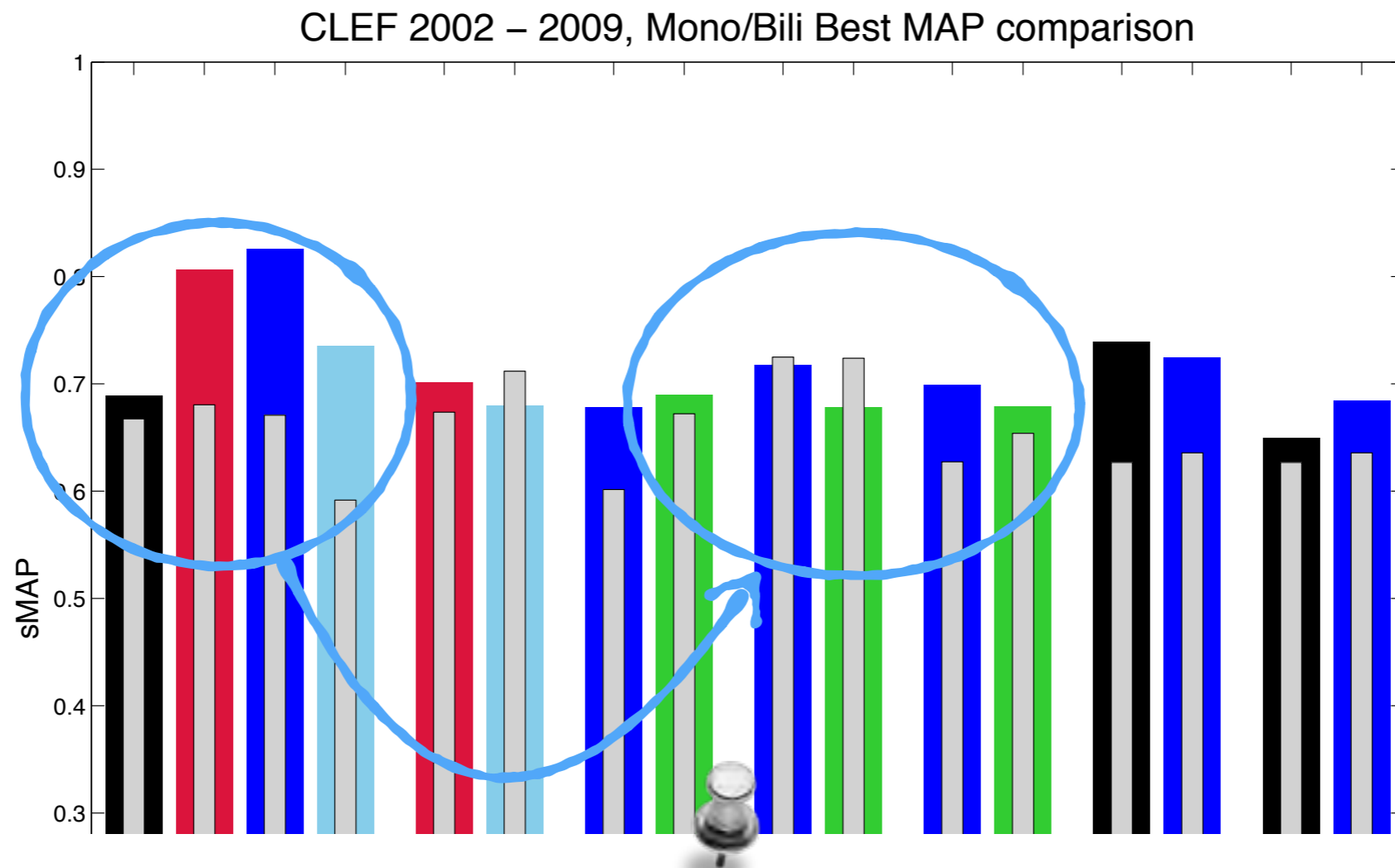
RQ3

Do monolingual systems have better performances than bilingual and multilingual systems?



RQ3

Do monolingual systems have better performances than bilingual and multilingual systems?



The gap between top monolingual and top bilingual systems is progressively reduced across the years and in several cases the trend is inverted

DE ES FR IT ES IT FR FI FR FI FR FI DE FR DE FR
AH-2002 AH-2003 AH-2004 AH-2005 AH-2006 TEL-2008 TEL-2009

RQ3

sMAP	Monolingual	Bilingual	Multilingual
Best	.8309	.7845	.8513
Median	.5344	.5165	.5173
Mean	.5054	.4898	.4914

Bilingual and multilingual systems have a similar median and mean sMAP even though they are slightly higher for the multilingual and both are exceeded by the monolingual systems.

It is interesting to note that the best system is the multilingual one that has a sMAP 8.52% higher than the top bilingual and 2.46% higher than the top monolingual system.

Wrapping-Up

- RQ1: Steady improvement of best systems and stable median performances taking into account the continuous growth of the community
- RQ2: CLEF had a significant impact in driving the improvement of bilingual systems by continuously stimulating the creation of new and better linguistic resources
- RQ3: Crossing the language barriers requires a big effort but pays off when you compare best systems



MATTERS

MATlab Toolkit for Evaluation of information Retrieval Systems

Find Out More

<http://matters.dei.unipd.it/>