CLEF 15th Birthday: What can we Learn 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

CLEF 2014 Conference and Labs of the Evaluation Forum

September 15th, 2014, London, UK

Outline

- Motivations

- Standardization

- Analysis of CLEF Ad-Hoc Campaigns

- Discussion and Future Works

Motivations

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

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

Our goal is to carry out a longitudinal study on the Ad-Hoc track of CLEF in order to understand its impact on monolingual, bilingual, and multilingual retrieval

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

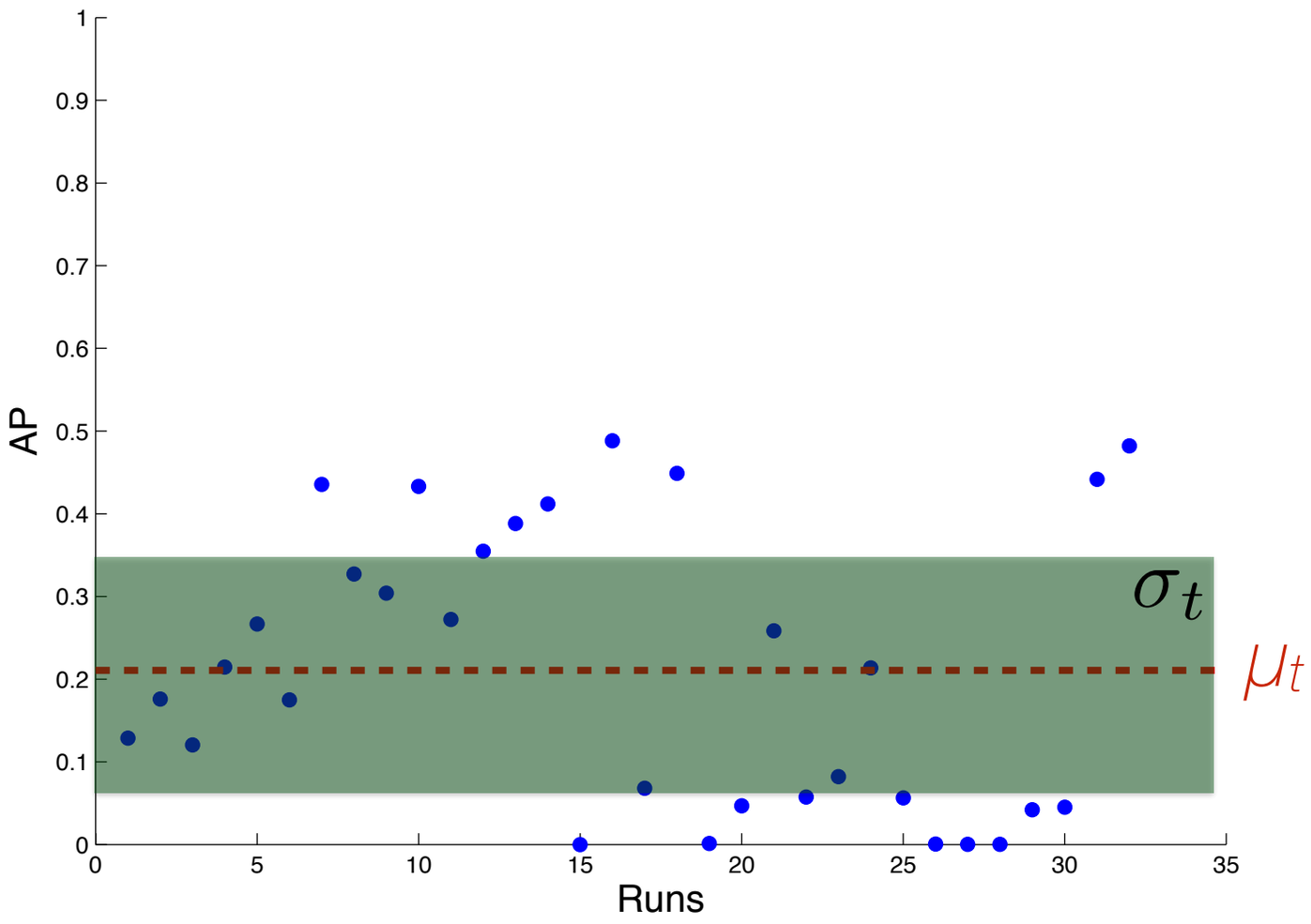
Inter-collection comparison between systems by limiting the effect of collections and by making system scores interpretable in themselves

– 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

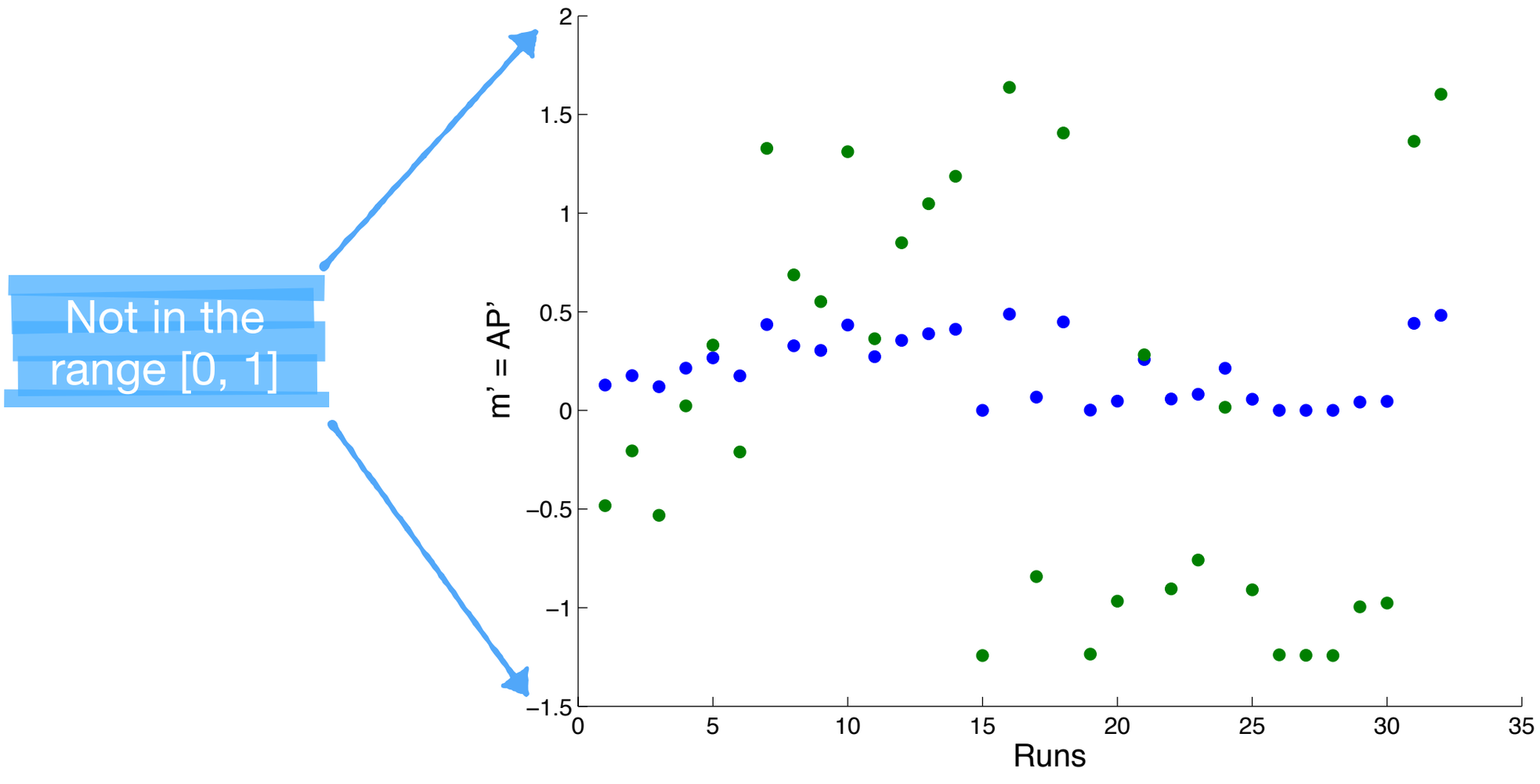
These are AP values of all the runs for topic 301 of CLEF Ad-Hoc bilingual English 2006



How standardization works

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

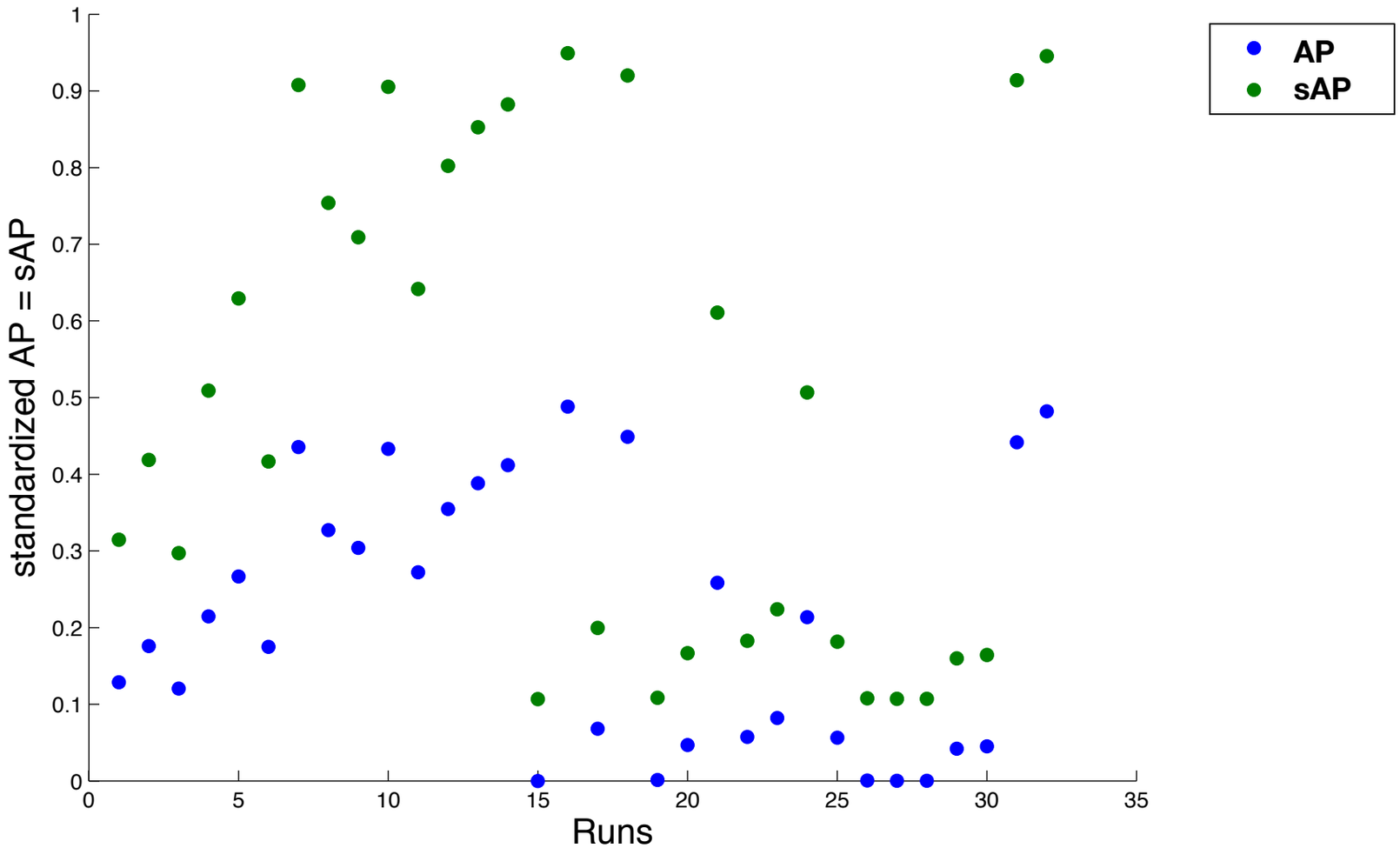
$$m' = \frac{m - \mu_t}{\sigma_t}$$



How standardization works

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

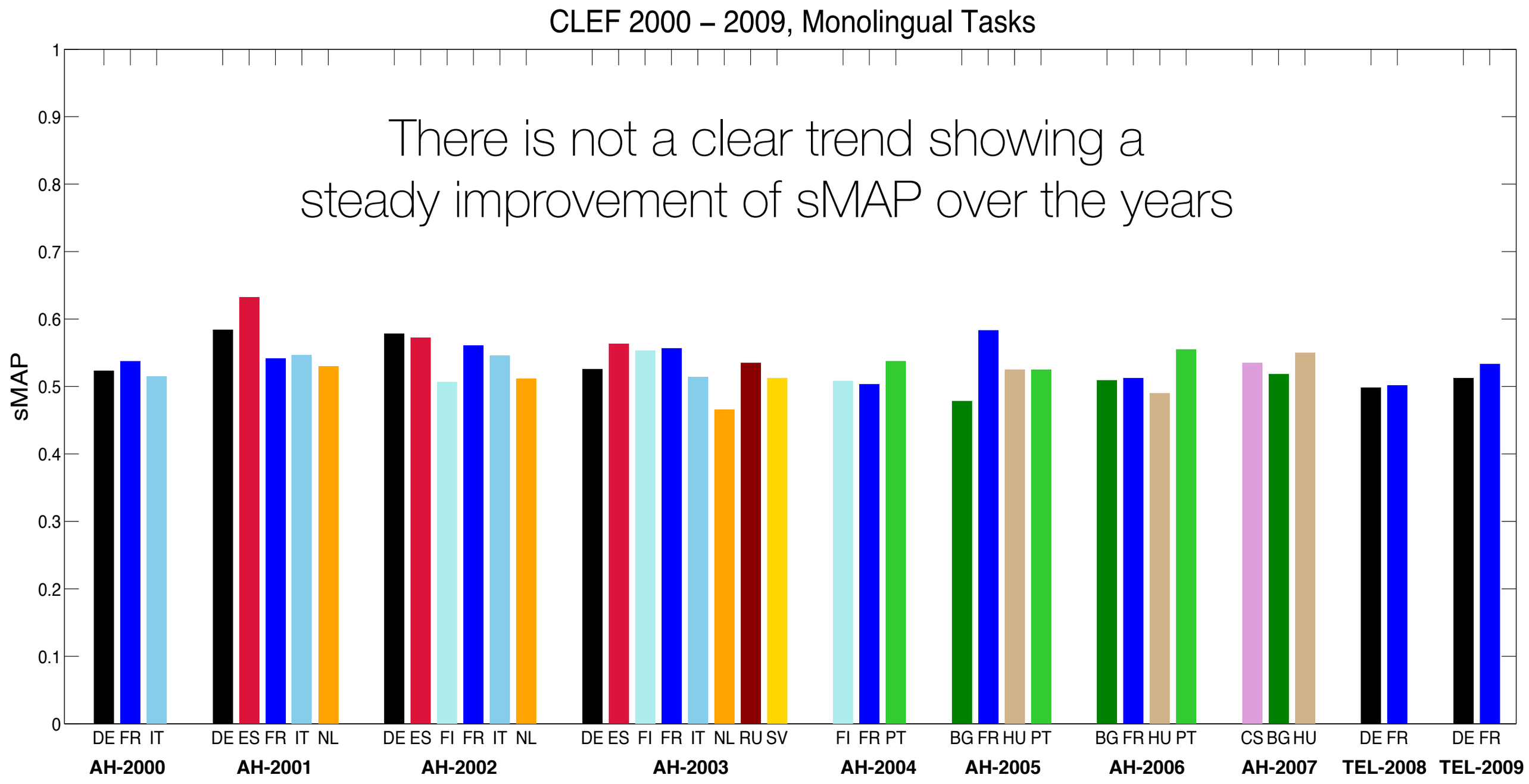
$$F_X(m') = \int_{-\infty}^{m'} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$



Analysis of CLEF Ad-Hoc Campaigns

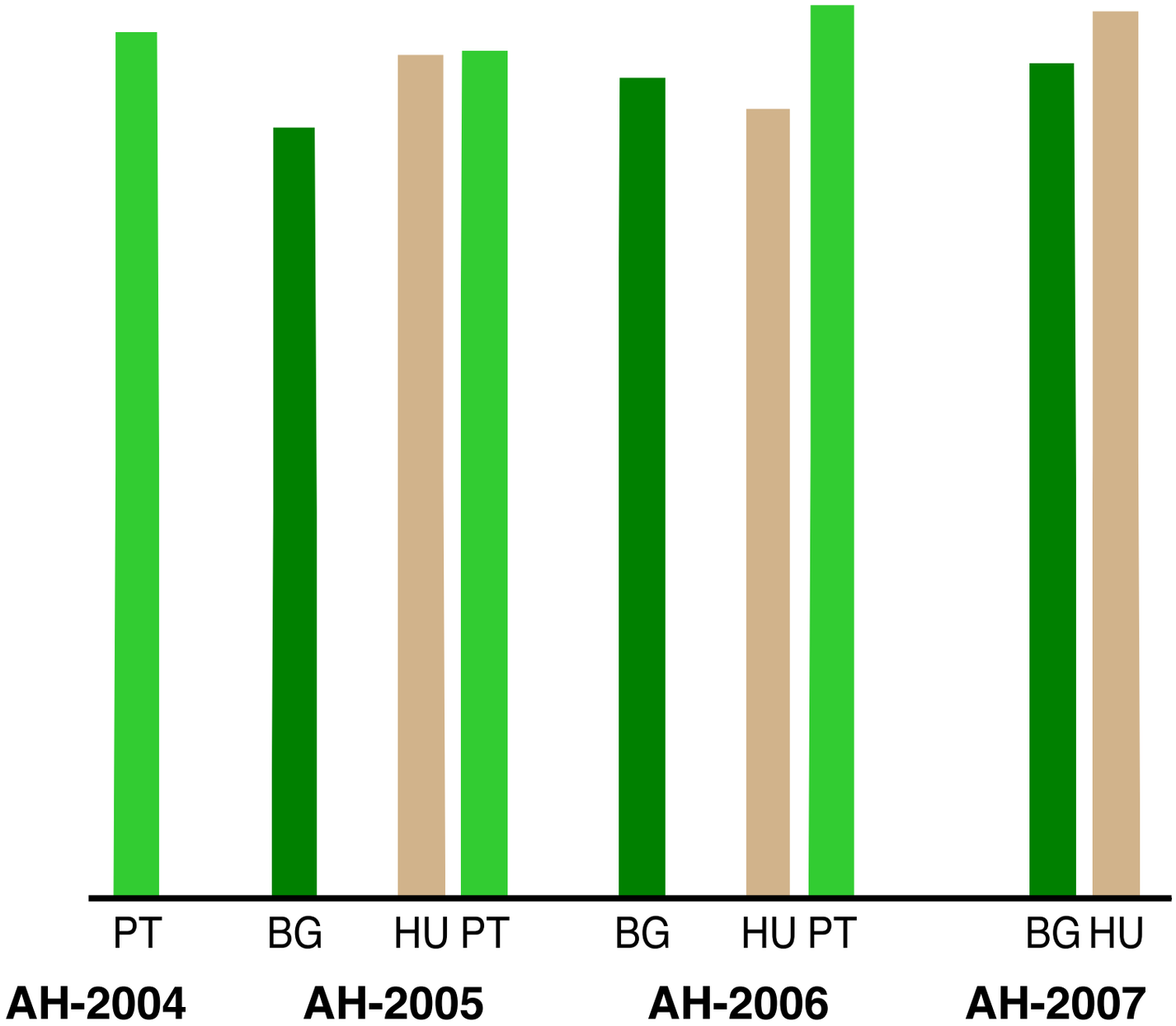
RQ 1

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



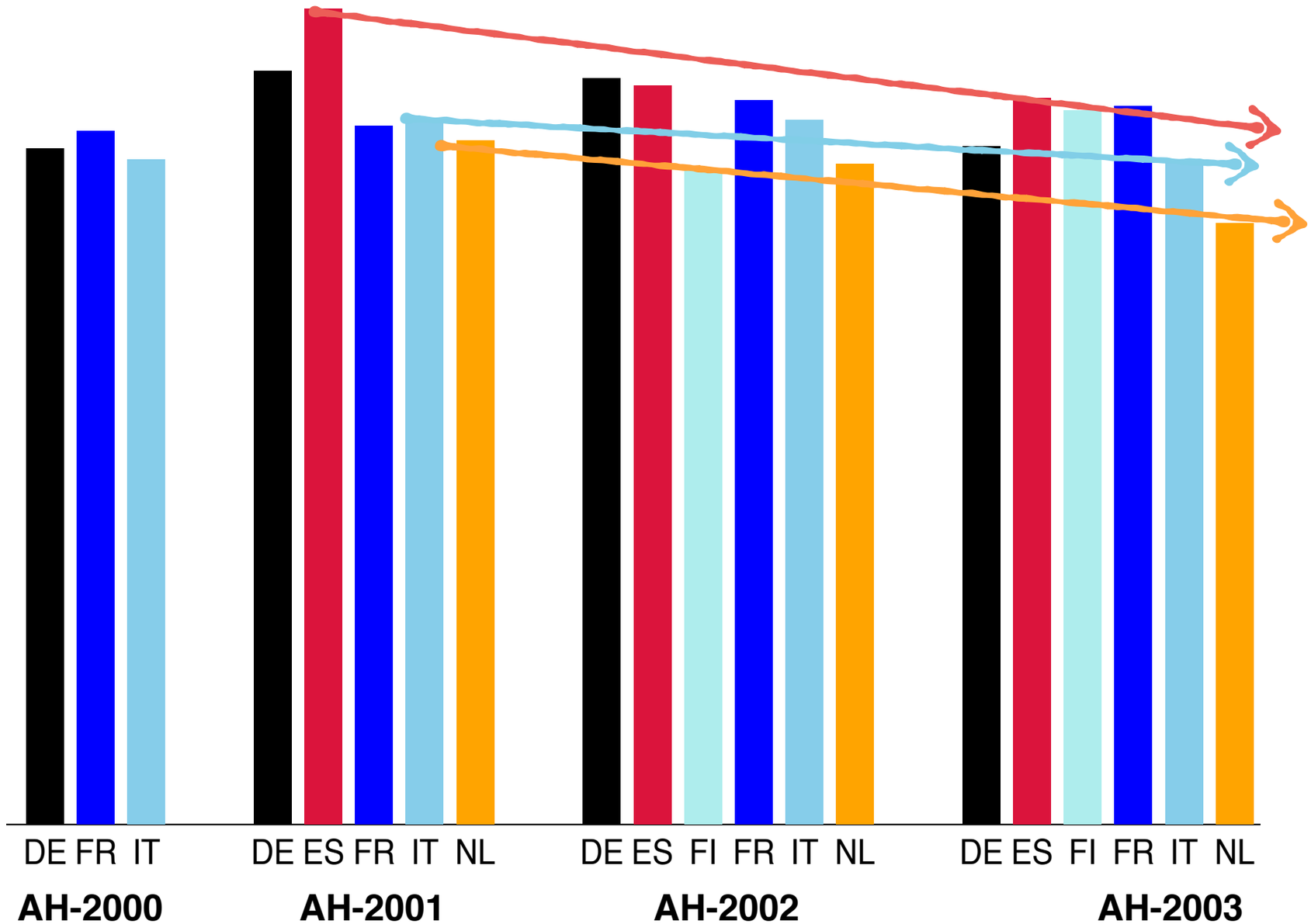
RQ 1: Do monolingual systems improve over the years?

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



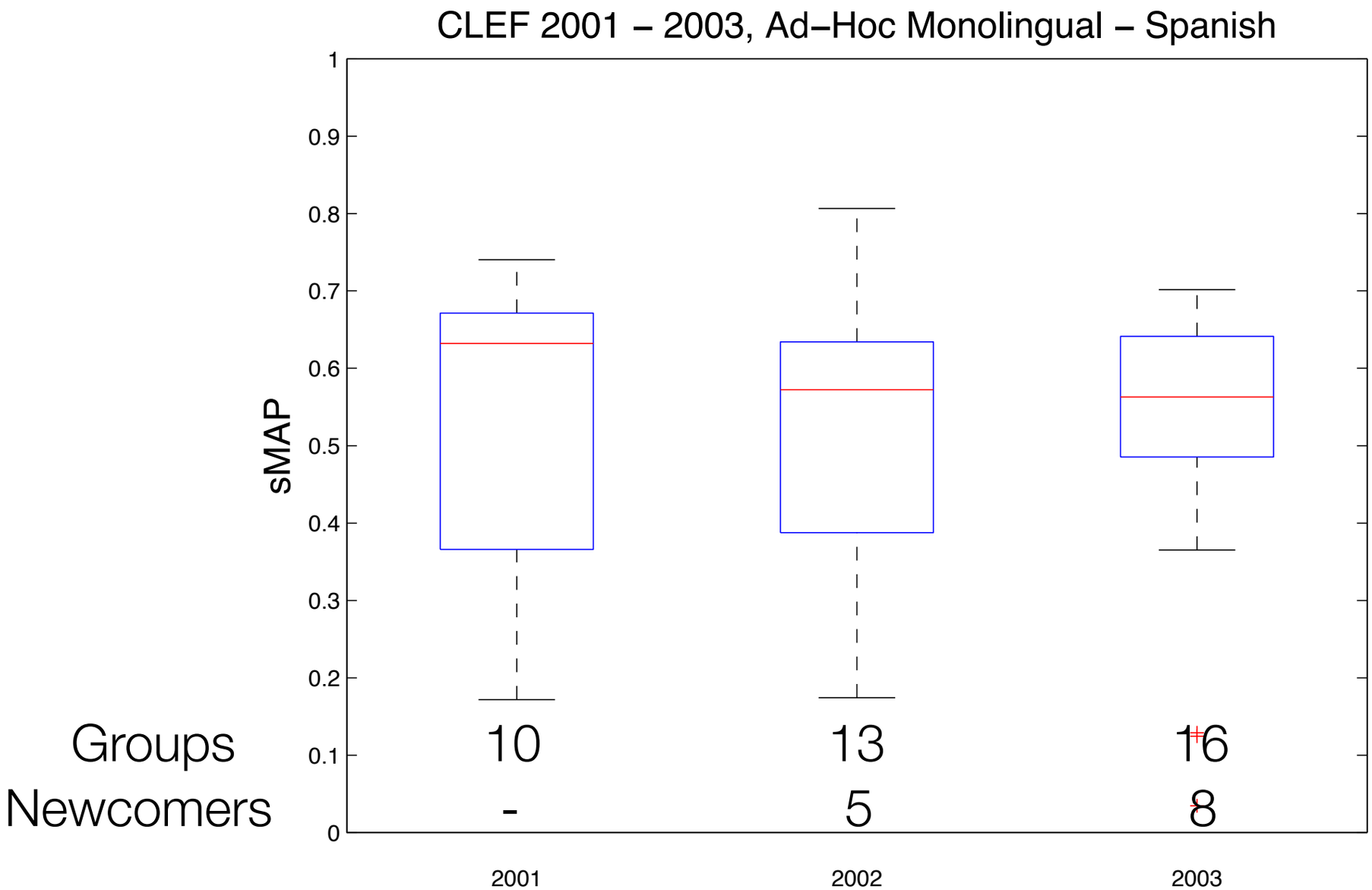
RQ 1: Do monolingual systems improve over the years?

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



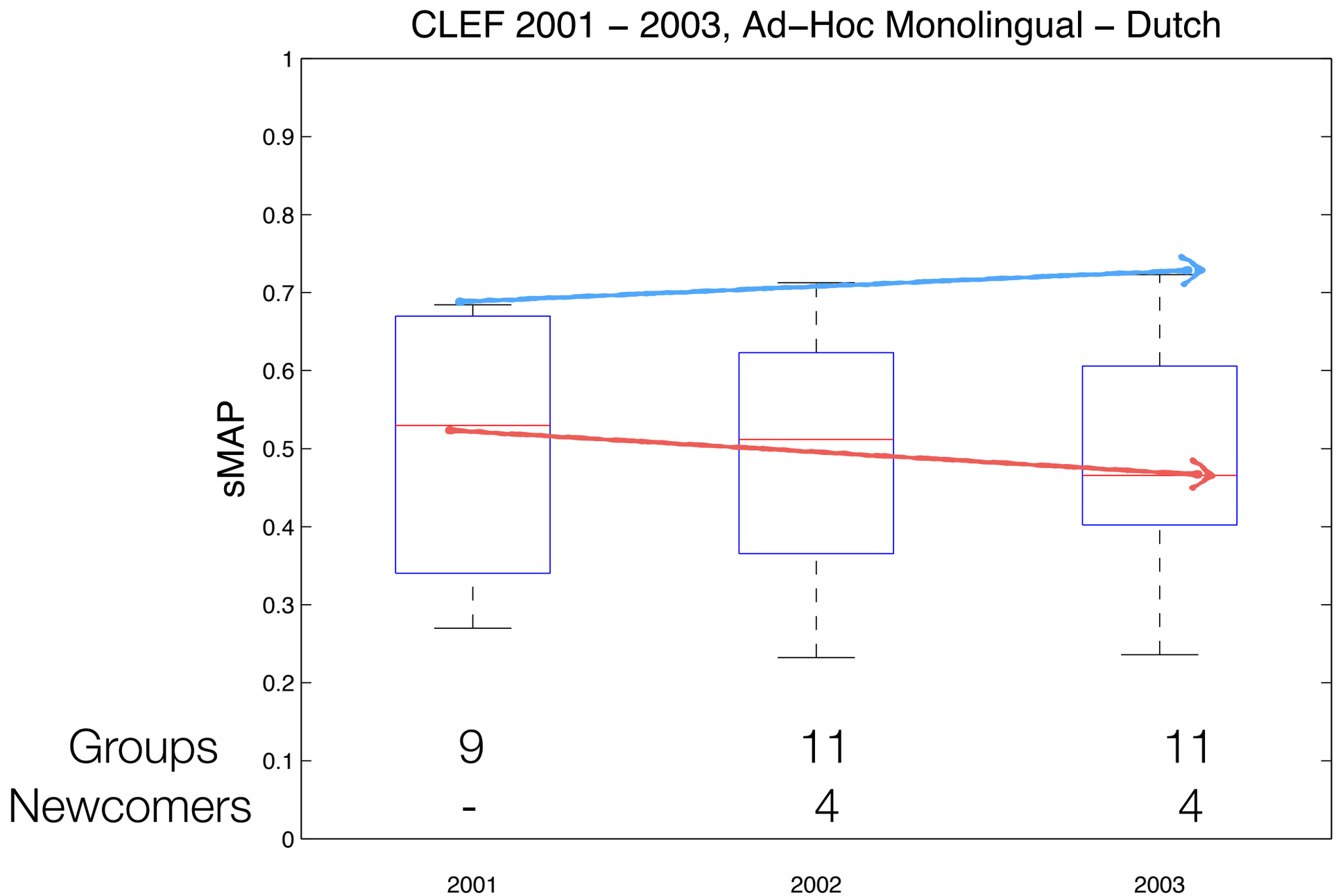
RQ 1: Do monolingual systems improve over the years?

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?

Overall, going beyond the raw numbers we can point out two positive effects:

Performances of best groups tend to steadily increase over time

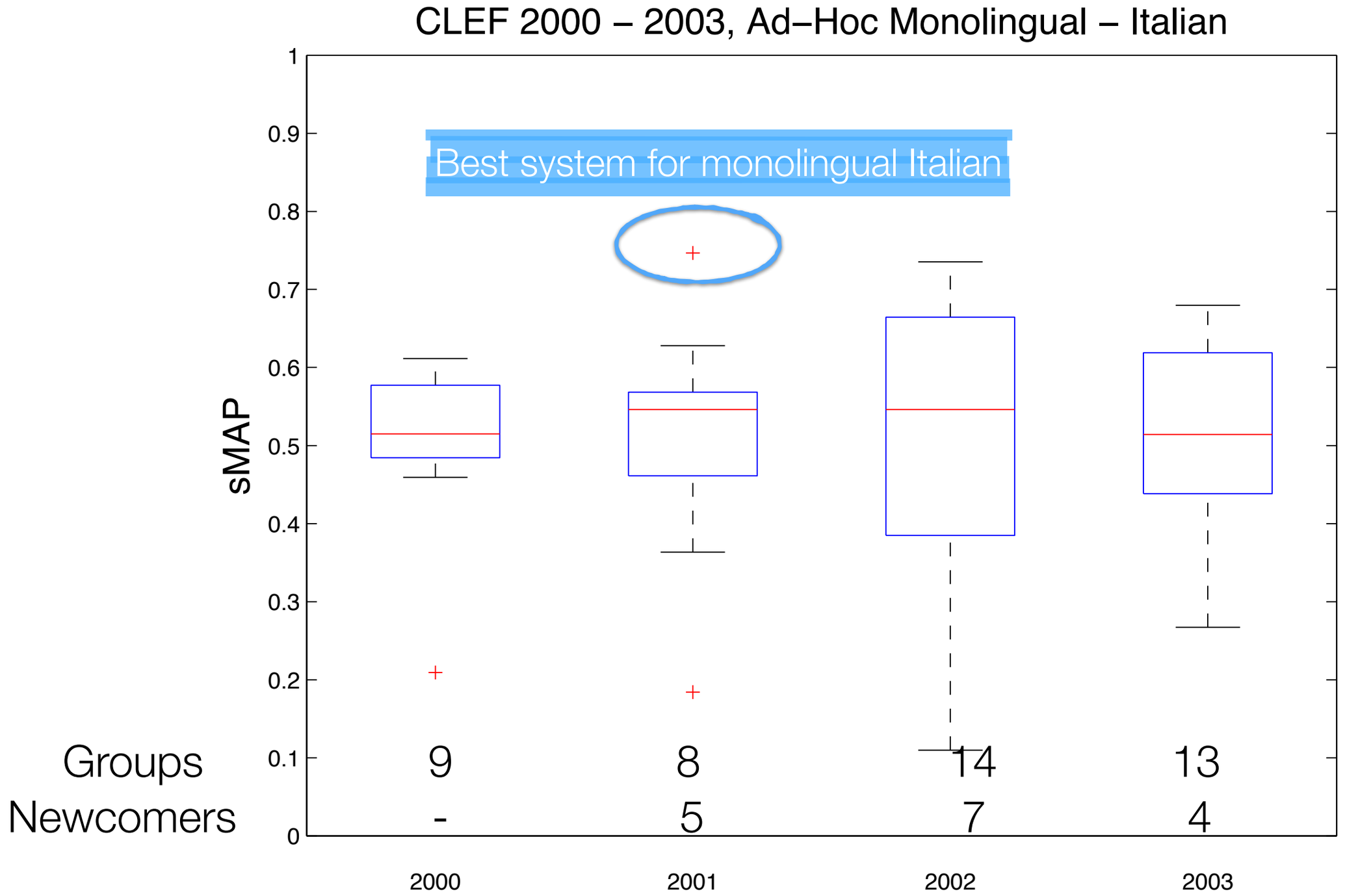
The CLEF community grows through time with newcomers year after year fostering the growth of new IR groups at the price of a contained decrease of median performances

RQ 1: Do monolingual systems improve over the years?

But… Are general trends enough to explain phenomena?

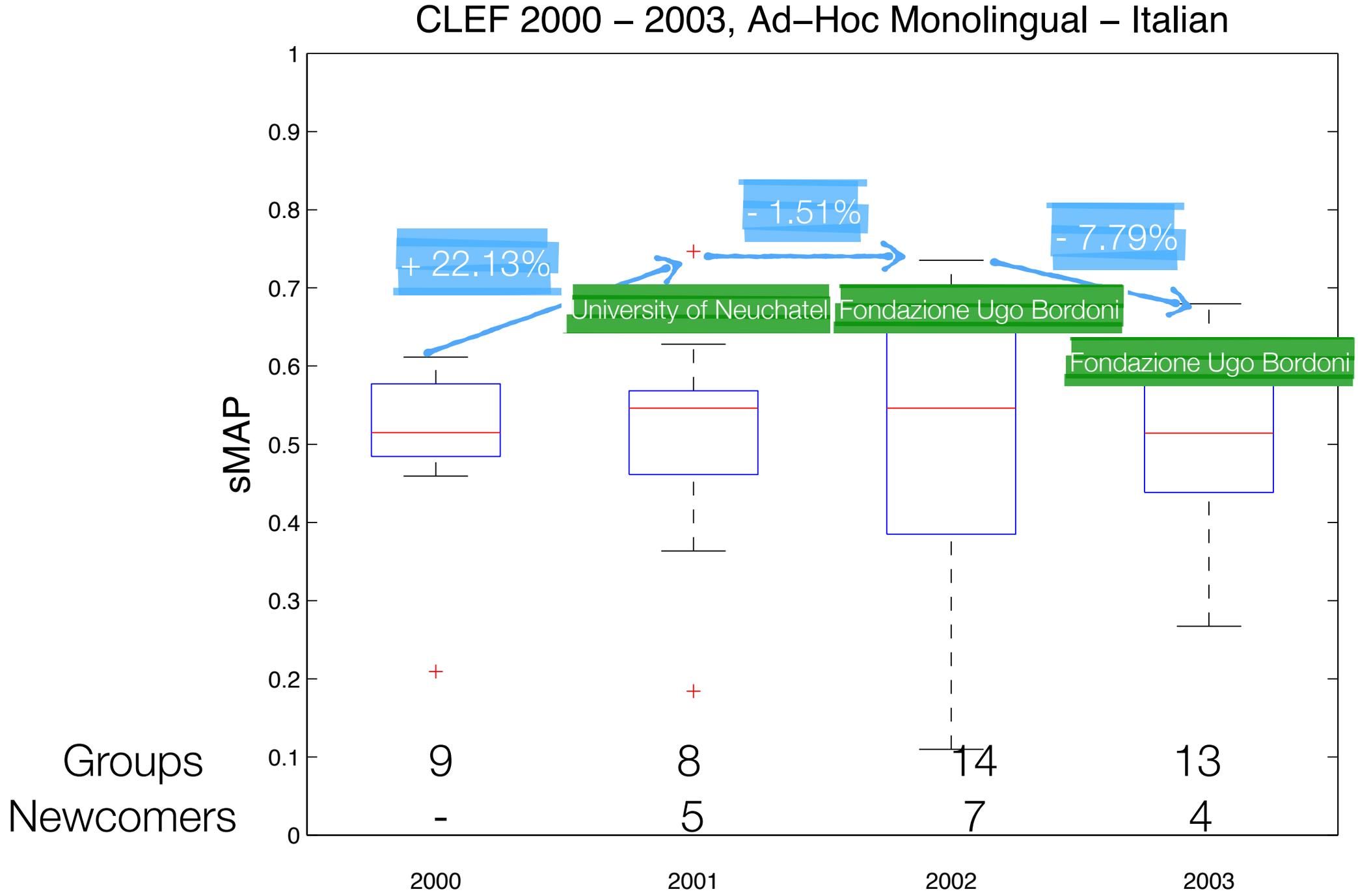
RQ 1: Do monolingual systems improve over the years?

Let's take a look to the Italian case

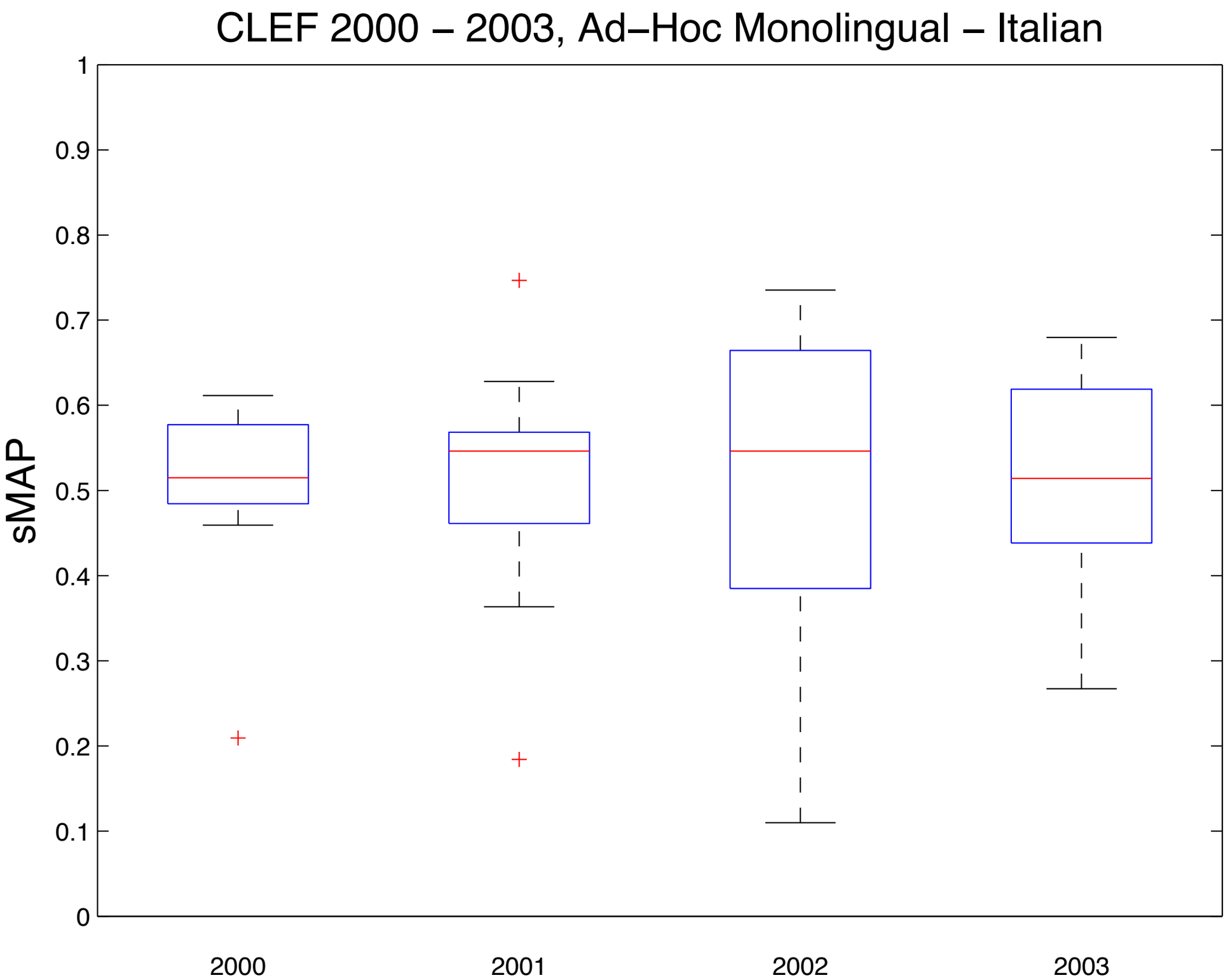


RQ 1: Do monolingual systems improve over the years?

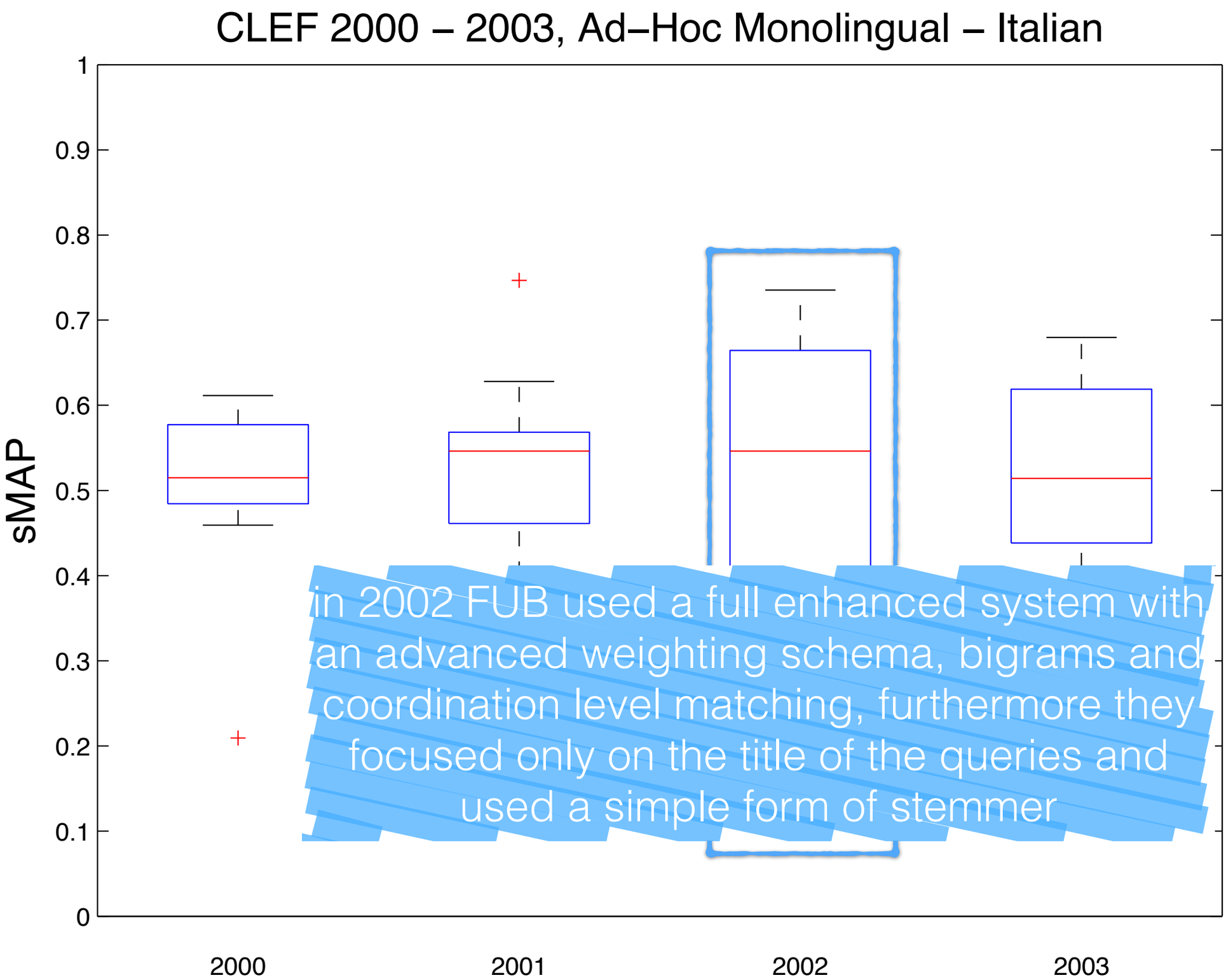
Let's take a look to the Italian case



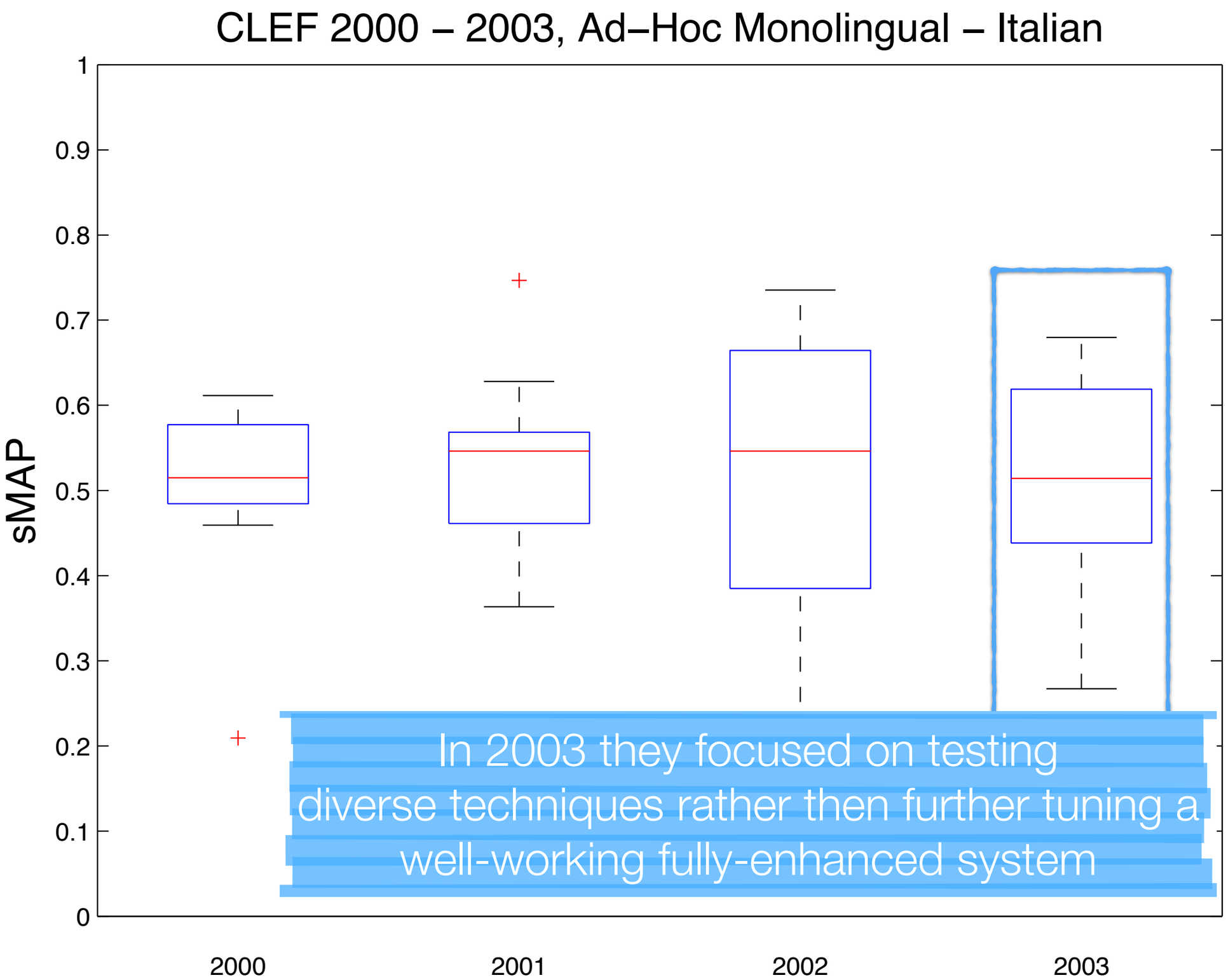
RQ 1: Do monolingual systems improve over the years?



RQ 1: Do monolingual systems improve over the years?



RQ 1: Do monolingual systems improve over the years?



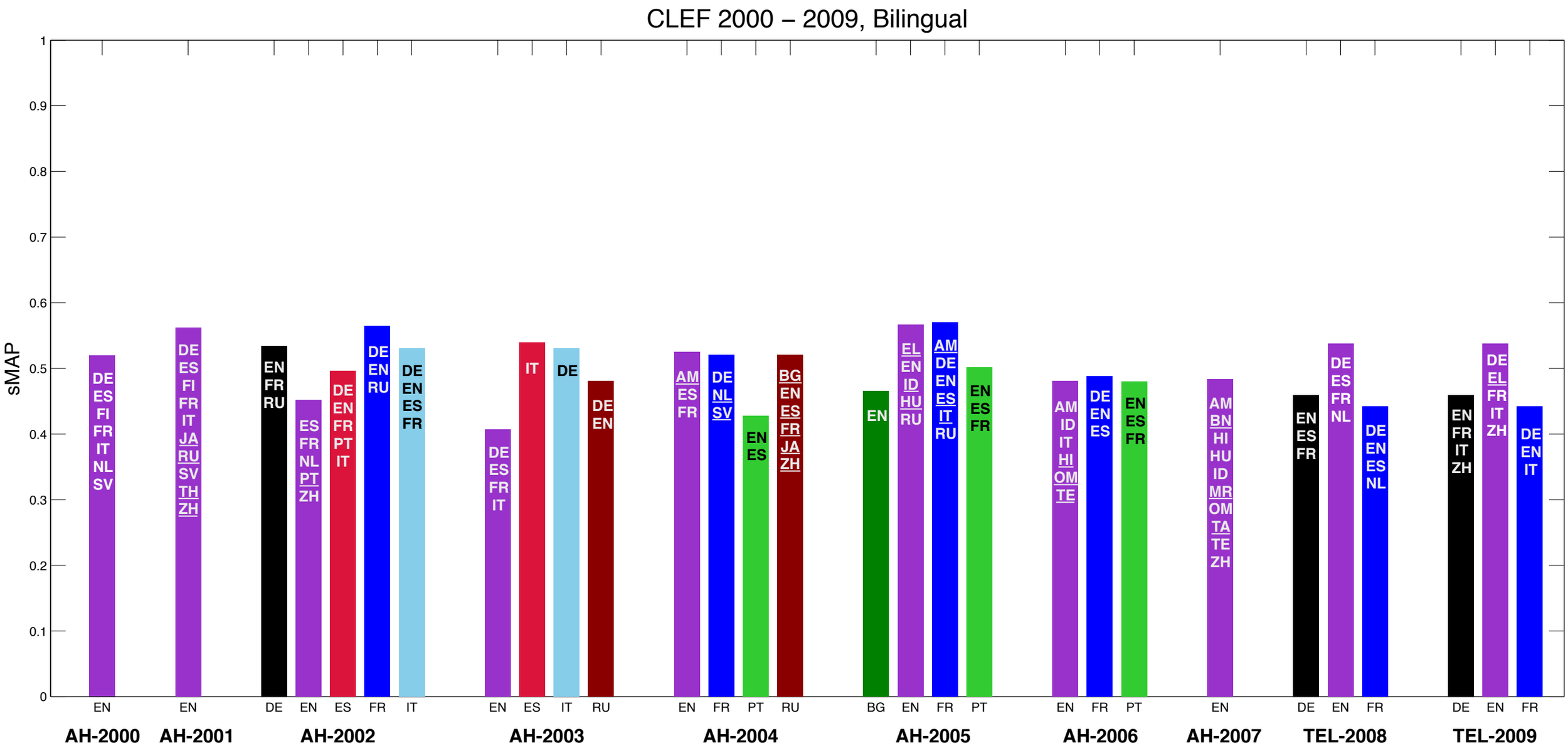
RQ 2

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

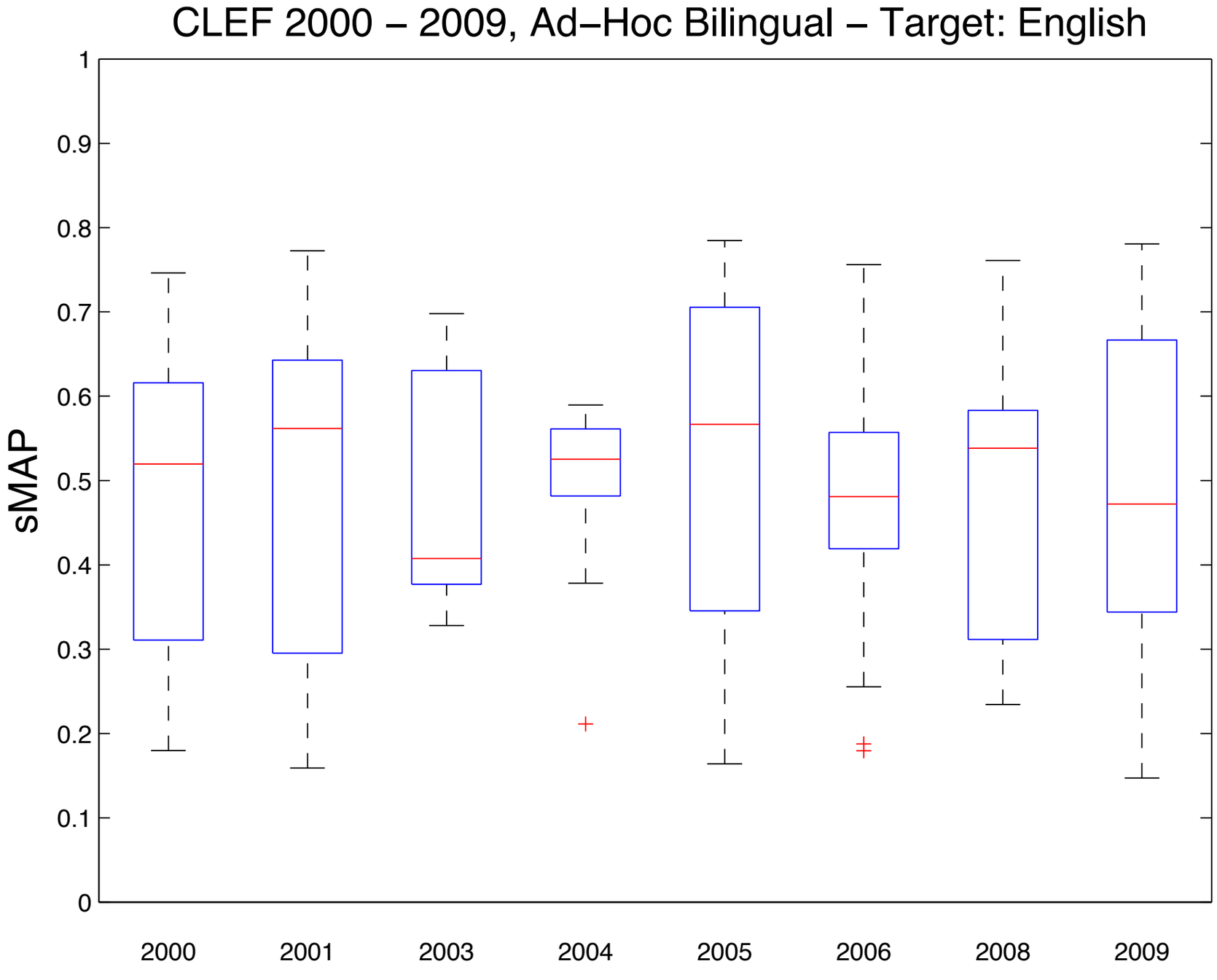
RQ 2

Do performances of bilingual systems increase over the years?



RQ 2

Do performances of bilingual systems increase over the years?



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%)
	2007	10(9)	67	.7746 (+2.47%)	.4835 (0.56%)
TEL Bili EN	2008	8(7)	24	.7611 (-1,74%)	.5382 (+11.31%)
	2009	10(7)	43	.7808 (2.59%)	.4719 (-12.32%)
AH Bili ES	2002	7(-)	16	**.6805** (-)	.4969 (-)
	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?

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%)
	2007	10(9)	67	.7746 (+2.47%)	.4835 (0.56%)
TEL Bili EN	2008	8(7)	24	.7611 (-1,74%)	.5382 (+11.31%)
	2009	10(7)	43	.7808 (2.59%)	.4719 (-12.32%)
AH Bili ES	2002	7(-)	16	**.6805** (-)	.4969 (-)
	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.**

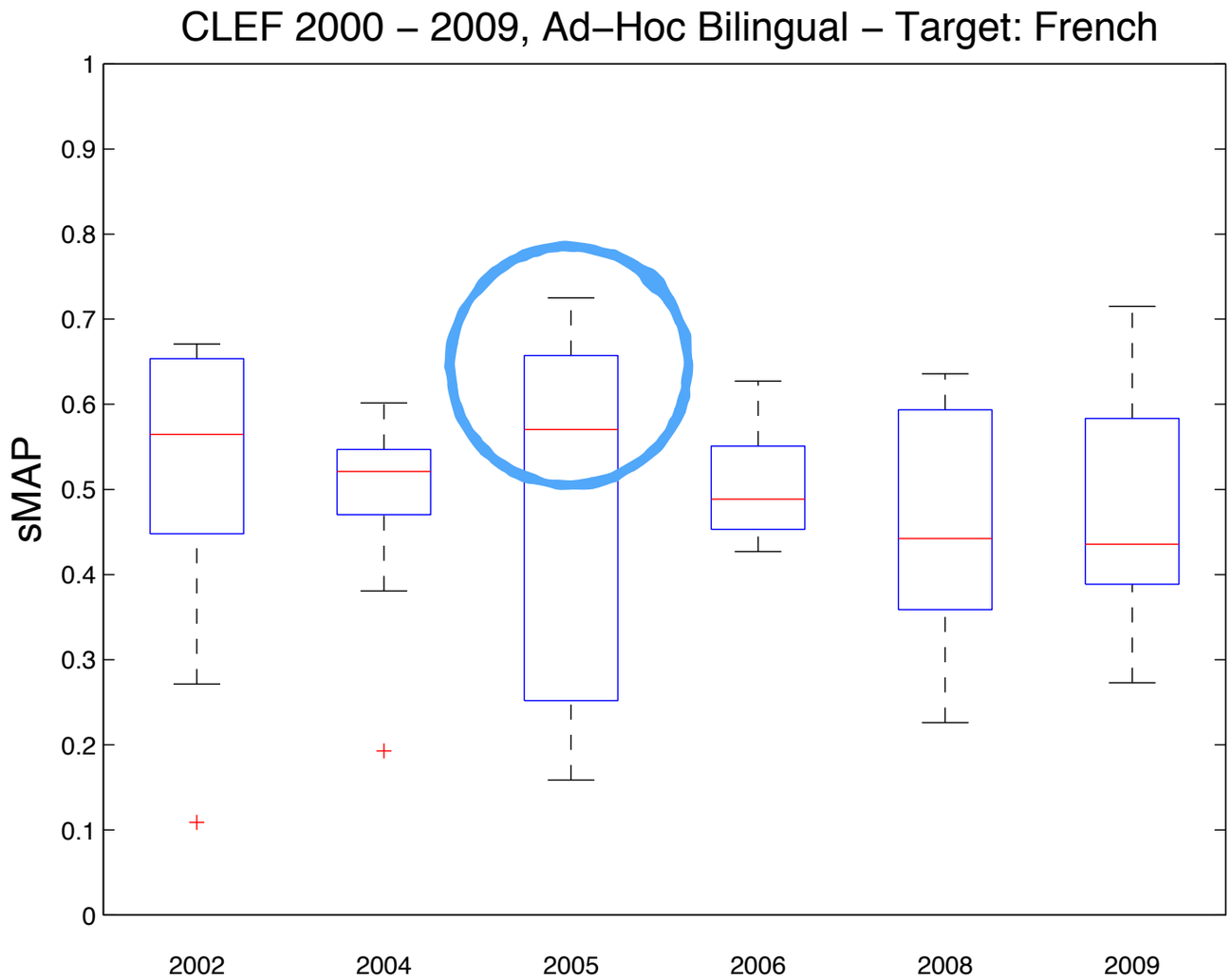
- This is an indicator of the **improvement of language resources** – e.g. dictionaries, external resources like Wikipedia and the use of semantic rather than syntactic resources
  - CLEF Continued Impact!

s	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

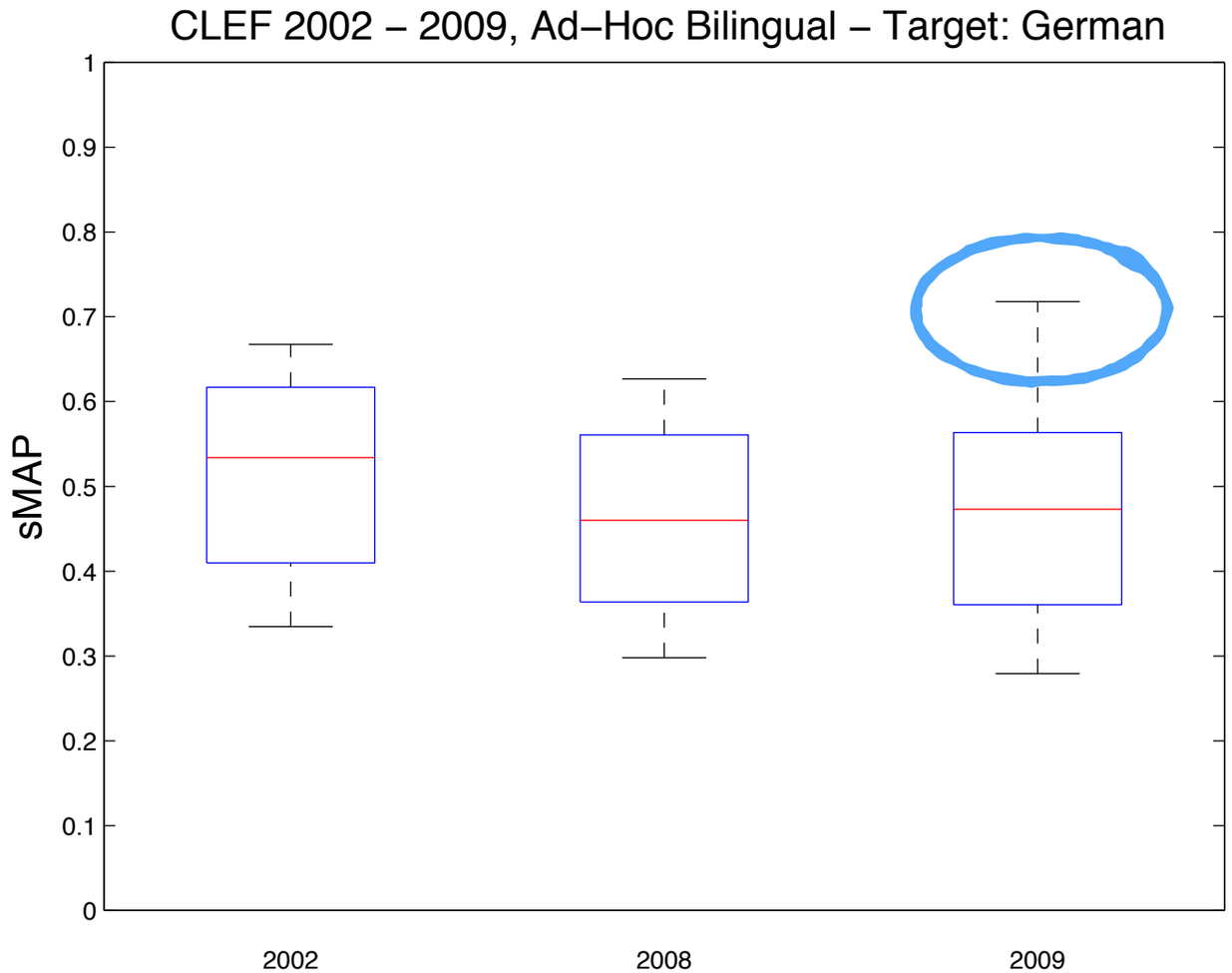
- The best bilingual system for the “X2FR” task exploited seven different machine translation systems, three bilingual dictionaries and ten freely available translation tools.



RQ 2: Do bilingual systems improve over the years?

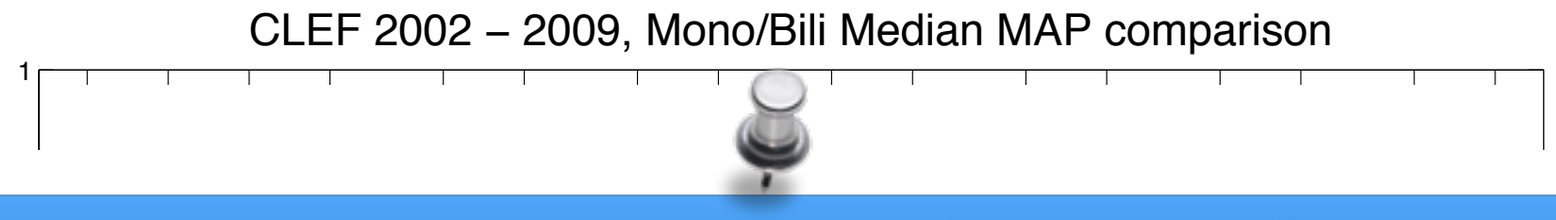
Effect of the improvement of language resources

- 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

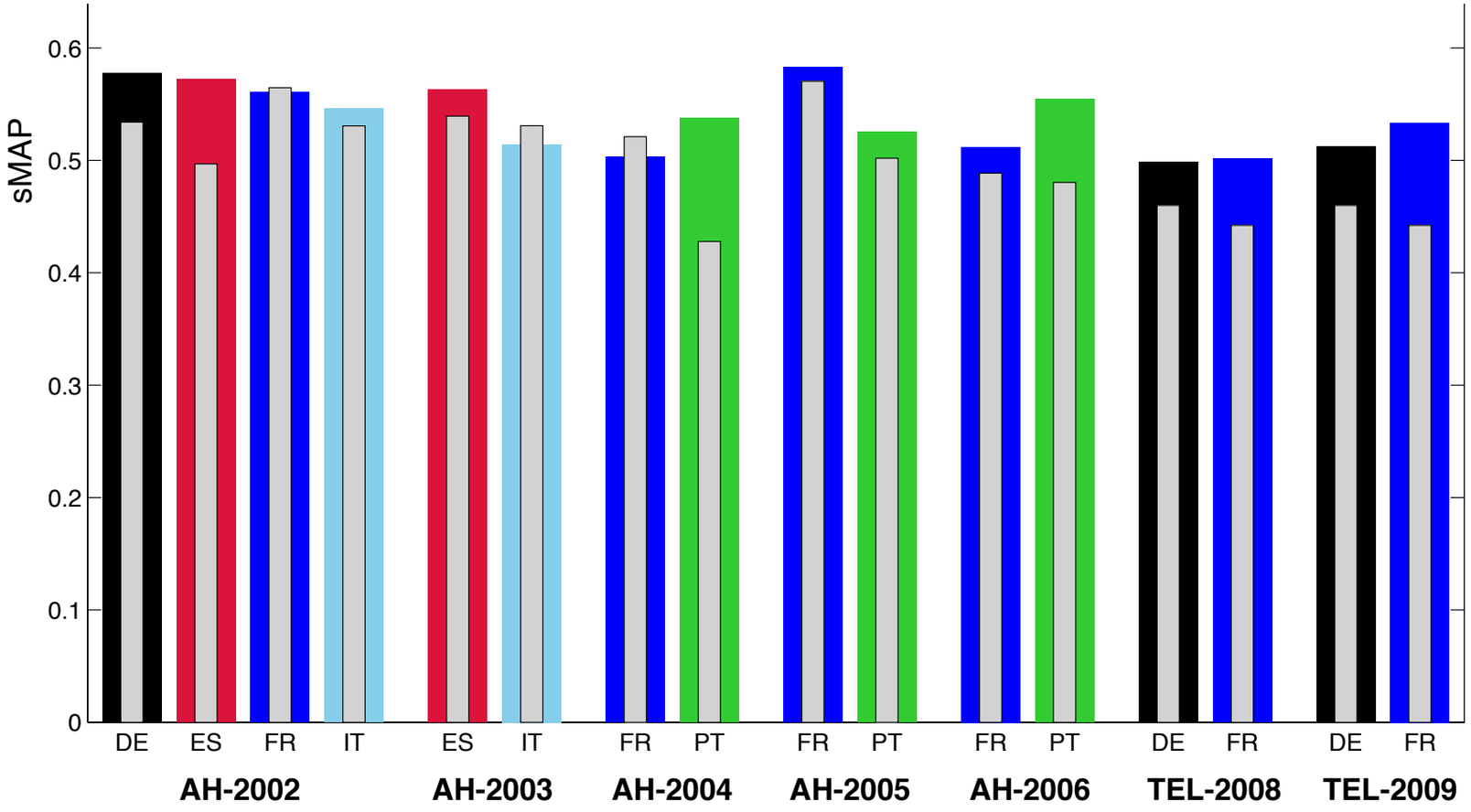


RQ3

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



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



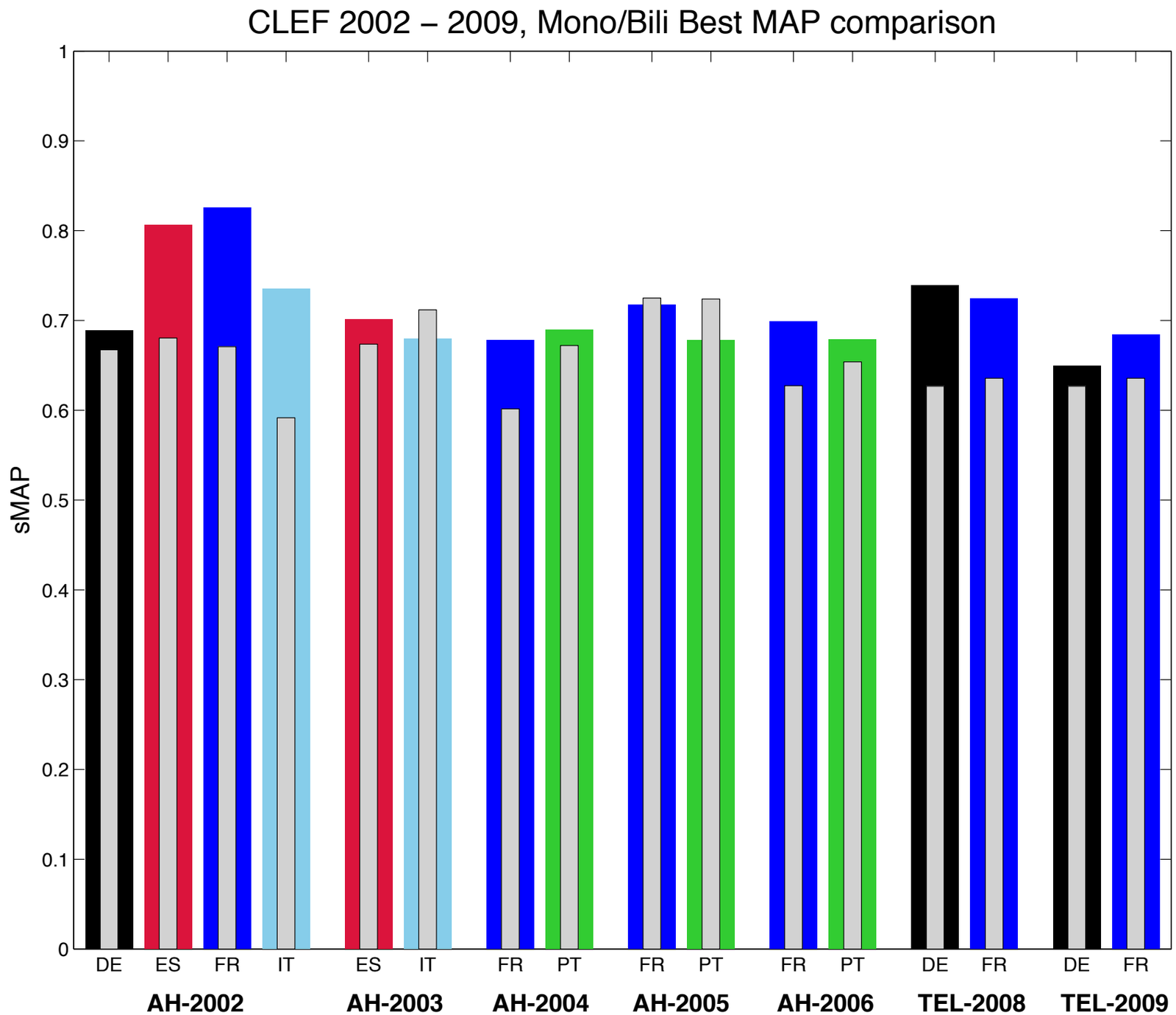
RQ3

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…

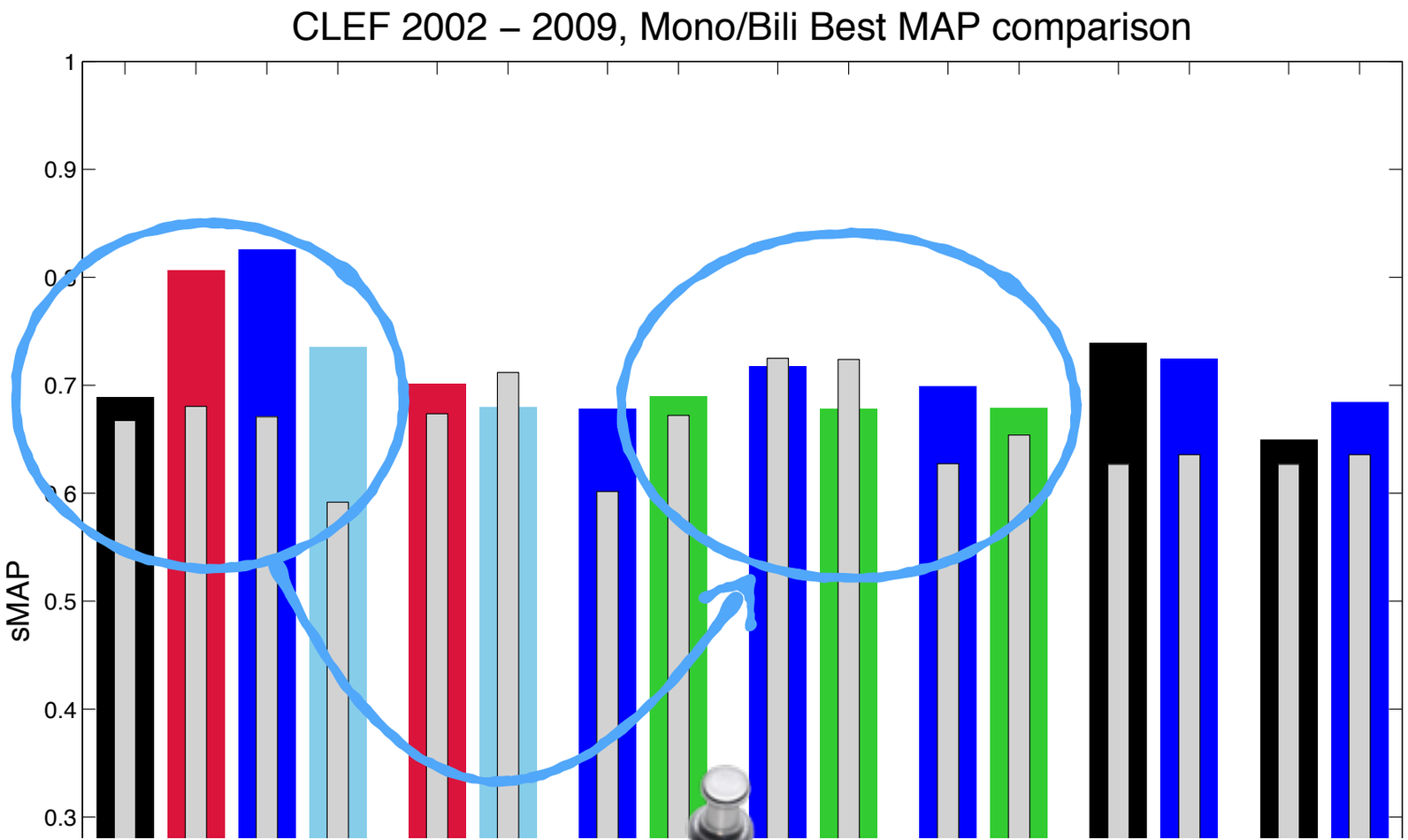
RQ3

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

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

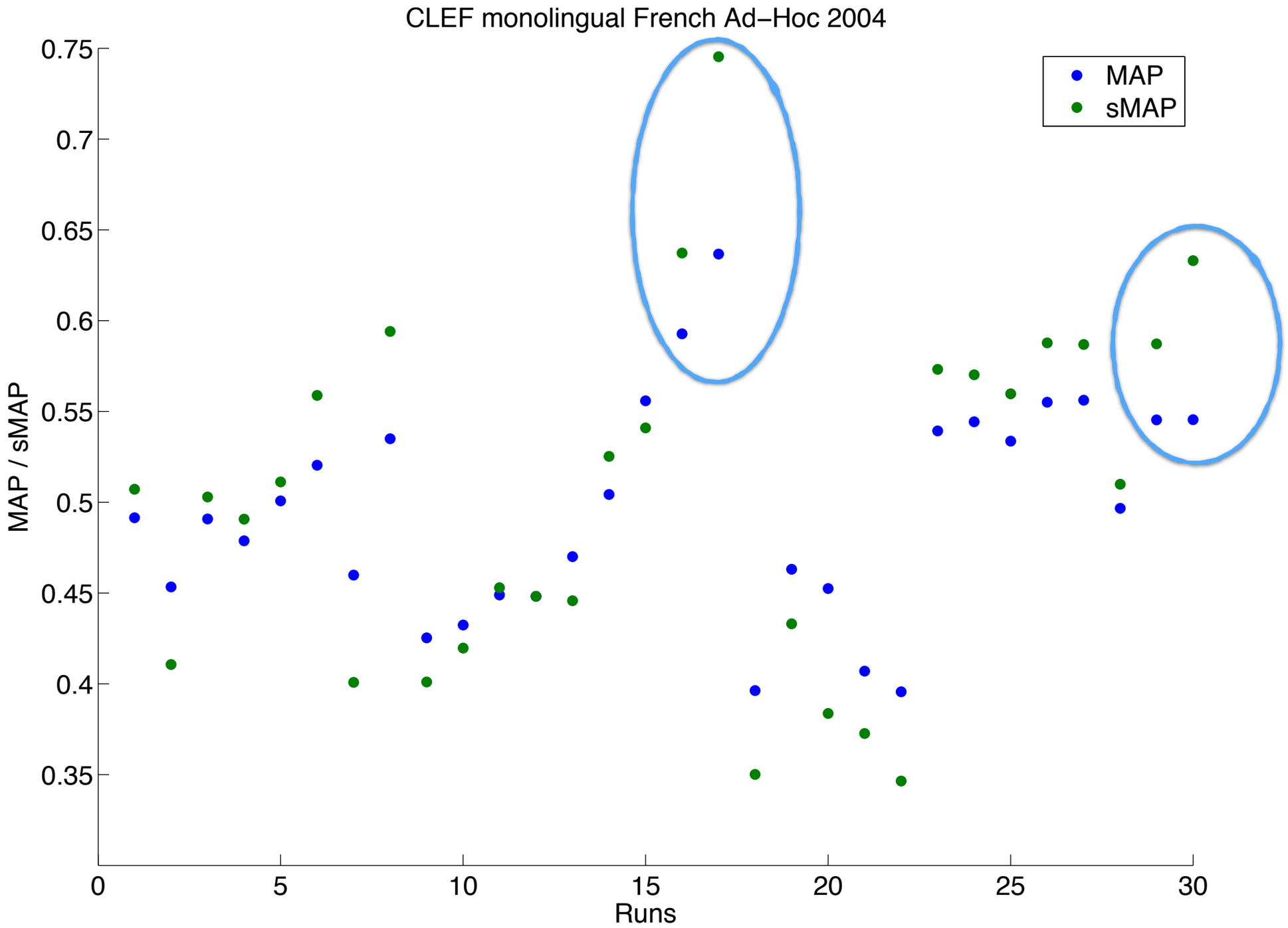
Discussion and Future Works

Something we must be aware of

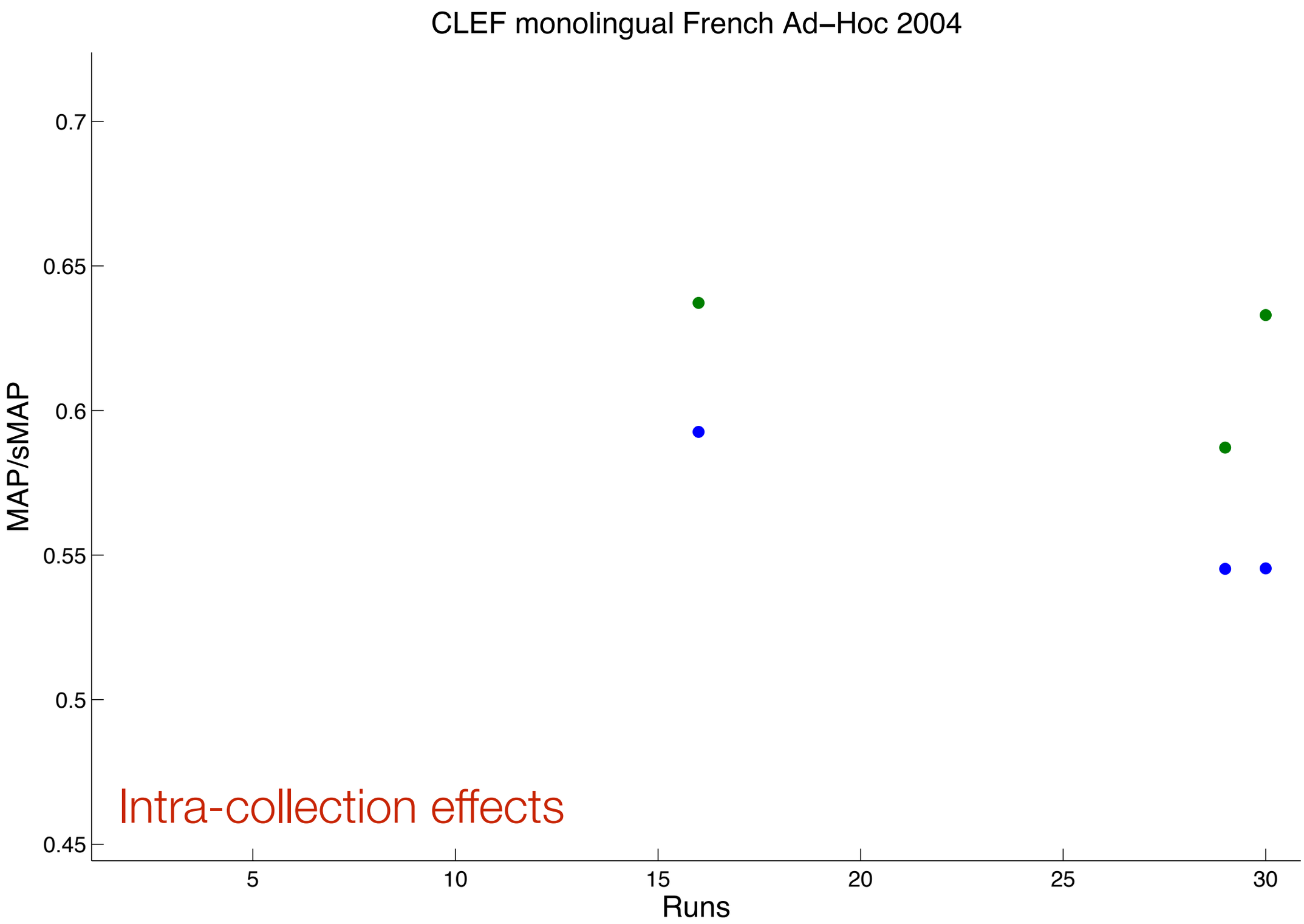
The effect of standardization of system performances could depend by the number of runs in the standardized collection: the higher the better

[Webber et al. 2008] reports that standardized AP over collections with at least 10 to 15 systems have a high Kendall's τ correlation with AP

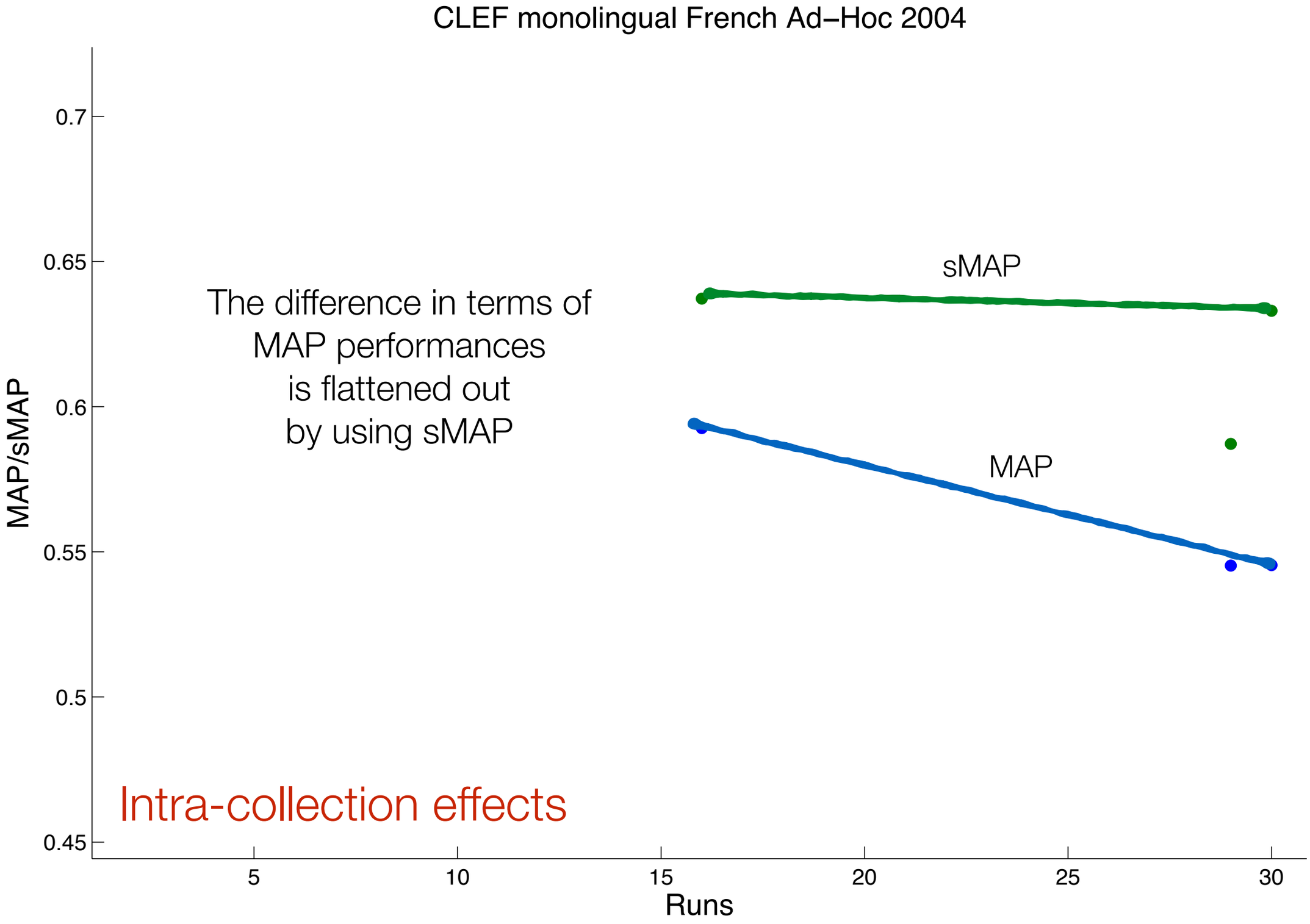
Something we must be aware of



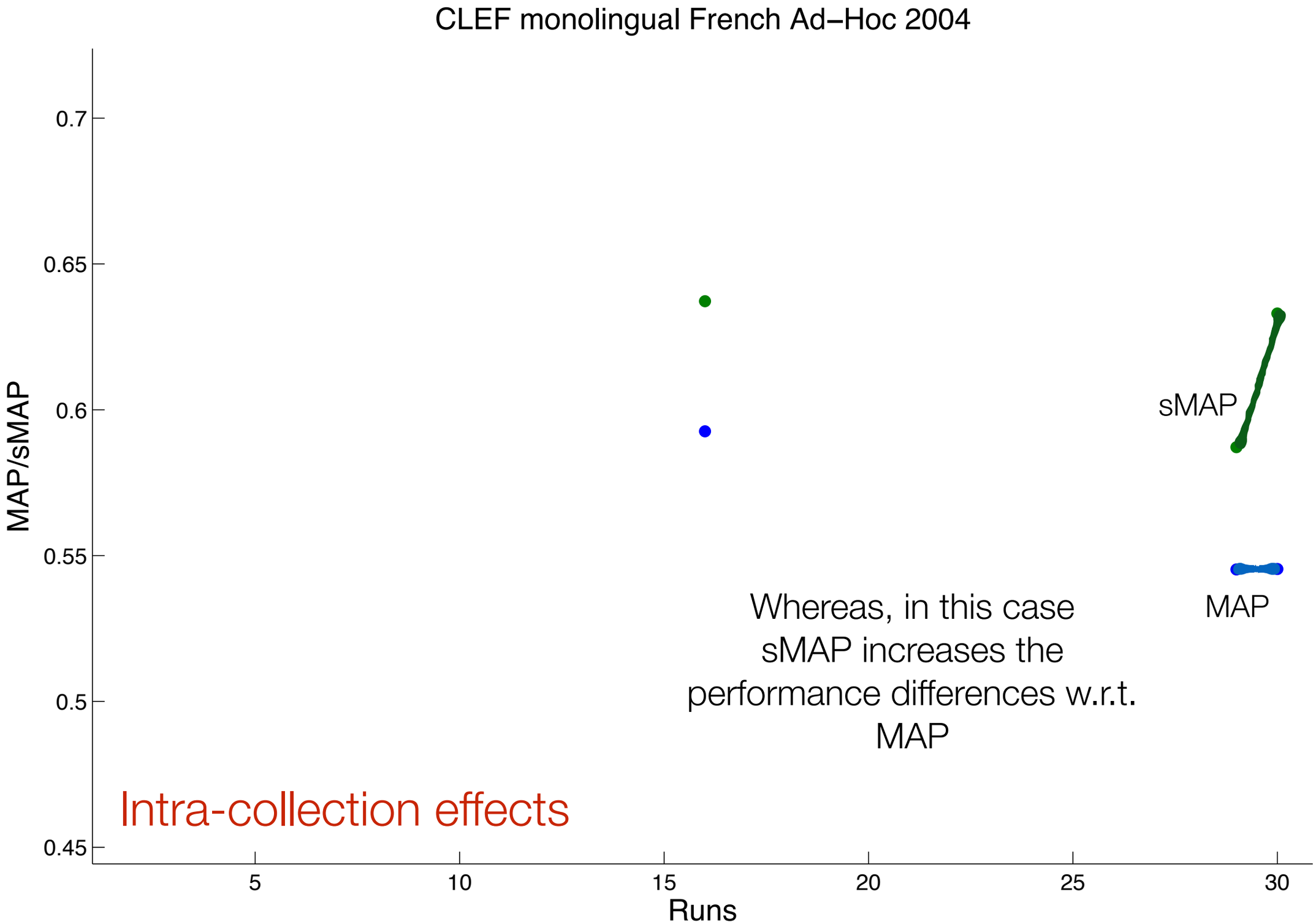
Something we must be aware of



Something we must be aware of



Something we must be aware of



A brief analysis with CLEF data

CLEF Ad-Hoc Bilingual English Collections

High correlation τ threshold = 0.9

	# Runs	Kendall's τ
2000	26	0.907
2001	55	0.912
2003	15	0.866
2004	11	0.855
2007	67	0.947

A brief analysis with CLEF data

CLEF Ad-Hoc Monolingual French Collections

High correlation τ threshold = 0.9

	# Runs	Kendall's τ
2000	10	0.866
2001	15	0.942
2002	16	0.950
2003	35	0.920
2004	38	0.888
2006	27	0.903

Future Works

- Apply standardization to other largely-adopted IR measures with the aim of analysing system performances from different perspectives

- Aggregate and analyse the systems on the basis of adopted retrieval techniques to better understand their impact on overall performances across the years

- Extend the analysis of bilingual and multilingual systems grouping them on a source and target language basis thus getting more insights into the role of language morphology and linguistic resources in cross-lingual IR

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

KEEP CALM AND ASK ME QUESTIONS