know-how

innovation

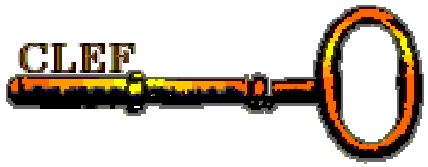
CLEF 2003
Overview of Results

Martin Braschler
Eurospider Information Technology AG
8006 Zürich, Switzerland
martin.braschler@eurospider.com

solution

eurospider
relevancy retrieval

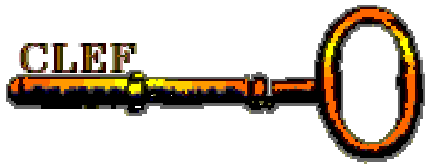
Outline



- Participants
- Experiment Details
- 4 Years of Growth
- Trends
- Effects
- Results
- Conclusions, Outlook

eurospider
relevancy retrieval

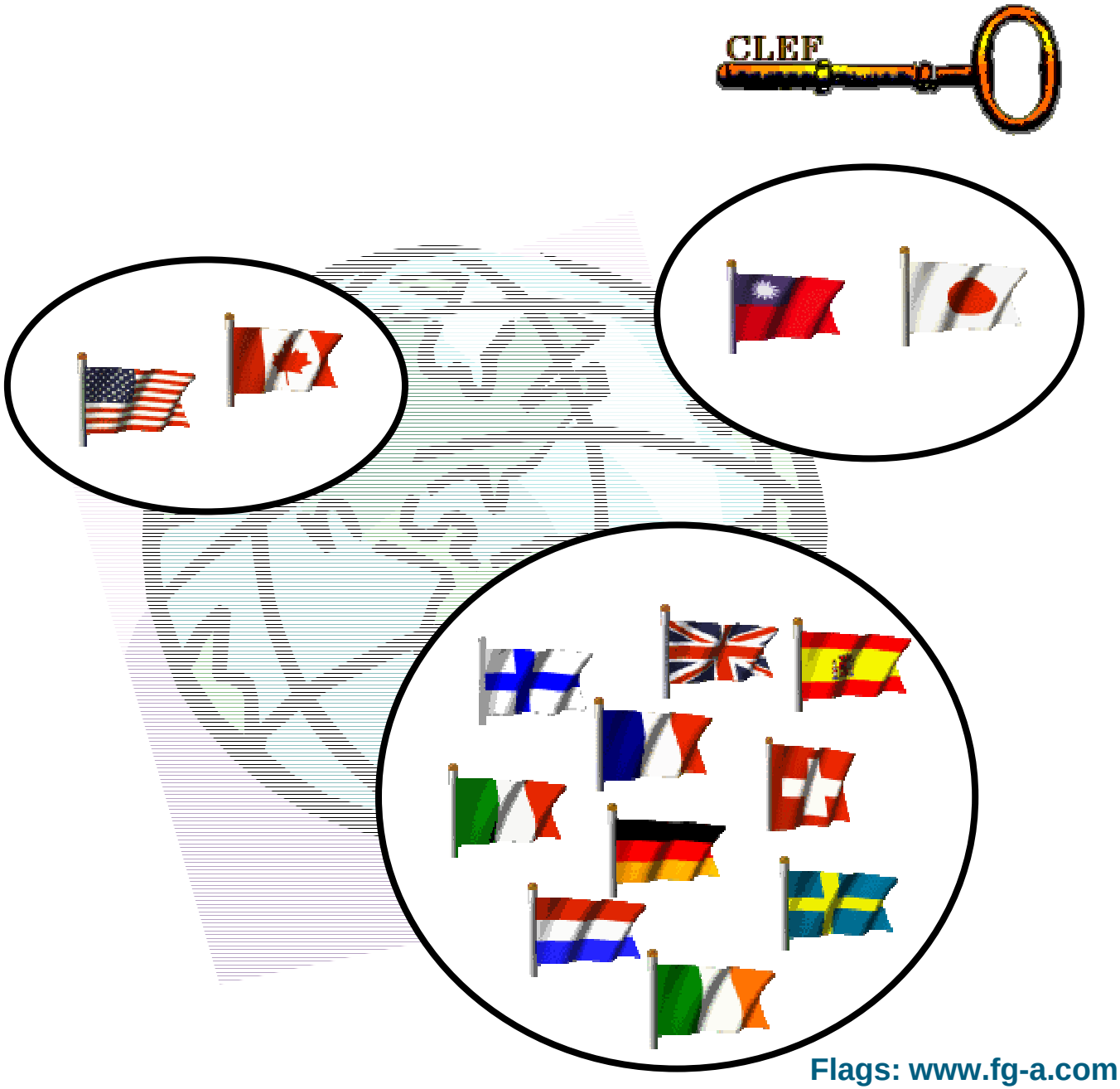
Participants



BBN/UMD (US)
CEA/LIC2M (FR)
CLIPS/IMAG (FR)
CMU (US) *
Clairvoyance Corp. (US) *
COLE Group/U La Coruna (ES) *
Daedalus (ES)
DFKI (DE)
DLTG U Limerick (IE)
ENEA/La Sapienza (IT)
Fernuni Hagen (DE)
Fondazione Ugo Bordoni (IT) *
Hummingbird (CA) **
IMS U Padova (IT) *
ISI U Southern Cal (US)
ITC-irst (IT) ***
JHU-APL (US) ***
Kermit (FR/UK)
Medialab (NL) **
NII (JP)
National Taiwan U (TW) **
OCE Tech. BV (NL) **
Ricoh (JP)
SICS (SV) **
SINAI/U Jaen (ES) **
Tagmatica (FR) *
U Alicante (ES) **
U Buffalo (US)
U Amsterdam (NL) **
U Exeter (UK) **
U Oviedo/AIC (ES)
U Hildesheim (DE) *
U Maryland (US) ***
U Montreal/RALI (CA) ***
U Neuchâtel (CH) **
U Sheffield (UK) ***
U Sunderland (UK)
U Surrey (UK)
U Tampere (FI) ***
U Twente (NL) ***
UC Berkeley (US) ***
UNED (ES) **

42 participants, 14 different countries.
(*/**/*** = one, two, three previous participations)

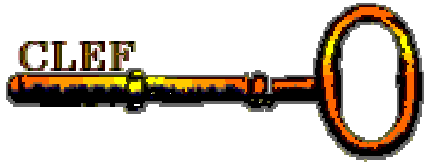
CLEF's Global Reach

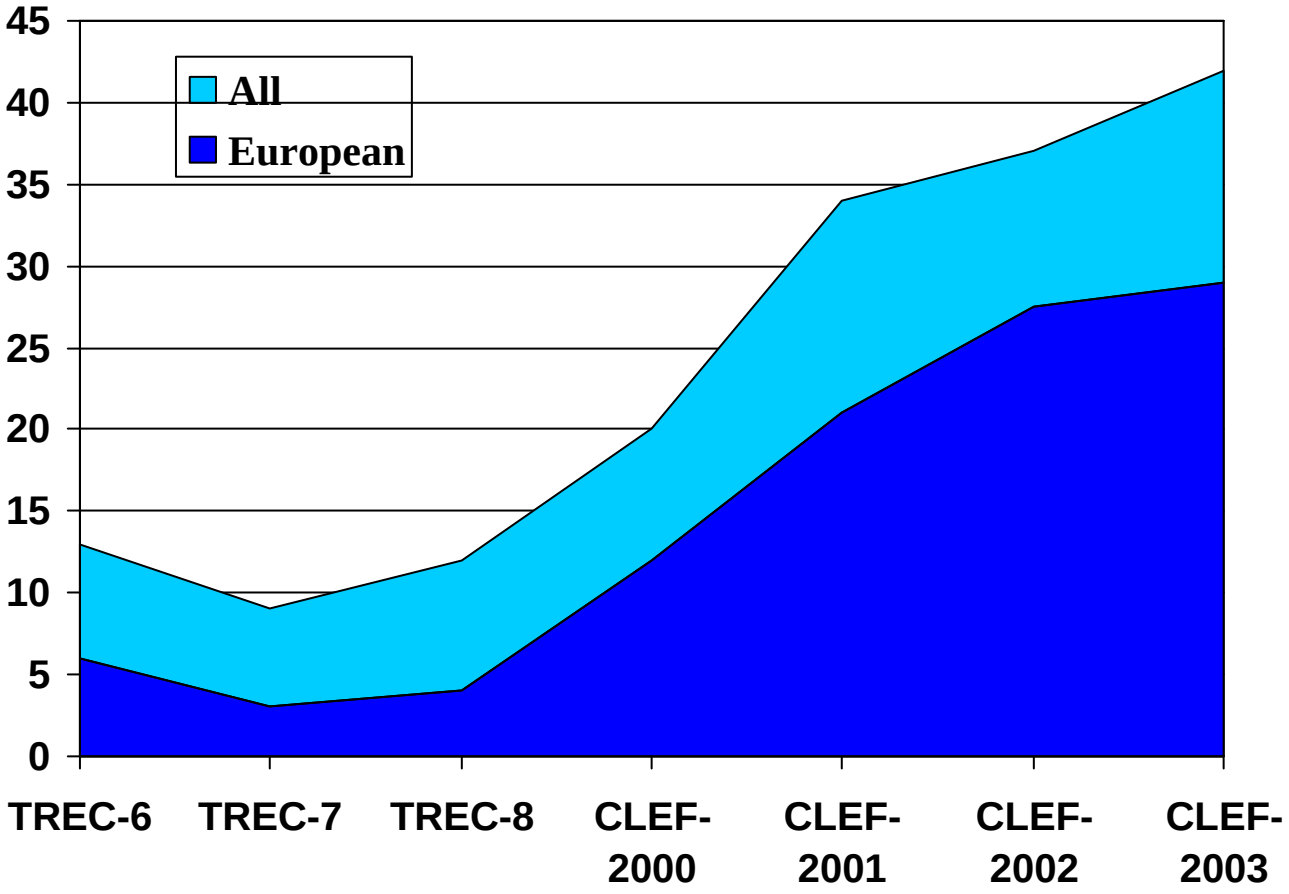


Flags: www.fg-a.com

eurospider
relevancy retrieval

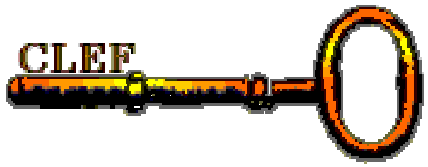
CLEF Growth
(Number of Participants)





eurospider
relevancy retrieval

The CLEF Multilingual Collection (Core Tracks)



	# part.	# lg.	# docs.	Size in MB	# assess.	# topics	# ass. per topic
CLEF 2003	33	9	1,611,178	4124	188,475	60 (37)	~3100
CLEF 2002	34	8	1,138,650	3011	140,043	50 (30)	~2900
CLEF 2001	31	6	940,487	2522	97,398	50	1948
CLEF 2000	20	4	368,763	1158	43,566	40	1089
TREC8 CLIR	12	4	698,773	1620	23,156	28	827
TREC8 AdHoc	41	1	528,155	1904	86,830	50	1736
TREC7 AdHoc	42+4	1	528,155	1904	~80,000	50	~1600

eurospider
relevancy retrieval

Tasks in CLEF 2002

- **Multilingual as “main task”: documents in 8 or 4 languages, topics in 10 languages**
- **Bilingual tasks: only some specific, “interesting” combinations**
  - **FI → DE, IT → ES, DE → IT, FR → NL**
  - **English as target language: only newcomers or special cases**
  - **Russian as target language: free choice of topic language**
- **Monolingual tasks: 8 target languages**
- **Domain-specific: GIRT (German and English docs.), bi- and monolingual, extra resources available**
- **Interactive track, QA, ImageCLEF, SDR: see special overview talks**

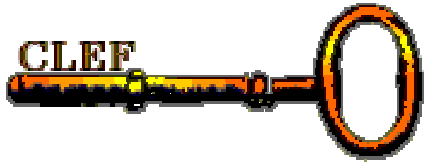
eurospider
relevancy retrieval

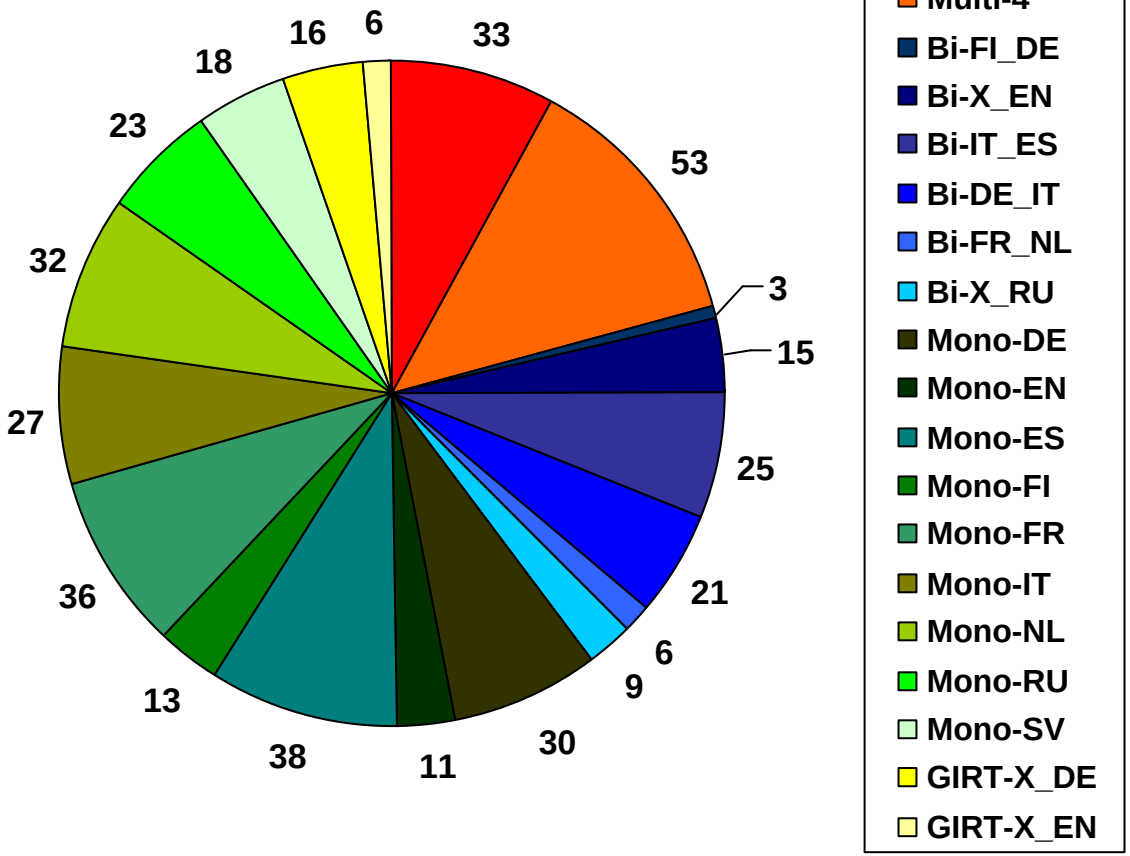
Details of Experiments

Track	*# Participants*	*# Runs/Experiments*
Multilingual-8	**7**	**33**
Multilingual-4	**14**	**53**
Bilingual to FI → DE	**2**	**3**
Bilingual to X → EN	**3**	**15**
Bilingual to IT → ES	**9**	**25**
Bilingual to DE → IT	**8**	**21**
Bilingual to FR → NL	**3**	**6**
Bilingual to X → RU	**2**	**9**
Monolingual DE	**13**	**30**
(Monolingual EN)	**(5)**	**11**
Monolingual ES	**16**	**38**
Monolingual FI	**7**	**13**
Monolingual FR	**16**	**36**
Monolingual IT	**13**	**27**
Monolingual NL	**11**	**32**
Monolingual RU	**5**	**23**
Monolingual SV	**8**	**18**
Domain-specific GIRT → DE	**4**	**16**
Domain-specific GIRT → EN	**2**	**6**
Interactive	**5**	
Question Answering	**8**	
Image Retrieval	**4**	
Spoken Document Retrieval	**4**	

eurospider
relevancy retrieval

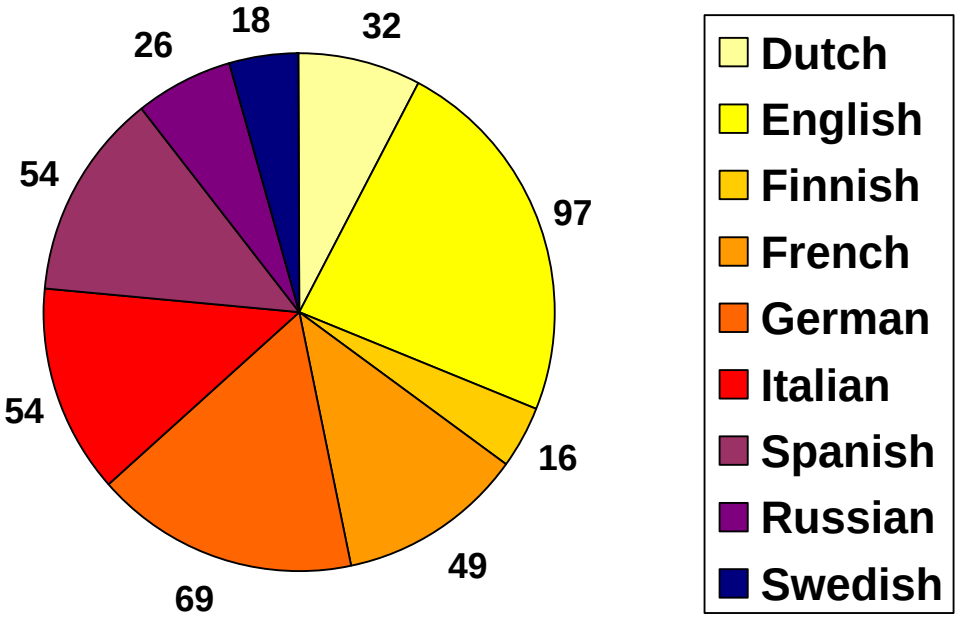
Runs per Task (Core Tracks)



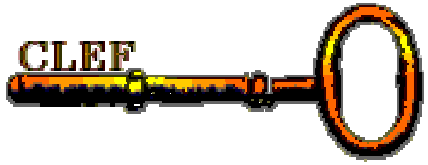


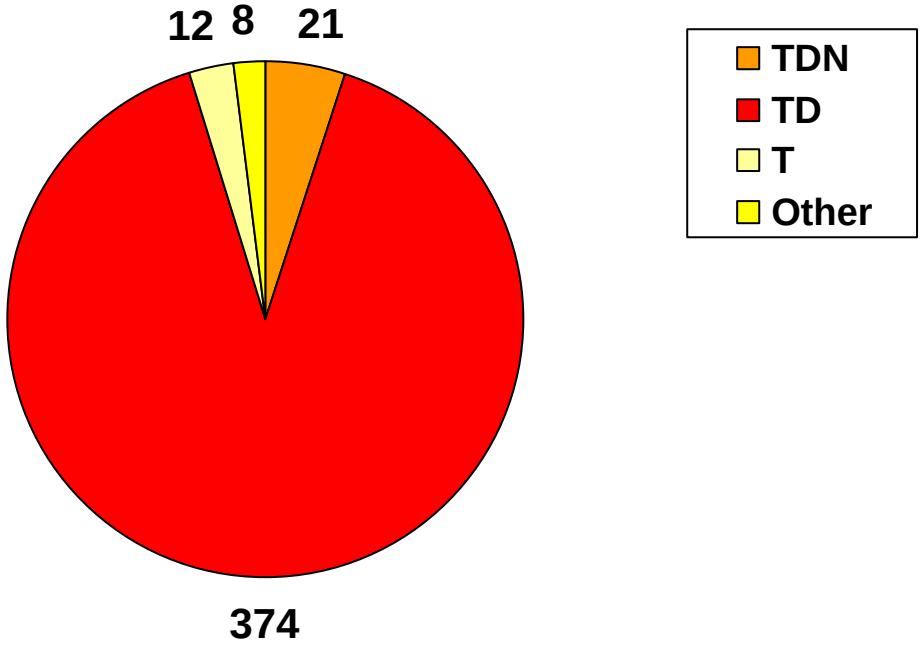
eurospider
relevancy retrieval

Runs per Topic Language (Core Tracks)



Topic Fields (Core Tracks)





eurospider
relevancy retrieval

Pooling

- **“Tool” to handle the size of relevance assessment work**
- **209 of 415 runs assessed**
- **Some tasks had all runs assessed: Bilingual to German and Russian, GIRT, Monolingual Finnish, Russian, Swedish**
- **Runs are pooled respecting nearly a dozen criteria:**
  - **participant’s preferences**
  - **“originality” (task, topic fields, languages, ..)**
  - **participant/task coverage**
  - **..**

eurospider
relevancy retrieval

Results from Pool Analysis

Pool testing

Simulation of "What would have happened if a group did not participate"?

Gives indication of reusability of test collection: are results of non-participants valid?

Mean absolute diff.	0.0008	Mean diff. in %	0.48%
Max absolute diff.	0.0030	Max diff. in %	1.76%
Standard deviation	0.0018	Standard dev. %	1.01%

- **Figures are calculated that show how much measures change for non-participants**
- **Values a bit higher for individual languages, espec. the "new" languages FI and SV**
- **Rankings are very stable! Figures compare very favorably to similar evaluations**

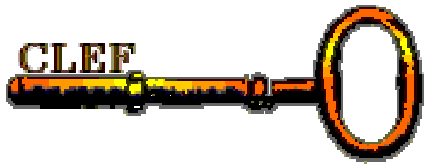
eurospider
relevancy retrieval

Preliminary Trends for CLEF-2003 (1)

- **A lot of detailed fine-tuning (per language, per weighting scheme, per transl. resource type)**
- **People think about ways to “scale” to new languages**
- **Merging is still a hot issue; however, no merging approach besides the simple ones has been widely adopted yet**
- **A few resources were really popular: Snowball stemmers, UniNE stopwordlists, some MT systems, “Freelang” dictionaries**
- **QT still rules**

eurospider
relevancy retrieval

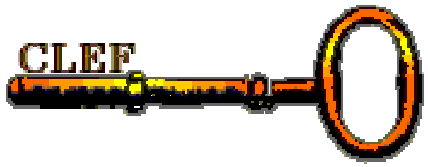
Preliminary Trends for CLEF-2003 (2)



- **Stemming and decompounding are still actively debated; maybe even more use of linguistics than before?**
- **Monolingual tracks were “hotly contested”, some show very similar performance among the top groups**
- **Bilingual tracks forced people to think about “inconvenient” language pairs**

eurospider
relevancy retrieval

CLEF-2003 vs. CLEF-2002



- **Many participants were back**
- **People try each other's ideas/methods:**
  - **collection-size based merging, 2step merging**
  - **(fast) document translation**
  - **compound splitting, stemmers**
- **Returning participants usually improve performance. ("Advantage for veteran groups")**
- **Scaling up to Multilingual-8 takes its time (?)**

eurospider
relevancy retrieval

"Effect" of CLEF in 2002 (recycled slide)

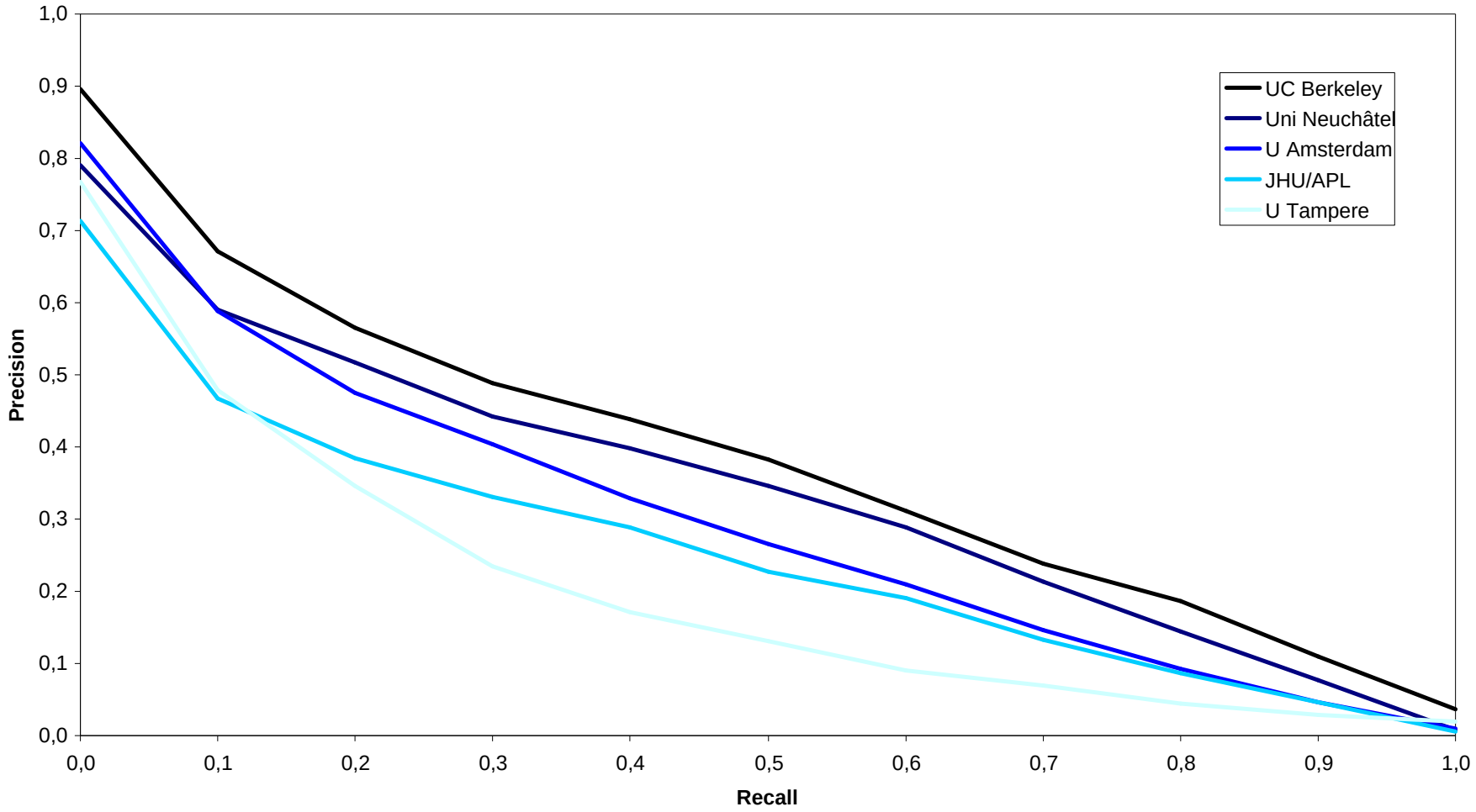
- **Number of European groups still growing (27,5!)**
- **Very sophisticated fine-tuning for individual languages**
  - **BUT: are we overtuning to characteristics of the CLEF collection?**
- **People show flexibility in adapting resources/ideas as they come along (architectures?)**
- **Participants move from monolingual → bilingual → multilingual**

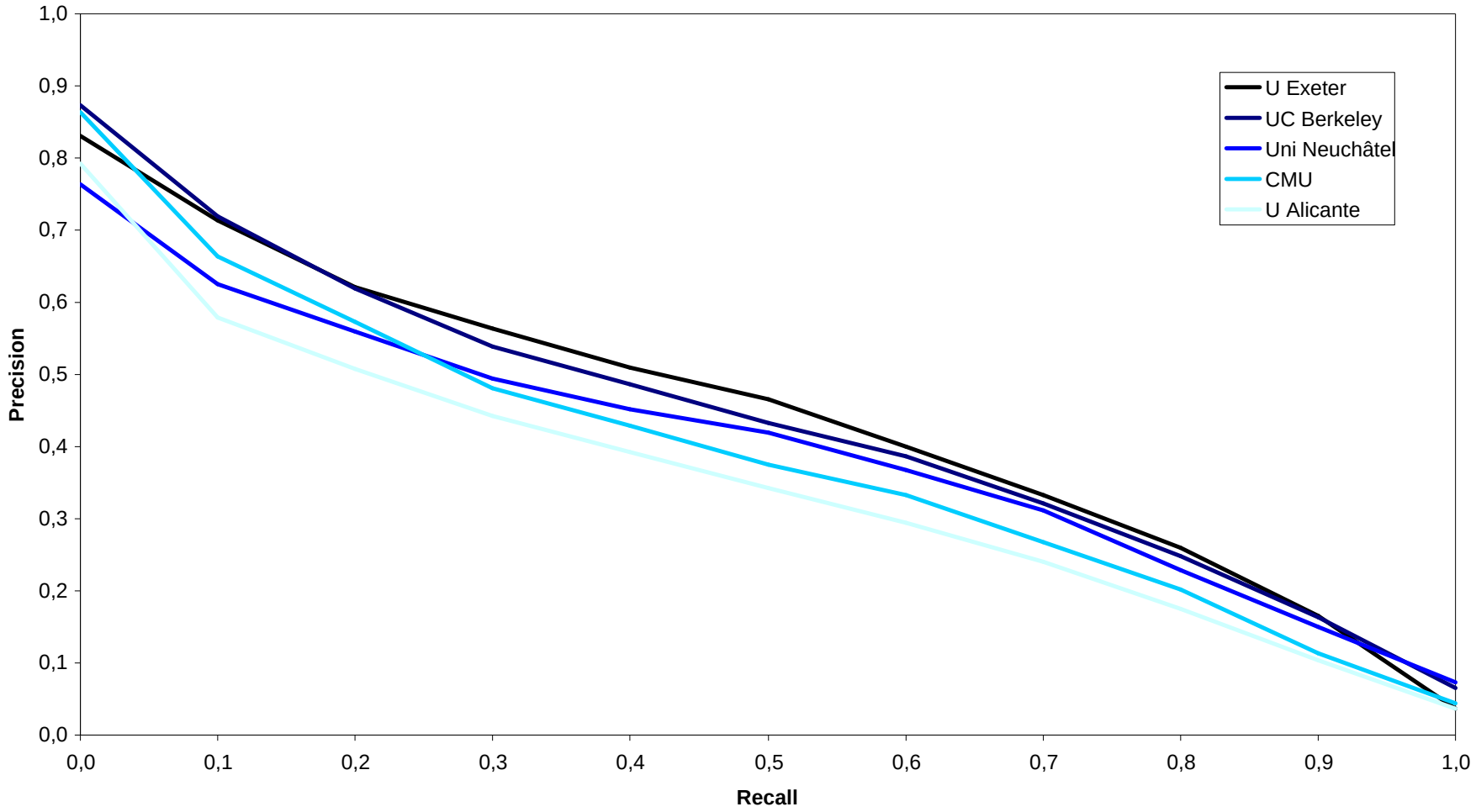
eurospider
relevancy retrieval

"Effect" of CLEF in 2003

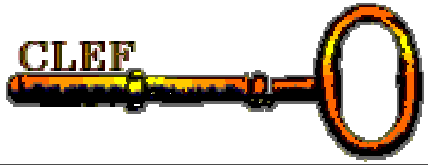
- **Number of European grows more slowly (29)**
- **Fine-tuning for individual languages, weighting schemes etc. has become a hot topic**
  - **The question remains: are we overtuning to characteristics of the CLEF collection?**
- **Some blueprints to "successful CLIR" have now been widely adopted**
  - **Are we headed towards a monoculture of CLIR systems?**
- **Multilingual-8 was dominated by veterans, but Multilingual-4 was very competitive**
- **Participants had to deal with "inconvenient" language pairs for bilingual; stimulating some interesting work**

eurospider
relevancy retrieval


CLEF 2003 Multilingual-8 Track - TD, Automatic
Precision
Recall
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0
UC Berkeley
Uni Neuchâtel
U Amsterdam
JHU/APL
U Tampere


CLEF 2003 Multilingual-4 Track - TD, Automatic
Precision
Recall
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0,0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0
U Exeter
UC Berkeley
Uni Neuchâtel
CMU
U Alicante

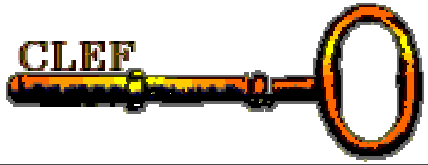
Bilingual Tasks



Task	Top Perf. (TD)	Diff. To 5th Place
Bilingual FI->DE	UC Berkeley	-
Bilingual X->EN	Daedalus	-
Bilingual IT->ES	U Alicante	+8.2%
Bilingual DE->IT	JHU/APL	+20.2%
Bilingual FR->NL	JHU/APL	-
Bilingual X->RU	UC Berkeley	-

eurospider
relevancy retrieval

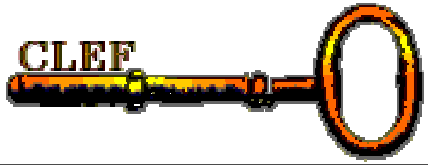
Monolingual Tasks



Task	Top Perf. (TD)	Diff. To 5th Place
Monol. DE	Hummingbird	+12.3%
Monol. ES	F. U. Bordoni	+7.3%
Monol. FI	Hummingbird	+17.2%
Monol. FR	U Neuchâtel	+2.4%
Monol. IT	F. U. Bordoni	+9.1%
Monol. NL	Hummingbird	+10.4%
Monol. RU	UC Berkeley	+28.0%
Monol. SV	UC Berkeley	+25.3%

eurospider
relevancy retrieval

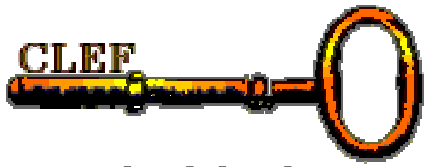
GIRT Tasks



Task	Top Perf. (TD)	Diff. To 5th Place
GIRT X->DE	UC Berkeley	-
GIRT X->EN	UC Berkeley	-

eurospider
relevancy retrieval

Conclusions and Outlook



- **Four years of CLEF campaigns are behind us, coupled with substantial growth**
- **CLIR as evaluated in the core tracks may be "matured"**
- **There is a lot of fine-tuning, BUT…**
- **Merging remains unsolved (?)**
- **How do we develop the core track to address the unresolved questions, but also open up new research challenges?**

eurospider
relevancy retrieval