Test Collection Construction: the CLEF Experience

Carol Peters – ISTI-CNR

Outline

- CLEF
  - Objectives, organisation
  - Evaluation methodology
  - CLEF test collections
- Lessons Learned

Cross-Language Evaluation Forum

Objectives

- Promote research and stimulate development of multilingual IR systems for European languages, through
  - Creation of evaluation infrastructure and organisation of regular evaluation campaigns for system testing
  - Building of an MLIA/CLIR research community
  - Construction of publicly available test-suites

Major Goal

- Encourage development of truly multilingual, multimodal systems

Cross-Language Evaluation Forum

Background

- Extension of CLIR track at TREC (1997-1999)
- Currently an activity of the DELOS Network of Excellence for Digital Libraries under FP6 – IST programme but …
- Mainly dependent on voluntary efforts
- Coordination is distributed:
  - National sites for each language in multilingual collection
  - Domain-experts responsible for work in individual tracks

CLEF 2004: Tracks

CLEF offers tracks designed to evaluate the performance of systems for:

- mono-, bi- and multilingual document retrieval on news collections (Ad-hoc)
- mono- and cross-lang. domain-specific retrieval (GIRT)

2001

- interactive cross-language retrieval (iCLEF)

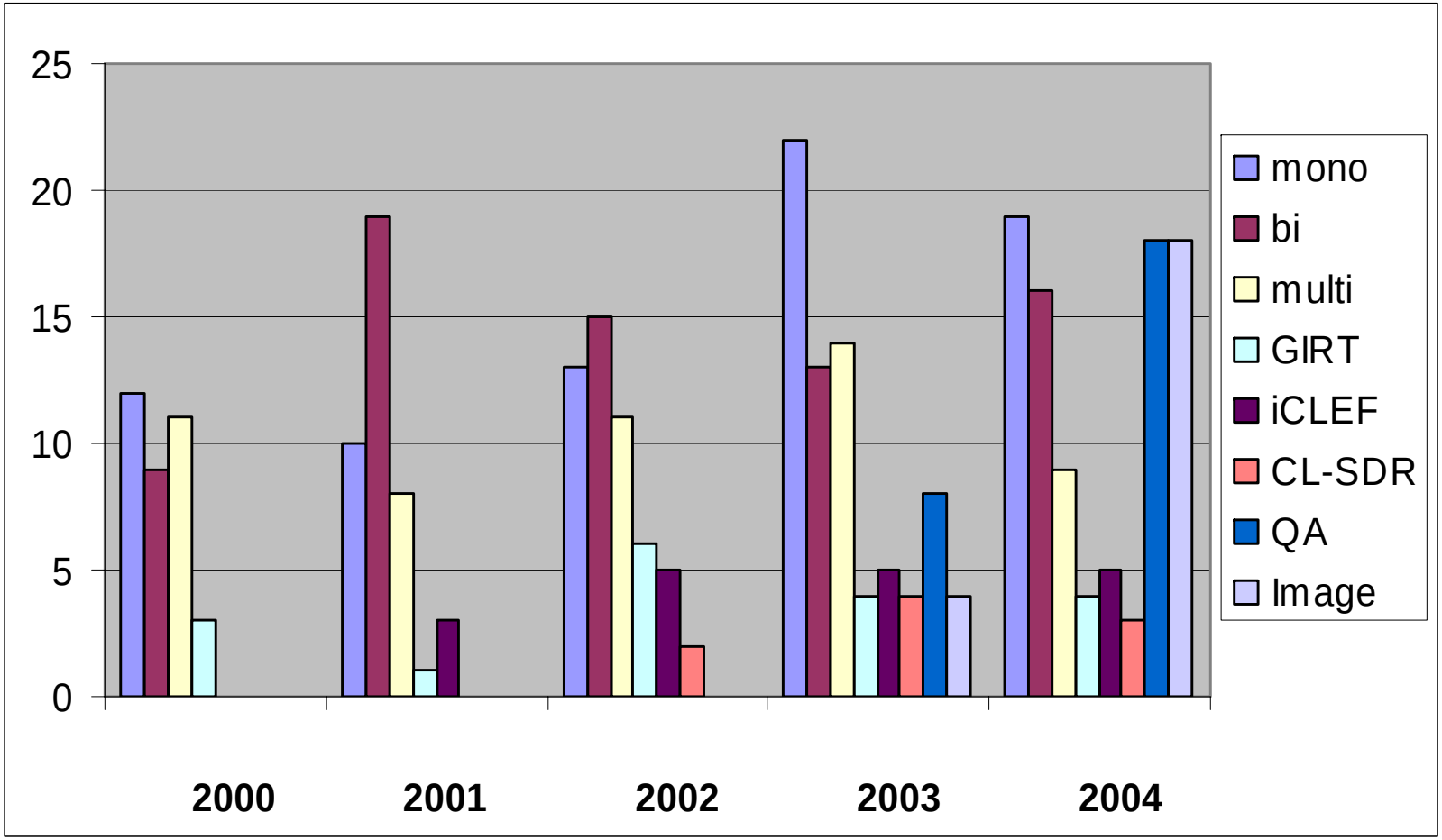
2002

- cross-lang. spoken doc. retrieval (CL-SDR)

2003

- multiple lang. question answering (QA@CLEF)
- cross-lang. retrieval on image collections (ImageCLEF)

CLEF 2000 – 2004
Shift in Focus



CLEF 2004: Coordination

- ISTI-CNR, Pisa, Italy *(Main Coordinators)*
- ITC-irst, Trento, Italy *(QA@CLEF, CL-SDR)*
- Inst. for Advanced Computer Studies, U. Maryland, USA *(iCLEF)*
- Dept. Computer Sci and Information Systems, U.Limerick, Ireland (*QA@CLEF)*
- Department of Information Studies, University of Sheffield, UK *(ImageCLEF)*
- Department of Information Studies, University of Tampere, Finland *(Ad-Hoc)*
- Eurospider Information Technology AG, Zürich, Switzerland *(Ad-Hoc, GIRT)*
- ELRA, Paris, France *(Ad-Hoc, QA@CLEF, Negotiations with Data Providers)*
- German Res, Centre for Artificial Intelligence, DFKI, Saarbrücken (*QA@CLEF)*
- Info & Language Processing Systems, U.Amsterdam, The Netherlands(*QA@CLEF)*
- InformationsZentrum Sozialwissenschaften, Bonn, Germany (*Ad-Hoc, GIRT)*
- LSI-UNED, Madrid, Spain *(iCLEF, QA@CLEF)*
- Linguateca Sintef, Oslo, Norway; U.Minho, Braga, Portugal *(Ad-hoc, QA@CLEF)*
- Linguistic Modelling Laboratory, Bulgarian Academy of Sciences (*QA@CLEF)*
- National Institute of Standards and Technology, USA *(Ad-hoc)*
- School of Computing, Dublin City University, Ireland (*CL-SDR)*
- University Hospitals of Geneva, Switzerland *(ImageCLEF)*

Evaluation in CLEF

CLEF follows the Cranfield tradition

- Laboratory testing of retrieval systems first done in Cranfield II experiment (1963)
  - fixed document and query sets
  - evaluation based on relevance judgments
  - relevance abstracted to topical similarity
- Laboratory tests less expensive/more diagnostic

BUT

- Cranfield tests used small collections and assessed relevance for whole collections
- TREC and CLEF have very big collection size - thus adopt pooling methodology

Organising an Evaluation Activity

- select control task(s)
- provide data to test and tune systems
- define protocol and metrics to be used in results assessment
- disseminate Calls for Participation

Aim is an objective comparison between systems and approaches and creation of

Effective, Reliable and Reusable Test Collections

Test Collection

- Set of **documents** - must be representative of task of interest; must be large
- Set of **topics** - statement of user needs from which system data structure (query) is extracted
- Sets of **relevance judgments** for each topic against the document set
- **Metrics** and **measures** for results analysis

CLEF 2004 created 6 different test collections

Cross-Language Test Collections

Consistency harder to obtain than for monolingual

- parallel or comparable document collections
- multiple assessors per topic creation and relevance assessment (for each language)
- must take care when comparing different language evaluations (e.g., cross run to mono baseline)

Pooling – when needed - harder to coordinate

- need to have large, diverse pools for all languages
- retrieval results are not balanced across languages

CLEF Document Collections: Text

Multilingual comparable corpus of over 1,800,000 news documents in 10 languages

Built up over the years – aim is representative sample of European languages

- different languages
- different subcollections per language

Raw data (from providers):

- in different file formats
- different internal structure
- different encodings

Multilingual Text Corpus: Data Format

- Everything in one consistent SGML/XML format (XML new in 2003 for non-Latin encodings)
- Data validates against a DTD without any warning or errors
- Data is formatted cleanly in ISO Latin 1 (or UTF-8, for Russian)
- As much of the source information as possible is retained, even parts not used directly for CLEF
- Readme files details special characteristics of subcollections – real world inconsistencies kept
- Participant's instructions detail permissible tags/parts of the documents

CLEF2004 Main Text Collection

used in Ad Hoc, QA and interactive tracks

Collection	Added in	Size (MB)	No. of Docs	Median Size of Docs. (Bytes)
Dutch: Algemeen Dagblad 94/95	2001	241	106483	1282
Dutch: NRC Handelsblad 94/95	2001	299	84121	2153
English: LA Times 94	2000	425	113005	2204
English: Glasgow Herald 95	2003	154	56472	2219
Finnish: Aamulehti late 94/95	2002	137	55344	1712
French: Le Monde 94	2000	158	44013	1994
French: ATS 94	2001	86	43178	1683
French: ATS 95	2003	88	42615	1715
German: Frankfurter Rundschau94	2000	320	139715	1598
German: Der Spiegel 94/95	2000	63	13979	1324
German: SDA 94	2001	144	71677	1672
German: SDA 95	2003	144	69438	1693
Italian: La Stampa 94	2000	193	58051	1915
Italian: AGZ 94	2001	86	50527	1454
Italian: AGZ 95	2003	85	48980	1474
Portuguese: Público 1994	2004	164	51751	NA
Portuguese: Público 1995	2004	176	55070	NA
Russian: Izvestia 95	2003	68	16761	NA
Spanish: EFE 94	2001	511	215738	2172
Spanish: EFE 95	2003	577	238307	2221
Swedish: TT 94/95	2002	352	142819	2171

Other Document Collections

- Structured documents
  - GIRT social science database.
    - 150,000 docs with pseudo-parallel DE/EN corpus; controlled vocabularies in DE-EN and DE-RU
  - Amaryllis database
- Image Collections
  - St Andrews University: 28,133 historic photographs
  - University Hospitals Geneva: 8,725 medical images
- Spoken Document Collection
  - TREC-8 and TREC-9 SDR tracks
  - CLEF 2005 MALACH collection holocaust archives

Example from St Andrews Historic Photographic Collection

Record Id: JV-A.000460
Short title: The Fountain, Alexandria
Long title: Alexandria. The Fountain
Location: Dunbartonshire, Scotland
Description: Street junction with large Ornate fountain with columns, surrounded by rails…..
Date: Registered 17 July 1934
Photographer: L.Valentine & Co
Categories: [columns unclassified][street lamps -ornate][electric stret lighting] [shepherds][shops][streetscapes]
Notes: JV-A460 jf/mb

CLEF 2004 Topics

- Queries for ad hoc tasks could be formulated from 50 topics in 14 languages (including Amharic, Bulgarian, Chinese, Japanese)
- 200 questions for QA tasks prepared in seven languages
- 50 short topics for cross-language spoken doc. track prepared in 6 languages
- Topics in twelve languages for 3 different image retrieval tasks involving text and content-based retrieval techniques
- iCLEF task used topics in English, Spanish and French

Topic Creation Criteria

- Topics must be created according to particular system features to be tested
- CLEF 2004 Ad Hoc Topics
  - Structured topics simulate query “input” for range of IR applications, keyword-style input as well as natural language formulations.
  - Features include people & place names, acronyms, terminology…
- CLEF 2004 QA@CLEF topics
  - 8 question types: Location, Manner, Measure, Object, Organization, Person, Time, Other
- ImageCLEF
  - Topics must test both translation and image retrieval

Topic Creation: Method

- Topics are created wrt the collection
- Image CLEF: St Andrews Collection
  - representative topic set to test capabilities of both translation and image retrieval
  - broad categories obtained from log files analysis, discussion with librarians and reference to a categorisation scheme for picture archives
- Image CLEF: University Hospitals Geneva
  - Radiologists selected preliminary set of representative images and case-notes; final selection by track coordinators

Example Topic for Ad Hoc Tasks

<top><num> C205 </num>
<PT-title> Ataques suicidas tamil </PT-title>
<PT-desc> Encontrar algumas informação sobre ataques bombistas suicidas dos Tigres Tamil ou acções kamikazes no Sri Lanka. </PT-desc>
<PT-narr> Apenas documentos sobre ataques bombistas suicidas por rebeldes tamil são relevantes; outras formas de ataque não são importantes. </PT-narr>
</top>

Example Topics for QA@CLEF2004

- 0001 En quelle année Thomas Mann a-t-il obtenu le Prix Nobel ? (Time)
- 0002 Quel est le directeur général de FIAT ? (Person)
- 0003 Quel Était le nom du parti politique d'Hitler ? (Other)
- 0004 Quel constructeur automobile a produit la "Beetle" ? (Organization)
- 0005 Comment est mort Jimi Hendrix ? (Manner)

Example Topic for Image CLEF2004

ImageCLEF – ad hoc task

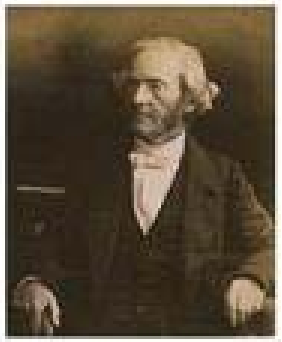
Example topics

<top>

<num> Number: 1 </num>

<title> Thomas Rodgerによる聖職者たちの自像画 </title>

</top>



Pictures of church ministers by Thomas Rodger

<top>

<num> Number: 2 </num>

<title> 1908صور لروما تحدث في أبريل </title>

</top>

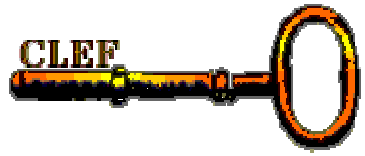
Pictures of Rome taken in April 1908

<top>

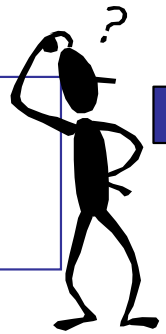
Relevance Assessments

- Relevance assessment in CLEF is performed in distributed mode by language/domain specialists
- Tight central coordination is needed
  - Ad hoc uses pooling system and binary judgements on relevance
  - ImageCLEF pooling, 3-way judgments, 3 sets of relevance judgements per topic/task
  - QA uses 4 values: correct/incorrect/unsupported/non-exact (measured accuracy and confidence weighted score)
  - iCLEF used same evaluation measures as QA

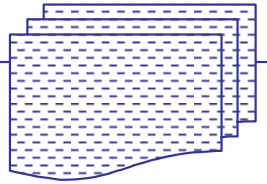
Using Pooling to Create Large Test Collections



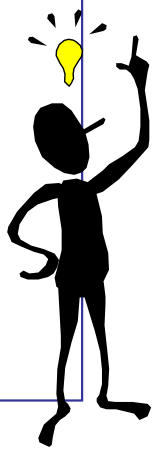
Assessors create topics.



A variety of different systems retrieve the top 1000 documents for each topic.

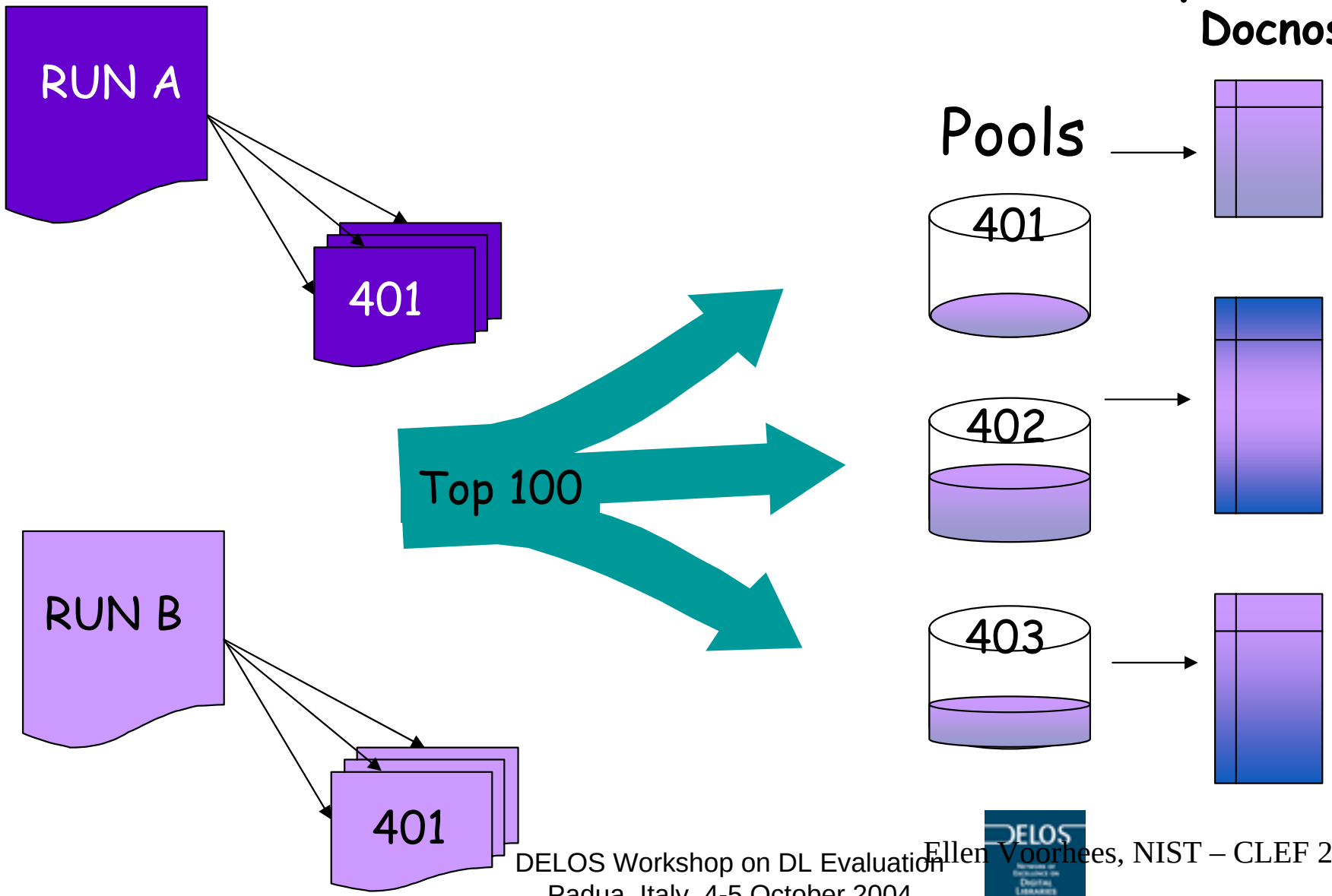


Form pools of unique documents from all submissions which the assessors judge for relevance.



Systems are evaluated using relevance judgments.

Creating the Pools



Creating the Pools in CLEF

Runs are pooled, respecting nearly a dozen criteria:

-participant's preferences

-"originality" (task, topic fields, languages, others…)

-participant/task coverage

-..

Results Analysis for Ad Hoc

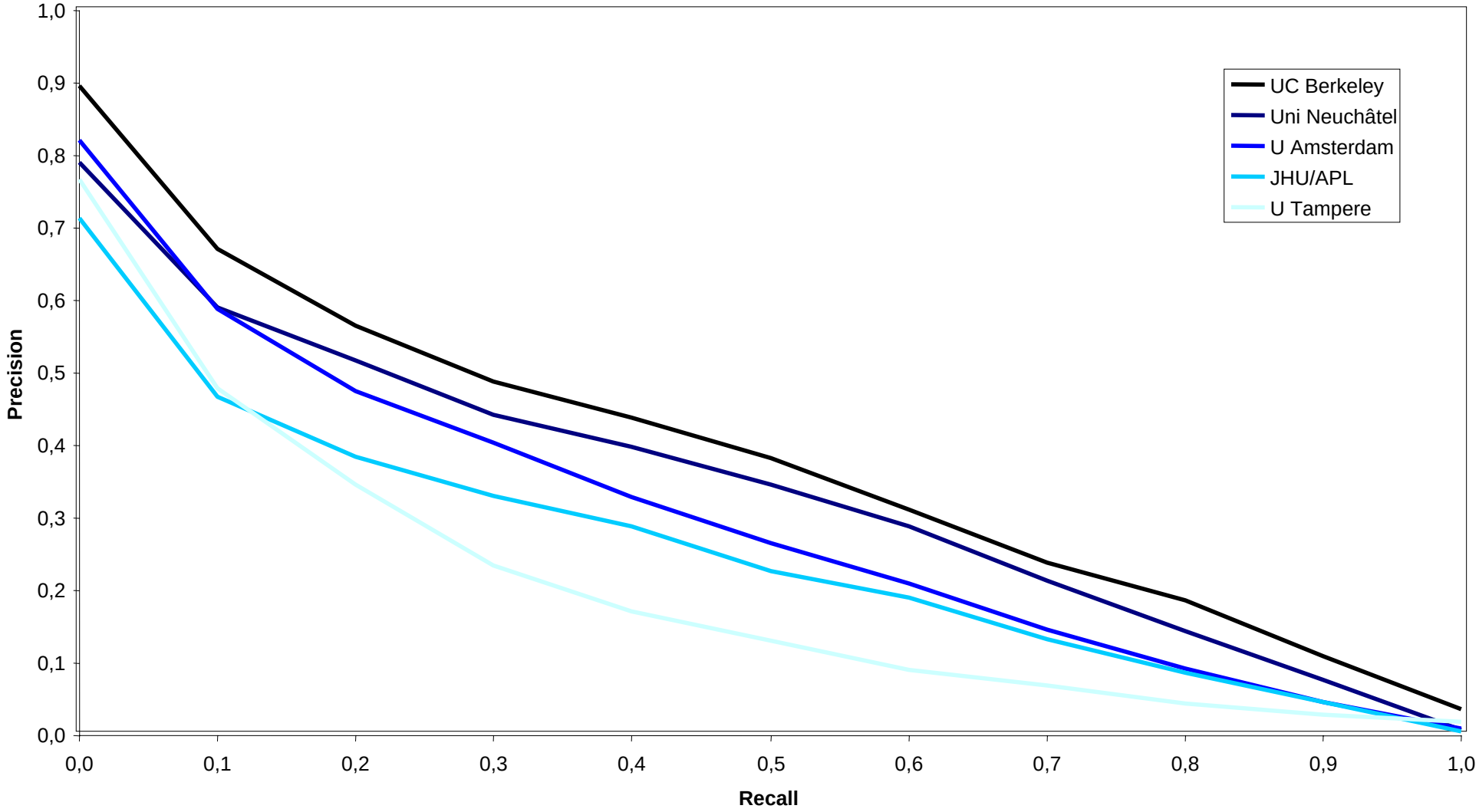
- Result processing (average precision figures etc)
- Statistical testing (ANOVA)
- Pool testing (unique relevant document tests, both for multilingual, and language-specific subsets)

Ad Hoc Results

For each experiment

- Recall/precision graph
- Av. precision for each query + graph with comparison to median performance
- Overall statistics
  - Total no. of relevant docs
  - Total no. of rel. docs. retrieved
  - Interpolated precision averages at specific recall levels
  - non.-interpolated av. precision over all queries
  - Precision nos, after specific nos. of documents
  - R-precision

Example of Recall-Precision Graph
CLEF 2003 Multilingual-8 Track - TD, Automatic



Results from Statistical Testing

- Difficult to find “exclusive groups” with statistically different performance – a fairly high difference in the measures is needed
- High variability across different queries (“easy” and “hard” queries, etc.)
- Possibility to use “joint” query sets from multiple years is very helpful for post-campaign experiments

Target Collection	Number of participants in the top statistical group
„Multi-8“	3/7
„Multi-4“	6/14
French	15/16

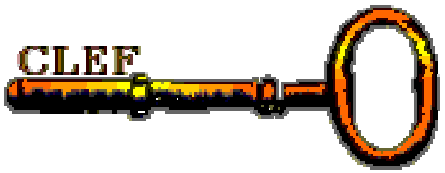
Results from Pool Analysis

- 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.	0.0005	Mean diff. in %	0.24%
Max absolute diff.	0.0014	Max diff. in %	0.77%
Standard deviation	0.0009	Standard dev. %	0.51%

- Figures are calculated that show how much measures change for non-participants
- Values a bit higher for individual languages, espec. the “newer” languages (e.g. DE: 0.29% vs. RU: 1.36%)
- Rankings are very stable! Figures compare very favorably to similar evaluations

From CLIR-TREC to CLEF
Growth in Test Collection



(Core Tracks)

	# part.	# lang	# docs.	MB	# assess.	# topics	# ass. per topic
CLEF 2004	24	10(5)	~1800,000	4473	114346	50(33)	~2287
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	12	4	698,773	1620	23,156	28	827

Effect of CLEF: Advance of State of the Art

- TREC6 CLIR:
  - French: 49%
  - German: 64%
- In contrast:
- CLEF 2003 Bilingual:
  - Spanish: 83%
  - Italian: 87%
  - Dutch: 82%
  - (even with restrictions in topic languages!)

Effect of CLEF (cont.)

- Careful work on individual languages, including fine-tuning: much more is known on (monolingual) IR in those languages now
- Some blueprints to “successful CLIR” have emerged as a direct consequence to CLEF
- “Inconvenient”, lesser spoken language pairs receive attention
- Research results using CLEF data now frequently cited in conferences/journals

Conclusions I

- need to have clear agreements with data providers
- need for particular care when constructing a (multilingual) collection in order to guarantee coherence and consistency (over languages)
  - Establish clear rules for topic creation/relevance assessment
- test collections are valuable resources – they must be reliable
  - make consistency checks

Conclusions II

- Test collection is expensive – go for reusability wherever possible
  - collections must be appropriate for task of interest BUT can be adapted to meet the requirements of other tasks
  - relevance assessments can be reused
- Test collections should be made publicly available for research and benchmarking