From CLEF to TrebleCLEF: the Evolution of the Cross-Language Evaluation Forum

Carol Peters - ISTI-CNR, Pisa, Italy
Nicola Ferro - University of Padua, Italy

NTCIR-7 Meeting
Tokyo, 16-19 December, 2008

Outline

- CLIR/MLIA System Evaluation
- Cross-Language Evaluation Forum
  - Objectives
  - Organisation
  - Activities
  - Results
- TrebleCLEF and the Future

CLIR/MLIA

1996 – First workshop on “Cross-Lingual Information Retrieval”, SIGIR, Zurich

1997 – Workshop on Cross-Language Text and Speech Retrieval, AAAI Spring Symposium Stanford

Grand Challenge: Fully multilingual, multimodal IR systems

- capable of processing a query in any medium and any language
- finding relevant information from a multilingual multimedia collection containing documents in any language and form,
- and presenting it in the style most likely to be useful to the user

CLIR/MLIA System Evaluation

In IR the role of an evaluation campaign is to support system development and testing and to **identify priority areas for research**

- First CLIR system evaluation campaigns begin in US and Japan: TREC (1997) and NTCIR (1998)
- CLIR evaluation in Europe: CLEF – extension of CLIR track at TREC (2000)
- Forum for Information Retrieval Evaluation, India (2008)

Cross Language Evaluation Forum

Objectives of CLEF

- Promote research and stimulate development of multilingual IR systems for European languages
- Build a MLIA/CLIR research community
- Construct publicly available test-suites

BY

- Creation of evaluation infrastructure and organisation of regular evaluation campaigns for system testing
- Designing tracks/tasks to meet emerging needs and to stimulate research in the”right” direction

Major Goal: Encourage development of truly multilingual, multimodal systems

CLEF Methodology

CLEF mainly based on Cranfield IR evaluation methodology

- Main focus on experiment comparability and performance evaluation
- Effectiveness of systems evaluated by analysis of representative sample search results

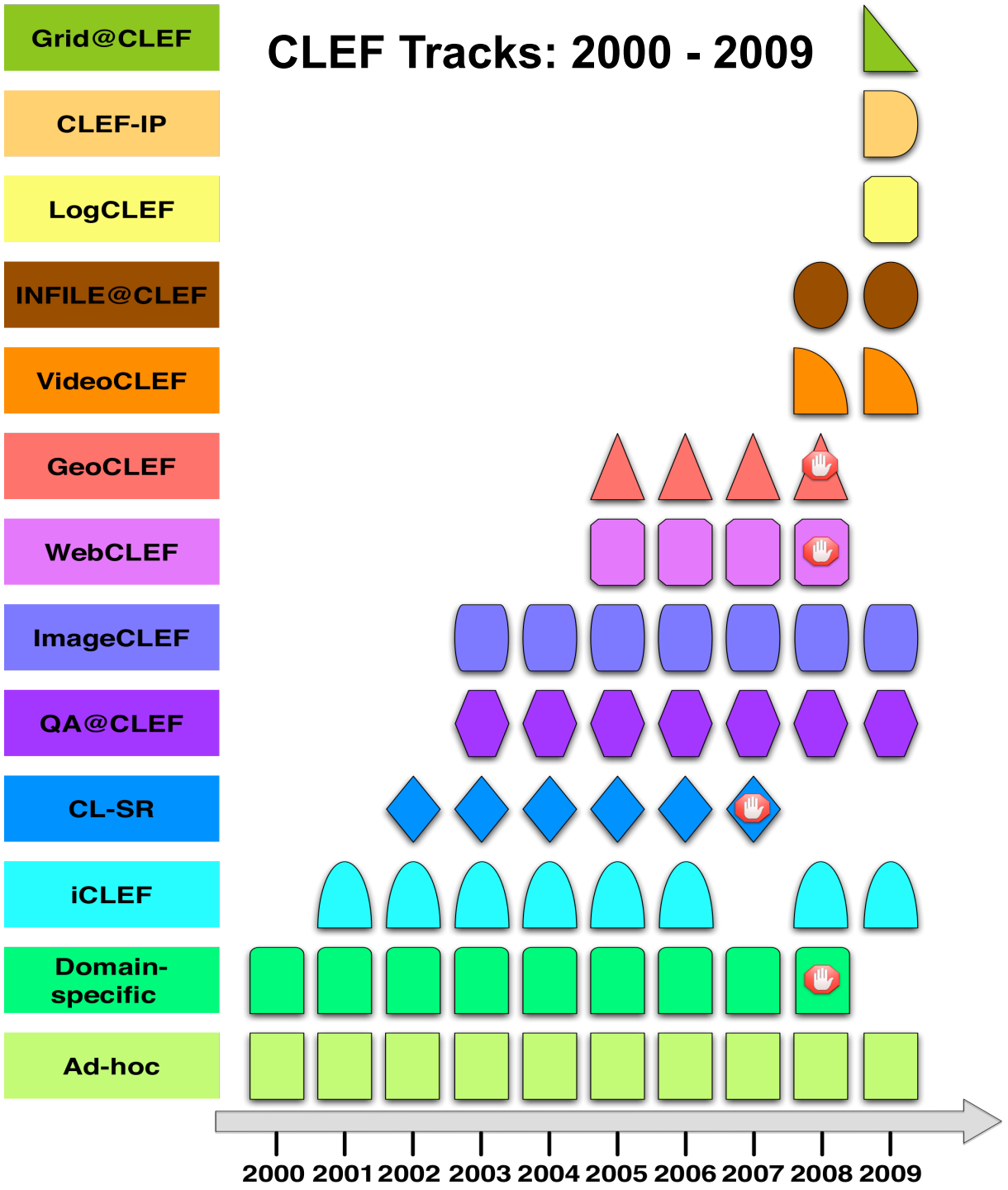
CLIR system evaluation is complex: integration of components and technologies

- need to evaluate single components
- need to evaluate overall system performance
- need to distinguish methodological aspects from linguistic knowledge

Influence of language and culture on usability of technology needs to be understood

Evolution of CLEF

CLEF 2000 Tracks	▪ mono-, bi- & multilingual text doc retrieval (Ad Hoc) ▪ mono- and cross-language information on structured scientific data (Domain-Specific)
CLEF 2001 New	▪ interactive cross-language retrieval (iCLEF)
CLEF 2002 New	▪cross-language spoken document retrieval (CL-SR)
CLEF 2003 New	▪ multiple language question answering (QA@CLEF) ▪ cross-language retrieval in image collections (ImageCLEF)
CLEF 2005 New	▪ multilingual retrieval of Web documents (WebCLEF) ▪ cross-language geographical retrieval (GeoCLEF)
CLEF 2008 New	▪ cross-language video retrieval (VideoCLEF) ▪ multilingual information filtering (INFILE@CLEF)
CLEF 2009 New	▪ intellectual property (CLEF-IP) ▪ log file analysis (LogCLEF) ▪ large-scale grid experiments (Grid@CLEF)


CLEF Tracks: 2000 - 2009
Grid@CLEF
CLEF-IP
LogCLEF
INFILE@CLEF
VideoCLEF
GeoCLEF
WebCLEF
ImageCLEF
QA@CLEF
CL-SR
iCLEF
Domain-specific
Ad-hoc
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009

CLEF Coordination

CLEF is Multilingual & MultiDisciplinary

Coordination is distributed over disciplines and over languages

- Expert Groups coordinate domain-specific activities
- Groups with native language competence coordinate language-specific activities

Supported by the EC IST & ICT programmes under unit for Digital Libraries

- 2000 – 2007 (mainly) DELOS
- 2008 – 2009 TrebleCLEF

Mainly run by voluntary efforts

CLEF Coordination

CLEF is coordinated by the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa
The following Institutions are contributing to the organisation of the different tracks of the CLEF 2008 campaign:

- Athena Research Center, Greece
- Business Information Systems, U. Applied Sciences Western Switzerland, Sierre, Switzerland
- Centre for Evaluation of Human Language & Multimodal Communication (CELCT), Italy
- Centruum vor Wiskunde en Informatica, Amsterdam,
- Computer Science Dept., U. Basque Country, Spain
- Computer Vision and Multimedia Lab, U. Geneva, CH
- Data Base Research Group, U. Tehran, Iran
- Dept. of Computer Science, U. Indonesia
- Dept. of Computer Science & Medical Informatics, RWTH Aachen U., Germany
- Dept. of Computer Science and Information Systems, U. Limerick, Ireland
- Dept. of Medical Informatics and Clinical Epidemiology, Oregon Health and Science U., USA
- Dept. of Information Engineering, U. Padua, Italy
- Dept. of Information Science, U. Hildesheim, Germany
- Dept. of Information Studies, U. Sheffield, UK
- Dept. Medical Informatics, U. Hospitals and University of Geneva, Switzerland
- Evaluations and Language Resources Distribution Agency, Paris, France
- German Centre Artificial Intelligence, DFKI
- GESIS- Social Science Information. Germany
- Information and Language Processing Systems, U. Amsterdam, The Netherlands
- Information Science, U. Groningen, NL
- Institute of Computer Aided Automation, Vienna University of Technology, Austria
- Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), Orsay, France
- U. Nacional de Educación a Distancia, Spain
- Linguateca, Sintef, Oslo, Norway
- Linguistic Modelling Lab., Bulgarian Acad Sci
- Microsoft Research Asia
- NIST, USA
- Research Computing Center of Moscow State U.
- Research Inst. Linguistics, Hungarian Acad. Sciences
- School of Computer Science and Mathematics, Victoria U., Australia
- School of Computing, DCU, Ireland
- TALP , U. Politècnica de Catalunya, Barcelona, Spain
- UC Data Archive and School of Information Management and Systems, UC Berkeley, USA
- U. "Alexandru Ioan Cuza", IASI, Romania

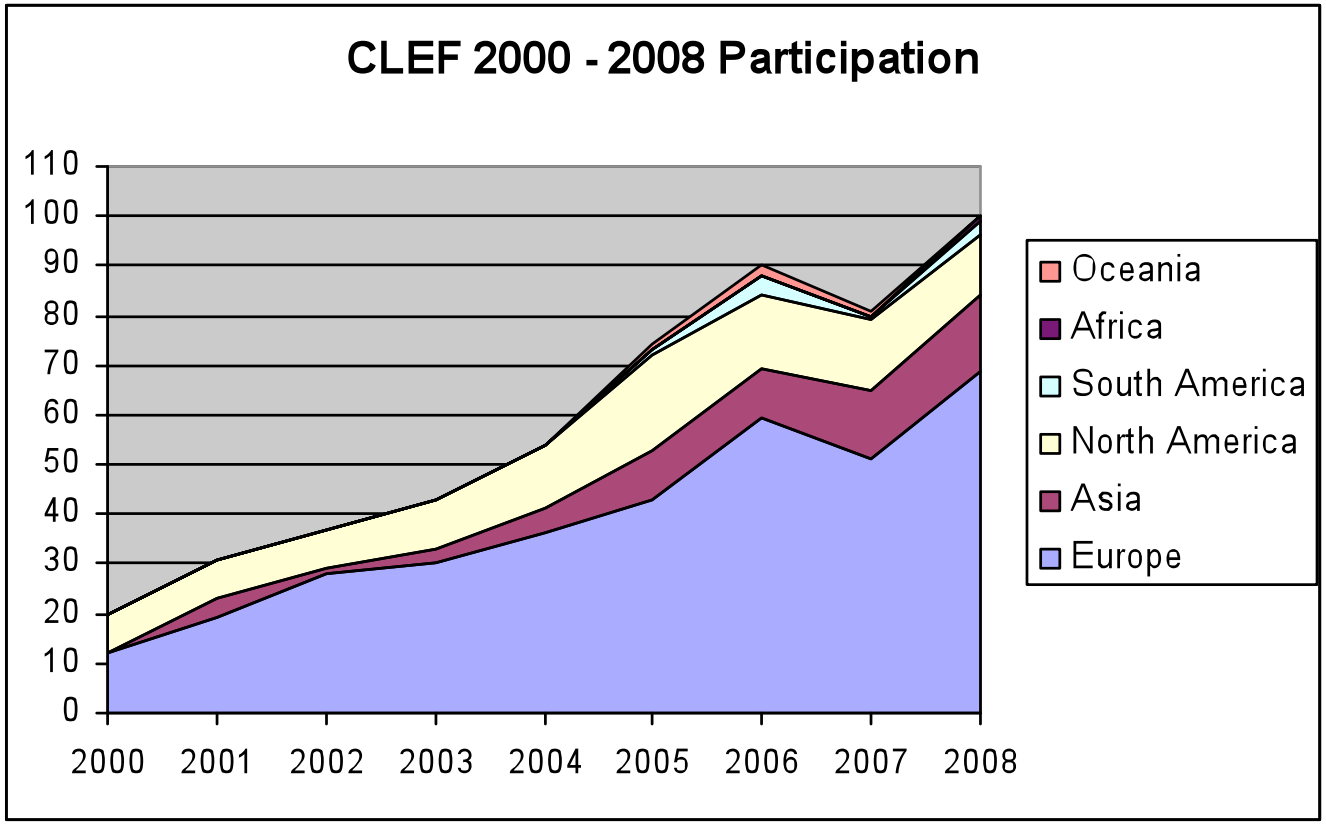
CLEF 2008: Track Coordinators

- **Ad Hoc**: Abolfazl AleAhmad, Hadi Amiri, Eneko Agirre, Giorgio Di Nunzio, Nicola Ferro, Thomas Mandl, Nicolas Moreau, Vivien Petras
- **Domain-Specific**: Vivien Petras, Stefan Baerisch
- **iCLEF:** Paul Clough, Julio Gonzalo, Jussi Karlgren
- **QA@CLEF**: Danilo Giampiccolo, Anselmo Peñas, Pamela Forner, Iñaki Alegria, Corina Forăscu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe, Erik Tjong Kim Sang, Alvaro Rodrigo, Jodi Turmo, Pere Comas, Sophie Rosset, Lori Lamel, Djamel Mostefa
- **ImageCLEF**: Allan Hanbury, Paul Clough, Thomas Arni, Mark Sanderson, Henning Müller, Thomas Deselaers, Thomas Deserno, Michael Grubinger, Jayashree Kalpathy–Cramer, and William Hersh
- **Web-CLEF**: Valentin Jijkoun and Maarten de Rijke
- **GeoCLEF:** Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Paula Carvalho
- **VideoCLEF:** Martha Larson, Gareth Jones
- **INFILE:** Djamel Mostefa
- **DIRECT:** Marco Dussin, Giorgio Di Nunzio, Nicola Ferro

CLEF 2008: Participating Groups

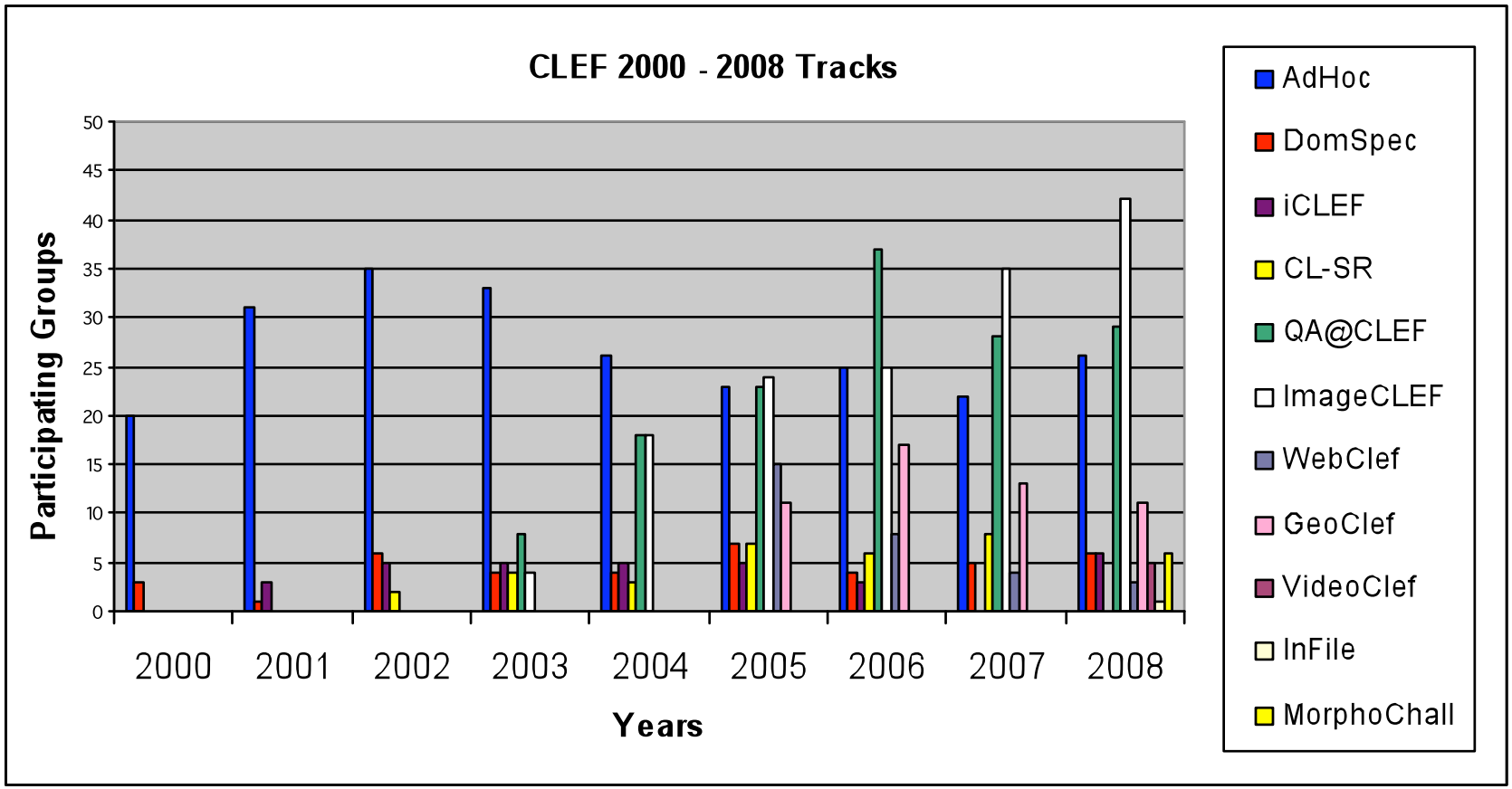
- Bulgarian Acad. Sci., Bulgaria
- Cal. State – San Marcos, USA
- CMU, USA
- CEA-LIST, France
- Charles U., Czech Rep.
- CWI, Netherlands
- DFKI, Germany
- Dublin City U., Ireland
- Hungarian Acad. Sci.
- IDIAP Research Inst-, Switzerland
- Imperial College, UK
- INAOE, Mexico
- Indian Statistical Inst., India
- INESC-ID (2), Portugal
- IIIT-Hyderabad, India
- IPAL-CNRS (IR2), Singapore
- IRIT/SIG Toulouse, France
- Johns Hopkins U., USA
- Know-Center, Austria
- Lab. LIG, France
- LIMSI-CNRS, France
- Linguateca-SINTEF, Norway
- LINA-Nantes, France
- LSIS-CNRS, France
- Macedonian & Slovenian U. Team, Macedonia/Slovenia
- Manchester Metropol. U., UK
- Microsoft Asia, China
- MIRACLE (2), Spain
- Nat. Inst. Informatics, Japan
- Nat. Inst. Health, USA
- Nat.Taiwan U., Taiwan
- Open Text Corp, Canada
- Open University, UK
- Oregon Health & Sci. U., USA
- Priberam Informatica, Portugal
- Research Inst. AI, Romania
- RWTH Aachen-HLT., Germany
- RWTH Aachen - Med.Inf., Germany
- SICS, Sweden
- SYNAPSE, France
- Tech.U. Chemnitz, Germany
- Tech. U. Darmstadt, Germany
- Tech U. Helsinki, Finland
- Tel Aviv U., Israel
- Telecom, Paris Tech, France
- TextMess, Spain
- U. & U.Hospitals Geneva, Switzerland
- U. Aberta, Portugal
- U. Alicante (2), Spain
- U. AI.I Cuza Iasi, Romania
- U. Amsterdam, Netherlands
- U. Banjaluka, Bosnia and Herzegovina
- U. Bari, Italy
- U. Basel, Switzerland
- U. Basque Country, Spain
- UC Berkeley, USA
- U. Complutense de Madrid, Spain
- U. Concordia –CLAC, Canada
- U. Cordoba, Argentina
- U. Evora, Portugal
- U. Federal do Rio Grande do Sul, Brasil
- U. Geneva, Switzerland
- U. Groningen, Netherlands
- U. Hagen, Germany
- U. Hildesheim, Germany
- U. Jaen, Spain
- U. Jean Monnet, France
- U. Karlsruhe, Germany
- U. Koblenz-Landau, Germany
- U. Lisbon, Portugal
- U. Makere, Uganda
- U. Maryland & US Gov.
- U. Meiji, Japan
- U. Nacional Colombia, Colombia
- UNED-LSI, Spain
- U. Neuchatel, Switzerland
- U. Ottawa, Canada
- U. Padova, Italy
- U. Peking, China
- U. Pittsburg
- UPMC-LIP6, France
- U. Politecnica Catalunya, Spain
- U. Politecnica Valencia, Spain
- U. Porto, Portugal
- U. Salamanca – REINA, Spain
- U. Sheffield, UK
- U. Tehran, Iran (7)
- U. Twente, Netherlands
- U. Tilberg, Netherlands
- U. Waseda, Japan
- U. Wolverhampton, UK
- Xerox SAS (CACAO), EU Project
- Xerox XRCE, Franc

CLEF: Trend in Participation



CLEF 2008: Europe = 69; N. America = 12; Asia = 15; S. America = 3; Africa = 1

CLEF 2000 – 2008 Participation per Track



CLEF System Evaluation

CLEF test collections: documents, topics/queries, relevance assessments

- Relevance assessments performed manually
- Pooling methodology adopted (depending on track)
- Consistency harder to obtain than for monolingual
  - 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)

CLEF Test Collections

2000

- News documents in 4 languages
- GIRT German Social Science database

2008

- CLEF multilingual comparable corpus of more than 3M news docs in 15 languages: BG,CZ,DE,EN,ES,EU,FI,FR,HU,IT,NL,RU,SV,PT and Persian
- The European Library Data in DE, EN, FR (>3M docs)
- GIRT-4 social science database in EN and DE, Russian ISISS collection; Cambridge Sociological Abstracts
- Online Flickr database
- IAPR TC-12 photo database (20,000 image, captions in EN, DE);
- ARRS Goldminer database (200,000 medical images)
- IRMA: 10,000 images for automatic medical image annotation
- INEX Wikipedia image collection (150,000 images)
- Very large multilingual collection of Web docs (EuroGov)
- Malach spontaneous speech collection – EN & CZ (Shoah archives)
- Dutch / English documentary TV videos
- Agence France Press (AFP) newswire in Arabic, French & English

CLEF System Evaluation

Experimental evaluation is a scientific activity and its outcome is very valuable scientific data

- Comparable experiments
- Performance measurements regarding the experiments
- Descriptive statistics about a collection of experiments
- Statistical tests for in-depth analysis of the experiments

The scientific data produced during an evaluation campaign should be archived, enriched, curated, preserved and properly cited to ensure future accessibility and reuse

Current evaluation methodology mainly focused on ensuring experiment reliability and comparability rather than modelling, organizing and managing the scientific data

DIRECT: Distributed IR Evaluation Campaign Tool

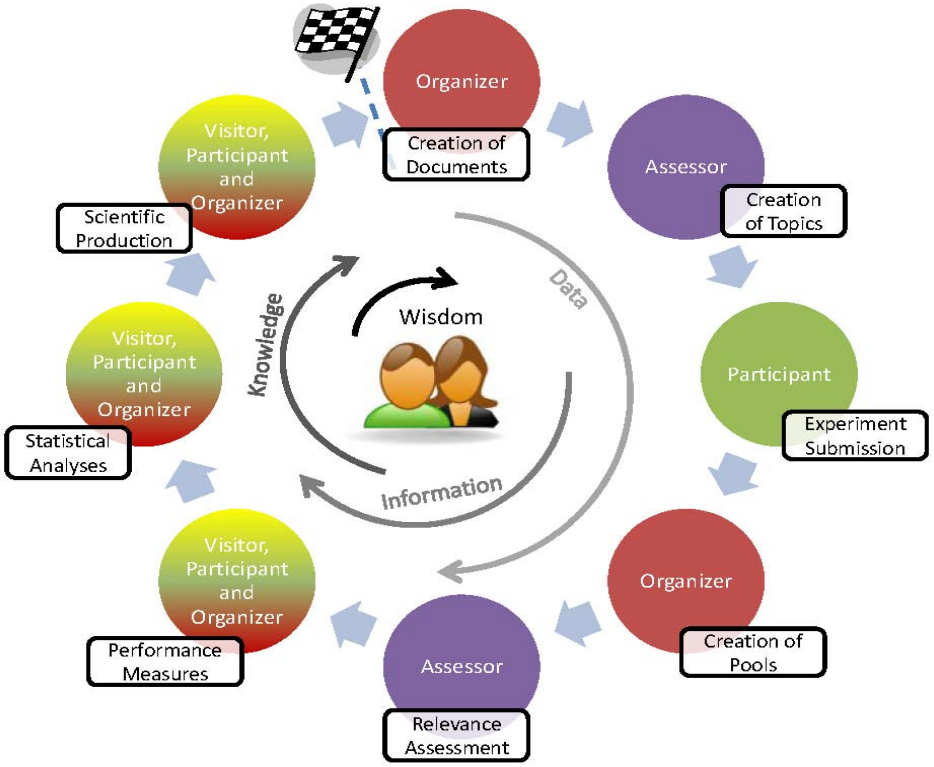
Main CLEF infrastructure is managed by the DIRECT DL system for **data curation** developed by Univ.Padua

DIRECT manages test data plus results submission and analyses for the ad hoc, question answering and geographic IR tracks and is responsible for:

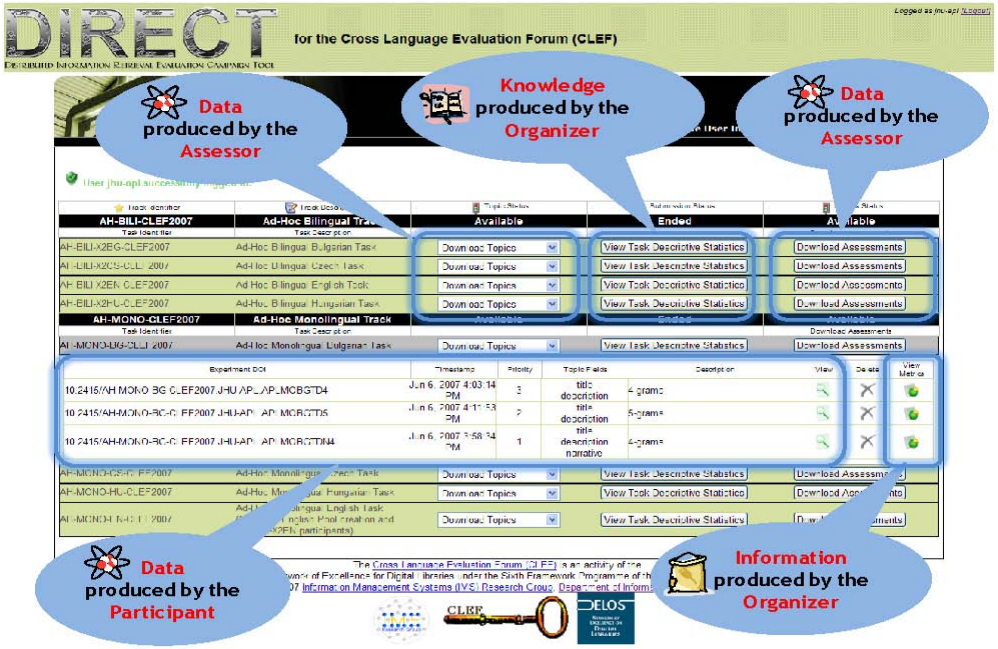
- track set-up, harvesting of documents, management of the registration of participants to tracks
- submission of experiments, collection of metadata about experiments, and their validation
- creation of document pools and management of relevance assessment
- provision of common statistical analysis tools for both organizers and participants in order to allow the comparison of the experiments
- provision of tools for producing reports and graphs on performance analyses

DIRECT@work in CLEF

Main Actors



http://direct.dei.unipd.it/



CLEF 2008 Tracks

- Multilingual textual document retrieval (Ad Hoc)
- Mono- and cross-language information retrieval on structured scientific data (Domain-Specific)
- Interactive cross-language retrieval (iCLEF)
- Multiple language question answering (QA@CLEF)
- Cross-language retrieval in image collections (ImageCLEF)
- Multilingual retrieval of web documents (WebCLEF)
- Cross-language geographical information retrieval GeoCLEF)

Pilots: Cross-language Video Retrieval (VideoCLEF)
Multilingual Information Filtering (INFILE)

CLEF 2008 Tracks

Ad-Hoc | iCLEF | QA@CLEF | ImageCLEF | WebCLEF | GeoCLEF | VideoCLEF

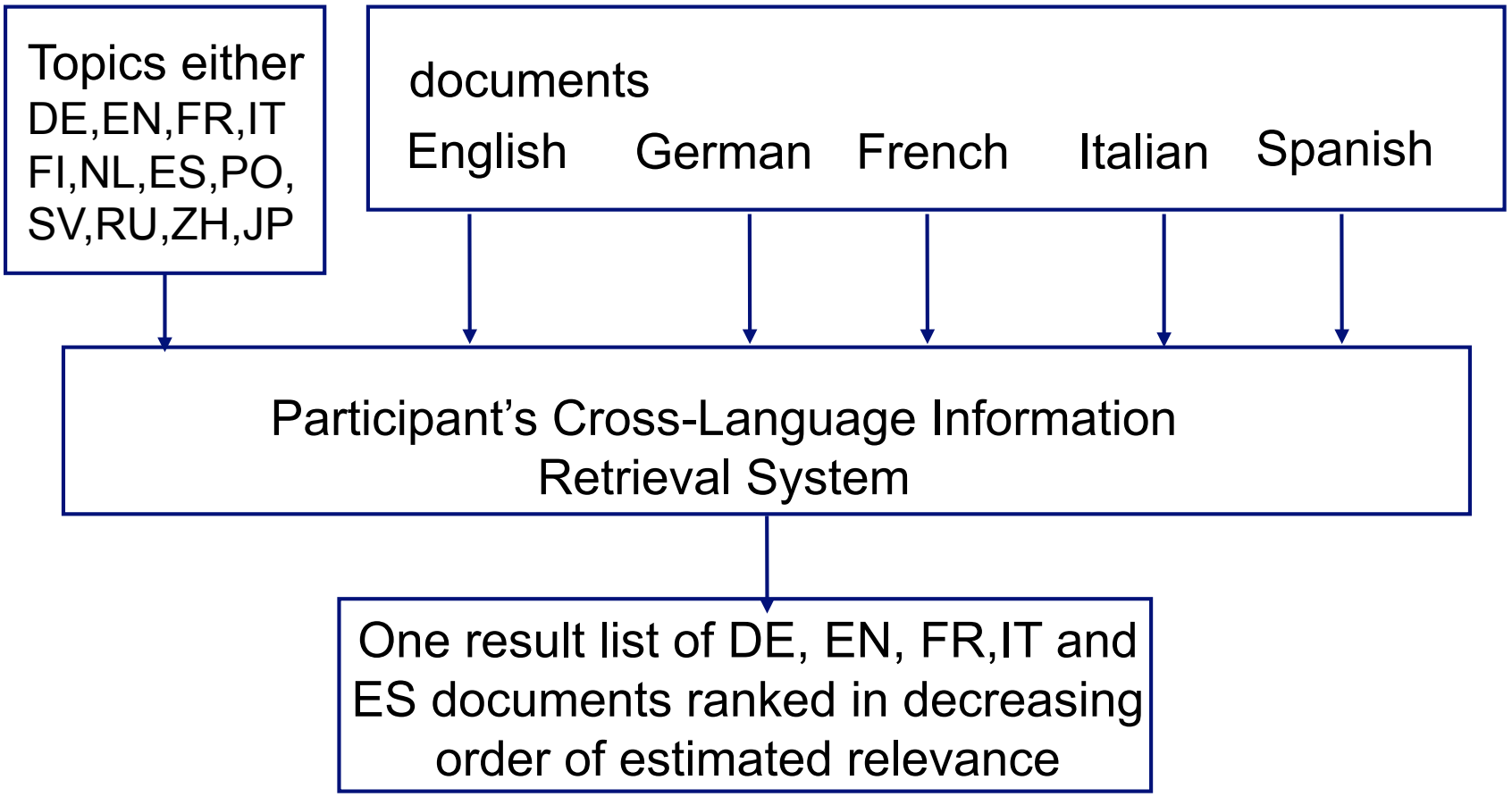
Promoting CLIR Research through Evaluation: AdHoc

- Aim: to promote development of mono and cross-language text retrieval systems
- AdHoc 2000-2007 European news collections: increasingly complex & diverse tasks
  - Monolingual – Bilingual – Multilingual
- Advanced Tasks – using previously built test collections
  - Multilingual 2 yrs on / merging
  - Robust – measuring stable performance

Ad Hoc: Importance of Monolingual IR

- Need to understand processing requirements of all languages to be queried, eg morphology, syntax, segmentation, special features
- Need to adopt best approach per languages
- CLEF test collection includes wide variety of European language types
  - Germanic: Dutch, English, German, Swedish
  - Romance: French, Italian, Portuguese, Spanish
  - Slavic: Russian, Bulgarian, Czech
  - Non-IndoEuropean: Ugro-Finnic – Finnish, Hungarian; and Basque
  - Plus Persian (Indo-Iranian)

Ad Hoc: Multilingual IR
CLEF 2002



Ad Hoc Track: Bilingual & Multilingual Tasks

- Tasks made increasingly difficult over the years
  - CLEF 2003 - 2 multilingual tasks
    - Small-multilingual: 4 “core” language (EN,ES,FR,DE)
    - Large-multilingual: 8 languages (+FI,IT,NL,SV)
  - Bilingual: “unusual” language combinations
    - IT -> ES FR -> NL
    - DE -> IT FI -> DE
    - x -> RU Newcomers only: x -> EN
  - CLEF 2007: Non-European topic languages
    - AM/ID/OR/ZH→ EN
    - BN/HI/MR/TA/TE→ EN

AdHoc	Monolingual	Bilingual	Multilingual
CLEF2000	DE;FR;IT	X→EN	X→DE;EN;FR;IT
CLEF2001	DE;ES;FR;IT;NL	X→EN, X→NL	X→DE;EN;ES;FR;IT
CLEF2002	DE;ES;FI;FR IT;NL;SV	X→DE;ES;FI;FR;IT;NL;SV X→EN(newcomer)	X→DE;EN;ES;FR;IT
CLEF2003	DE;ES;FI;FR IT;NL;RU;SV	IT→ES;DE→IT FR→NL;FI→DE X→RU;X→EN	X→DE;EN;ES;FR X→DE;EN;ES;FI FR;IT;NL;SV
CLEF2004	FI;FR;RU;PT	ES/FR/IT/RU→FI DE/FI/NL/SV→FR X→RU;X→EN	X→FI;FR;RU;PT
CLEF2005	BG;FR;HU;PT	X→ BG;FR;HU;PT EX →EN	Multi8 2yrson Multi8 merge
CLEF2006	BG;FR;HU;PT	X→ BG;FR;HU;PT X →EN	ROBUST:X→DE;EN;ES; FR;NL
CLEF2007	BG, CZ, HU ROBUST: EN;FR;PT	X→ BG;CZ;HU; AM/ID/OR/ZH→ EN BN/HI/MR/TA/TE→ EN ROBUST: X→EN;FR;PT	
CLEF2008	FA TEL: DE; EN; FR ROBUST: WSD EN	EN→FA TEL: x→DE;EN;FR ROBUST: WSD Es →EN	

Ad Hoc: Results

Comparing bilingual results with monolingual baselines:

- TREC-6, 1997:
  - EN→FR: 49% of best monolingual French system
  - EN→DE: 64% of best monolingual German system
- CLEF 2002:
  - EN→FR: 83,4% of best monolingual French system
  - EN→DE: 85,6% of best monolingual German system
- CLEF 2003 enforced the use of “unusual” language pairs:
  - IT→ES: 83% of best monolingual Spanish IR system
  - DE→IT: 87% of best monolingual Italian IR system
  - FR→NL: 82% of best monolingual Dutch IR system
- CLEF2005 :
  - X -> FR: 85% of best monolingual French IR system
  - X -> PT: 88% of best monolingual Portuguese IR system
  - X -> BG: 74% of best monolingual Bulgarian IR system
  - X -> HU: 73% of best monolingual Hungarian IR system

Figures for FR and PT reflect state-of-the-art
Room for improvement for “new” languages

CLEF 2005: Multi-8 Two-Yrs-on

- Test collection used in 2003
- Docs in 8 languages: DE,EN,ES,FI,FR,IT,NL,SV
- 2 Objectives:
  - check improvement in system performance over time
  - focus on problem of merging results form different collections/languages
- Findings: participating groups
  - top performing submissions to Multilingual 2-Yrs-On and Merging tasks are both higher than the best submission to CLEF 2003 task
  - there is scope for further improvement in multilingual IR from focused exploration of merging techniques.

Ad Hoc: Robust Task

Robustness in multilingual retrieval

- Emphasizes importance of stable performance instead of high average performance
- Stable performance over all topics instead of high average performance
- Stable performance over different languages
- Uses existing test collections for English, French, Portuguese

Various Approaches

- Different expansion techniques
- Heuristic to determine hard topics on training set
- Test with other evaluation measures
- Experiments with fusion techniques

Trends in Ad Hoc

- Most traditional approaches to CLIR tested: n-gram indexing, machine translation, machine readable bilingual dictionaries, multilingual ontologies, pivot languages
- Corpus-based approaches less popular
- Query translation is dominant but some doc. translation
- Experiments with adaption to „new" languages
- Many groups using free resources
- Usual issues examined: word-sense disambiguation, out-of-dictionary vocabulary, ways to apply relevance feedback, results merging
- In monolingual task: development of new or adaption of existing stemmers or morphological analysers
- Recently, increasing use of external resources, e.g. Wikipedia

Ad Hoc: CLEF 2008

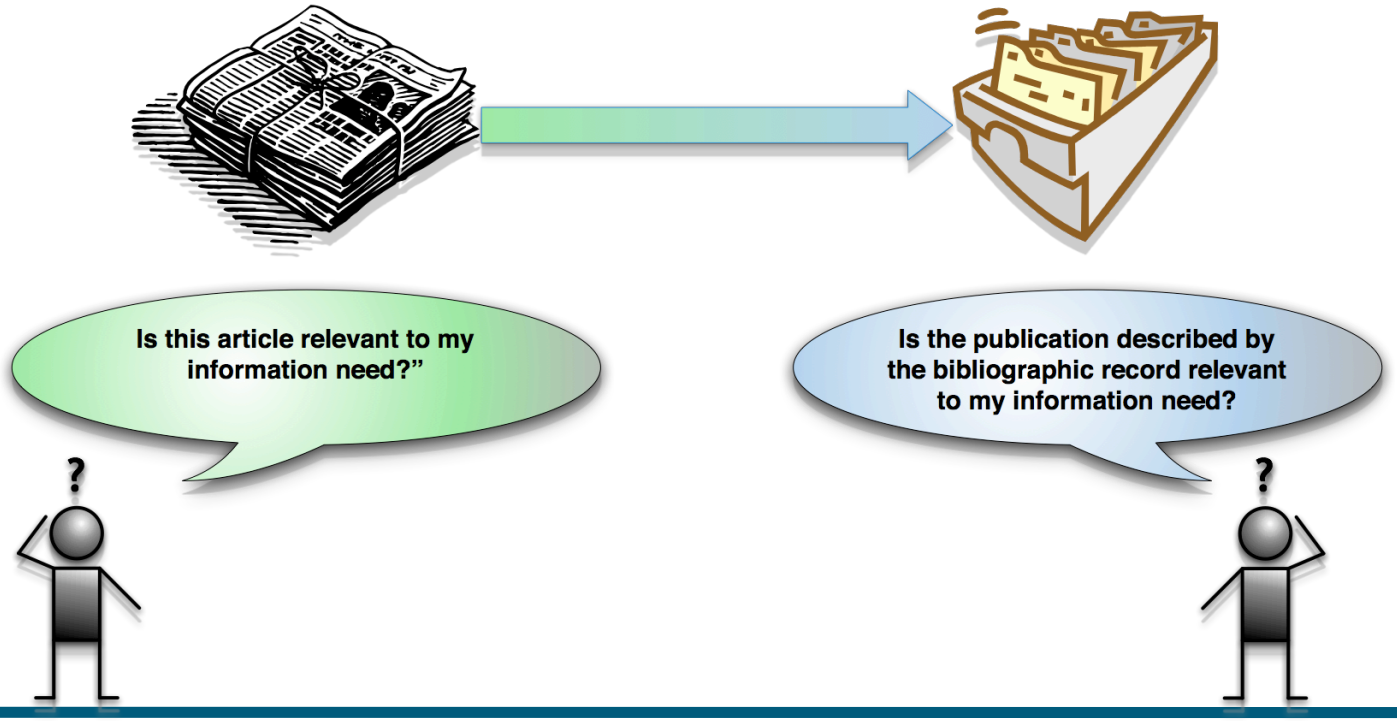
Focus on three different issues:

- **real scenario**: document retrieval from multilingual and sparse catalogue records to meet actual user needs
- **linguistic resources**: “exotic languages” (Indian languages, Persian, maybe Turkish) to favour the creation of new experimental collections and the growth of regional IR communities
- **advanced language processing**: robust and WSD to strengthen system performances

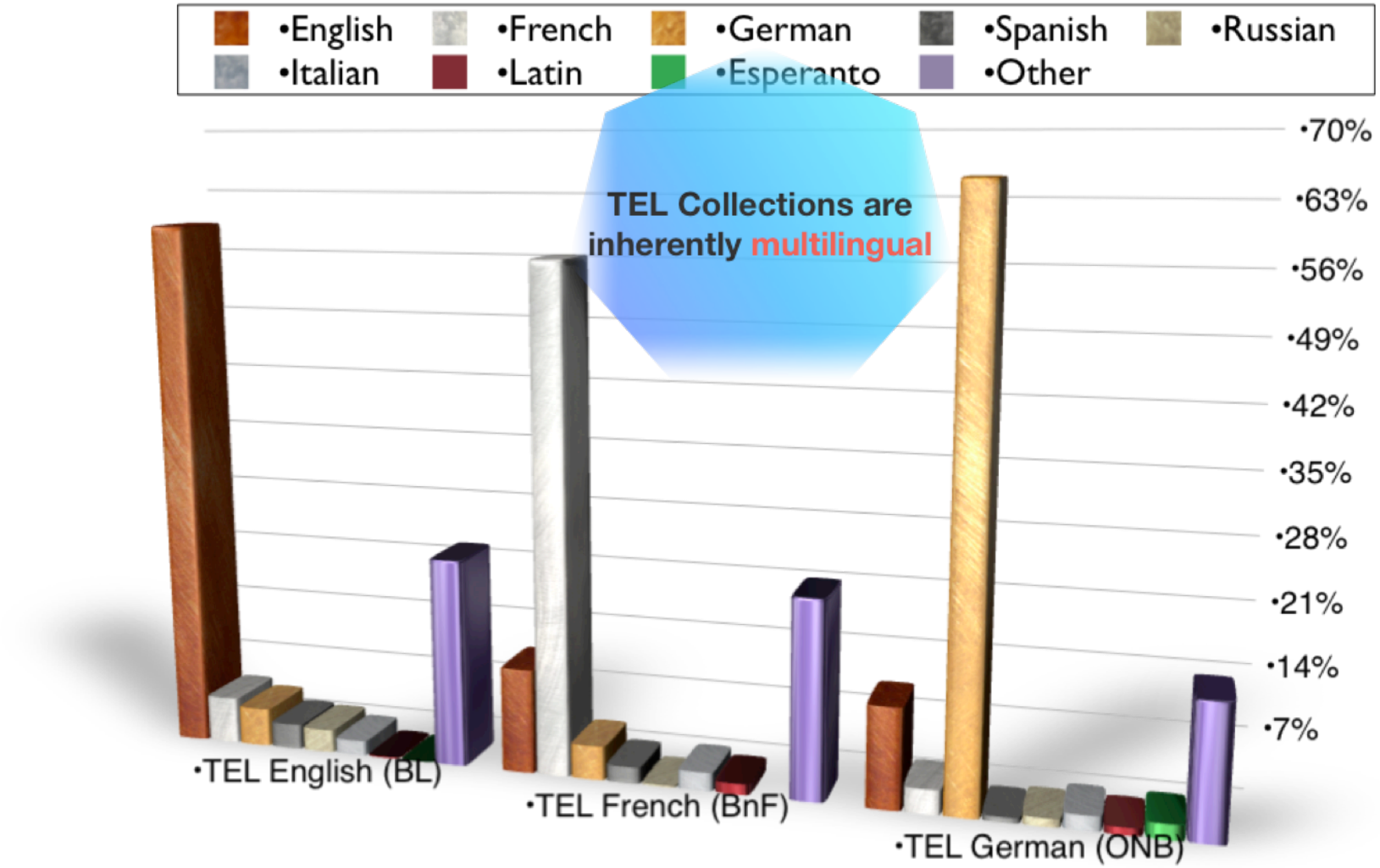
Ad-hoc TEL Task

Real world task

- Search and retrieve relevant items from collections of library catalog cards, which are surrogates for documents held by libraries
- Sparse and inherently multilingual data
- Monolingual and bilingual tasks

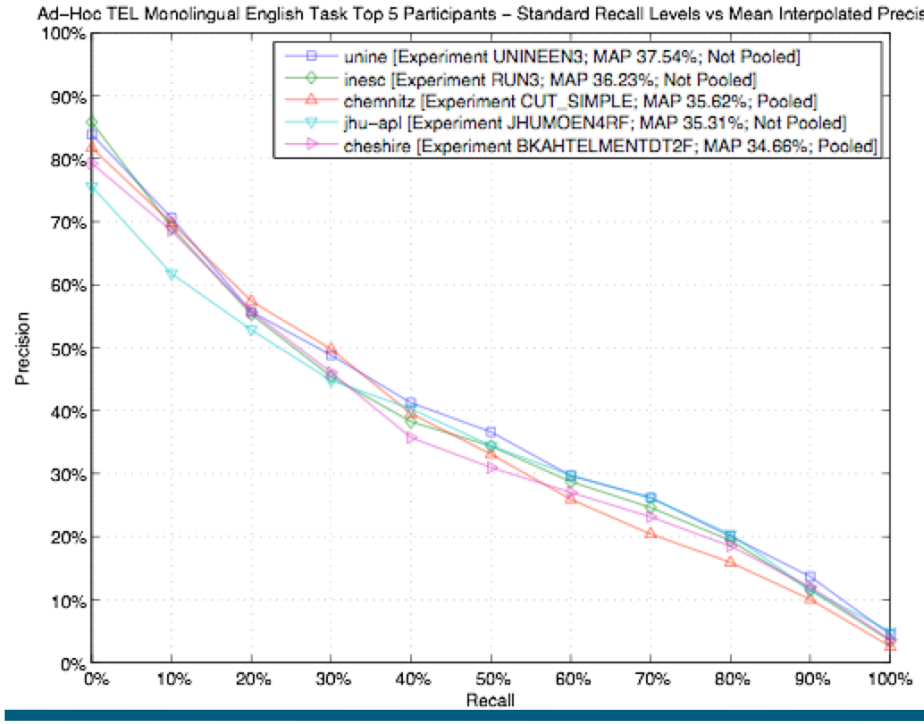


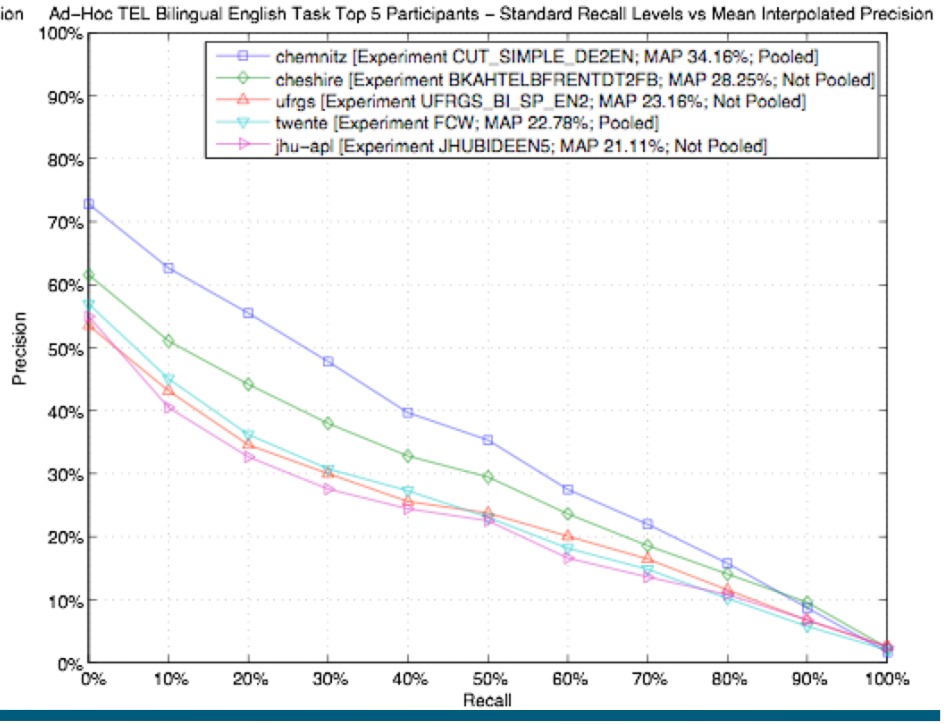
TEL Collections: Distribution of the Languages



TEL English

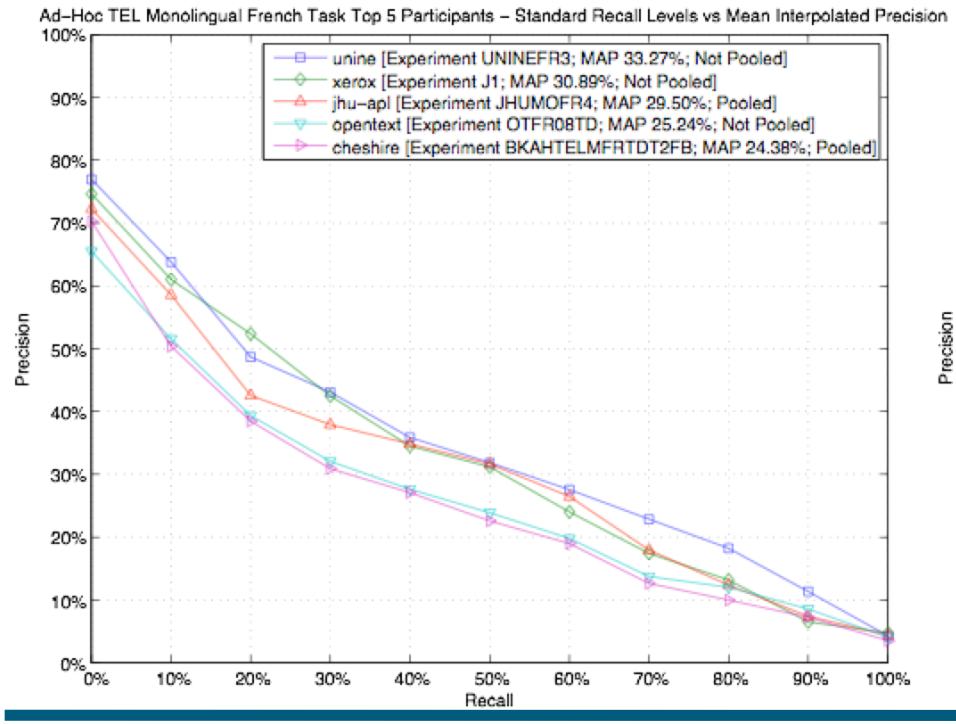
Bilingual is 91% of monolingual

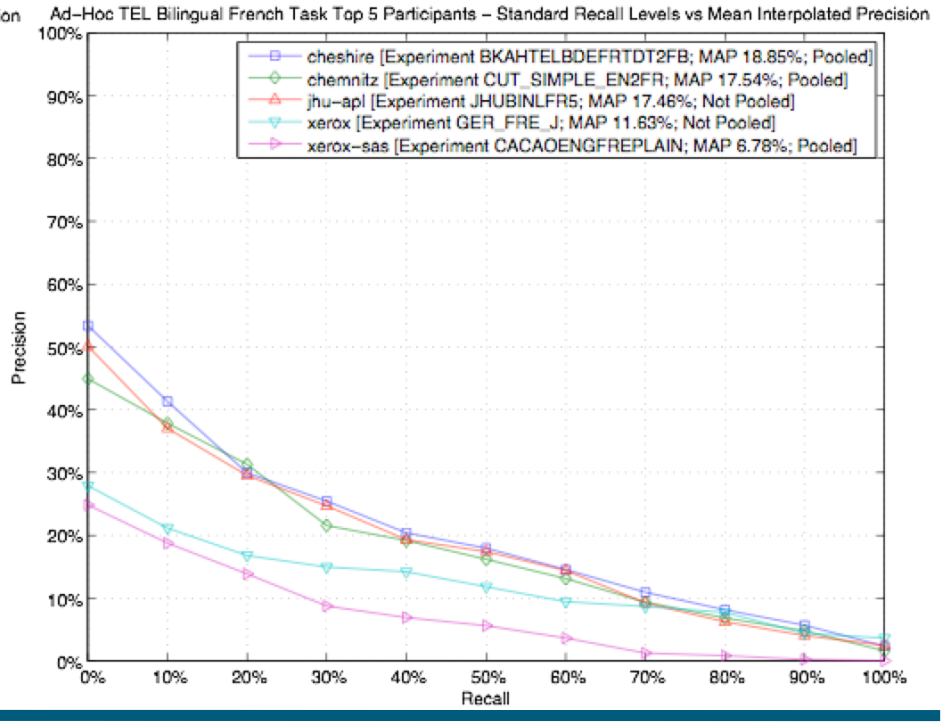




TEL French

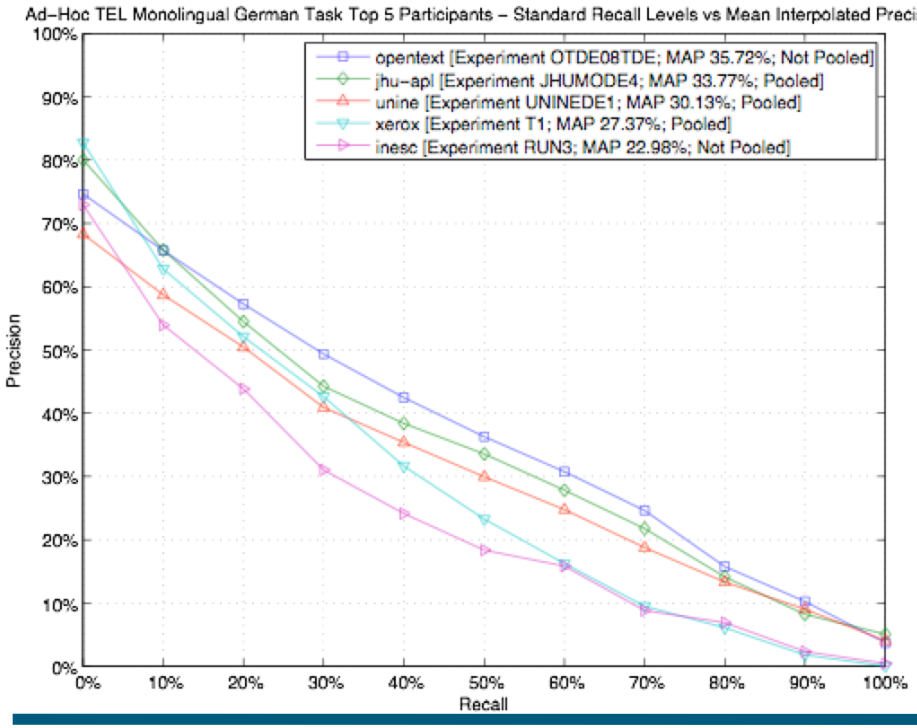
Bilingual is 57% of monolingual

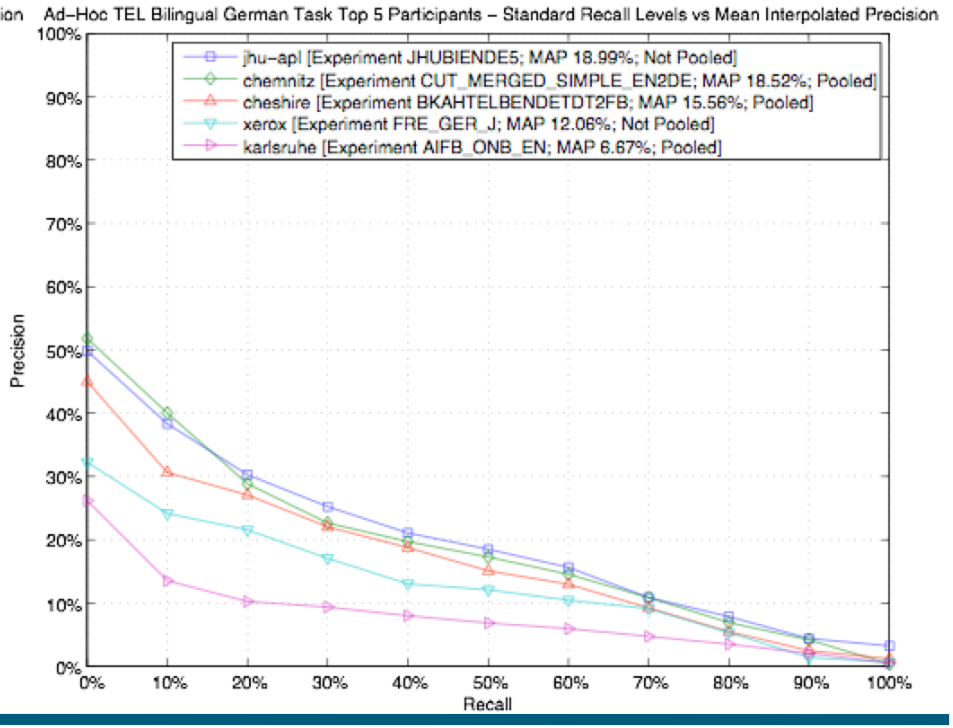




TEL German

Bilingual is
53%
of monolingual





Ad-hoc: Persian Task

- For the first time, a non-European language target collection is part of the CLEF corpus
- Persian uses **challenging script**, which is a modified version of the Arabic alphabet with elision of short vowels and is written from right to left
- Persian **morphology is complex** and makes extensive use of suffixes and compounding
- Task organized together with the **Data Base Research Group (DBRG) of the University of Tehran** which provided the Hamshahri corpus
- Both **monolingual and bilingual** tasks offered

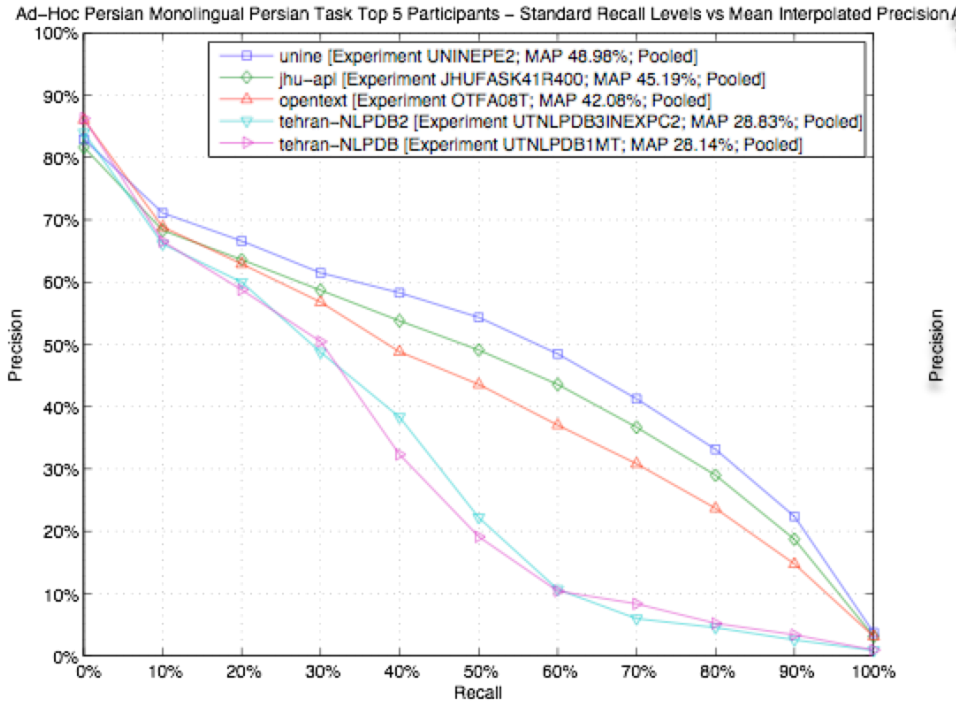
Persian Collection

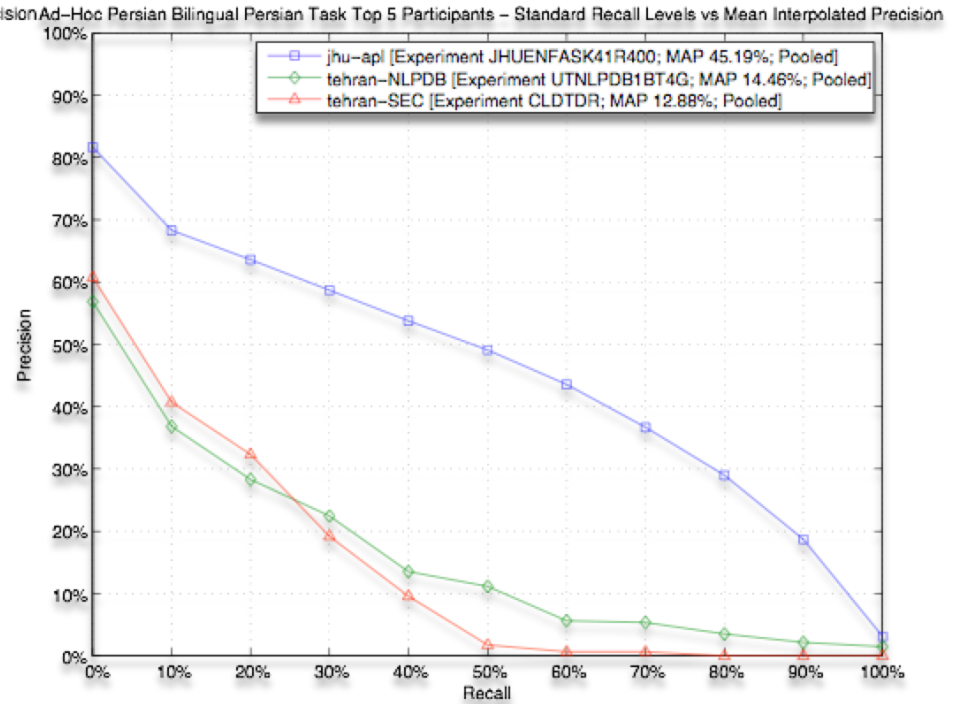
- The **Hamshahri corpus** is a newspaper corpus with news articles from 1996 to 2002, made available by the DBRG of University of Teheran (http://ece.ut.ac.ir/dbrg/hamshahri/)
- News article are categorized both in Persian and English
- It consists of:
  - size: 628,471,252 bytes
  - items:166,774 documents

<DOC>
<DOCID>H-750405-266S1</DOCID>
<DOCNO>H-750405-266S1</DOCNO>
<DATE>1996-06-25</DATE>
<CAT xml:lang="fa">ادب و هنر</CAT>
<CAT xml:lang="en">Literature and Art</CAT>
<TEXT> معرفي كتاب جديد مارکز گابريل گارسيا مارکز، نويسنده کلمبيائ و برنده جايزه نوبل ادبيات، روز
14 کتاب جديد خود تحت عنوان شرح حال يك اَدم ربائ را در ماديريد معرفي کرد. کتاب
25 مي (ارديبهشت)
مذکور شرح حال خانم مارگا پاکن (Pachon Marga) و همسرش اليرتو ويلاميزا (Villamiza Alberto) است .
خانم پاکن توسط قاچاقچيان کلمبيا ربوده شده بود و همسرش در طول شش ماه براي آزادي وي مبارزه کرد و
شخصا با پابلو اسکوبا (رئيس معروف قاچاقچيان کلمبيائ) مذاکره کرد. گارسيا مارکز درروز اول سفر خود
به ماديريد با فيليپ گونزالس (دوست شخصي اوو نخست وزير سابق اسپانيا) ملاقات کرد . </TEXT>
</DOC>

Persian

Bilingual is
92%
of monolingual





Ad-hoc: Robust WSD Task

- Idea: Provide English documents and topics (LA94 GH95) with automatically annotated word senses (WordNet)
- Participants explore how the word senses (plus the semantic information in wordnets) can be used in (CL)IR
- 10 Groups participated
- Monolingual: ENG → ENG;
  - Best GMAP results with WSD
  - Several top scoring teams report improvements in MAP and GMAP using WSD
- Bilingual: ES→ENG
  - Best results without WSD
  - Use WordNet as the sole translation resource
  - Several teams report improvements in MAP and GMAP

Ad-hoc 2008: First Conclusions

- Encouraging participation in the various tasks and interesting results have been achieved
- The experience gained this year will be very useful to further tune the tasks (e.g. only 100 docs retrieved by Persian groups)
- Robust WSD: ample room for further exploration
- TEL Task:
  - traditional IR approaches seem to work well and achieve good results
  - only two groups have exploited the inherent multilinguality of the data
  - almost no group has exploited the semi-structured nature of the data or used the subject headings

We need to do more

CLEF 2008 Tracks

Ad-Hoc | iCLEF | QA@CLEF | ImageCLEF | WebCLEF | GeoCLEF | VideoCLEF

Promoting CLIR Research through Evaluation: iCLEF

Interactive CLIR – iCLEF (from 2001)

- Cross-Lang. IR from a user-inclusive perspective
  - Interactive document selection/query formulation
  - How can interaction with user help a QA system
- "Difficult" track to run
- CLEF 2007 & 2008: task based on Flickr database: images with textual comments, captions, and titles in many languages

iCLEF 2008: Changes

- 2006: Move from news collections to images in a multilingual social network context (Flickr)
- 2006: Move from canned information needs to more naturalistic scenarios
- 2008: Lower threshold of entry for test subjects and experimenters alike
- 2008: Move from system design towards log analysis

iCLEF 2008: Task

- Test collection: **Flickr image set** (> 100M images with annotations in several languages)
- Search task: given a raw image, find it in Flickr (image is annotated in any of EN,ES,FR,NL,DE,IT)
- **Single search interface** available to all web users, registration (with language profile) required
- **Game-like features**: the more images you find, the higher your rank
- Task for iCLEF groups: **Log analysis**

Flickling - An online game for searching Flickr across language boundaries - Mozilla Firefox

Archivo Editar Ver Historial Marcadores Herramientas Ayuda

http://soporte1.lsi.uned.es/flickling/index.html?locale=es#{"start_index"%3A0%2C "num_results"%3A20%2C "base_lang"%3A2%2C "is_tag_search"%3

Más visitados Comenzar a usar Firef... Últimas noticias Posgrados (UNED) El analista Blogs ELPAIS.com - M... cursos MOODLE Gmail - Inbox (872) - j... ELPAIS.com - Última H... eBay.e

Google web search Buscar Marcadores PageRank Corrector ortográfico Traducir Enviar a web search

julio.gonzalo@

Flickling

An online game for searching Flickr across language boundaries

Encuentra esta imagen

Me rindo

monolingüe | multilingüe

atardecer Buscar

escribo en Español

traduce mi consulta a DE EN ES FR IT NL

Traducciones

Español	*atardecer*
Alemán	abend
Inglés	eve
Francés	soir
Italiano	sera
Holandés	avond

Resultados **1-20** de **500** para ***atardecer*** *(Mostrar consulta a Flickr)*

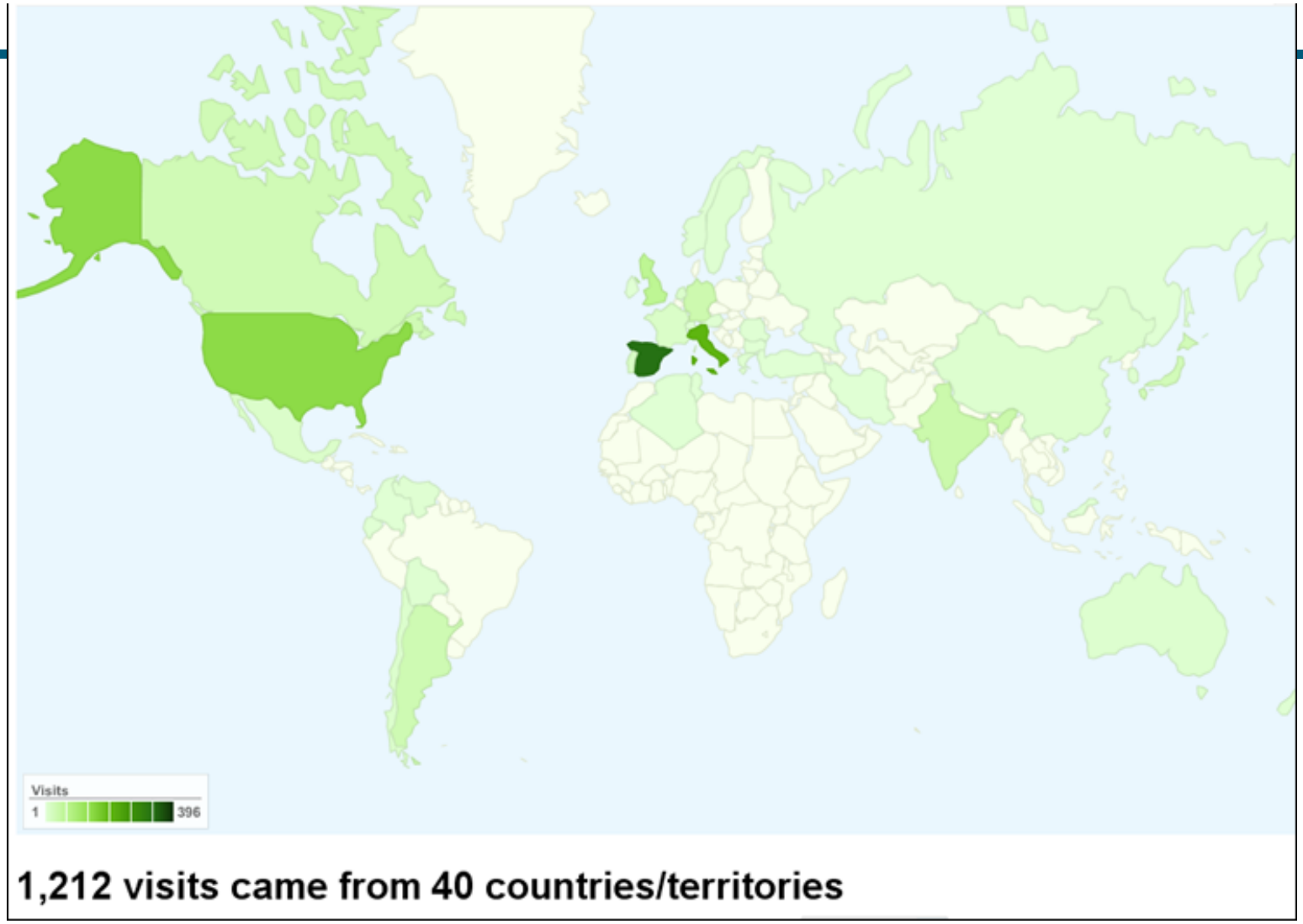
Tal vez quieras intentar con: sunset, sol, mar, playa, cielo, sea, nubes, sky, sun, beach... [mostrar todo]

moroccan sunset

red, brussels, rot, night, rouge, **abend**, bruxelles, **avond**, **soir**, brüssel... [mostrar todo]

At the **evening** of my life!

world, life, sunset, evening, bravo, alone, **soir**, coucherdesoleil, vie, seuls... [mostrar todo]



- 300 participants, 230 active:
- researchers, students, photo buffs

iCLEF 2008: Results

- Truly reusable data set (first time in iCLEF!)
  - > 5,000 complete search sessions recorded
  - > 5,000 post-search and post-experience questionnaires
  - > 100 queries covering six (target) languages
  - > 200 active users from 40 countries
- Quantification of the differences (in success, behaviour, satisfaction) between different user profiles (active, passive, unknown) and search settings (mono, bi, multilingual)
- Six groups submitted results (4 log analysis, 2 observational studies)

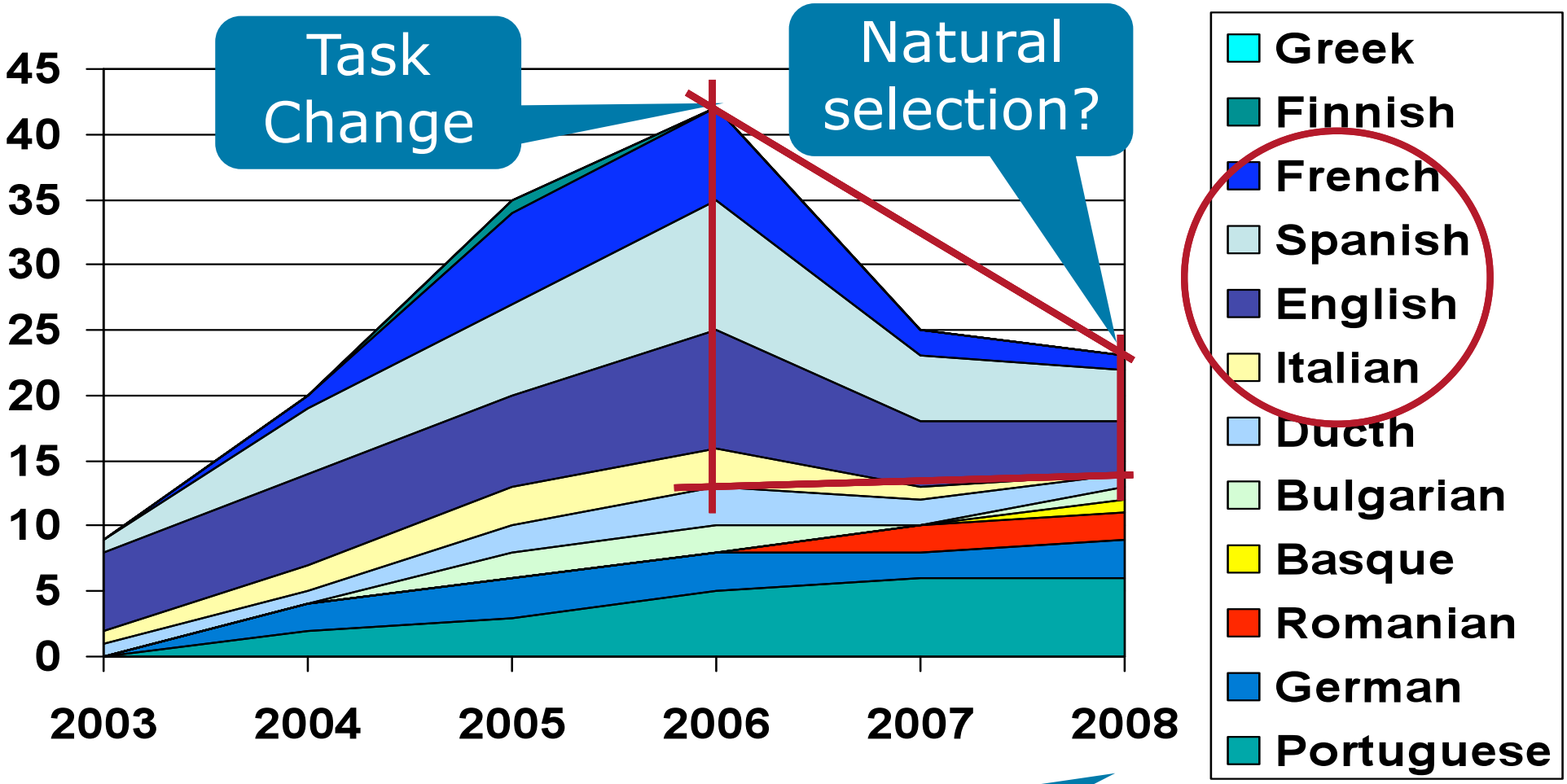
CLEF 2008 Tracks

Ad-Hoc | iCLEF | QA@CLEF | Image CLEF | Web CLEF | Geo CLEF | Video CLEF

Promoting CLIR Research through Evaluation: QA@CLEF

	2003	2004	2005	2006	2007	2008
Target languages	3	7	8	9	10	11
Collections	News 1994		+News 1995		+Wikipedia Nov. 2006	
Type of questions	200 Factoid		+ Temporal restrictions + Definitions	- Type of question + Lists	+ Linked questions + Closed lists	
Supporting information	Doc.			Snippet		
Pilots and Exercises		Temporal restrictions Lists		AVE Real Time WiQA	AVE QAST	AVE QAST WSDQA

Drop in Groups per Target Collection



QA@CLEF2008: Conclusions

- Less participants per language
  - Poor comparison
  - Change methodology: one task for all
- Critics to collections
  - Easier to find questions with IR in wikipedia
  - No user model
  - Change collection
- QA proposal for 2009 (ResPubliQA)
  - New collection: European treaties
  - Simplify the task: close to passage retrieval
  - Work on developing realistic use scenarios

CLEF 2008 Tracks

Ad-Hoc | iCLEF | QA@CLEF | **Image CLEF** | Web CLEF | Geo CLEF | Video CLEF

Promoting CLIR Research through Evaluation:ImageCLEF

Objectives of ImageCLEF

- initiate & promote research in cross lang. image retrieval

Began in 2003 as pilot experiment

- in 2008, 45 groups submitted results

- Retrieval methods
  - concept-based: abstracted features assigned to the image (e.g. captions, metadata etc.)
  - content-based: using primitive features based on pixels which form the contents of an image

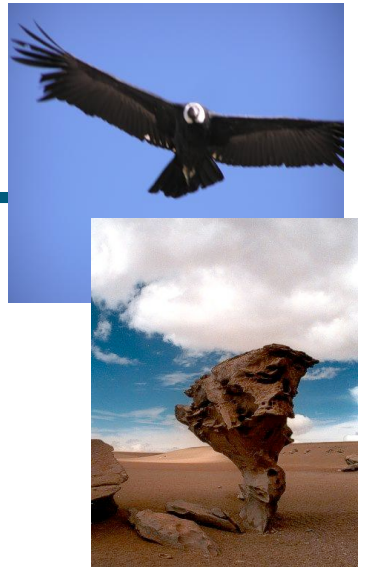
Cross-language image retrieval

- retrieval based on visual features is language-independent
- language of associated texts should have minimal affect on their usefulness for retrieval

ImageCLEF 2008: Tasks

- Photographic retrieval task
  - Aimed at promoting diversity
- Automatic concept detection task
  - Using a simple hierarchy of objects
- Wikipedia retrieval task
  - Image retrieval task using a larger-scale collection of heterogeneous Wikipedia images with semi-structured annotations
- Medical hierarchical image classification/ annotation task
- Ad-hoc retrieval of documents
  - Using scientific literature sources including images

Photo Retrieval 2008

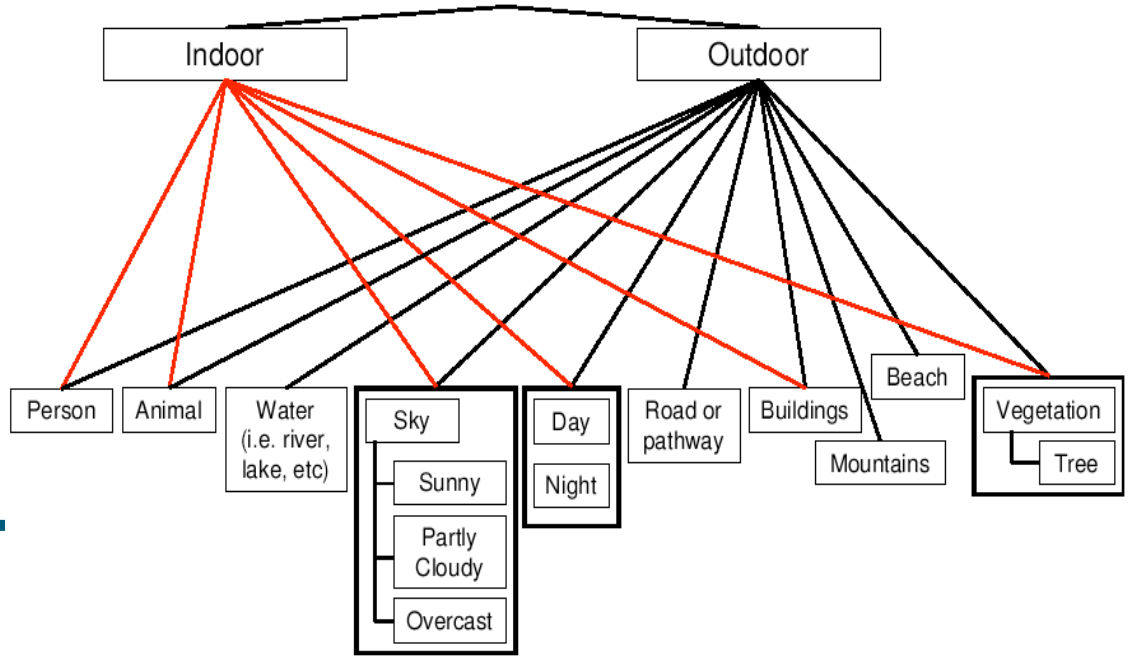


- Promote diversity in retrieval
  - Evaluated using Cluster Recall
- Very strong participation
  - Most participants used two stage process: perform ad-hoc retrieval; then cluster results
- Analysis of results showed
  - Standard retrieval does not promote diversity
  - Choice of language negligible for results
  - Combining content and concept-based methods gives best results

Dimensions	Type	2008		2007		2006	
		Runs	Groups	Runs	Groups	Runs	Groups
Annotation language	EN	514	24	271	17	137	2
	RND	495	2	32	2		
Modality	Text Only	404	22	167	15	121	2
	Mixed (text and image)	605	19	255	13	21	1
	Image Only	33	11	52	12		
Run type	Manual	3	1	19	3		
	Automatic	1039	25	455	19	142	2

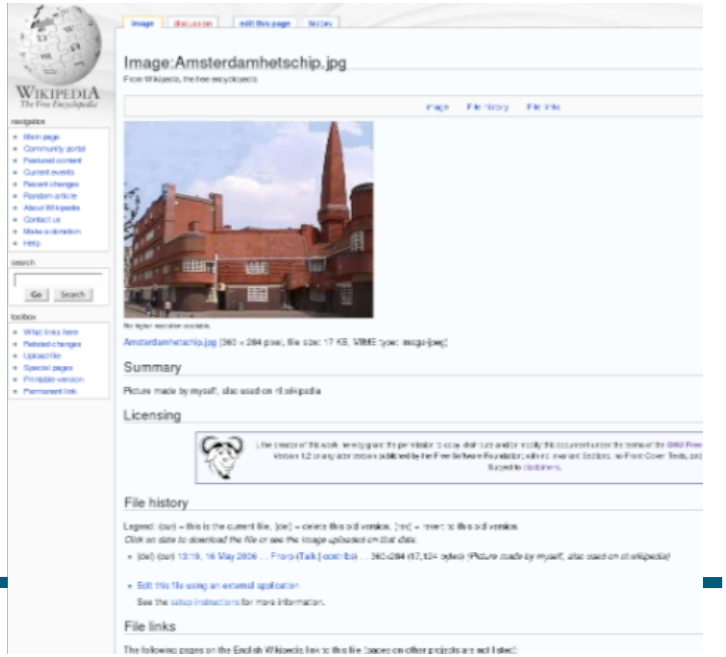
Visual Concept Detection Task

- Small hierarchy of concepts for annotation
- Purely visual concept detection works well
- Local features such as SIFT outperform other techniques
- Link with photo retrieval, but only used by a single group



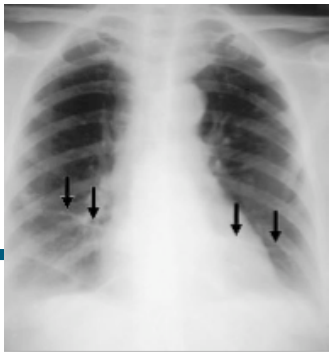
WikipediaMM Retrieval Task

- Semi-Structured annotation together with images
  - This year annotation and topics in English
- Not all topics contained images
  - Bias against visual retrieval
- Text retrieval works well
  - Visual concepts can improve overall performance
- Participants are judges



Medical Task

- Images and full-text articles of Radiology/ Radiographics (thanks to the RSNA!)
  - Captions of the figures with detailed information on the figures, subfigures
  - The kind of data that clinicians search
- Detailed search tasks as used may not be the most common for diagnosis, rather teaching
- More adapted for text retrieval, image analysis has to be done with care
  - Visual retrieval can improve early precision



Medical Annotation Task

- Again a hierarchy of classes for visual classification
  - Distribution of classes in training and test data not equal
  - Forced to use confidence on a hierarchy level
- Local features outperform global ones
- Machine learning techniques are key to success
- Results of past years published in special issue

CLEF 2008 Tracks

Ad-Hoc | iCLEF | QA@CLEF | Image CLEF | **Web CLEF** | Geo CLEF | Video CLEF

Promoting CLIR Research through Evaluation: WebCLEF

- Launched as a known-item search task in 2005, repeated in 2006
  - Resources created used for a number of purposes
- In 2007 a multilingual information synthesis task
  - For a given topic, systems extract important snippets from web pages
  - Topics and assessments created by participants
  - Few participants: task too difficult/too heavy
- In 2008, similar but simpler task
  - User model: knowledgable person writing survey article using only online sources in specified list of languages
  - Very disappointing participation

CLEF 2008 Tracks

Ad-Hoc | iCLEF | QA@CLEF | Image CLEF | Web CLEF | **Geo CLEF** | Video CLEF

Promoting CLIR Research through Evaluation: GeoCLEF

- Aim: to evaluate retrieval of multilingual documents with an emphasis on geographic search:
  - "find me news stories about riots near Dublin"
- Many documents contains geo-references expressed in multiple languages
- Standard IR systems (and evaluations) pay little attention to spatial aspects of queries and documents
  - Four editions
  - Document languages: English, German, Portuguese
  - 100 Topics: English, German, Portuguese
  - Monolingual and bilingual ad-hoc retrieval tasks

GeoCLEF 2008 Results

Best systems in mono-lingual and most competitive tasks (many runs) use specific geo reasoning

- named-entity recognition using Wikipedia
- NER Topic parsing (event part and geographic part)
- Geographic ontology (using geographic taxonomies such as GeoNames, World Gazetteer)
- query expansion using geographic ontology

For most other tasks (esp. bi-lingual), the best systems use no specific geo components

- Standard approaches like BM25 and blind relevance feedback also work well on Geographic IR

CLEF 2008 Tracks

Ad-Hoc | iCLEF | QA@CLEF | ImageCLEF | WebCLEF | GeoCLEF | VideoCLEF

Promoting CLIR Research through Evaluation: VideoCLEF

- Promote research on intelligent access to multimedia content in a multilingual environment
- Encourage exploitation of multimodal information streams: speech transcripts, video content, metadata, …
- Develop and evaluate multilingual video analysis tasks
- Extend the recent Cross-Language Speech Retrieval tracks into new challenges

  - 50 dual language videos (30 hours) from The Netherlands Institute for Sound and Vision
  - Videos are episodes of Dutch television documentaries
  - Dutch is the main language; English is embedded language
  - Dutch language archival metadata
    - ♣ **Speech recognition transcripts** in MPEG-7 by U. Twente
    - ♣ Shot-level keyframes supplied by Dublin City University

Main Achievements

- Stimulation of research activity in new, previously unexplored areas
- Study and implementation of evaluation methodologies for diverse types of cross-language IR systems
- Creation of a large set of empirical data about multilingual information access from the user perspective
- Quantitative and qualitative evidence with respect to best practice in cross-language system development
- Creation of reusable test collections for system benchmarking
- Building of a strong, multidisciplinary research community

Treble-CLEF

The CLEF research results have led to development of a new generation of multilingual retrieval system prototypes

BUT lack of technology transfer

CLEF 2008 – 2009 sponsored by 7FP within TrebleCLEF Coordination Action

Treble-CLEF extends the CLEF activity by:

- continuing to promote MLIA R&D via evaluation campaigns;
- providing a consistent training activity: tutorials, workshops, summer school;
- producing best practice guidelines for system implementation;
- providing resources to encourage the multilingual system development

www.trebleclef.eu

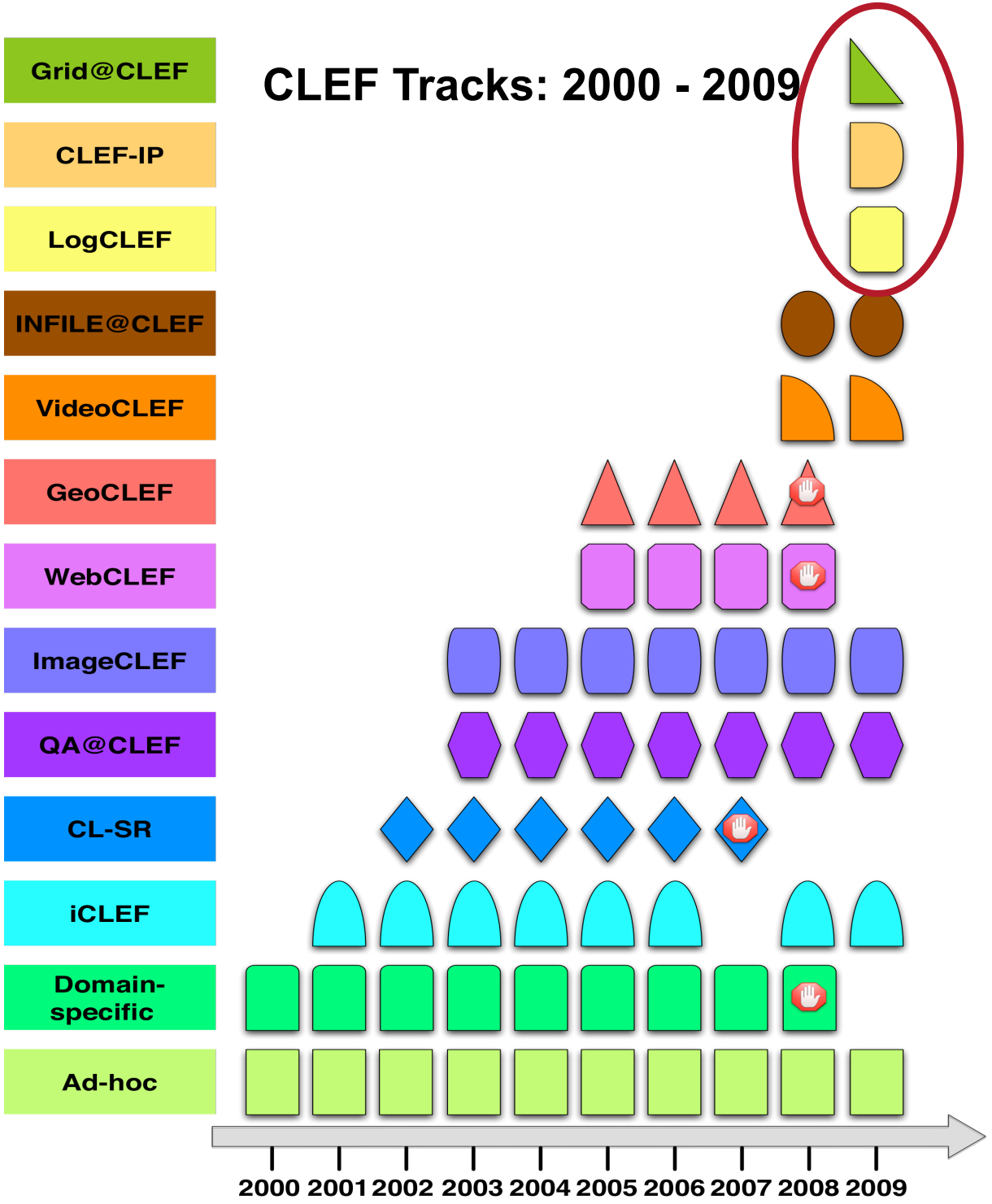
Approach

- Evaluation
  - test collections and laboratory evaluation
  - user evaluation and modelling
  - log analysis
- Best Practices & Guidelines
  - system-oriented aspects of MLIA applications
  - collaborative user studies
  - user-oriented aspects of MLIA interfaces
- Dissemination and Training
  - tutorials
  - workshops
  - summer school

TrebleCLEF & CLEF

Within TrebleCLEF **CLEF** will continue to promote R&D of multilingual, multimodal information access functionality with particular focus on user needs & in-depth results analysis:

- user modeling, e.g. the requirements of different classes of users when querying multilingual information sources
- results presentation, e.g. how can results be presented in the most useful and comprehensible way to the user
- language-specific experimentation, e.g. looking at differences across languages in order to derive best practices for each language


CLEF Tracks: 2000 - 2009
Grid@CLEF
CLEF-IP
LogCLEF
INFILE@CLEF
VideoCLEF
GeoCLEF
WebCLEF
ImageCLEF
QA@CLEF
CL-SR
iCLEF
Domain-specific
Ad-hoc
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009

CLEF 2009: New Tracks

- **Intellectual Property (CLEF-IP)**
  - Search tasks on more than 1M patent documents from European patent office in English, French, and German
- **Log File Analysis (LogCLEF)**
  - Analysis of queries as expression of user behaviour. Goal is to analyse and classify queries in order to improve search systems.
  - Logs from The European Library (TEL) will be used
- **Grid@CLEF**
  - Experiments designed to improve our understanding of MLIA systems and their behaviour with respect to languages

Grid@CLEF: Background

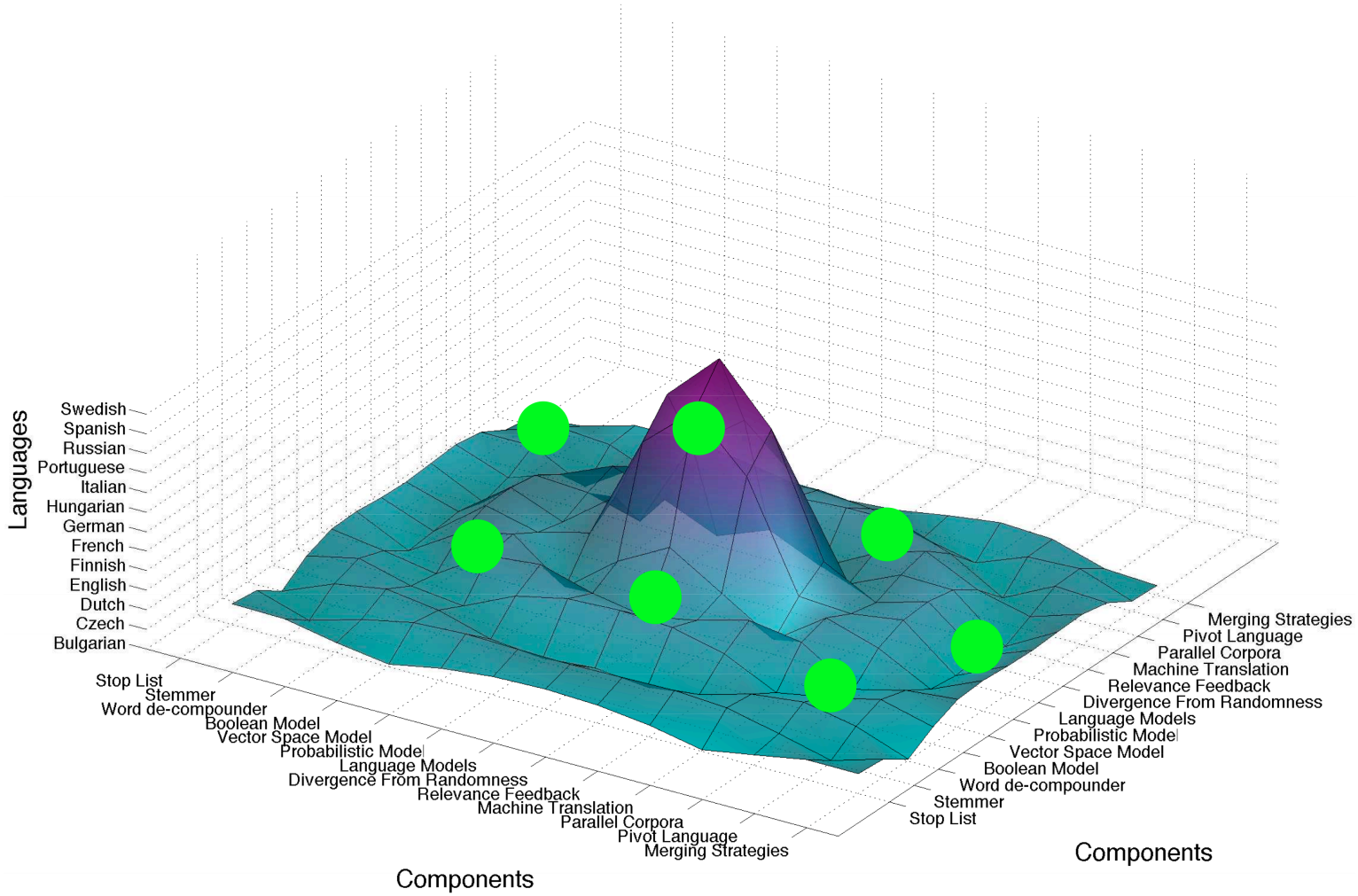
- The CLEF research community has been outstanding and very active in designing, developing, and testing MLIA methods and techniques, constantly improving the performances of such components

BUT

- Do we really know how MLIA components behave with respect to languages?

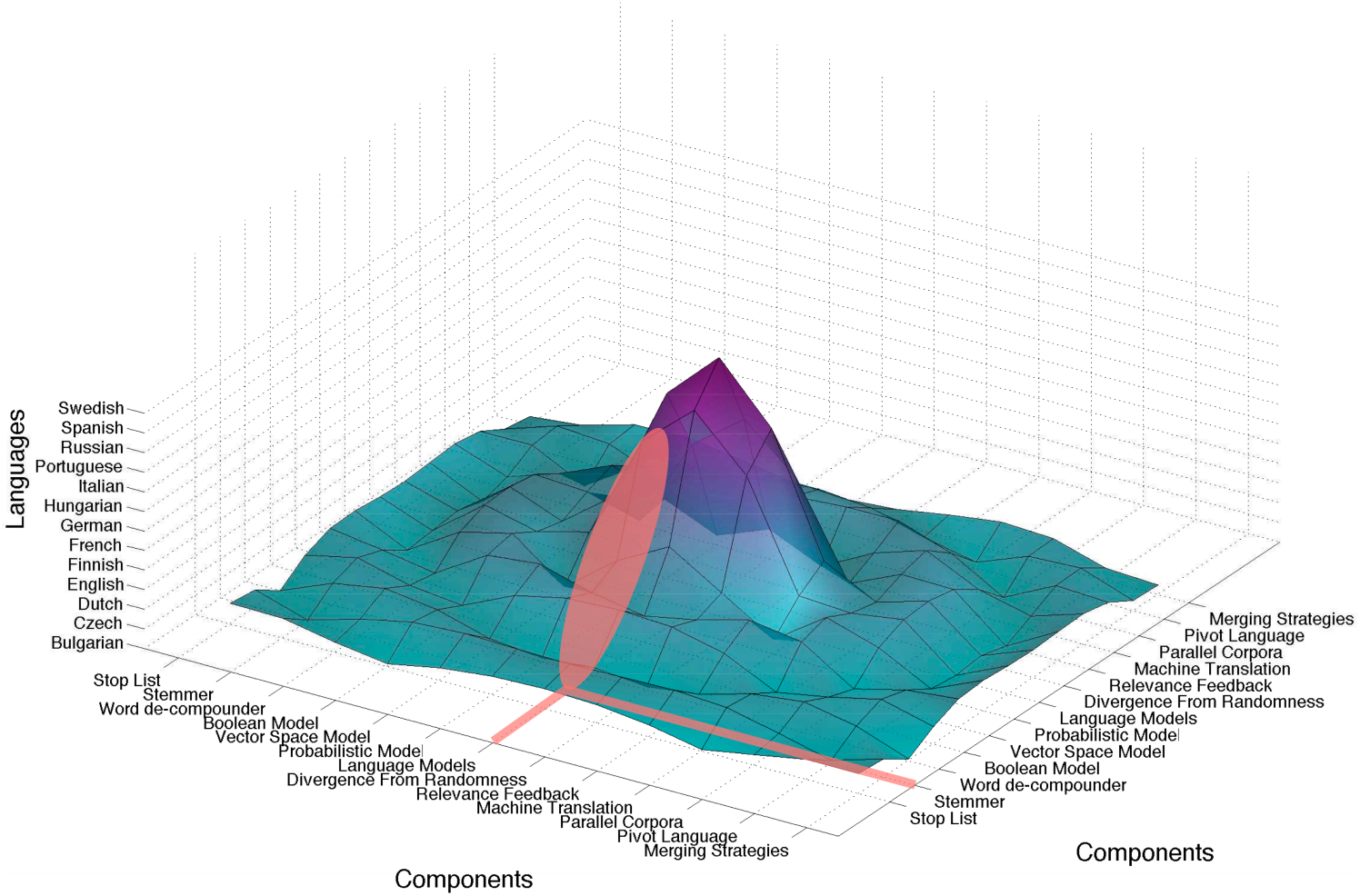
- Do we have a deep comprehension of how these components interact together when the language changes?

Grid@CLEF: Where we are?



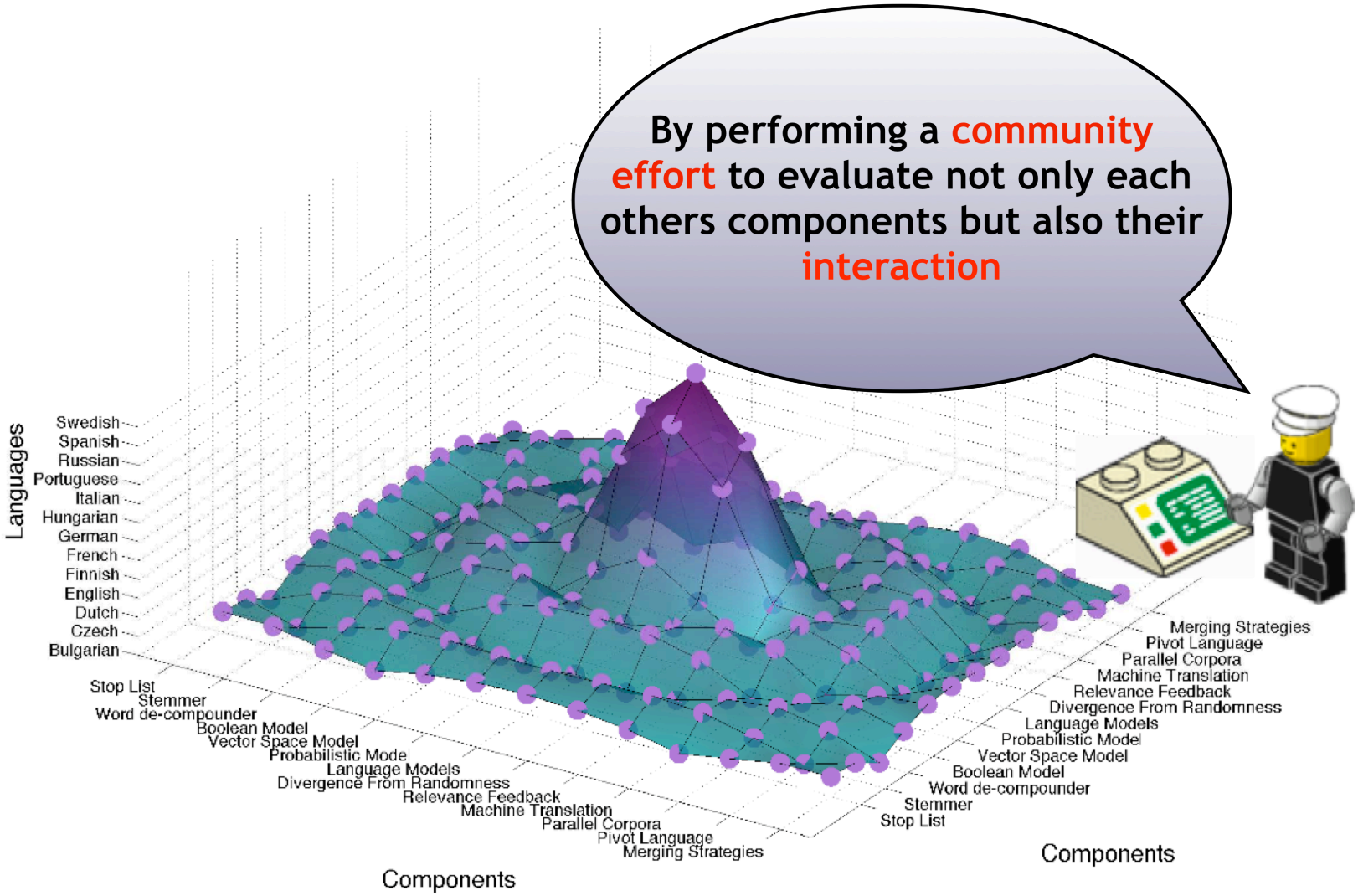
NTCIR-7 Meeting
Tokyo, 16-19 December, 2008

Grid@CLEF: Where we are?



NTCIR-7 Meeting
Tokyo, 16-19 December, 2008

Grid@CLEF: How Can We Get There?



NTCIR-7 Meeting
Tokyo, 16-19 December, 2008

Grid@CLEF: Approach

- **Re-use** the **resources** and **experimental collections** currently available in CLEF
- Select a **core set of components** to be tested (stop lists, stemmers, IR models, ...)
- Design a very controlled environment to clearly isolate relevant factors, i.e. **behaviour across languages** and **interaction of components**
- Two modalities of participation:
  - **island mode**: each group works on its own and by complying with the experimental protocol puts its own dots on the grid
  - **archipelago mode**: groups will participate in a framework to plug-in and connect their components in order to study their interaction
- **Comparative analysis** of the results

Summing Up

- Importance of Test Collection Creation
- How best to make the data freely available
- Distinguish between language-specific and language independent issues
- Need to understand complex interaction between topics, systems & data
- Don’t forget the User
- Cruciality of success / failure analysis
- Resource sharing / Community Building

Points for Discussion

- What are the current pressing research issues?
- How to model / study multicultural issues
- What new tasks/evaluation methodologies are needed to address more advanced information requirements?
- How can we best reduce the gap between research and application communities?

TrebleCLEF Survey

Language Resources for MLIA: Existing Resources and Best Practices

Aim of the Survey is to collect information on the current needs of MLIA system developers in terms of applications, resources, evaluation activities

Compile the questionnaire online at

www.trebleclef.eu/clef

NTCIR-7 Meeting
Tokyo, 16-19 December, 2008