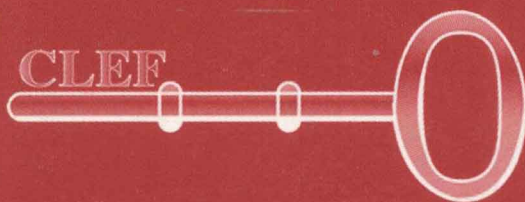


Carol Peters Paul Clough  
Julio Gonzalo Gareth J. F. Jones  
Michael Kluck Bernardo Magnini (Eds.)

LNCS 3491

# Multilingual Information Access for Text, Speech and Images

5th Workshop of the  
Cross-Language Evaluation Forum, CLEF 2004  
Bath, UK, September 2004, Revised Selected Papers



Carol Peters Paul Clough  
Julio Gonzalo Gareth J.F. Jones  
Michael Kluck Bernardo Magnini (Eds.)

# Multilingual Information Access for Text, Speech and Images

5th Workshop of the  
Cross-Language Evaluation Forum, CLEF 2004  
Bath, UK, September 15-17, 2004  
Revised Selected Papers



Springer

## Volume Editors

Carol Peters

Istituto di Scienza e Tecnologie dell'Informazione  
Consiglio Nazionale delle Ricerche (ISTI-CNR)  
Via G. Moruzzi 1, 56124 Pisa, Italy  
E-mail: carol.peters@isti.cnr.it

Paul Clough

University of Sheffield, Department of Information Studies  
Western Bank, Sheffield, S10 2TN, UK  
E-mail: p.d.clough@sheffield.ac.uk

Julio Gonzalo

Universidad Nacional de Educación a Distancia  
Departamento de Lenguajes y Sistemas Infomáticos  
Juan del Rosal, 16, 28040 Madrid, Spain  
E-mail: julio@lsi.uned.es

Gareth J.F. Jones

Dublin City University, School of Computing  
Glasnevin, Dublin 9, Ireland, UK  
E-mail: gareth.jones@computing.dcu.ie

Michael Kluck

Informationszentrum Sozialwissenschaften (IZ)  
Lennéstr. 30, 53113 Bonn, Germany  
E-mail: michaelkluck@web.de

Bernardo Magnini

Centro per la Ricerca Scientifica e Tecnologica (ITC-Irst)  
Via Sommarive 18, 38050 Povo (TN), Italy  
E-mail: magnini@itc.it

Library of Congress Control Number: 2005928950

CR Subject Classification (1998): H.3, I.2, H.4

ISSN 0302-9743

ISBN-10 3-540-27420-0 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-27420-9 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2005

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 11519645 06/3142 5 4 3 2 1 0

## Preface

The fifth campaign of the Cross-Language Evaluation Forum (CLEF) for European languages was held from January to September 2004. Participation in the CLEF campaigns has increased each year and CLEF 2004 was no exception: 55 groups submitted results for one or more of the different tracks compared with 42 groups in the previous year. CLEF 2004 also marked a breaking point with respect to previous campaigns. The focus was no longer mainly concentrated on multilingual document retrieval as in previous years but was diversified to include different kinds of text retrieval across languages (e.g., exact answers in the question-answering track) and retrieval on different kinds of media (i.e., not just plain text but collections containing image and speech as well). In addition, increasing attention was given to issues that regard system usability and user satisfaction with tasks to measure the effectiveness of interactive systems or system components being included in both the cross-language question answering and image retrieval tasks with the collaboration of the coordinators of the interactive track.

The campaign culminated in a two-and-a-half-day workshop held in Bath, UK, 15–17 September, immediately following the 8th European Conference on Digital Libraries. The workshop was attended by nearly 100 researchers and system developers. In addition to presentations by participants in the campaign, talks included reports on the activities of the NTCIR evaluation initiative for Asian languages, and on industrial experience in building cross-language applications. The final session consisted of a panel in which the members attempted to analyze the current organization of the CLEF campaigns in depth, discussing whether CLEF is working on the right problems, choosing its investments wisely, and giving sufficient attention to the user perspective. Suggestions for the CLEF 2005 campaign included multilingual Web retrieval and a cross-language Geographic Information Retrieval track.

CLEF 2004 was conducted as an activity of the DELOS Network of Excellence on Digital Libraries, within the framework of the Information Society Technologies programme of the European Commission. These post-campaign proceedings were prepared with the assistance of the Center for the Evaluation of Language and Communication Technologies (CELCT), Trento, Italy. The support of DELOS and CELCT is gratefully acknowledged. We should also like to thank the other members of the CLEF Steering Committee for their assistance in the coordination of this event.

April 2005

Carol Peters  
Paul Clough  
Julio Gonzalo  
Gareth J.F. Jones  
Michael Kluck  
Bernardo Magnini

# CLEF 2004 Coordination

CLEF is coordinated by the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, ISTI-CNR, Pisa.

The following institutions contributed to the organization of the different tracks of the CLEF 2004 campaign:

- Center for the Evaluation of Language and Communication Technologies (CELCT), Trento, Italy
- Centro per la Ricerca Scientifica e Tecnologica, Istituto Trentino di Cultura, Trento, Italy
- College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, USA
- Department of Computer Science and Information Systems, University of Limerick, Ireland
- Department of Information Studies, University of Sheffield, UK
- Department of Information Studies, University of Tampere, Finland
- Eurospider Information Technology AG, Zürich, Switzerland
- Evaluations and Language Resources Distribution Agency, ELDA, Paris, France
- German Research Centre for Artificial Intelligence, DFKI, Saarbrücken, Germany
- Information and Language Processing Systems, University of Amsterdam, Netherlands
- Informationszentrum Sozialwissenschaften, Bonn, Germany
- Lenguajes y Sistemás Informáticos, Universidad Nacional de Educación a Distancia, Madrid, Spain
- Linguatca, Sintef, Oslo, Norway
- Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Sofia, Bulgaria
- National Institute of Standards and Technology, Gaithersburg MD, USA
- School of Computing, Dublin City University, Ireland
- University and University Hospitals of Geneva, Switzerland

## CLEF 2004 Steering Committee

- Maristella Agosti, University of Padua, Italy
- Eija Airio, University of Tampere, Finland
- Martin Braschler, Eurospider Information Technologies, Switzerland
- Hsin-Hsi Chen, National Taiwan University, Taipei, Taiwan
- Khalid Choukri, Evaluations and Language Resources Distribution Agency, ELDA, Paris, France
- Paul Clough, University of Sheffield, UK
- David A. Evans, Clairvoyance Corporation, USA
- Marcello Federico, ITC-irst, Trento, Italy
- Christian Fluhr, CEA-LIST, Fontenay-aux-Roses, France
- Norbert Fuhr, University of Duisburg, Germany
- Frederic C. Gey, UC Berkeley, USA
- Julio Gonzalo, LSI-UNED, Madrid, Spain
- Donna Harman, National Institute of Standards and Technology, USA
- Gareth Jones, Dublin City University, Ireland
- Franciska de Jong, University of Twente, Netherlands
- Noriko Kando, National Institute of Informatics, Tokyo, Japan
- Jussi Karlgren, Swedish Institute of Computer Science, Sweden
- Michael Kluck, Informationszentrum Sozialwissenschaften Bonn, Germany
- Natalia Loukachevitch, Moscow State University, Russia
- Bernardo Magnini, ITC-irst, Trento, Italy
- Paul McNamee, Johns Hopkins University, USA
- Henning Müller, University and University Hospitals, Geneva, Switzerland
- Douglas W. Oard, University of Maryland, USA
- Maarten de Rijke, University of Amsterdam, Netherlands
- Jacques Savoy, University of Neuchâtel, Switzerland
- Peter Schäuble, Eurospider Information Technologies, Switzerland
- Richard Sutcliffe, University of Limerick, Ireland
- Hans Uszkoreit, German Research Center for Artificial Intelligence (DFKI), Germany
- Felisa Verdejo, LSI-UNED, Madrid, Spain
- José Luis Vicedo, University of Alicante, Spain
- Ellen Voorhees, National Institute of Standards and Technology, USA
- Christa Womser-Hacker, University of Hildesheim, Germany

# Table of Contents

What Happened in CLEF 2004? <i>Carol Peters</i> .....	1
<b>Part I. Ad Hoc Text Retrieval Tracks</b>	
CLEF 2004: Ad Hoc Track Overview and Results Analysis <i>Martin Braschler, Giorgio M. Di Nunzio, Nicola Ferro, Carol Peters</i> .....	10
<b>Mainly Cross-Language Experiments</b>	
Selection and Merging Strategies for Multilingual Information Retrieval <i>Jacques Savoy, Pierre-Yves Berger</i> .....	27
Using Surface-Syntactic Parser and Deviation from Randomness <i>Jean-Pierre Chevallet, Gilles Sérasset</i> .....	38
Cross-Language Retrieval Using HAIRCUT at CLEF 2004 <i>Paul McNamee, James Mayfield</i> .....	50
Experiments on Statistical Approaches to Compensate for Limited Linguistic Resources <i>Giorgio M. Di Nunzio, Nicola Ferro, Nicola Orio</i> .....	60
Application of Variable Length <i>N</i> -Gram Vectors to Monolingual and Bilingual Information Retrieval <i>Daniel Gayo-Avello, Darío Álvarez-Gutiérrez, José Gayo-Avello</i> .....	73
Integrating New Languages in a Multilingual Search System Based on a Deep Linguistic Analysis <i>Romarc Besançon, Olivier Ferret, Christian Fluhr</i> .....	83
IR-n r2: Using Normalized Passages <i>Fernando Llopis, Rafael Muñoz, Rafael M. Terol, Elisa Noguera</i> .....	90
Using COTS Search Engines and Custom Query Strategies at CLEF <i>David Nadeau, Mario Jarmasz, Caroline Barrière, George Foster, Claude St-Jacques</i> .....	100

Report on Thomson Legal and Regulatory Experiments at CLEF-2004 <i>Isabelle Moulinier, Ken Williams</i> .....	110
Effective Translation, Tokenization and Combination for Cross-Lingual Retrieval <i>Jaap Kamps, Sisay Fissaha Adafre, Maarten de Rijke</i> .....	123
Two-Stage Refinement of Transitive Query Translation with English Disambiguation for Cross-Language Information Retrieval: An Experiment at CLEF 2004 <i>Kazuaki Kishida, Noriko Kando, Kuang-Hua Chen</i> .....	135
Dictionary-Based Amharic – English Information Retrieval <i>Atelach Alemu Argaw, Lars Asker, Rickard Cöster, Jussi Karlgren</i> .....	143
Dynamic Lexica for Query Translation <i>Jussi Karlgren, Magnus Sahlgren, Timo Järvinen, Rickard Cöster</i> ....	150
SINAI at CLEF 2004: Using Machine Translation Resources with a Mixed 2-Step RSV Merging Algorithm <i>Fernando Martínez-Santiago, Miguel A. García-Cumbreras, Manuel C. Díaz-Galiano, L. Alfonso Ureña</i> .....	156
Mono- and Crosslingual Retrieval Experiments at the University of Hildesheim <i>René Hackl, Thomas Mandl, Christa Womser-Hacker</i> .....	165
University of Chicago at CLEF2004: Cross-Language Text and Spoken Document Retrieval <i>Gina-Anne Levow, Irina Matveeva</i> .....	170
UB at CLEF2004: Cross Language Information Retrieval Using Statistical Language Models <i>Miguel E. Ruiz, Munirathnam Srikanth</i> .....	180
MIRACLE’s Hybrid Approach to Bilingual and Monolingual Information Retrieval <i>José M. Goñi-Menoyo, José C. González, José Luis Martínez-Fernández, Julio Villena</i> .....	188
Searching a Russian Document Collection Using English, Chinese and Japanese Queries <i>Fredric C. Gey</i> .....	200



Dublin City University at CLEF 2004: Experiments in Monolingual, Bilingual and Multilingual Retrieval <i>Gareth J.F. Jones, Michael Burke, John Judge, Anna Khasin, Adenike Lam-Adesina, Joachim Wagner</i> .....	207
--	-----

## Monolingual Experiments

Finnish, Portuguese and Russian Retrieval with Hummingbird SearchServer <sup>TM</sup> at CLEF 2004 <i>Stephen Tomlinson</i> .....	221
--	-----

Data Fusion for Effective European Monolingual Information Retrieval <i>Jacques Savoy</i> .....	233
--	-----

The XLDB Group at CLEF 2004 <i>Nuno Cardoso, Mário J. Silva, Miguel Costa</i> .....	245
--	-----

The University of Glasgow at CLEF 2004: French Monolingual Information Retrieval with Terrier <i>Christina Lioma, Ben He, Vassilis Plachouras, Iadh Ounis</i> .....	253
--	-----

## Part II. Domain-Specific Document Retrieval

The Domain-Specific Track in CLEF 2004: Overview of the Results and Remarks on the Assessment Process <i>Michael Kluck</i> .....	260
---	-----

University of Hagen at CLEF 2004: Indexing and Translating Concepts for the GIRT Task <i>Johannes Leveling, Sven Hartrumpf</i> .....	271
---	-----

IRIT at CLEF 2004: The English GIRT Task <i>Mustapha Baziz, Mohand Boughanem, Nathalie Aussenac-Gilles</i> .....	283
---	-----

Ricoh at CLEF 2004 <i>Yuichi Kojima</i> .....	292
--	-----

GIRT and the Use of Subject Metadata for Retrieval <i>Vivien Petras</i> .....	298
--	-----

### Part III. Interactive Cross-Language Information Retrieval

iCLEF 2004 Track Overview: Pilot Experiments in Interactive Cross-Language Question Answering <i>Julio Gonzalo, Douglas W. Oard</i> .....	310
Interactive Cross-Language Question Answering: Searching Passages Versus Searching Documents <i>Fernando López-Ostenero, Julio Gonzalo, Víctor Peinado, Felisa Verdejo</i> .....	323
Improving Interaction with the User in Cross-Language Question Answering Through Relevant Domains and Syntactic Semantic Patterns <i>Borja Navarro, Lorenza Moreno, Sonia Vázquez, Fernando Llopis, Andrés Montoyo, Miguel Ángel Varó</i> .....	334
Cooperation, Bookmarking, and Thesaurus in Interactive Bilingual Question Answering <i>Preben Hansen, Jussi Karlgren, Magnus Sahlgren</i> .....	343
Summarization Design for Interactive Cross-Language Question Answering <i>Daqing He, Jianqiang Wang, Jun Luo, Douglas W. Oard</i> .....	348
Interactive and Bilingual Question Answering Using Term Suggestion and Passage Retrieval <i>Carlos G. Figuerola, Angel F. Zazo, José L. Alonso Berrocal, Emilio Rodríguez Vázquez de Aldana</i> .....	363

### Part IV. Multiple Language Question Answering

Overview of the CLEF 2004 Multilingual Question Answering Track <i>Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov, Richard F.E. Sutcliffe</i> .....	371
A Question Answering System for French <i>Laura Perret</i> .....	392
Cross-Language French-English Question Answering Using the DLT System at CLEF 2004 <i>Richard F.E. Sutcliffe, Igal Gabbay, Michael Mulcahy, Aoife O’Gorman</i> .....	404

Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question/Answering System <i>Günter Neumann, Bogdan Sacaleanu</i> .....	411
Making Stone Soup: Evaluating a Recall-Oriented Multi-stream Question Answering System for Dutch <i>David Ahn, Valentin Jijkoun, Karin Müller, Maarten de Rijke, Stefan Schlobach, Gilad Mishne</i> .....	423
The DIOGENE Question Answering System at CLEF-2004 <i>Hristo Tanev, Matteo Negri, Bernardo Magnini, Milen Kouylekov</i> .....	435
Cross-Lingual Question Answering Using Off-the-Shelf Machine Translation <i>Kisuh Ahn, Beatrice Alex, Johan Bos, Tiphaine Dalmás, Jochen L. Leidner, Matthew B. Smillie</i> .....	446
Bulgarian-English Question Answering: Adaptation of Language Resources <i>Petya Osenova, Alexander Simov, Kiril Simov, Hristo Tanev, Milen Kouylekov</i> .....	458
Answering French Questions in English by Exploiting Results from Several Sources of Information <i>Brigitte Grau, Gabriel Illouz, Laura Monceaux, Isabelle Robba, Anne Vilnat, Guillaume Bourdil, Faïza Elkateb-Gara, Olivier Ferret, Benoît Mathieu</i> .....	470
Finnish as Source Language in Bilingual Question Answering <i>Lili Aunimo, Reeta Kuuskoski, Juha Makkonen</i> .....	482
miraQA: Experiments with Learning Answer Context Patterns from the Web <i>César de Pablo-Sánchez, José Luis Martínez-Fernández, Paloma Martínez, Julio Villena</i> .....	494
Question Answering for Spanish Supported by Lexical Context Annotation <i>Manuel Pérez-Coutiño, Tamar Solorio, Manuel Montes-y-Gómez, Aurelio López-López, Luis Villaseñor-Pineda</i> .....	502
Question Answering Using Sentence Parsing and Semantic Network Matching <i>Sven Hartrumpf</i> .....	512

First Evaluation of Esfinge – A Question Answering System for Portuguese  
*Luís Costa* ..... 522

University of Évora in QA@CLEF-2004  
*Paulo Quaresma, Luís Quintano, Irene Rodrigues, José Saias, Pedro Salgueiro* ..... 534

COLE Experiments at QA@CLEF 2004 Spanish Monolingual Track  
*Enrique Méndez Díaz, Jesús Vilares Ferro, David Cabrero Souto* ..... 544

Does English Help Question Answering in Spanish?  
*José Luis Vicedo, Maximiliano Saiz, Rubén Izquierdo, Fernando Llopis* ..... 552

The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints  
*Daniel Ferrés, Samir Kanaan, Alicia Ageno, Edgar González, Horacio Rodríguez, Mihai Surdeanu, Jordi Turmo* ..... 557

ILC-UniPI Italian QA  
*Francesca Bertagna, Luminita Chiran, Maria Simi* ..... 569

Question Answering Pilot Task at CLEF 2004  
*Jesús Herrera, Anselmo Peñas, Felisa Verdejo* ..... 581

Evaluation of Complex Temporal Questions in CLEF-QA  
*Estela Saquete, José Luis Vicedo, Patricio Martínez-Barco, Rafael Muñoz, Fernando Llopis* ..... 591

**Part V. Cross-Language Retrieval in Image Collections**

The CLEF 2004 Cross-Language Image Retrieval Track  
*Paul Clough, Henning Müller, Mark Sanderson* ..... 597

Caption and Query Translation for Cross-Language Image Retrieval  
*Paul Clough* ..... 614

Pattern-Based Image Retrieval with Constraints and Preferences on ImageCLEF 2004  
*Maximiliano Saiz, José Luis Vicedo, Rubén Izquierdo* ..... 626

How to Visually Retrieve Images from the St. Andrews Collection Using GIFT <i>Henning Müller, Antoine Geissbühler</i> . . . . .	633
UNED at ImageCLEF 2004: Detecting Named Entities and Noun Phrases for Automatic Query Expansion and Structuring <i>Víctor Peinado, Javier Artiles, Fernando López-Ostenero, Julio Gonzalo, Felisa Verdejo</i> . . . . .	643
Dublin City University at CLEF 2004: Experiments with the ImageCLEF St. Andrew's Collection <i>Gareth J.F. Jones, Declan Groves, Anna Khasin, Adenike Lam-Adesina, Bart Mellebeek, Andy Way</i> . . . . .	653
From Text to Image: Generating Visual Query for Image Retrieval <i>Wen-Cheng Lin, Yih-Chen Chang, Hsin-Hsi Chen</i> . . . . .	664
Toward Cross-Language and Cross-Media Image Retrieval <i>Carmen Alvarez, Ahmed Id Oumohmed, Max Mignotte, Jian-Yun Nie</i> . . . . .	676
FIRE – Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation <i>Thomas Deselaers, Daniel Keysers, Hermann Ney</i> . . . . .	688
MIRACLE Approach to ImageCLEF 2004: Merging Textual and Content-Based Image Retrieval <i>José Luis Martínez-Fernández, Ana García-Serrano, Julio Villena, Víctor Méndez-Sáenz</i> . . . . .	699
Cross-Media Feedback Strategies: Merging Text and Image Information to Improve Image Retrieval <i>Romaric Besançon, Patrick Hède, Pierre-Alain Moellic, Christian Fluhr</i> . . . . .	709
ImageCLEF 2004: Combining Image and Multi-lingual Search for Medical Image Retrieval <i>Henning Müller, Antoine Geissbühler, Patrick Ruch</i> . . . . .	718
Multi-modal Information Retrieval Using FINT <i>Menno van Zaanen, Guido de Croon</i> . . . . .	728
Medical Image Retrieval Using Texture, Locality and Colour <i>Peter Howarth, Alexei Yavlinsky, Daniel Heesch, Stefan Rürger</i> . . . . .	740

SMIRE: Similar Medical Image Retrieval Engine <i>Pei-Cheng Cheng, Been-Chian Chien, Hao-Ren Ke, Wei-Pang Yang</i> .....	750
A Probabilistic Approach to Medical Image Retrieval <i>Koen Lubbers, Arjen P. de Vries, Theo Huibers, Paul van der Vet</i> ...	761
UB at CLEF2004 Cross Language Medical Image Retrieval <i>Miguel E. Ruiz, Munirathnam Srikanth</i> .....	773
Content-Based Queries on the CasImage Database Within the IRMA Framework <i>Christian Thies, Mark Oliver Güld, Benedikt Fischer, Thomas M. Lehmann</i> .....	781
Comparison and Combination of Textual and Visual Features for Interactive Cross-Language Image Retrieval <i>Pei-Cheng Cheng, Jen-Yuan Yeh, Hao-Ren Ke, Been-Chian Chien, Wei-Pang Yang</i> .....	793
MSU at ImageCLEF: Cross Language and Interactive Image Retrieval <i>Vineet Bansal, Chen Zhang, Joyce Y. Chai, Rong Jin</i> .....	805
<b>Part VI. Cross-Language Spoken Document Retrieval</b>	
CLEF 2004 Cross-Language Spoken Document Retrieval Track <i>Marcello Federico, Nicola Bertoldi, Gina-Anne Levow, Gareth J.F. Jones</i> .....	816
<b>Part VII. Issues in CLIR and in Evaluation</b>	
The Key to the First CLEF with Portuguese: Topics, Questions and Answers in CHAVE <i>Diana Santos, Paulo Rocha</i> .....	821
How Do Named Entities Contribute to Retrieval Effectiveness? <i>Thomas Mandl, Christa Womser-Hacker</i> .....	833
<b>Author Index</b> .....	843

# What Happened in CLEF 2004?

Carol Peters

Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Pisa, Italy  
carol.peters@isti.cnr.it

**Abstract.** The organization of the CLEF 2004 evaluation campaign is described and details are provided concerning the tracks, test collections and participation. Information on new activities for CLEF 2005 is also given.

## 1 Introduction

This volume reports the results of the fifth in a series of annual system evaluation campaigns organised by the Cross-Language Evaluation Forum (CLEF)<sup>1</sup>. The main objectives of CLEF are (i) to provide an infrastructure that facilitates testing of all kinds of multilingual information access systems – from monolingual retrieval for multiple languages to the implementation of complete multilingual multimedia search services, and (ii) to construct test-suites of reusable data that can be used for benchmarking purposes. These objectives are achieved through the organisation of evaluation campaigns that culminate each year in a workshop in which the groups that participated in the campaign can report and discuss their experiments. An additional aim of CLEF is to encourage contacts between the R&D and the application communities and promote the industrial take-up of research results.

The main features of the 2004 campaign are briefly outlined below in order to provide the necessary background to the experiments reported in these post-campaign proceedings.

## 2 Tracks and Tasks in CLEF 2004

In recent years, CLEF distinguished between the core tracks, which were those offered regularly each campaign (the monolingual, bilingual, multilingual and domain-specific tracks), and additional tracks, which were organised on an experimental basis with the objective of identifying new requirements and appropriate methodologies for their testing in a cross-language context. This distinction no longer held in 2004. The interactive track, run since 2001, was finally recognised as part of the main activity, and the great success of the pilot tracks in CLEF 2003, and in particular the cross-language question answering and image retrieval activities, led to their inclusion as regular tracks. This meant that the scope of CLEF 2004 was considerably widened with respect to previous years, with much attention being given to tasks involving information extraction (question answering) and retrieval from multimedia.

---

<sup>1</sup> CLEF 2004 is included in the activities of the DELOS Network of Excellence on Digital Libraries, funded by the Sixth Framework Programme of the European Commission. DELOS is an “old” friend of CLEF, having promoted the first two campaigns in 2000 and 2001. For information on DELOS, see [www.delos.info](http://www.delos.info).

CLEF 2004 thus offered six tracks designed to evaluate the performance of systems for:

- mono-, bi- and multilingual document retrieval on news collections (Ad hoc)
- mono- and cross-language domain-specific retrieval (GIRT)
- interactive cross-language retrieval (iCLEF)
- multiple language question answering (QA@CLEF)
- cross-language retrieval on image collections (ImageCLEF)
- cross-language spoken document retrieval (CL-SDR)

The organisation of each of these tracks and the results obtained are described and commented in the track overviews at the beginning of each section of this volume.

### 3 CLEF 2004 Test Collections

CLEF campaigns adopt a comparative evaluation approach in which system performance is measured using appropriate test collections. The test collections consist of sets of sample query statements often called “topics”, document collections, and relevance judgments determining the set of relevant documents in a collection for a given query statement. All language dependent tasks such as topic/question creation and relevance assessment are performed in a distributed setting by native speakers. Rules are established and a tight central coordination is maintained in order to ensure consistency and coherency of topic and relevance judgment sets over the different collections, languages and tracks.

Five separate document collections were used in CLEF 2004:

- CLEF multilingual comparable corpus
- GIRT social science database
- St Andrews historical photographic archive
- CasImage radiological medical database
- Speech transcriptions supplied by TREC

The main CLEF collection is the multilingual comparable corpus of newspaper and news agency documents. In 2004, this collection was used by the Ad hoc, iCLEF and QA@CLEF tracks. The multilingual corpus increases in size and coverage every year as new collections and languages are added. In 2004 it contained nearly 1.8 million news documents from the same time period (1994-1995) in ten languages: Dutch, English, Finnish, French, German, Italian, Portuguese<sup>2</sup>, Russian, Spanish and Swedish. Table 1 gives the main specifics of this collection and Table 2 shows which part of this collection was used in which track in 2004.

However, news media have characteristics which may not hold true for other genres: e.g. wide use of proper nouns (names and places), association of date stamps, particular style of writing and a rapid evolution of general-purpose vocabulary. Certain features may facilitate access and retrieval, others may hinder it. For this reason, a separate document collection is included for systems tuned for domain-specific tasks.

---

<sup>2</sup> Portuguese was a new addition in 2004; Hungarian and Bulgarian newspaper collections are being added for 2005.



**Table 1.** Sources and dimensions of the main CLEF 2004 multilingual document collection

Collection	Added in	Size (MB)	No. of Docs	Median Size of Docs. (Bytes)	Median Size of Docs. (Tokens) <sup>3</sup>	Median Size of Docs. (Features)
Dutch: Algemeen Dagblad 94/95	2001	241	106483	1282	166	112
Dutch: NRC Handelsblad 94/95	2001	299	84121	2153	354	203
English: LA Times 94	2000	425	113005	2204	421	246
English: Glasgow Herald 95	2003	154	56472	2219	343	202
Finnish: Aamulehti late 94/95	2002	137	55344	1712	217	150
French: Le Monde 94	2000	158	44013	1994	361	213
French: ATS 94	2001	86	43178	1683	227	137
French: ATS 95	2003	88	42615	1715	234	140
German: Frankfurter Rundschau94	2000	320	139715	1598	225	161
German: Der Spiegel 94/95	2000	63	13979	1324	213	160
German: SDA 94	2001	144	71677	1672	186	131
German: SDA 95	2003	144	69438	1693	188	132
Italian: La Stampa 94	2000	193	58051	1915	435	268
Italian: AGZ 94	2001	86	50527	1454	187	129
Italian: AGZ 95	2003	85	48980	1474	192	132
Portuguese: Público 94	2004	164	51751	NA	NA	NA
Portuguese: Público 95	2004	176	55070	NA	NA	NA
Russian: Izvestia 95	2003	68	16761	NA	NA	NA
Spanish: EFE 94	2001	511	215738	2172	290	171
Spanish: EFE 95	2003	577	238307	2221	299	175
Swedish: TT 94/95	2002	352	142819	2171	183	121

SDA/ATS/AGZ = Schweizerische Depeschagentur (Swiss News Agency)

EFE = Agencia EFE S.A (Spanish News Agency)

TT = Tidningarnas Telegrambyrå (Swedish newspaper)

NA = Not Available at this moment

<sup>3</sup> The number of tokens extracted from each document can vary slightly across systems, depending on the respective definition of what constitutes a token. Consequently, the number of tokens and features given in this table are approximations and may differ from actual implemented systems.