CLEF - IST-2000-31002
Cross-Language Evaluation Forum

Results Interpretation and Report (Campaign 2)

D4.2.2

Deliverable Type: REPORT
Number: D4.2.2
Nature: Public
Contractual Date of Delivery: month 27
Actual Date of Delivery: month 27
Task WP4.2

Name of responsible: Eurospider Information Technology AG (EIT)

Authors: Martin Braschler, EIT

contact info
martin.braschler@eurospider.com

Abstract

Deliverable 4.2.2 presents an overview of the results obtained by the participants in the CLEF 2003 campaign, and discusses the validity of the results based on measures of statistical significance and the stability of the evaluation.

Keyword List

Information Retrieval, Evaluation Results, CLIR, Statistical Testing, Relevance Assessments, Pool Validity

Executive Summary

In this deliverable, we present an overview of the results obtained by the participants at the CLEF 2003 campaign, and discuss the validity of the results based on measures of statistical significance and the stability of the evaluation.

CLEF 2003 has seen a large increase in the number of experiments submitted by the participants, as compared with earlier CLEF campaigns. We also note that CLEF continues to attract a large percentage of participants from Europe. We report both on the main characteristics of the experiments and on trends seen in the approaches and methodologies used by the participants. With over 400 experiments submitted for the core tracks alone, an exhaustive report on all individual results is by necessity outside the scope of this report, but all the main results are summarized and analyzed.

Careful statistical analysis allows a potential generalization of claims based on findings inside CLEF. We describe how to carry out such statistical analysis and give results for the main task in CLEF, the multilingual track.

CLEF relies heavily on relevance assessments for the calculation of its performance measures. To ensure validity of the results published by CLEF, we investigate the quality of the relevance assessments by computing their coverage.

The importance of statistical analysis and analysis of the validity of relevance assessments should not be underestimated, since the reusability of the testing resources built by CLEF is central to the success of the project.

Main findings of the 2003 campaign are:

1. A lot of detailed fine-tuning for the characteristics of specific languages has taken place.
2. The most widely spoken languages on offer are also the most frequently used languages in the CLEF campaign. For these languages, the monolingual retrieval tasks were hotly contested, with many systems showing similar performance.
3. Combination systems, i.e. systems that combine multiple approaches to translation and/or retrieval, continue to be popular for the multilingual track. Consequently, participants study the question of how to merge results from multiple translation or retrieval sources very actively.
4. The investigation into the quality of the relevance assessment pools shows that they are very stable, and that the test collection should therefore be well suited for later post-hoc experiments. This also ensures that results as published by CLEF should be valid within the inherent limitations of the testing methodology.

From the standpoint of the increased number of experiments, and considering the many different offerings in the core and additional tracks, which were taken up eagerly, and the quality of the resulting test collection, we judge the 2003 campaign a big success.

Version	Date	Status	Notes
1.1	15 December 2003	Draft	Distributed to partners
2.1	17 December 2003	Final	Public

Table of Contents

Executive Summary 2
Table of Contents 3
1 Introduction 4
2 Overview of Results 4
2.1 Participants 4
2.2 Collection, Tracks and Tasks 5
2.3 Experiments and their Characteristics 6
2.4 Main Trends 9
2.5 The Results 9
3 Statistical Testing 13
4 Pool Quality and Result Validity 17
5 Conclusions 20
References 21

1 Introduction

The campaigns organized as centerpiece of the CLEF project build on work carried out in earlier years inside the TREC campaigns [9], organized in the United States by the National Institute of Standards and Technology (1997-1999), and on work carried out as part of the DELOS Network of Excellence under the banner "CLEF" (2000-2001) [2], [3]. In this sense, much of this work is a continuation of earlier efforts, and it is possible to draw comparisons to the campaigns organized as part of these earlier activities. The deliverable does this freely, where appropriate. A similar report in the form of deliverable D4.2.1 was published following the CLEF 2002 campaign. The present deliverable borrows freely from this earlier report where appropriate, to avoid tedious cross-referencing between the two.

This deliverable reports on general characteristics of the experiments submitted for CLEF, such as the participants and the languages/topic fields used, as well as on trends observed in the work by the participating groups. It therefore provides in a sense a summary of the campaign from a technical standpoint, but it also gives a snapshot of what participants are working on, and therefore an overview of the state of research as far as the participants in CLEF are concerned. For reasons of practicality, only a brief summary of the hundreds of different results obtained by the participants can be given, since an exhaustive listing is well beyond the scope of this report and would duplicate efforts by the participants themselves. Interested readers can refer to the complete working notes of the CLEF workshop, which contains nearly 300 pages of individual result listings [11], plus the descriptions by the participants of their own systems and experiments.

CLEF follows a well-defined “laboratory setting” methodology that uses a limited set of "constructed" information needs as a representation of queries from real users of CLIR system. These representations are known as topics, based on which the queries for the individual systems are formulated. Consequently, we must investigate how the results generalize beyond this laboratory setting. Statistical analysis provides us with tools for this task. We describe how to carry out statistical analysis on CLEF results and present results for the main, multilingual task.

CLEF relies heavily on relevance assessments to compute the published performance measures. We investigate the quality of the relevance assessments by carrying out an analysis of their coverage (completeness).

The deliverable is structured in three main sections, Sections 2-4, giving details on the experiments (Section 2), on statistical analysis (Section 3), and on the quality of the relevance assessments (Section 4). Conclusions are given in Section 5.

2 Overview of Results

2.1 Participants

In all, 42 participants from 14 different countries participated in one or more activities offered under the CLEF umbrella. This represents an increase from the total number of 37 participants in the 2002 campaign, and a substantial growth compared to the first CLEF campaign, which attracted 20 participants. However, even more pronounced growth occurred in the amount of data that was submitted by the participants and processed by the CLEF consortium: a total of 415 experiments were submitted for the main tracks[1]. Many of the participants had already

[1] Additionally, a substantial number of experiments for the additional tracks were submitted, which are not included in this total, because, while carried out under the CLEF umbrella, they are not part of the official project work as defined by the technical annex and not funded as part of the CLEF project.

taken part in earlier CLEF campaigns or in related activities, such as TREC (North America) and/or NTCIR (East Asia). However, there were also a healthy number of newcomers. The number of European groups is substantial, and they are now clearly in the majority (28 out of 42, showing the importance of CLEF in fostering interest in CLIR research in Europe – European participation in the TREC7 (1998) CLIR track was a meager 3 groups!) (see Table 1).

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

Table 1. Participants in CLEF 2003. One star (*) denotes a participant that has taken part in any one previous campaign (2000 to 2002), two stars (**) denote participants that have taken part in two previous campaigns, while participants marked with three stars (***) have submitted work to all three previous campaigns.

2.2 Collection, Tracks and Tasks

For 2003, the CLEF consortium again expanded the test collection used for the experiments in every respect: more documents (+40%), more languages covered (9, with Russian being new) and most importantly, more relevance assessments (+35% more). Aspects of the additional documents and languages are covered in deliverable 2.3.2 "Multilingual Collection for Campaign 2", consigned month 18 [5]. Relevance assessment procedures are detailed in deliverable 3.2.2 "Test Collection Report for Campaign 2" [7].

For the 2003 campaign, CLEF tracks and tasks were structured as follows:

1. Multilingual Track: In a change from 2002, two different multilingual tasks were offered. It was strongly felt that the multilingual track should include languages that were introduced in 2002 and earlier, such as Dutch, Finnish and Swedish, which so far had been left outside this track. However, a total of eight languages was felt to be too demanding for some participants, especially those that recently joined the CLEF

campaigns, and also potentially detrimental for exploration of some research questions that are not centered around the handling of a maximum number of languages. Consequently, the multilingual track was split into two tasks: the Multilingual-8 tasks, which consisted of searching a document collection containing documents each written in one of eight languages, and the Multilingual-4 task, which restricted the document collection to four core languages. A grand total of nearly 1.6 million documents in the languages Dutch, English, Finnish, French, German, Italian, Spanish and Swedish made up the multilingual collection. The multilingual track was the "main" activity of the campaign. Participants had a free choice of 9 topic languages.

2. Bilingual Track: A few hand-picked pairs of languages were offered as bilingual tasks. The pairs were selected to represent different (research) challenges:
    a. Finnish to German, as a pair covering the Uralic and Germanic languages, with both languages rich in compound words,
    b. Italian to Spanish, as a pair of closely related Romance languages, potentially opening the possibility for language-independent approaches,
    c. German to Italian as a pair of widely used languages covering both the Germanic and Romance groups, and
    d. French to Dutch, to cater for a traditionally strong community of Dutch groups participating in the CLEF campaign.
    e. In addition, bilingual retrieval from any topic language to English was offered specifically for newcomers to allow them participation without the need to immediately adapt their systems to new languages, plus bilingual retrieval to Russian from any language, since Russian was newly introduced for the 2003 campaign.
3. Monolingual Track. Choice of 8 topic languages (DE, ES, FI, FR, IT, NL, RU, SV). Documents in same language as topic language.

CLEF de-emphasizes retrieval on English language documents (only included in the multilingual track and for newcomers in the bilingual track), as it is already covered in the TREC evaluation campaigns.

CLEF 2003 also offered domain-specific retrieval in the form of the GIRT track, with a choice of three topic languages (DE, EN, RU). Retrieval takes place on German and English abstracts and documents from the domain of social sciences.

2.3 Experiments and their Characteristics

A total of 415 experiments were officially submitted for the core tracks. This is an increase of more than 45% compared to CLEF 2002, making the 2003 campaign by far the largest undertaking by the CLEF project so far. Submissions were divided among the tracks as follows (Table 2, Figure 1):

Track	#Participants	#Runs/Experiments
Multilingual-8	7	33
Multilingual-4	14	53
Bilingual FI->DE	2	3
Bilingual X->EN	3	15
Bilingual IT->ES	9	25
Bilingual DE->IT	8	21
Bilingual FR->NL	3	6
Bilingual 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	

Table 2. Different tracks/tasks, and the respective number of particpants/experiments.

This is a fairly even distribution, both in terms of the tasks and the languages covered (see also Table 3). The number of participants working on the two multilingual tasks is a big success. We did not expect to have 7 groups working on as many as eight languages simultaneously, and on 14 groups tackling the smaller multilingual track. Obviously, it is very difficult to "steer" the distribution of participants with regard to the tasks, since this distribution reflects the participants' interest. It is therefore no surprise that some of the bilingual pairs were somewhat less popular. It must also be considered that there was a limit on the number of experiments that could be submitted by any one group [2], in order to avoid an overload of the campaign. Forcing groups thus to set priorities and potentially drop some experiments may have hurt some tasks more than others. Therefore, the fact that nearly all tasks/combinations are well represented is encouraging. The participation in the domain-specific tasks was somewhat below what we had hoped, but this is in line with earlier experiences which showed that while a lot of interest is initially expressed by many participants, groups tend to drop these tasks when they run out of resources for their experiments.

CLEF has offered a number of additional tracks in 2003, namely the Interactive track, the Question Answering track, the Image Retrieval track and the Spoken Document Retrieval track. While under the CLEF umbrella, these were not funded by the EC project and are not analyzed further here.

[2] Which was a very liberal maximum of 45 experiments by any one group. However, there were also limits to the number of experiments for any one specific task, meaning that to reach the overall maximum, groups had to work on different tracks and tasks.

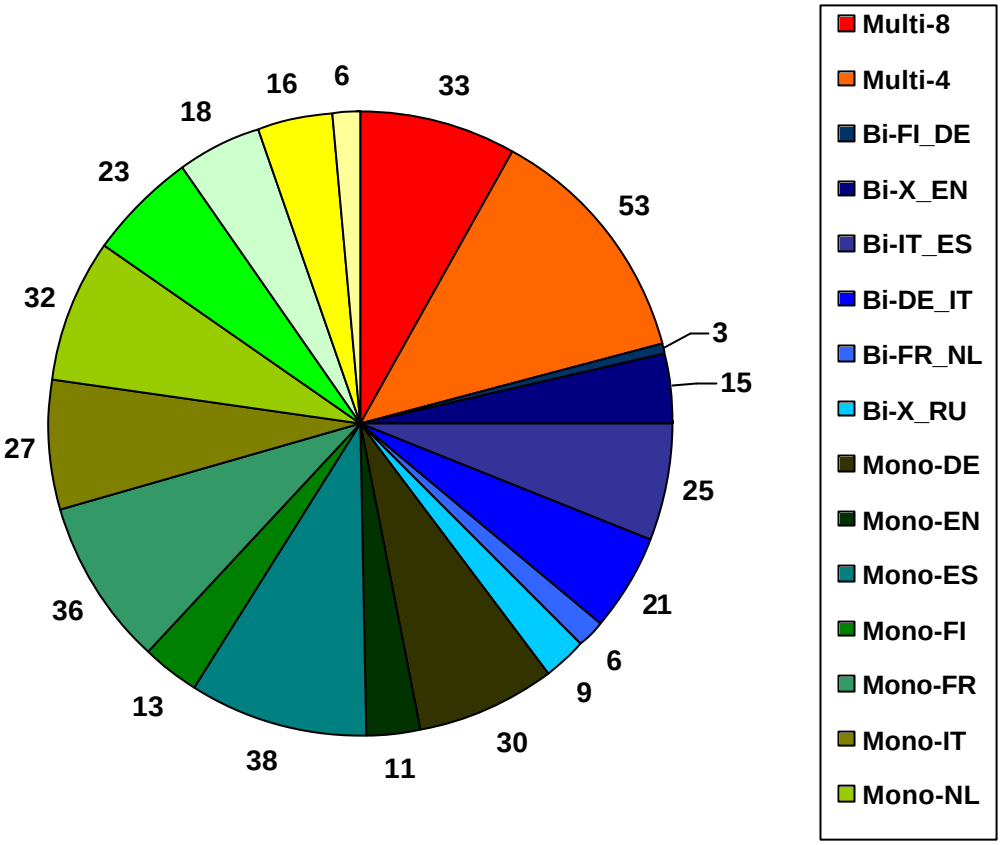

Figure 1. Distribution of the experiments across different tracks/tasks.

CLEF offers the choice of using short, medium-length and long queries for all experiments. All three choices were used by participants, with the medium-length queries dominating (participants were required to submit at least one experiment per task using medium length in order to boost comparability across sites) (Table 4). Queries could be constructed either manually or automatically out of the statements of information need (topics structured in title, description and narrative fields) distributed by the CLEF organization. The overwhelming majority of participants used automatic query construction [6,7].

Topic Language	# Experiments
DE German	69
EN English	97
ES Spanish	54
FI Finnish	16
FR French	49
IT Italian	54
NL Dutch	32
RU Russian	26
SV Swedish	18

Table 3. Distribution of tasks across topic (query) languages.

Topic fields	# Experiments
TDN – long queries	21
TD – medium-length queries	374
T – short queries	12
Other	8

Table 4. Different topic fields used for query construction. T=title, D=description, N=narrative.

2.4 Main Trends

With CLEF building on earlier campaigns organized both under the same name and under other umbrellas (TREC in North America, NTCIR in East Asia), there are participants that have worked on this type of evaluation for several years. Therefore, CLEF acts as a "trendsetter", and methods that work well one year are adopted eagerly by other participants in following campaigns. This is clearly a valuable contribution that CLEF plays in distributing successful ideas.

For the 2003 campaign, we discern the following main trends:

- Participants spent a lot of effort on detailed fine-tuning per language, per weighting scheme, and per translation resource type
- Groups thought about (generic) ways to "scale" up to new languages
- Merging of different (intermediate) retrieval results continued to be a hot issue; however, no merging approach besides the simple ones has been widely adopted yet. Methods that have been adopted by groups include collection size-based merging and 2-step merging.
- A few resources were very popular, among them the "Snowball" stemmers, stopword lists by Université de Neuchâtel, some machine translation systems, dictionaries by "Freelang" and others.
- Query translation is still the favorite choice to cross the language barrier.
- Stemming and decompounding are still actively debated; a slightly increased use of linguistics can be discerned.
- Monolingual tracks were "hotly contested", for some (especially the most frequently used) languages, very similar performance was obtained among the top groups
- The new definition of the bilingual tasks forced people to think about "inconvenient" language pairs, stimulating some of the most original work.
- Returning participants usually improve performance. ("Advantage for veteran groups"). This is especially true for the large "multilingual-8" task, where veteran groups dominated. It seems that scaling up to this many languages takes its time. The "multilingual-4" task was very competitive.
- Some blueprints to "successful CLIR" seem now to be in place, and some of the "older" systems resemble each other. There is a trend towards systems combining different types of translation resources. The question arises if we are headed towards a monoculture of CLIR systems.

2.5 The Results

The individual results of the participants are reported in detail in the CLEF 2003 Working Notes [11] distributed to the participants at the CLEF workshop in Trondheim, Norway and are also available on the CLEF website. The focus of this report and the number of

experiments submitted make it impossible to provide exhaustive lists of all individual results in this deliverable. In the following, we summarize the results for the multilingual, bilingual and monolingual track briefly.

Multilingual Track

The multilingual track is the hardest task to complete in CLEF and is therefore the main focus of the activities. This year, the track has been divided into two tasks, the Multilingual-8 and Multilingual-4 task. Seven groups submitted 33 runs to the Multilingual-8 task, a very encouraging number considering the difficulties in handling so many languages simultaneously. Figure 2 shows the best entries of the five top performing groups in terms of average precision figures. Only entries using the title+description topic field combination were used for this comparison. Multilingual-4, the smaller task, had double the number of participants, namely fourteen. These groups submitted a grand total of 53 runs for the task (Figure 3).

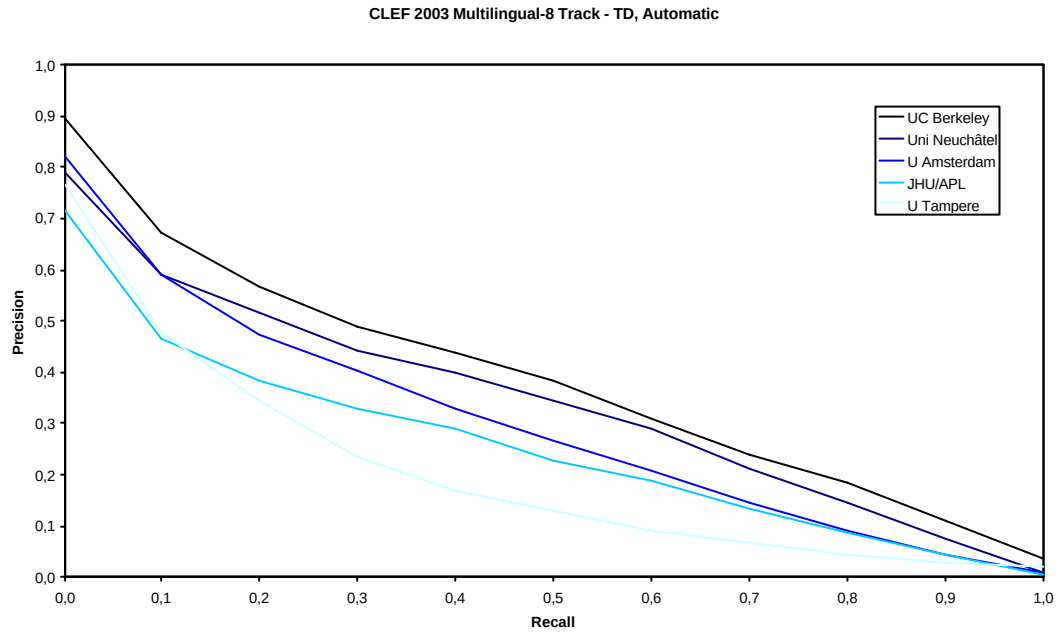

Figure 2. Best performing entries of the top five participants for the large Multilingual-8 task. Shown is the precision/recall curve, giving precision values at varying levels of recall. Only experiments using the title+description topic fields are included.

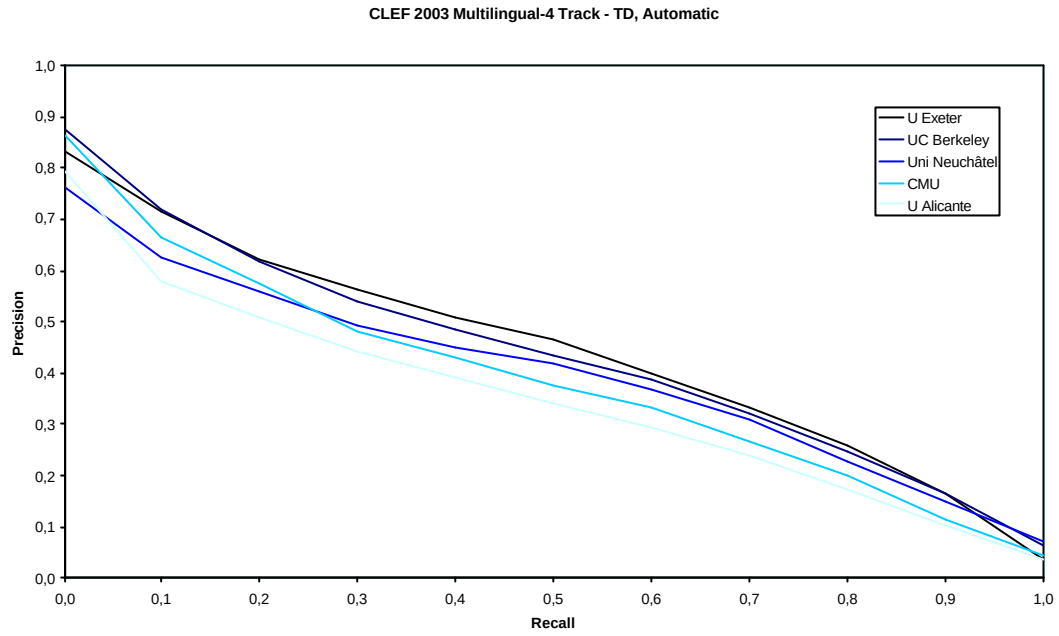

Figure 3. Best performing entries of the top five participants for the Multilingual-4 task. Shown is the precision/recall curve, giving precision values at varying levels of recall. Only experiments using the title+description topic fields are included.

As can be seen, there is more difference in the top performances for the Multilingual-8 track than for Multilingual-4. Clearly, long-time participants had an advantage for the larger task. The results for multilingual-4 are very close, showing that groups have a good understanding of how to tune their systems well to the most popular languages. The top entries for the large multilingual task used elaborate combination approaches that helps them handle the difficulties of the languages.

Bilingual Track

The 2003 campaign offered a newly defined bilingual track that was structured in four subtasks related to specific language pairs, one subtask for newcomers only (bilingual retrieval to English) and one subtask for bilingual retrieval to Russian. This was a departure from 2002, where the CLEF consortium responded to numerous requests from participants and opened the bilingual track to all eight target languages (DE, EN; ES, FI, FR, IT, NL, SV; EN for newcomers or under special conditions only). While allowing for added flexibility in testing the systems on the participant's part, this decision made comparing different bilingual experiments somewhat harder, since experiments on different target languages use different document sets. It was therefore necessary to investigate eight different result sets, one for each target language.

The introduction of specific language pairs led to a larger number of participants per pair. Table 5 shows the best entries by the top five performing participants for each target language, including only runs using the mandatory title+description topic field combination.

Target Language	1st	2nd	3rd	4th	5th
Biling FI->DE	UC Berkeley	JHU/APL			
Biling X->EN	Daedalus	IMS/U Padua			
Biling IT->ES	U Alicante	UC Berkeley	CMU	IRST	JHU/APL
Biling DE->IT	JHU/APL	U Exeter	CMU	UC Berkeley	U Amsterdam
Biling FR->NL	JHU/APL	U Amsterdam	UC Berkeley		
Biling X->RU	UC Berkeley	U Amsterdam			

Table 5. Best entries for the bilingual track. Shown are at most the top five participants for each target language (title+description topic fields only).

Monolingual Track

The CLEF 2003 campaign offered monolingual retrieval for all target languages besides English. Again, Table 6 summarizes the best entries of the top five performing groups for the title+description topic field combination.

Target Language	1st	2nd	3rd	4th	5th
DE German	Hummingbird	UC Berkeley	U Exeter	U Neuchâtel	U Amsterdam
ES Spanish	F. U. Bordoni	U Neuchâtel	IRST	Hummingbird	Ricoh/USL
FI Finnish	Hummingbird	UC Berkeley	JHU/APL	U Neuchâtel	U Amsterdam
FR French	U Neuchâtel	Hummingbird	F. U. Bordoni	IRST	UC Berkeley
IT Italian	F. U. Bordoni	UC Berkeley	IRST	Ricoh/USL	U Neuchâtel
NL Dutch	Hummingbird	UC Berkeley	U Amsterdam	U Neuchâtel	JHU/APL
RU Russian	UC Berkeley	JHU/APL	U Neuchâtel	Hummingbird	U Amsterdam
SV Swedish	UC Berkeley	U Neuchâtel	JHU/APL	U Amsterdam	Hummingbird

Table 6. Best entries for the monolingual track. Shown are the top five participants for each target language (title+description topic fields only).

As clearly seen in Table 7 the differences between the top performers for some of the most popular languages, which have also been introduced early in the campaigns, are quite small. This phenomenon is most pronounced for French, where the difference between the top performing group and the 5th placed group is only 2.4%! For the more recently added languages, differences are still larger, and thus the potential for substantial improvements in the next campaigns may be larger. An exception of some sort to these conclusions is German, which has been adopted by the campaign as one of the target languages from the beginning, but where the difference is slightly larger than for French, Italian and Spanish. We attribute this to the decompounding problem, which typically is more resource intensive than stemming and which seems to pose some challenges to which groups must adapt.

Task	Diff. To 5th Place
Monolingual DE	+12.3%
Monolingual ES	+7.3%
Monolingual FI	+17.2%
Monolingual FR	+2.4%
Monolingual IT	+9.1%
Monolingual NL	+10.4%
Monolingual RU	+28.0%
Monolingual SV	+25.3%

Table 7. Percentual difference between the best performing experiment of the top placed group and the best performing experiment of the fifth placed group.

Domain-specific

As already stated, we had less entries for the domain-specific tracks than for the other tracks. We again give a summary of the best entries of the top performing groups for the title+description topic field combination (Table 8).

Track	1st	2nd	3rd	4th	5th
GIRT->DE	UC Berkeley	U Amsterdam	FU Hagen	ENEA	
GIRT->EN	UC Berkeley	ENEA			

Table 8. Best entries for the domain-specific tracks. Shown are the top five participants for each target language (title+description topic fields only).

The smaller number of participants makes it difficult to draw overall conclusions. Indeed, the performance obtained by the groups was very dissimilar, probably due to a mixture of monolingual and bilingual experiments and due to the different degree of tuning for the characteristic of the domain-specific data. A detailed description of the different experiments for the domain-specific tracks can be found in the CLEF 2003 working notes [11].

3 Statistical Testing

CLEF uses, for reasons of practicality, a limited number of queries (60 in 2003; up from an initial 40 in 2000 and 50 in 2001-2002), which are intended to represent a more or less appropriate sample of the population of all possible queries that users would want to ask from the collection. When the goal is to validate how well results can be expected to hold beyond this particular set of queries, statistical testing can help determine what differences between runs appear to be real as opposed to differences that are due to sampling variation. As with all statistical testing, conclusions will be qualified by an error probability, which was chosen to be 0.05 in the following.

Using the IR-STAT-PAK tool [1], a statistical analysis of the results for the multilingual track was carried out for the first time after the 2001 campaign. We have repeated this analysis in 2002, and expanded it for 2003. The tool provides an Analysis of Variance (ANOVA), which is the parametric test of choice in such situations but requires that some assumptions concerning the data be checked. Hull [10] provides details of these; in particular, the scores in question should be approximately normally distributed and their variance has to be approximately the same for all runs. IR-STAT-PAK uses the Hartley test to verify the equality of variances. In the case of the CLEF multilingual collection, it indicates that the assumption is violated. For such cases, the program offers an arcsine transformation,

$$f(x) = \arcsin\left(\sqrt{x}\right)$$

which Tague-Sutcliffe [13] recommends for use with Precision/Recall measures, and which we have therefore applied.

The ANOVA test proper only determines if there is at least one pair of runs that exhibit a statistical difference. Following a significant ANOVA, various comparison procedures can be employed to investigate significant differences. IR-STAT-PAK uses the Tukey T test for grouping the runs.

One way to present the overall results is in tabular form, which we chose for the following presentation of the analysis of the Multilingual-8 and Multilingua-4 tasks.
Looking at the result (Table 8), all runs that are included in the same group (denoted by "X") do not have significantly different performance. All runs scoring below a certain group perform significantly worse than at least the top entry of that group. Likewise, all runs scoring above a certain group perform significantly better than at least the bottom entry in that group. To determine all runs that perform significantly worse than a certain run, determine the rightmost group that includes the run. All runs scoring below the bottom entry of that group are significantly worse. Conversely, to determine all runs that perform significantly better than a given run, determine the leftmost group that includes the run. All those runs that score better than the top entry of that group perform significantly better.

As mentioned, it is well-known that it is fairly difficult to detect statistically significant differences between retrieval runs based on 60 queries. While 60 queries remains a good choice based on practicality for doing relevance assessments, statistical testing would be one of the areas to benefit most from having additional topics. This fact is addressed by the measures taken to ensure stability of at least part of the document collection across different campaigns, which allows participants to run their system on aggregate sets of queries for post-hoc experiments.

For the 2003 campaign, we conducted statistical analysis of the “pools of experiments” for all target collections of the multilingual, bilingual and monolingual tasks: the eight language multilingual collection (Table 9), the four language multilingual collection (Table 10), and the eight non-English monolingual collections (Table 11). We do not report numbers for the English monolingual and domain-specific collections, as there were too few experiments to report a consistent picture.

For the 2003 campaign, the picture is somewhat less clear than in 2002, where we observed a fairly clear division of runs into performance groups for the multilingual track.
For the multilingual-8 task, there is some division between the entries by the top performing groups UC Berkeley, Université de Neuchâtel and University of Amsterdam compared to the rest of the group. This division is a bit less clear between U Amsterdam and JHU/APL as the third and fourth group, respectively, but fairly pronounced between UC Berkeley, U Neuchâtel and the other groups from U Tampere downwards. As mentioned before, veteran groups have submitted the best performing experiments.

Multilingual-4 was very competitive, with many groups obviously already having a good understanding of the languages involved. This leads to a fairly continuous field of performances, with no clear drop-offs between groups. The top six groups submitted at least one experiment whose performance difference is not statistically significant with regard to the top performing entry (by University of Exeter).

Arcsine-transformed average precision values	Run ID									
0.63941	UniNEml	X								
0.63699	bkmul8en3	X								
0.59807	bkmul8en2	X	X							
0.58374	bkmul8en1	X	X	X						
0.57940	UniNEml4	X	X	X						
0.57933	UniNEml1	X	X	X						
0.54374	UniNEml2	X	X	X	X					
0.54219	UniNEml3	X	X	X	X					
0.53290	UAmsC03EnM8SS4G	X	X	X	X					
0.53080	UAmsC03EnM84GiSb	X	X	X	X					
0.51692	UAmsC03EnM8SS4G6		X	X	X	X				
0.47626	UAmsC03EnM84Gr			X	X	X	X			
0.47082	UAmsC03EnM84Gr6			X	X	X	X			
0.46960	aplmuen8b			X	X	X	X			
0.46030	aplmuen8a				X	X	X			
0.41480	UTAmul1					X	X	X		
0.41167	UTAmul4					X	X	X		
0.41052	UTAmul5					X	X	X		
0.41036	UTAmul2					X	X	X		
0.40982	UTAmul3					X	X	X		
0.39399	uja03LargeRRPrf						X	X		
0.37298	uja03LargeRSV2m						X	X	X	
0.37126	uja03LargeRR						X	X	X	
0.33118	UBENmultirf3							X	X	
0.33014	uja03LargeRSV2							X	X	
0.32183	UBENmultirf1							X	X	
0.30814	UBENmultirf2							X	X	
0.26865	UBESmultirf3								X	
0.26865	UBESmultishort2								X	
0.26667	UBESmultirf1								X	
0.26417	UBESmultirf2								X	
0.14714	UBENmultishort3									X
0.14631	UBENmultishort2									X

Table 9. Results of statistical analysis (ANOVA) on the experiments submitted for the large Multilingual-8 track. All experiments, regardless of topic language or topic fields, are included. Results are therefore only valid for comparison of individual pairs of runs, and not in terms of absolute performance.

Arcsine-transformed avg. precision values	Run ID															
0.70441	exemulttc	X														
0.69861	bkmul4en3	X														
0.69533	UniNEms	X														
0.67936	bkmul4en2	X	X													
0.65612	UniNEms4	X	X	X												
0.64085	cmuM4lowfbre	X	X	X	X											
0.63730	bkmul4en1	X	X	X	X											
0.63625	UniNEms1	X	X	X	X											
0.63352	cmuM4fbre	X	X	X	X											
0.63250	UniNEms2	X	X	X	X											
0.63069	UniNEms3	X	X	X	X											
0.60997	exemult4s	X	X	X	X	X										
0.60953	exemult4p	X	X	X	X	X										
0.60488	exemult4u	X	X	X	X	X										
0.59983	IRn-m-exp-sp	X	X	X	X	X										
0.59953	IRn-mi-exp-sp	X	X	X	X	X	X									
0.59253	cmuM4lowfb	X	X	X	X	X	X	X								
0.58839	IRn-m-exp-nsp	X	X	X	X	X	X	X								
0.58441	UAmsC03EnM44GiSb	X	X	X	X	X	X	X								
0.58138	exemult4d	X	X	X	X	X	X	X								
0.57950	UAmsC03EnM4SS4G	X	X	X	X	X	X	X								
0.55475	IRSTen2xx_1		X	X	X	X	X	X	X							
0.55253	IRn-m-nexp-sp		X	X	X	X	X	X	X							
0.54866	IRSTen2xx_2		X	X	X	X	X	X	X							
0.54757	IRSTen2xx_3			X	X	X	X	X	X							
0.54495	IRn-m-nexp-nsp			X	X	X	X	X	X							
0.54285	IRSTen2xx_4			X	X	X	X	X	X							
0.53832	UHImlt4R2			X	X	X	X	X	X							
0.53664	UAmsC03EnM44Gr			X	X	X	X	X	X							
0.53051	cmuM4fb			X	X	X	X	X	X							
0.52895	spxxQTordirect			X	X	X	X	X	X							
0.52671	aplmuen4a			X	X	X	X	X	X	X						
0.51804	UHImlt4R1				X	X	X	X	X	X						
0.51127	aplmuen4b				X	X	X	X	X	X						
0.48759	spxxQTor3					X	X	X	X	X	X					
0.46800	uja03ShortRRPrf						X	X	X	X	X	X				
0.46458	spxxQTdoc							X	X	X	X	X	X			
0.44014	uja03ShortRR								X	X	X	X	X	X		
0.42893	uja03ShortRSV2m								X	X	X	X	X	X		
0.42679	spxxQTorall								X	X	X	X	X	X		
0.39626	xrce_ML4_run2									X	X	X	X	X		
0.36998	uja03ShortRSV2										X	X	X	X	X	
0.34886	NTUm4TopnTpCw											X	X	X	X	
0.33893	NTUm4TopnLinear											X	X	X	X	X
0.33333	NTUm4TopnTp												X	X	X	X
0.33148	NTUm4Topn													X	X	X
0.31548	lic2mes1													X	X	X
0.26438	lic2mde1														X	X
0.25501	spxxQTorallrr														X	X
0.24534	lic2mfr2														X	X
0.24140	lic2mfr1														X	X
0.21585	lic2men1															X
0.00017	kcca300															X

Table 10. Results of statistical analysis (ANOVA) on the experiments submitted for the Multilingual-4 track. All experiments, regardless of topic language or topic fields, are included. Results are therefore only valid for comparison of individual pairs of runs, and not in terms of absolute performance.

In addition to the two multilingual tasks, we have also examined non-English monolingual target collections. These analyses include both the respective monolingual runs, but also the bilingual runs to that target language, i.e. the German analysis contains both German monolingual and Finnish->German bilingual experiments. The fact that the monolingual tasks were so competitive this year, and that many groups submitted experiments with very similar performance, also reflects in this analysis, with practically all groups submitting at least one experiment with a performance difference that is not statistically significant from the top performing experiment. Note, however, that experiments of very different character are mixed in this analysis.

Target collection	Number of groups in the "top group" of the statistical analysis/total number of groups
"DE" German	10/13
"ES" Spanish	15/18
"FI" Finnish	6/7
"FR" French	15/16
"IT" Italian	13/16
"NL" Dutch	9/11
"RU" Russian	5/5
"SV" Swedish	4/8

Table 11. Results of statistical analysis (ANOVA) on the experiments submitted for the individual monolingual subcollections. Shown is the ratio of groups that submitted at least one experiment with a performance difference that is not statistically significant compared to the top performance compared to the total number of groups submitting experiments for that target collection.

4 Pool Quality and Result Validity

The results reported in the CLEF campaigns rely heavily on the concept of judging the relevance of documents with respect to given topics. The relevance of a document is judged by human assessors, making this a costly undertaking. These relevance assessments are then used for the calculation of the recall/precision figures that underlie the graphs and figures presented to the participants.

Their central importance for the calculation of many popular evaluation measures means that relevance assessments are not without critics. Generally, concerns mentioned focus mostly on two aspects: the "quality" and the "coverage" ("completeness") of the assessments. The first concern stems from the subjective nature of relevance, which can lead to disagreements between different assessors or even when the same assessor judges a document twice. Such disagreements can emerge from, among other things, personal bias of the judge, or a lack of understanding of the topics and documents. There is no "solution" for obtaining universal relevance judgments. Rather, researchers that rely upon the results from an evaluation campaign such as CLEF have to be aware of this issue and its implications. Numerous studies have analyzed the impact of disagreement in judging on the validity of evaluation results. These studies generally conclude that as long as sufficient consistency is maintained during judging, the ranking and comparison of systems is stable even if the absolute performance values calculated on the basis of the assessments change. The quality and consistency of the assessments in CLEF is ensured by following a well-proven

methodology based on TREC experience. More details of relevance assessment processes can be found in [12] and in deliverables 3.2.1 [6] and 3.2.2 [7].

The problem of coverage arises from practical considerations in the production of the relevance assessments. While it is comparatively easy to judge a substantial part of the top-ranked results submitted by participants, it is much harder to judge the documents that were not part of any of the submitted result sets, since the number of such documents is usually far greater than that of the documents retrieved in result sets. This is especially the case with today's large test collections. Judging the non-retrieved documents is necessary to calculate some evaluation measures such as recall.

In order to keep costs manageable, only documents included and highly ranked in at least one result set are judged for relevance (with the union of all judged result sets forming a "document pool"). This implies that some relevant documents potentially go undetected if they are not retrieved by any of the participating systems. The assertion is that a sufficient number of diverse systems will turn up most relevant documents this way. Figures calculated based on these "limited" assessments are then a good approximation of theoretical figures based on complete assessments. A potential problem is the usability of the resulting test collection for the evaluation of a system that did not contribute to this "pool of judged documents". If such a system retrieves a substantial number of unjudged documents that are relevant, but went undetected, it is unfairly penalized when calculating the evaluation measures. An investigation into whether the assessments for the CLEF multilingual collection provide sufficient coverage follows below.

One way to analyze the coverage of the relevance judgments is by focusing on the "unique relevant documents" [14]. For this purpose, a unique relevant document is defined as a document that was judged relevant with respect to a specific topic, but that would not have been part of the pool of judged documents had a certain group not participated in the evaluation, i.e., only one group retrieved the document with a score high enough to have it included in the judgment pool. This addresses the concern that systems not directly participating in the evaluation are unfairly penalized. Subtracting relevant documents only found by a certain group, and then reevaluating the results for this group, simulates the scenario that this group was a non-participant. The smaller the change in performance that is observed, the higher is the probability that the relevance assessments are sufficiently complete.

This kind of analysis has been run by the CLEF consortium since the 2000 campaign for the multilingual track. In 2002, we have expanded the analysis to include an investigation of the subcollections formed by the individual target languages. A total of $n+1$ sets of relevance assessments are used: the original set, and n sets that are built by taking away the relevant documents uniquely found by one specific participant. The results for every experiment are then recomputed using the set without the group-specific relevant documents. We chose the same analysis for 2003. The key figures obtained after rerunning the evaluations can be found in Table 12.

Multilingual-8	Mean difference	Max difference	StdDev difference
Absolute	0.0005	0.0014	0.0009
Percentage	0.24	0.77	0.5128

Multilingual-4[3]	Mean difference	Max difference	StdDev difference
Absolute	0.0007	0.0025	0.0016
Percentage	0.33	2.06	0.7677

DE German	Mean difference	Max difference	StdDev difference
Absolute	0.0013	0.0038	0.0026
Percentage	0.29	1.04	0.6439

EN English	Mean difference	Max difference	StdDev difference
Absolute	0.0021	0.0052	0.0033
Percentage	0.73	2.09	1.3477

ES Spanish[4]	Mean difference	Max difference	StdDev difference
Absolute	0.0022	0.0109	0.0050
Percentage	0.52	3.11	1.2057

FI Finnish	Mean difference	Max difference	StdDev difference
Absolute	0.0004	0.0011	0.0008
Percentage	0.09	0.31	0.1897

FR French	Mean difference	Max difference	StdDev difference
Absolute	0.0009	0.0070	0.0020
Percentage	0.19	1.58	0.4345

IT Italian	Mean difference	Max difference	StdDev difference
Absolute	0.0010	0.0096	0.0025
Percentage	0.34	3.03	0.7907

NL Dutch	Mean difference	Max difference	StdDev difference
Absolute	0.0016	0.0082	0.0033
Percentage	0.38	2.03	0.8024

RU Russian	Mean difference	Max difference	StdDev difference
Absolute	0.0040	0.0139	0.0053
Percentage	1.36	5.92	1.9356

[3] One experiment that was an extreme outlier in terms of performance was removed before calculation of the Multilingual-4 figures to avoid a non-representative skew in the numbers.

[4] Two experiments that were extreme outliers in terms of performance were removed before calculation of the Spanish figures to avoid a non-representative skew in the numbers.

SV Swedish	Mean difference	Max difference	StdDev difference
Absolute	0.0023	0.0073	0.0053
Percentage	0.59	2.02	1.3455

Table 12. Key values of the pool quality analysis: mean and maximum change in average precision when removing the pool contribution of one participant, and associated standard deviation.

The quality of a document pool can therefore be judged by the mean performance difference in terms of average precision that is obtained if the pool had been missing the contribution of a specific group. This difference should be as small as possible, indicating that the pool is "sufficiently exhaustive" and that adding more documents to the pool, such as documents found by an additional participant, does not substantially influence results and/or rankings. As we also found for all previous campaigns, the pool used for the multilingual tasks is very stable. The maximum change in performance scores is 0.77% for the Multilingual-4, and 2.06% for the Multilingual-4 task. These small differences influence only direct comparisons between systems that have practically identical performance, and where the original performance differences are not considered significant in any case. The value of the multilingual pool for reuse in post-hoc experiments should thus be assured, and the validity of the results reported by CLEF should be given within the inherent limits of interpretation (restricted set of queries, characteristics of evaluation measure and others).

The pools for individual target languages are smaller, since they are restricted to the document set of that language. Only runs for that language, and therefore a smaller number than for the multilingual pool, contributed. It is therefore not surprising that differences found for the individual languages tend to be somewhat higher than for the multilingual pool. We feel, however, that they are still comfortably within acceptable limits, and they do indeed compare favorably with numbers reported for comparable collections in the past [2], [3]. Not surprisingly, the pool for Russian is the least stable, owing to being introduced newly and late in the campaign, plus having fewer contributions than other languages. We had issues with a few outliers that obfuscate the measures somewhat, but we believe that all pools should be of comparable quality across the other languages.

5 Conclusions

We have reported on the results obtained for the 2003 campaign and their interpretation. CLEF 2003 experienced substantial growth in the number of experiments submitted. This deliverable summarizes the main characteristics of the 415 experiments submitted for the campaign, and discusses trends observed and the main results. Statistical significance analysis has been conducted for all subcollections formed by the individual languages, as well as for the multilingual tasks, where we provide detailed tables. Lastly, we investigate the validity of the results by analyzing the completeness of the relevance assessment pools, which is critical for calculating the performance measures used by CLEF.

In summary, we can conclude that, much as for 2002, people adopt each other's ideas and methods across campaigns, and that those returning groups that have the experience to build complex combination systems have performed well in the main, multilingual tasks. More than ever, for the monolingual track we observe that good performance in a wide variety of target languages requires careful fine tuning for all these languages. The monolingual tasks were extremely competitive this year, with many groups obtaining good performance results.

The core tracks in CLEF seem to have “matured” considerably. A challenge will be to determine how to adapt them in the future to continue stimulating new research challenges for the CLIR field.

Statistical analysis allows to qualify and better interpret the results as published by CLEF. As evidenced by an analysis of the experiments that we present, fairly large performance differences are needed to reach a level of statistical significance. This is especially true for the monolingual tasks. For this kind of testing, having a maximum number of queries available is of great benefit. The CLEF consortium strives for stability in the test collections to allow post-hoc experiments with combined resources from several campaigns for this reason.
Finally, the results published by CLEF are only as good as the data they build on. We investigate the quality of the relevance assessments by investigating their completeness through pool quality evaluation. We find that the CLEF relevance assessments seem to be very stable, making them suitable for reuse in post-hoc experiments, and further validating the results published during the campaigns.

References

[1] Blustein, J.: IR STAT PAK. URL: http://www.csd.uwo.ca/~jamie/IRSP-overview.html
[2] Braschler, M.: CLEF 2000 – Overview of Results. In Peters, C. (Ed.) Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. Pages 89-101
[3] Braschler, M.: CLEF 2001 – Overview of Results. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds): Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, pages 9-26, 2002.
[4] CLEF Consortium: Deliverable 2.3.1, Multilingual Collection for Campaign 1.
[5] CLEF Consortium: Deliverable 2.3.2, Multilingual Collection for Campaign 2.
[6] CLEF Consortium: Deliverable 3.2.1, Test Collection Report for Campaign 1.
[7] CLEF Consortium: Deliverable 3.2.2, Test Collection Report for Campaign 2.
[8] CLEF Consortium: Deliverable 4.2.1, Result Interpretation and Report (Campaign 2).
[9] Harman, D., Braschler, M., Hess, M., Kluck, M., Peters, C., Schäuble, P., Sheridan P.: CLIR Evaluation at TREC. In Peters, C. (Ed.) Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. Pages 7-23.
[10] Hull, D. A: Using Statistical Testing in the Evaluation of Retrieval Experiments. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburg, USA, 1993.
[11] Peters, C. (Ed).: Results of the CLEF 2003 Cross-Language System Evaluation Campaign. Working Notes for the CLEF 2003 Workshop.
[12] Peters, C., Braschler, M.: European research letter: Cross-language system evaluation: The CLEF campaigns, In Journal of the American Society for Information Science and Technology, Volume 52, Issue 12, pages 1067-1072
[13] Tague-Sutcliffe, J., Blustein, J.: A Statistical Analysis of the TREC-3 Data. In Proceedings of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-226. Page 385ff.
[14] Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998)