CLEF 2005: Ad Hoc Track Overview

Giorgio M. Di Nunzio[1], Nicola Ferro[1], Gareth J.F. Jones[2], and Carol Peters[3]

[1] Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it
[2] School of Computing, Dublin City University, Ireland
gjones@computing.dcu.ie
[3] ISTI-CNR, Area di Ricerca – 56124 Pisa – Italy
carol.peters@isti.cnr.it

Abstract. We describe the objectives and organization of the CLEF 2005 ad hoc track and discuss the main characteristics of the tasks offered to test monolingual, bilingual, and multilingual textual document retrieval. The performance achieved for each task is presented and a statistical analysis of results is given. The mono- and bilingual tasks followed the pattern of previous years but included target collections for two new-to-CLEF languages: Bulgarian and Hungarian. The multilingual tasks concentrated on exploring the reuse of existing test collections from an earlier CLEF campaign. The objectives were to attempt to measure progress in multilingual information retrieval by comparing the results for CLEF 2005 submissions with those of participants in earlier workshops, and also to encourage participants to explore multilingual list merging techniques.

1 Introduction

The ad hoc retrieval track is generally considered to be the core track in the *Cross-Language Evaluation Forum (CLEF)*. The aim of this track is to promote the development of monolingual and cross-language textual document retrieval systems. As in past years, the CLEF 2005 ad hoc track was structured in three tasks, testing systems for monolingual (querying and finding documents in one language), bilingual (querying in one language and finding documents in another language) and multilingual (querying in one language and finding documents in multiple languages) retrieval, thus helping groups to make the progression from simple to more complex tasks. The document collections used were taken from the CLEF multilingual comparable corpus of news documents.

The **Monolingual** and **Bilingual** tasks were principally offered for Bulgarian, French, Hungarian and Portuguese target collections. Additionally, in the bilingual task only, newcomers (i.e. groups that had not previously participated in a CLEF cross-language task) or groups using a "new-to-CLEF" query language could choose to search the English document collection. The aim in all cases was to retrieve relevant documents from the chosen target collection and submit the results in a ranked list.

The **Multilingual** task was based on the CLEF 2003 multilingual-8 test collection which contained news documents in eight languages: Dutch, English, French, German, Italian, Russian, Spanish, and Swedish. There were two subtasks: a traditional multilingual retrieval task (Multi-8 Two-Years-On), and a new task focusing only on the multilingual results merging problem using standard sets of ranked retrieval output (Multi-8 Merging Only). One of the goals for the first task was to see whether it is possible to measure progress over time in multilingual system performance at CLEF by reusing a test collection created in a previous campaign. In running the merging only task our aim was to encourage participation by researchers interested in exploring the multilingual merging problem without the need to build retrieval systems for the document languages.

In this paper we describe the track setup, the evaluation methodology and the participation in the different tasks (Section 2) and present the main characteristics of the experiments and show the results (Sections 3 - 5). The final section provides a brief summing up. For information on the various approaches and resources used by the groups participating in this track and the issues they focused on, we refer the reader to the other papers in this section of the proceedings.

2 Track Setup

The ad hoc track in CLEF adopts a corpus-based, automatic scoring method for the assessment of system performance, based on ideas first introduced in the Cranfield experiments [1] in the late 1960s. The test collection used consists of a set of "topics" describing information needs and a collection of documents to be searched to find those documents that satisfy these information needs. Evaluation of system performance is then done by judging the documents retrieved in response to a topic with respect to their relevance, and computing the recall and precision measures. The distinguishing feature of CLEF is that it applies this evaluation paradigm in a multilingual setting. This means that the criteria normally adopted to create a test collection, consisting of suitable documents, sample queries and relevance assessments, have been adapted to satisfy the particular requirements of the multilingual context. All language dependent tasks such as topic creation and relevance judgment 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.

2.1 Test Collection

This year, for the first time, separate test collections were used in the ad hoc track: the monolingual and bilingual tasks were based on document collections in Bulgarian, English, French, Hungarian and Portuguese with new topics and relevance assessments, whereas the two multilingual tasks reused a test collection - documents, topics and relevance assessments - created in CLEF 2003.

Documents The document collections used for the CLEF 2005 ad hoc tasks are part of the CLEF multilingual corpus of news documents described in the Introduction to these Proceedings.

In the monolingual and bilingual tasks, the English, French and Portuguese collections consisted of national newspapers and news agencies for the period 1994 and 1995. Different variants were used for each language. Thus, for English we had both US and British newspapers, for French we had a national newspaper of France plus Swiss French news agencies, and for Portuguese we had national newspapers from both Portugal and Brazil. This meant that, for each language, there were significant differences in orthography and lexicon over the sub-collections. This is a real world situation and system components, i.e. stemmers, translation resources, etc., should be sufficiently robust to handle such variants. The Bulgarian and Hungarian collections used in these tasks were new in CLEF 2005 and consisted of national newspapers for the year 2002[4]. This meant that the collections we used in the ad hoc mono- and bilingual tasks this year were not all for the same time period. This had important consequences on topic creation. For the multilingual tasks we reused the CLEF 2003 multilingual document collection. This consisted of news documents for 1994-95 in the eight languages listed above.

Topics Topics in CLEF are structured statements representing information needs; the systems use the topics to derive their queries. Each topic consists of three parts: a brief "title" statement; a one-sentence "description"; a more complex "narrative" specifying the relevance assessment criteria.

Sets of 50 topics were created for the CLEF 2005 ad hoc mono- and bilingual tasks. One of the decisions taken early on in the organization of the CLEF ad hoc tracks was that the same set of topics would be used to query all collections, whatever the task. There are a number of reasons for this: it makes it easier to compare results over different collections, it means that there is a single master set that is rendered in all query languages, and a single set of relevance assessments for each language is sufficient for all tasks. However, the fact that the collections used in the CLEF 2005 ad hoc mono- and bilingual tasks were from two different time periods (1994-1995 and 2002) made topic creation particularly difficult. It was not possible to create time-dependent topics that referred to particular date-specific events as all topics had to refer to events that could have been reported in any of the collections, regardless of the dates. This meant that the CLEF 2005 topic set is somewhat different from the sets of previous years as the topics tend to be of broad coverage. However, it was difficult to construct topics that would find a limited number of relevant documents in each collection, and a - probably excessive - number of topics used for the 2005 mono- and bilingual tasks have a very large number of relevant documents. Although we have not analyzed in-depth the possible impact of this fact on results calculation, we suspect that it has meant that the 2005 ad hoc test collection is less

[4] It proved impossible to find national newspapers in electronic form for 1994 and/or 1995 in these languages.

effective in "discriminating" between the performance of different systems. For this reason, we subsequently decided to create separate test collections for the two different time-periods for the CLEF 2006 ad hoc mono- and bilingual tasks.

For the multilingual task, the CLEF 2003 topic sets of 60 topics were used. For CLEF 2005 these were divided into two sets: 20 topics for training and 40 for testing. Topics were potentially available in all the original languages for the CLEF 2003 tasks. For CLEF 2005 participants variously chose to use English, Dutch and Spanish language topics.

Below we give an example of the English version of a typical CLEF topic:

```
<top> <num> C254 </num>
<EN-title> Earthquake Damage </EN-title>
<EN-desc> Find documents describing damage to property or persons caused
by an earthquake and specifying the area affected.</EN-desc>
<EN-narr> Relevant documents will provide details on damage to buildings
and material goods or injuries to people as a result of an earthquake.
The geographical location (e.g. country, region, city) affected by the
earthquake must also be mentioned.</EN-narr>
</top>
```

2.2 Participation Guidelines

To carry out the retrieval tasks of the CLEF campaign, systems have to build supporting data structures. Allowable data structures include any new structures built automatically (such as inverted files, thesauri, conceptual networks, etc.) or manually (such as thesauri, synonym lists, knowledge bases, rules, etc.) from the documents. They may not, however, be modified in response to the topics, e.g. by adding topic words that are not already in the dictionaries used by their systems in order to extend coverage.

Some CLEF data collections contain manually assigned, controlled or uncontrolled index terms. The use of such terms has been limited to specific experiments that have to be declared as "manual" runs.

Topics can be converted into queries that a system can execute in many different ways. CLEF strongly encourages groups to determine what constitutes a base run for their experiments and to include these runs (officially or unofficially) to allow useful interpretations of the results. Unofficial runs are those not submitted to CLEF but evaluated using the trec_eval package. This year we have used the new package written by Chris Buckley for the *Text REtrieval Conference (TREC)* (trec_eval 7.3) and available from the TREC website.

As a consequence of limited evaluation resources, a maximum of 4 runs for each multilingual task and a maximum of 12 runs overall for the bilingual tasks, including all language combinations, was accepted. The number of runs for the monolingual task was limited to 12 runs. No more than 4 runs were allowed for any individual language combination. Overall, participants were allowed to submit at most 32 runs in total for the multilingual, bilingual and monolingual tasks.

2.3 Relevance Assessment

The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead approximate recall values are calculated using pooling techniques. The results submitted by the groups participating in the ad hoc tasks are used to form a pool of documents for each topic and language by collecting the highly ranked documents from all submissions. This pool is then used for subsequent relevance judgments. The stability of pools constructed in this way and their reliability for post-campaign experiments is discussed in [2] with respect to the CLEF 2003 pools. After calculating the effectiveness measures, the results are analyzed and run statistics produced and distributed. New pools were formed in CLEF 2005 for the runs submitted for the mono- and bilingual tasks and the relevance assessments were performed by native speakers. The multilingual tasks used the original pools and relevance assessments from CLEF 2003.

The individual results for all official ad hoc experiments in CLEF 2005 are given in the Appendix at the end of the on-line Working Notes prepared for the Workshop [3]. They are discussed below in Sections 3, 4 and 5, for the mono-, bi-, and multilingual tasks, respectively.

2.4 Result Calculation

Evaluation campaigns such as TREC and CLEF are based on the belief that the effectiveness of *Information Retrieval Systems (IRSs)* can be objectively evaluated by an analysis of a representative set of sample search results. For this, effectiveness measures are calculated based on the results submitted by the participant and the relevance assessments. Popular measures usually adopted for exercises of this type are Recall and Precision. Details on how they are calculated for CLEF are given in [4].

2.5 Participants and Experiments

As shown in Table 1, a total of 23 groups from 15 different countries submitted results for one or more of the ad hoc tasks - a slight decrease on the 26 participants of last year. A total of 254 experiments were submitted, nearly the same as the 250 experiments of 2004. Thus, there is a slight increase in the average number of submitted runs per participant: from 9.6 runs/participant of 2004 to 11 runs/participant of this year.

Participants were required to submit at least one title+description ("TD") run per task in order to increase comparability between experiments. The large majority of runs (188 out of 254, 74.02%) used this combination of topic fields, 54 (21.27%) used all fields, 10 (3.94%) used the title field, and only 2 (0.79%) used the description field. The majority of experiments were conducted using automatic query construction. A breakdown into the separate tasks is shown in Table 2(a).

Table 1. CLEF 2005 ad hoc participants – new groups are indicated by *.

Part.icipant	Institution	Country
alicante	U. Alicante - Comp.Sci	Spain
buffalo	SUNY at Buffalo - Informatics	USA
clips	CLIPS-IMAG Grenoble	France
cmu	Carnegie Mellon U.- Lang.Tec.	USA
cocri	ENSM St. Etienne	France*
dcu	Dublin City U. - Comp.Sci.	Ireland
depok	U.Indonesia - Comp.Sci	Indonesia*
dsv-stockholm	U.Stockholm, NLP	Sweden
hildesheim	U.Hildesheim - Inf.Sci	Germany
hummingbird	Hummingbird Core Tech.	Canada
ilps	U.Amsterdam - Informatics	The Netherlands
isi-unige	U.Geneva - Inf.Systems	Switzerland*
jaen	U.Jaen - Intell.Systems	Spain
JHU/apl	Johns Hopkins U.- App.Physics	USA
miracle	Daedalus & Madrid Univs	Spain
msu-nivc	Moscow State U.- Computing	Russia*
sics	Swedish Inst. for Comp.Sci	Sweden
tlr	Thomson Legal Regulatory	USA
u.budapest	Budapest U. Tech. & Econom	Hungary*
u.glasgow	U.Glasgow - IR	UK
u.surugadai	U.Surugadai - Cultural Inf.	Japan
unine	U.Neuchatel - Informatics	Switzerland
xldb	U.Lisbon - Informatics	Portugal

Thirteen different topic languages were used in the ad hoc experiments - the Dutch run was in the multilingual tasks and used the CLEF 2003 topics. As always, the most popular language for queries was English, and French was second. Note that Bulgarian and Hungarian, the new collections added this year, were quite popular as new monolingual tasks - Hungarian was also used in one case as a topic language in a bilingual run. The number of runs per topic language is shown in Table 2(b).

3 Monolingual Experiments

Monolingual retrieval was offered for Bulgarian, French, Hungarian, and Portuguese. As can be seen from Table 2(a), the number of participants and runs for each language was quite similar, with the exception of Bulgarian, which had a slightly smaller participation. This year just 5 groups out of 16 (31.25%) submitted monolingual runs only (down from ten groups last year), and just one of these groups was a first time participant in CLEF. This is in contrast with previous years where many new groups only participated in monolingual experiments. This year, most of the groups submitting monolingual runs were doing this as part of their bilingual or multilingual system testing activity.

Table 2. Breakdown of experiments into tracks and topic languages.

(a) Number of experiments per track, participant.

Track	# Part.	# Runs
AH-2-years-on	4	21
AH-Merging	3	20
AH-Bilingual-X2BG	4	12
AH-Bilingual-X2FR	9	31
AH-Bilingual-X2HU	3	7
AH-Bilingual-X2PT	8	28
AH-Bilingual-X2EN	4	13
AH-Monolingual-BG	7	20
AH-Monolingual-FR	12	38
AH-Monolingual-HU	10	32
AH-Monolingual-PT	9	32
Total		**254**

(b) List of experiments by topic language.

Topic Lang.		# Runs
EN	English	118
FR	French	42
HU	Hungarian	33
PT	Portuguese	33
BG	Bulgarian	32
ES	Spanish	20
ID	Indonesian	18
DE	German	15
AM	Amharic	8
GR	Greek	4
IT	Italian	3
RU	Russian	3
NL	Dutch	1
Total		**254**

Table 3 shows the top five groups for each target collection, ordered by mean average precision. The table reports: the short name of the participating group; the mean average precision achieved by the run; the run identifier, specifying whether the run has participated in the pool or not, and the page in Appendix A of the Working Notes [3] containing all figures and graphs for this run; and the performance difference between the first and the last participant. The pages of Appendix A containing the overview graphs are indicated under the name of the sub-task. Table 3 regards runs using title + description fields only (the mandatory run).

All the groups in the top five had participated in previous editions of CLEF. Both pooled and not pooled runs are included in the best entries for each track. It can be noted that the trend observed in the previous editions of CLEF is confirmed: differences for top performers for tracks with languages introduced in past campaigns are small: in particular only 5.35% in the case of French (French monolingual has been offered in CLEF since 2000) and 7.55% in the case of Portuguese, which was introduced in 2004. However, for the new languages, Bulgarian and Hungarian, the differences are much greater, in the order of 25%, showing that there should be room for improvement if these languages are offered in future campaigns.

A main focus in the monolingual tasks was the development of new or the adaptation of existing stemmers and/or morphological analysers for the "new" CLEF languages.

Figures from 1 to 4 compare the performances of the top participants of the Monolingual Bulgarian, French, Hungarian, Portuguese tasks.

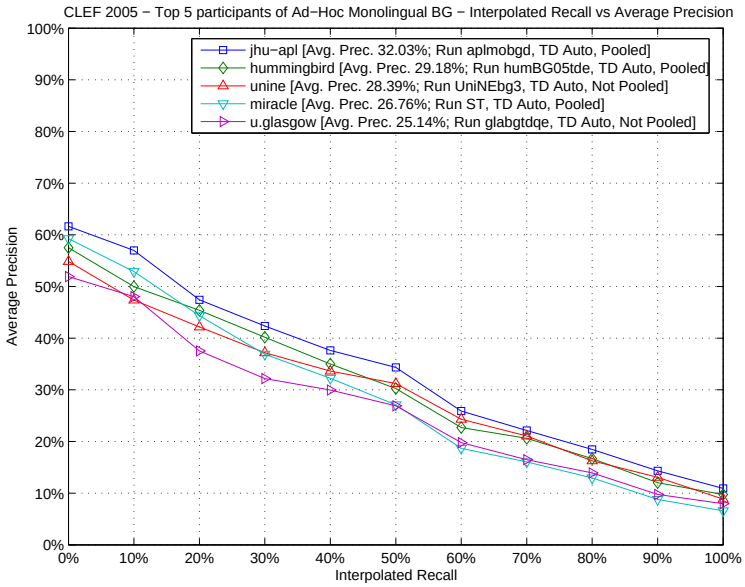

Fig. 1. Monolingual Bulgarian

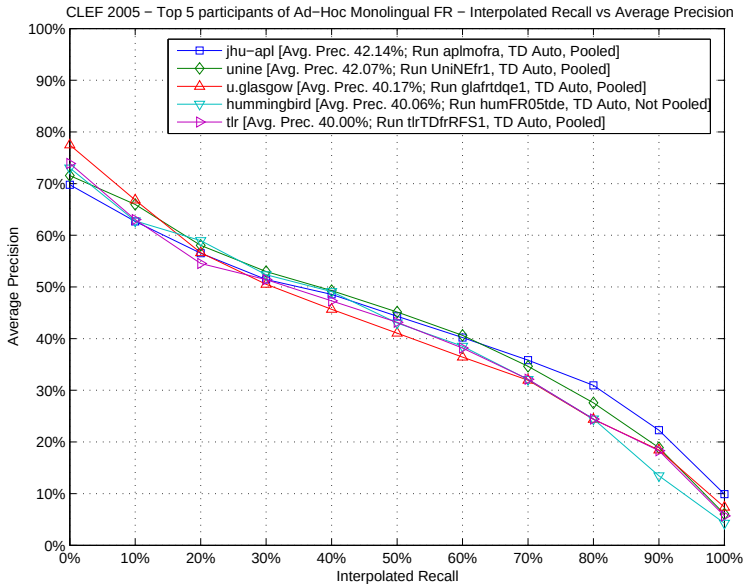

Fig. 2. Monolingual French

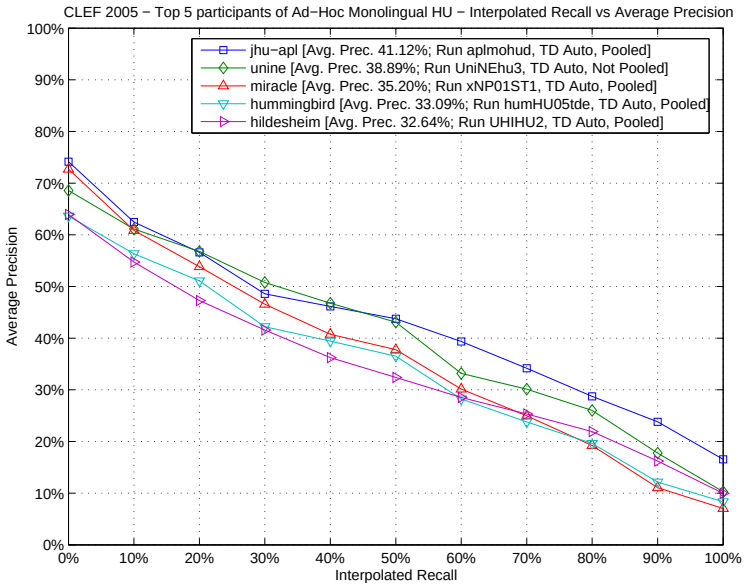

Fig. 3. Monolingual Hungarian

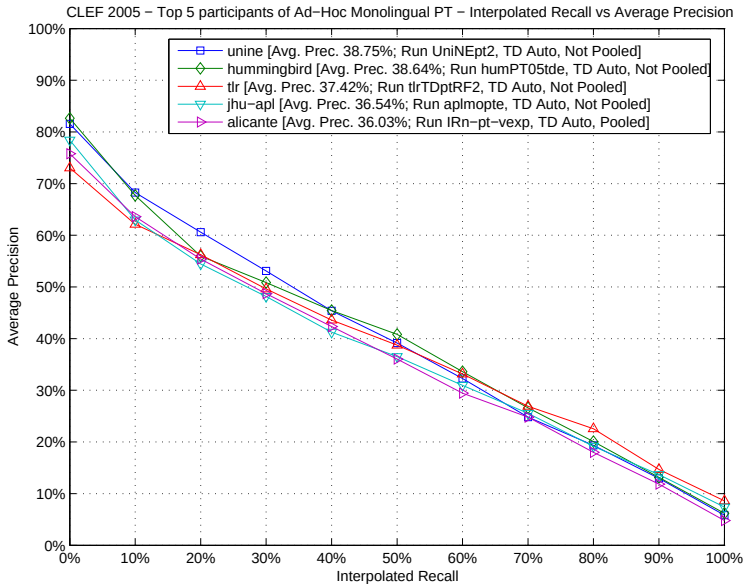

Fig. 4. Monolingual Portuguese

Table 3. Best entries for the monolingual track.

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Bulgarian (A.45–A.46)	jhu/apl 32.03% aplmobgd pooled (A.232)	hummingbird 29.18% humBG05tde pooled (A.230)	unine 28.39% UniNEbg3 not pooled (A.242)	miracle 26.76% ST pooled (A.235)	u.glasgow 25.14% glabgtdqe not pooled (A.239)	1st vs 5th 27.41%
French (A.49–A.50)	jhu/apl 42.14% aplmofra pooled (A.261)	unine 42.07% UniNEfr1 pooled (A.278)	u.glasgow 40.17% glafrtdqe1 pooled (A.275)	hummingbird 40.06% humFR05tde not pooled (A.260)	tlr 40.00% tlrTDfrRFS1 pooled (A.273)	1st vs 5th 5.35%
Hungarian (A.53–A.54)	jhu/apl 41.12% aplmohud pooled (A.294)	unine 38.89% UniNEhu3 not pooled (A.312)	miracle 35.20% xNP01ST1 pooled (A.297)	hummingbird 33.09% humHU05tde pooled (A.288)	hildesheim 32.64% UHIHU2 pooled (A.285)	1st vs 5th 25.98%
Portuguese (A.57–A.58)	unine 38.75% UniNEpt2 pooled (A.338)	hummingbird 38.64% humPT05tde not pooled (A.322)	tlr 37.42% tlrTDptRF2 not pooled (A.332)	jhu-apl 36.54% aplmopte not pooled (A.326)	alicante 36.03% IRn?pt?vexp pooled (A.314)	1st vs 5th 7.55%

4 Bilingual Experiments

The bilingual task was structured in four subtasks (X → BG, FR, HU or PT target collection) plus, as usual, an additional subtask with English as a target language restricted to newcomers to a CLEF cross-language task or to groups using unusual or new topic languages (Amharic, Greek, Indonesian, and Hungarian).

Table 4 shows the best results for this task for runs using the title+description topic fields. The performance difference between the best and the last (up to 5) placed groups is given (in terms of average precision. Again both pooled and non pooled runs are included in the best entries for each track, with the exception of Bilingual X → EN.

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2005:

- 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.

Similarly to monolingual, this is an interesting result. Whereas, the figures for French and Portuguese reflect those of recent literature [5], for the new languages where there has been little *Cross Language Information Retrieval (CLIR)*

Table 4. Best entries for the bilingual task.

Track	Participant Rank					
	1st	**2nd**	**3rd**	**4th**	**5th**	**Diff.**
Bulgarian (A.25–A.26)	miracle 23.55% ENXST pooled (A.135)	unine 13.99% UniNEbibg3 not pooled (A.143)	u.glasgow 12.04% glaenbgtd pooled (A.136)	jhu/apl 9.59% aplbienbge pooled (A.133)		1st vs 4th 145.57%
French (A.33–A.34)	alicante 35.90% IRn-enfr-vexp not pooled	unine 34.67% UniNEbifr2 not pooled	hildesheim 34.65% UHIENFR2 not pooled	jhu/apl 34.42% aplbienfrc pooled	miracle 30.76% ENSST not pooled	1st vs 5th 16.71%
Hungarian (A.37–A.38)	miracle 30.16% ENMST not pooled	unine 28.82% UniNEbihu3 not pooled	jhu/apl 24.58% aplbienhue not pooled			1st vs 3rd 22.70%
Portuguese (A.41–A.42)	unine 34.04% UniNEbipt1 pooled (A.216)	jhu/apl 31.85% aplbiesptb not pooled (A.204)	miracle 31.06% ESAST not pooled (A.209)	alicante 29.18% IRn-enpt-vexp not pooled (A.197)	tlr 23.58% tlrTDfr2ptRFS1 pooled (A.212)	1st vs 5th 44.36%
English (A.29–A.30)	jhu/apl 33.13% aplbiidena pooled (A.152)	u.glasgow 29.35% glagrentdqe pooled (A.156)	depok 12.85% UI-TD10 pooled (A.146)			1st vs 3rd 157.82%

system experience and testing so far it can be seen that, there is much room for improvement. It is interesting to note that when CLIR system evaluation began in 1997 at TREC-6 the best CLIR systems had the following results:

- EN → FR: 49% of best monolingual French IR system;
- EN → DE: 64% of best monolingual German IR system.

Figures 5 to 9 compare the performances of the top participants of the Bilingual tasks with the following target languages: Bulgarian, French, Hungarian, Portuguese, and English. Although, as usual, English was by far the most popular language for queries, some less common and interesting query to target language pairs were tried, e.g. Amharic, Spanish and German to French, and French to Portuguese.

From the reports of the groups that participated in the bilingual ad hoc tasks, it appears that the CLEF 2005 experiments provide a good overview of most of the traditional approaches to CLIR when matching between query and target collection, including n-gram indexing, machine translation, machine-readable bilingual dictionaries, multilingual ontologies, pivot languages, query and document translation - perhaps corpus-based approaches were less used than in previous years continuing a trend first noticed in CLEF 2004. Veteran

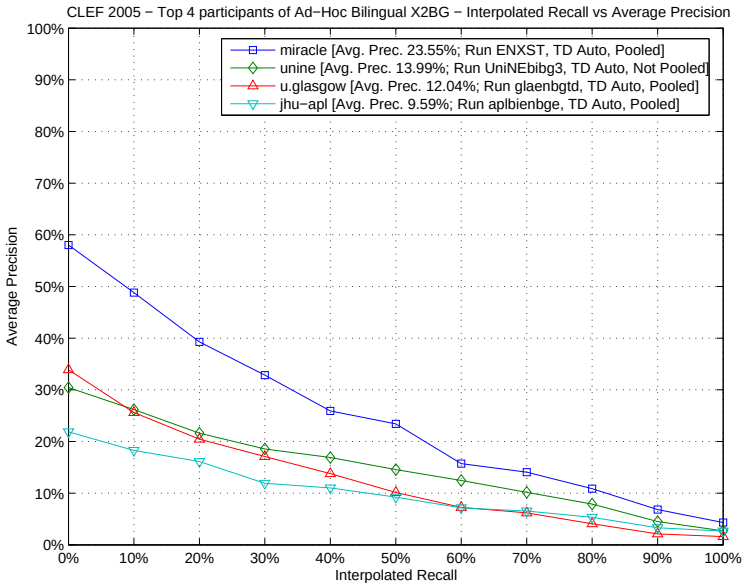

Fig. 5. Bilingual Bulgarian

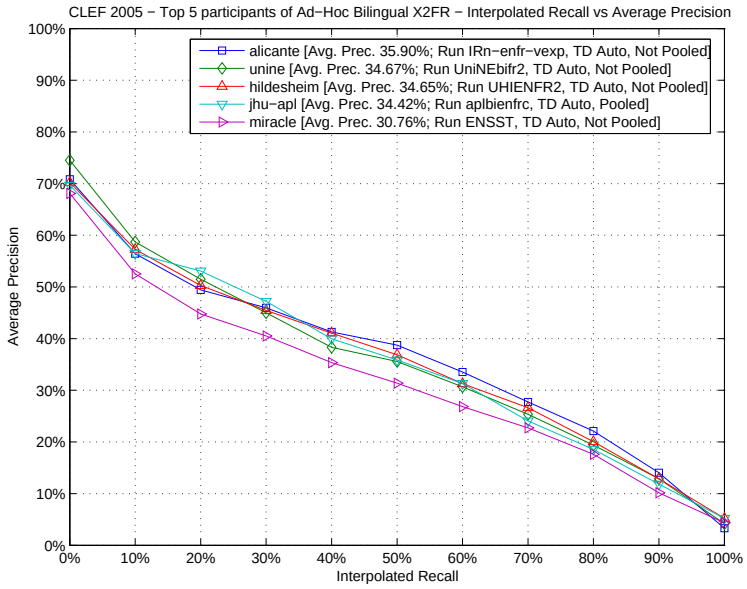

Fig. 6. Bilingual French

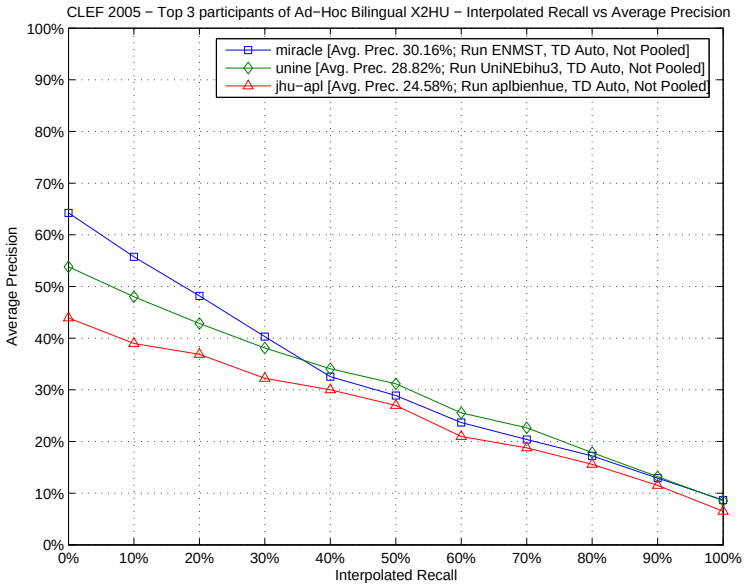

Fig. 7. Bilingual Hungarian

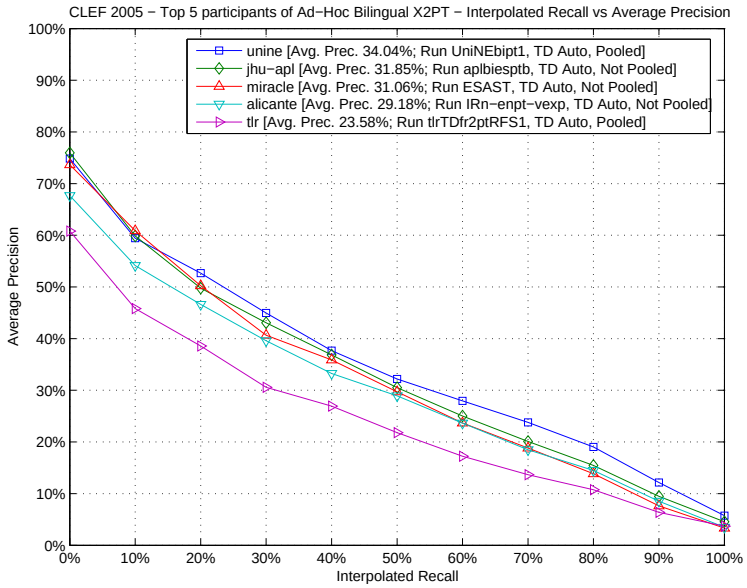

Fig. 8. Bilingual Portuguese

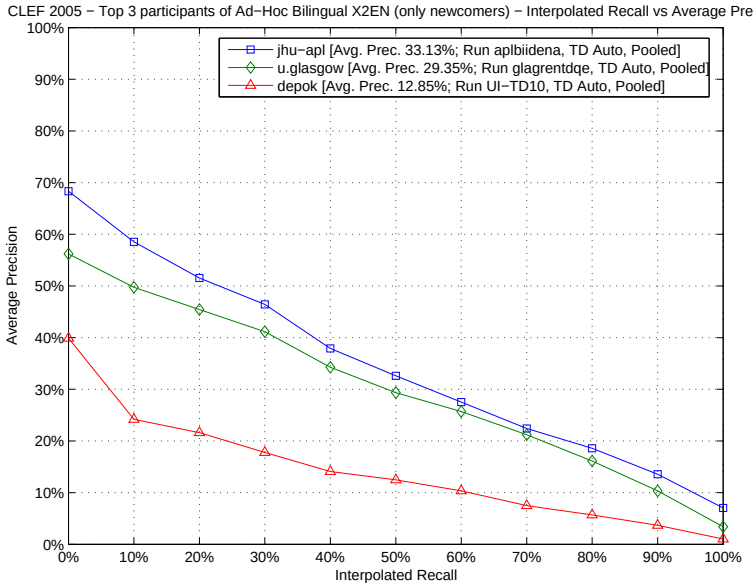

Fig. 9. Bilingual English

groups were mainly concerned with fine tuning and optimizing strategies already tried in previous years. The issues examined were the usual ones: word-sense disambiguation, out-of-dictionary vocabulary, ways to apply relevance feedback, results merging, etc.

5 Multilingual Experiments

Table 5 shows results for the best entries for the multilingual tasks. The table reports: the short name of the participating group; the mean average precision achieved by the run; the run identifier; the page in Appendix A of the Working Notes [3] containing all figures and graphs for this run; the performance difference between the first and the last participant. The pages of Appendix A containing the overview graphs are indicated under the name of the sub-task.

Table 5 shows runs using title + description fields only (the mandatory run). The first row of the table shows the results of the top 5 group submissions of the CLEF 2003 Multi-8 task for comparison with the 2-Years-On and Merging tasks of this year. Additional rows for each task show the difference in the MAP for this run compared to the best performing run at this rank in the original CLEF 2003 Multi-8 task.

Since the CLEF 2005 multilingual tasks used only 40 topics of the original 60 topics of the 2003 as the test set (topics 161 to 200), while the first 20 topics

Table 5. Best entries for the multilingual task.

Track	Participant Rank					
	1st	**2nd**	**3rd**	**4th**	**5th**	**Diff.**
CLEF 2003	UC Berkeley 38.77% bkmul8en3 pooled	U. Neuchatel 35.69% UniNEml1 not pooled	U. Amsterdam 29.62% UAmsC03EnM8SS4G not pooled	jhu/apl 25.29% aplmuen8b not pooled	U. Tampere 18.95% UTAmul1 pooled	1st vs 5th 104.59%
2 Years On **(A.17–A.18)**	Cmu 44.93% adhocM5Trntes not pooled (A.93) +15.89%	jaen 29.57% UJAPRFRSV2RR not pooled (A.101) -17.34%	miracle 26.06% esml9XstiSTp not pooled (A.110) -12.02%	isi-unige 10.33% AUTOEN not pooled (A.96) -59.15%		1st vs 4th 334.95%
Merging **(A.21–A.22)**	Cmu 41.19% UNET150w05test – (A.118) +6.24%	dcu 32.86% dcu.Prositqgm2 – (A.121) -7.93%	Jaen 30.37% UJAMENEDFRR – (A.129) +2.53%			1st vs 3rd 35.63%

(topics 141 to 160) were used as a training set, the average precision of the original 2003 runs was recomputed for the 40 test topics used this year. These revised MAP figures are reported in Table 5. These figures are thus slightly different from the original results which appear in the CLEF 2003 proceedings [2] which were calculated for the original set of 60 topics., although the ranking of these runs remains unchanged.

It can be seen from Table 5 that the performance difference between the first and the last participant for the 2-Years-On track is much greater (nearly 3 times) than the corresponding difference in 2003, even if the task performed in these two tracks is the same. On the other hand, the performance difference for the Merging track is nearly one third of the corresponding difference in 2003: it seems that merging the results of the run reduces the gap between the best and the last performer, even though there is still a considerable difference (35.63%), if compared to the small differences between the results for the most popular monolingual languages, e.g. 5.35% of monolingual French. We can note that the top participant of the 2-Years-On task achieves a 15.89% performance improvement with respect to the top participant of CLEF 2003 Multi-8. On the other hand, the fourth participant of the 2-Years-On task has a 59.15% decrease in performance with respect to the fourth participant of CLEF 2003 Multi-8. Similarly, we can note that the top participant of the Merging track achieves a 6.24% performance improvement with respect to the top participant of 2003.

In general, we can note that for the 2-Years-On task there is a performance improvement only for the top participant, while the performances deteriorate quickly for the other participants with respect to 2003. On the other hand, for the Merging task the performance improvement of the top participant with respect to 2003 is less than in the case of the 2-Years-On task. There is also less variation between the submissions for the Merging task than seen in the earlier 2003 runs. This is probably due to the fact that the participants were using the same ranked lists, and that the variation in performance arises only from the merging strategies adopted.

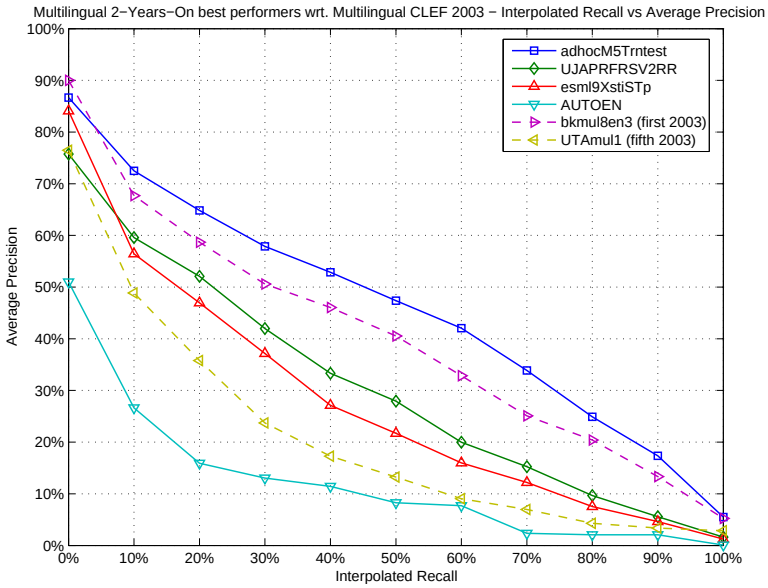

Fig. 10. Interpolated Recall vs Average Precision. Comparison between Multilingual 2-Years-On and CLEF 2003 Multilingual-8.

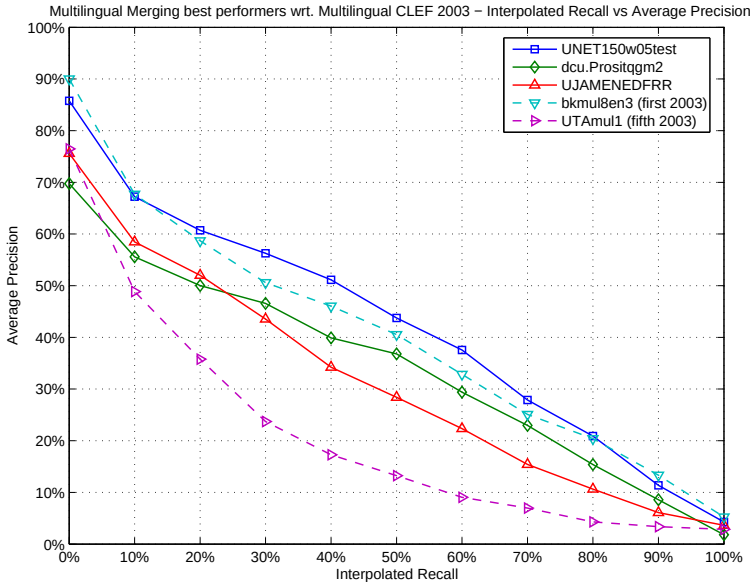

Fig. 11. Interpolated Recall vs Average Precision. Comparison between Multilingual Merging and CLEF 2003 Multilingual-8.

Figure 10 compares the performances in terms of the precision at different document cut-off values of the top participants of the 2-Years-On task with respect to the top and the fifth performer of CLEF 2003 Multilingual-8. Figure 11 shows corresponding results for the Multilingual Merging task. Trends in these figures are similar to those seen in Table 5. The top performing submissions for the Multilingual 2-Years-On and Merging tasks are both clearly higher than the best submission to the CLEF 2003 task. The variation between submissions for 2-Years-On is also greater than that observed for the Merging only task.

The multilingual tasks at CLEF 2005 were intended to assess whether reuse of the CLEF 2003 Multi-8 task data could give an indication of progress in multilingual information retrieval and to provide common sets of ranked lists to enable specific exploration of merging strategies for multilingual information retrieval. The submissions to these tasks show that multilingual performance can indeed be improved beyond that reported at CLEF 2003 both when performing the complete retrieval process and when merging ranked result lists generated by other groups. The initial running of this task suggests that there is scope for further improvement in multilingual information retrieval from exploiting ongoing improvements in information retrieval methods, but also from focused exploration of merging techniques.

6 Statistical Testing

For reasons of practicality, the CLEF 2005 multilingual track used a limited number of queries (40), which are intended to represent a more or less appropriate sample 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 to determine what differences between runs appear to be real as opposed to differences that are due to sampling issues. We aim to identify runs with results that are significantly different from the results of other runs. "Significantly different" in this context means that the difference between the performance scores for the runs in question appears greater than what might be expected by pure chance. As with all statistical testing, conclusions will be qualified by an error probability, which was chosen to be 0.05 in the following. We have designed our analysis to follow closely the methodology used by similar analyses carried out for TREC [6].

We used the MATLAB Statistics Toolbox 5.0.1 this year, which provides the necessary functionality plus some additional functions and utilities. We use the *ANalysis Of VAriance (ANOVA)* test. ANOVA makes some assumptions concerning the data be checked. Hull [6] 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. Two tests for goodness of fit to a normal distribution were chosen using the MATLAB statistical toolbox: the Lilliefors test [7] and the Jarque-Bera test [8]. In the case of the CLEF tasks under analysis, both tests indicate that the assumption of normality is violated for most of the data samples (in this case the runs for each participant).

Table 6. Lilliefors test for each track with (LL) and without Tague-Sutcliffe arcsin transformation (LL & TS). Jarque-Bera test for each track with (JB) and without Tague-Sutcliffe arcsin transformation (JB & TS).

Track	**LL**	**LL & TS**	**JB**	**JB & TS**
2 Years On	8/21	17/21	13/21	19/21
Merging	8/20	15/20	13/20	18/20
Bilingual Bulgarian	0/12	1/12	0/12	5/12
Bilingual English	12/31	24/31	21/31	25/31
Bilingual French	6/31	19/31	19/31	22/31
Bilingual Hungarian	0/7	5/7	1/7	5/7
Bilingual Portuguese	9/28	19/28	10/28	19/28
Monolingual Bulgarian	4/20	17/20	14/20	19/20
Monolingual French	12/28	38/38	30/28	38/38
Monolingual Hungarian	2/32	17/32	12/32	26/32
Monolingual Portuguese	24/32	30/32	27/32	28/32

In such cases, a transformation of data should be performed. The transformation for measures that range from 0 to 1 is the arcsin-root transformation:

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

which Tague-Sutcliffe [9] recommends for use with precision/recall measures.

Table 6 shows the results of the Lilliefors test before and after applying the Tague-Sutcliffe transformation. After the transformation the analysis of the normality of samples distribution improves significantly, with the exception of the bilingual Bulgarian. Each entry shows the number of experiments whose performance distribution can be considered drawn from a Gaussian distribution, with respect to the total number of experiment of the track. The value of alpha for this test was set to 5%. The same table shows also the same analysis with respect to the Jarque-Bera test. The value of alpha for this test was set to 5%. The difficulty to transform the data into normally distributed samples derives from the original distribution of run performances which tend towards zero within the interval [0,1].

Figure 12 presents a boxplot graph providing a more detailed analysis of the above mentioned phenomenon for the bilingual task with Bulgarian target collection. As can be seen, the distribution of the average precision for the different experiments is skewed, and this helps to explain the deviation from the normality. Moreover, the data distribution tends towards low performances, which confirms the difficulty of dealing with new languages.

The following tables, from Table 7 to Table 17, summarize the results of this test. All experiments, regardless the 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. Each table shows the overall results where all the runs that are included in the same group do not have a significantly different performance. All runs scoring below a certain group perform significantly worse than at least the top entry of the group. Likewise all the runs scoring above a

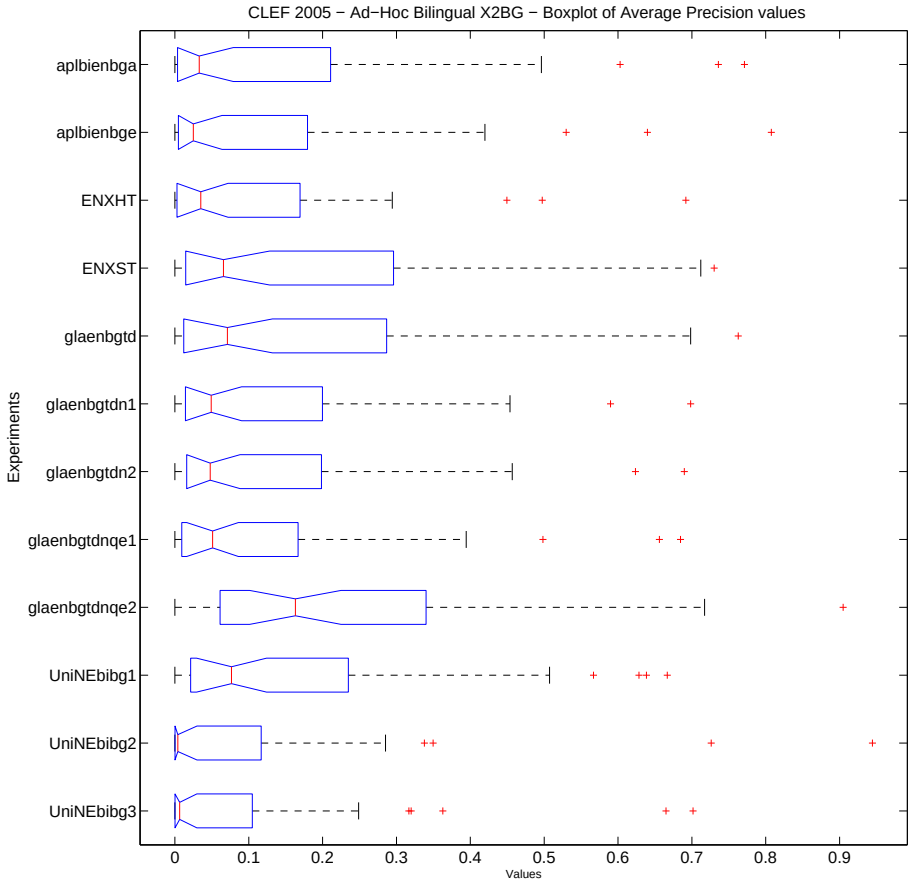

Fig. 12. Boxplot analysis of the bilingual task with Bulgarian target collection.

certain group perform significantly better than at least the bottom entry in that group.

It is well-known that it is fairly difficult to detect statistically significant differences between retrieval results based on 40 queries [9,10]. While 40 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.

Table 7. Monolingual Bulgarian. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups				
0.5687	aplmobgd	X				
0.5568	aplmobgc	X				
0.5343	humBG05tde	X	X			
0.5206	UniNEbg3	X	X	X		
0.5191	aplmobge	X	X	X		
0.5172	UniNEbg1	X	X	X		
0.5120	ST	X	X	X	X	
0.5120	humBG05td	X	X	X	X	
0.4937	UniNEbg2	X	X	X	X	X
0.4874	humBG05t	X	X	X	X	X
0.4742	glabgtdqe	X	X	X	X	X
0.4619	glabgtdnqe	X	X	X	X	X
0.4275	glabgtdn		X	X	X	X
0.4154	r1SR		X	X	X	X
0.4091	UHIBG2			X	X	X
0.3974	UHIBG1				X	X
0.3939	BGHT				X	X
0.3844	IRn-bu-vnexp					X
0.3775	IRn-bu-fexp					X
0.3755	IRn-bu-vexp					X

Table 8. Monolingual French. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups							
0.6821	UniNEfr1	X							
0.6779	aplmofra	X	X						
0.6691	aplmofrb	X	X	X					
0.6686	UniNEfr3	X	X	X					
0.6648	UniNEfr2	X	X	X					
0.6609	tlrTDfrRFS1	X	X	X					
0.6598	humFR05tde	X	X	X					
0.6581	glafrtdqe1	X	X	X					
0.6459	aHRSR	X	X	X	X				
0.6444	SrgdMono01	X	X	X	X				
0.6359	UHIFR2	X	X	X	X				
0.6328	UHIFR1	X	X	X	X				
0.6315	tlrTDfr3	X	X	X	X				
0.6279	aplmofre	X	X	X	X				
0.6276	aHRSRxNP01HR1	X	X	X	X				
0.6271	aplmofrc	X	X	X	X				
0.6265	humFR05td	X	X	X	X				
0.6251	aHTST	X	X	X	X				
0.6240	glafrtdqe2	X	X	X	X				
0.6002	IRn-fr-vexp	X	X	X	X	X			
0.5862	IRn-fr-fexp	X	X	X	X	X	X		
0.5779	sics-fr-k	X	X	X	X	X	X	X	
0.5672	sics-fr-b		X	X	X	X	X	X	
0.5653	glafrtdn		X	X	X	X	X	X	
0.5640	sics-fr-van			X	X	X	X	X	
0.5421	IRn-fr-vnexp				X	X	X	X	X
0.5418	humFR05t				X	X	X	X	X
0.4991	UHIFR4					X	X	X	X
0.4929	UHIFR3					X	X	X	X
0.4872	xNP01r1SR1					X	X	X	X
0.4754	RIMfuzzLemme080					X	X	X	X
0.4704	RIMfuzzLemme050					X	X	X	X
0.4685	RIMfuzzTD050						X	X	X
0.4313	CLIPS05FR0							X	X	X
0.4056	RIMfuzzET050							X	X	X
0.4054	RIMfuzzET020								X	X
0.3413	CLIPS05FR1									X
0.3209	CLIPS05FR2									X

Table 9. Monolingual Hungarian. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups							
0.6853	aplmohud	X							
0.6844	aplmohuc	X							
0.6834	aplmohue	X							
0.6571	UniNEhu3	X	X						
0.6284	UniNEhu1	X	X	X					
0.6103	xNP01ST1	X	X	X					
0.6012	aHTSTxNP01ST1	X	X	X					
0.5974	UniNEhu2	X	X	X					
0.5896	humHU05tde	X	X	X	X				
0.5786	UHIHU2	X	X	X	X	X			
0.5721	IRn-hu-vexp	X	X	X	X	X	X		
0.5659	qfstfs	X	X	X	X	X	X		
0.5634	qfirststemtall	X	X	X	X	X	X		
0.5606	IRn-hu-vnexp	X	X	X	X	X	X	X	
0.5587	humHU05td	X	X	X	X	X	X	X	X
0.5575	IRn-hu-fexp	X	X	X	X	X	X	X	X
0.5514	UHIHU1		X	X	X	X	X	X	X
0.5361	tlrTDhuSC		X	X	X	X	X	X	X
0.5321	tlrTDhuE		X	X	X	X	X	X	X
0.5200	UAmsMoHu1AnH			X	X	X	X	X	X
0.5193	HUHT			X	X	X	X	X	X
0.5180	qalltall			X	X	X	X	X	X
0.5179	UAmsMoHu3AnL			X	X	X	X	X	X
0.5027	humHU05t			X	X	X	X	X	X
0.4670	UAmsMoHu2AnG				X	X	X	X	X
0.4423	qnostemtfirsstem					X	X	X	X
0.4395	UAmsMoHu4AnV						X	X	X
0.4289	qnostemtnostem							X	X
0.4282	qnostemtall								X
0.2685	glahutdnqe								X
0.2685	glahutdqe								X
0.2592	glahutdn								X

Table 10. Monolingual Portuguese. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups						
0.6573	humPT05tde	X						
0.6562	UniNEpt2	X						
0.6483	UniNEpt1	X	X					
0.6453	tlrTDptRF2	X	X					
0.6413	SR	X	X					
0.6336	xNP01SR1	X	X	X				
0.6323	aplmopte	X	X	X				
0.6289	humPT05td	X	X	X				
0.6269	IRn-pt-vexp	X	X	X				
0.6261	aplmoptc	X	X	X				
0.6257	UniNEpt3	X	X	X				
0.6257	tlrTDptRFS1	X	X	X				
0.6202	tlrTDpt3	X	X	X				
0.6165	ST	X	X	X				
0.6090	IRn-pt-fexp	X	X	X				
0.5983	IRn-pt-vnexp	X	X	X				
0.5816	UBmono-pt-rf2	X	X	X	X			
0.5792	UBmono-pt-rf1	X	X	X	X			
0.5788	UBmono-pt-comb1	X	X	X	X			
0.5777	UBmono-pt-rf3	X	X	X	X			
0.5770	aplmoptd	X	X	X	X			
0.5614	aplmopta	X	X	X	X			
0.5556	humPT05t	X	X	X	X			
0.5394	XLDBTumba01		X	X	X	X		
0.5217	aSRr1SR			X	X	X	X	
0.4860	glapttdqe				X	X	X	X
0.4832	XLDBTumba05				X	X	X	X
0.4826	glapttdnqe				X	X	X	X
0.4427	glapttdn					X	X	X
0.4127	XLDBTumba02						X	X
0.4071	XLDBTumba09							X
0.3942	XLDBTumba06							X

Table 11. Bilingual target Bulgarian. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups		
0.4608	ENXST	X		
0.3618	glaenbgtdnqe1	X	X	
0.3548	ENXHT	X	X	
0.3470	glaenbgtdnqe2	X	X	
0.3077	glaenbgtdn1		X	X
0.3000	glaenbgtdn2		X	X
0.2944	UniNEbibg3		X	X
0.2846	glaenbgtd		X	X
0.2711	UniNEbibg2		X	X
0.2598	UniNEbibg1		X	X
0.2111	aplbienbge			X
0.1951	aplbienbga			X

Table 12. Bilingual target French. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups				
0.6002	IRn-enfr-vexp	X				
0.5961	UniNEbifr2	X				
0.5958	UHIENFR2	X				
0.5950	UniNEbifr3	X				
0.5857	UniNEbifr1	X				
0.5804	aplbienfrc	X	X			
0.5789	UHIENFR1	X	X			
0.5543	ENSxNP01SR1	X	X			
0.5537	ESSxNP01SR1	X	X			
0.5448	ENSST	X	X	X		
0.5319	IRn-enfr-vnexp	X	X	X		
0.5256	ESSST	X	X	X		
0.5249	IRn-enfr-fexp	X	X	X		
0.5048	ESSxNP01HR1	X	X	X		
0.5011	glaitfrtdnqe	X	X	X		
0.5007	ENSxNP01HR1	X	X	X		
0.4847	UHIRUFR1	X	X	X		
0.4758	SrgdMgE03	X	X	X		
0.4731	SrgdQT04	X	X	X		
0.4644	SrgdMgG02	X	X	X		
0.4362	glaitfrtdn		X	X	X	
0.4078	glaitfrtd			X	X	
0.3065	SrgdDT05				X	X
0.1693	CLIPS05DEFR0					X
0.1341	CLIPS05ESFR0					X
0.1337	CLIPS05DEFR					X
0.1257	CLIPS05EFR					X
0.1226	ds-am-fr-da-s					X
0.1224	ds-am-fr-nonda-s					X
0.1004	ds-am-fr-nonda-l					X
0.0898	ds-am-fr-da-l					X

Table 13. Bilingual target Hungarian. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups	
0.5448	aplbienhua	X	
0.5377	aplbienhue	X	
0.5097	UniNEbihu2	X	X
0.5004	UniNEbihu1	X	X
0.4385	UniNEbihu3	X	X
0.4346	ENMxNP01ST1	X	X
0.4098	ENMST		X

Table 14. Bilingual target Portuguese. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups									
0.5943	UniNEbipt1	X									
0.5927	ESASR	X									
0.5828	ESAxNP01SR1	X	X								
0.5808	aplbiesptb	X	X								
0.5714	ESAST	X	X	X							
0.5632	UniNEbipt2	X	X	X							
0.5630	UniNEbipt3	X	X	X							
0.5567	aplbienptb	X	X	X							
0.5366	IRn-enpt-vexp	X	X	X							
0.5334	IRn-enpt-fexp	X	X	X	X						
0.5078	IRn-enpt-fexpfl	X	X	X	X	X					
0.4943	IRn-enpt-vnexp	X	X	X	X	X					
0.4640	tlrTDfr2ptRFS1		X	X	X	X	X				
0.4514	tlrTDfr2pt3			X	X	X	X	X			
0.4132	ENSSR				X	X	X	X			
0.4109	ENSxNP01SR1					X	X	X	X		
0.4024	ENSST					X	X	X	X		
0.3751	UBbi-en-pt-t2						X	X	X		
0.3741	UBbi-en-pt-comb2						X	X	X		
0.3740	UBbi-en-pt-t1						X	X	X		
0.3449	UBbi-en-pt-comb1							X	X	X	
0.3073	glaespttdnqe								X	X	X
0.2448	glaespttdn									X	X	X
0.2202	glaespttd										X	X
0.1389	XLDBTumba03											X
0.1373	XLDBTumba04											X
0.1344	XLDBTumba08											X
0.1239	XLDBTumba07											X

Table 15. Bilingual target English. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups								
0.6952	IRn-en-vexp	X								
0.6844	IRn-en-fexp	X								
0.6757	IRn-en-vnexp	X								
0.6755	UBmono-en-3	X								
0.6673	prise2	X								
0.6668	UAmsC05EnEnStmLM	X								
0.6606	UAmsC05EnEnStm	X								
0.6548	UBmono-en-2	X								
0.6475	UBmono-en-1	X	X							
0.6310	aplbiidend	X	X							
0.6183	UAmsC05EnEn4Gr	X	X							
0.6016	UAmsC05EnEnWrdLM	X	X							
0.5972	prise4	X	X							
0.5736	aplbiidena	X	X	X						
0.5689	prise1	X	X	X						
0.5574	prise3	X	X	X						
0.5095	glagrentdqe		X	X	X					
0.4526	cirGHLAru2en			X	X	X				
0.4498	aplbigrena			X	X	X				
0.4457	glagrentdn			X	X	X				
0.4076	cirGHLAen2en100				X	X	X			
0.3973	cirGHLAen2en110				X	X	X	X		
0.3874	cirGHLAen2en150				X	X	X	X	X	
0.3777	aplbihuena				X	X	X	X	X	
0.3337	cirGHLAen2en152					X	X	X	X	X
0.2973	UI-TD10						X	X	X	X
0.2738	UI-TD20						X	X	X	X
0.2683	UI-TITLE20							X	X	X
0.2541	UI-TITLE10								X	X
0.2324	UI-DESC10									X
0.2275	UI-DESC20									X

Table 16. Multilingual Merging. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups						
0.6786	UNET150w05test	X						
0.6615	UNET15w05test	X	X					
0.6549	UNEC150test	X	X	X				
0.6448	UNEC1000test	X	X	X	X			
0.5996	HBC1000test	X	X	X	X	X		
0.5687	dcu.Prositqgm2		X	X	X	X	X	
0.5641	dcu.Prositqgm1		X	X	X	X	X	
0.5604	dcu.Prositqgt		X	X	X	X	X	
0.5512	UJAMENEDFRR			X	X	X	X	
0.5501	HBC150test				X	X	X	
0.5495	dcu.Prositqgp				X	X	X	
0.5446	HBT150w05test				X	X	X	
0.5397	UJAMENEDF					X	X	
0.5326	UJAMENEOK					X	X	
0.5326	UJAMENEOKRR					X	X	
0.4882	HBT15w05test						X	X
0.4277	dcu.hump							X	X
0.4147	dcu.humm1							X	X
0.3985	dcu.humm2							X	X
0.3764	dcu.humt								X

Table 17. Multilingual 2 Years On. The table shows the Tukey T Test. The table reports the results of statistical analysis (two-way ANOVA) on the experiments.

Arcsin-transformed avg. prec. values	Run ID	Groups						
0.7247	adhocM3Trntest	X						
0.7184	adhocM4Trntest	X	X					
0.7046	adhocM5Trntest	X	X					
0.6992	adhocM5w1test	X	X					
0.5834	frml9XntfSRp		X	X				
0.5576	enml0XSRpHL			X	X			
0.5391	UJAPRFRSV2RR			X	X			
0.5357	UJAUARSV2RR			X	X			
0.5356	UJARSV2RR			X	X			
0.5310	UJARSV2			X	X			
0.5258	esml9XnteSRp			X	X			
0.4975	esml9XstiSTp			X	X	X		
0.4946	enmlXSRpA			X	X	X		
0.4841	enmlSTpHL			X	X	X		
0.4469	enmlSTpH			X	X	X	X	
0.4224	FEEDBCKEN				X	X	X	X	
0.3626	ADJUSTEN					X	X	X	X
0.3225	ADJUSTSP						X	X	X
0.3137	ADJUSTFR						X	X	X
0.3073	ADJUSTDU							X	X
0.2617	AUTOEN								X

7 Conclusions

We have reported the results of the ad hoc cross-language text document retrieval track at CLEF 2005. This track is considered to be central to CLEF as for many groups it is the first track in which they participate and provides them with them an opportunity to test their systems and compare performance between monolingual and cross-language runs, before perhaps moving on to more complex system development and subsequent evaluation. However, the track is certainly not just aimed at beginners. It also gives groups the possibility to measure advances in system performance over time. In addition, each year, we also include a task aimed at examining particular aspects of cross-language text retrieval. This year, the focus was on multilingual retrieval with our Multi-8 2-years-on and Multi-8 merging tasks.

The ad hoc track in CLEF 2006 offers the same target languages for the main mono- and bilingual tasks as in 2005 but has two additional focuses. Groups are encouraged to use non-European languages as topic languages in the bilingual task. Among others, we are offering Amharic, Hindi, Indonesian, Oromo, and Telugu. In addition, we have set up the "robust task" with the objective of providing the more expert groups with the chance to do in-depth failure analysis. At the time of writing, participation in these two particular tasks is encouraging. For more information, see our website[5].

Finally, it should be remembered that, although over the years we vary the topic and target languages offered in the track, all participating groups also have the possibility of accessing and using the test collections that have been created in previous years for all of the twelve languages included in the CLEF multilingual test collection. This test collection should soon be made publicly available on the *Evaluations and Language resources Distribution Agency (ELDA)* catalog[6].

References

1. Cleverdon, C.W.: The Cranfield Tests on Index Languages Devices. In Spack Jones, K., Willett, P., eds.: Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, California, USA (1997) 47–60
2. Braschler, M.: CLEF 2003 - Overview of results. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003) Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany (2004) 44–63
3. Di Nunzio, G.M., Ferro, N.: Appendix A. Results of the Core Tracks and Domain-Specific Tracks. In Peters, C., Quochi, V., eds.: Working Notes for the CLEF 2005 Workshop, `http://www.clef-campaign.org/2005/working_notes/workingnotes2005/appendix_a.pdf` [last visited 2006, February 28] (2005)
4. Braschler, M., Peters, C.: CLEF 2003 Methodology and Metrics. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Comparative Evaluation of Multilingual

[5] `http://www.clef-campaign.org/`

[6] `http://www.elda.org/`

Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003) Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany (2004) 7–20
5. Gonzalo, J., Peters, C.: The Impact of Evaluation on Multilingual Text Retrieval. In Baeza-Yates, R., Ziviani, N., Marchionini, G., Moffat, A., Tait, J., eds.: Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), ACM Press, New York, USA (2005) 603–604
6. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In Korfhage, R., Rasmussen, E., Willett, P., eds.: Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993), ACM Press, New York, USA (1993) 329–338
7. Conover, W.J.: Practical Nonparametric Statistics. 1st edn. John Wiley and Sons, New York, USA (1971)
8. Judge, G.G., Hill, R.C., Griffiths, W.E., Lütkepohl, H., Lee, T.C.: Introduction to the Theory and Practice of Econometrics. 2nd edn. John Wiley and Sons, New York, USA (1988)
9. Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. In Spack Jones, K., Willett, P., eds.: Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, California, USA (1997) 205–216
10. Voorhees, E.M., Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error. In Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R., Zobel, J., eds.: Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998), ACM Press, New York, USA (1998) 307–314