CLEF 2006: Ad Hoc Track Overview

Giorgio M. Di Nunzio[1], Nicola Ferro[1], Thomas Mandl[2], and Carol Peters[3]

[1] Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it
[2] Information Science, University of Hildesheim – Germany
mandl@uni-hildesheim.de
[3] ISTI-CNR, Area di Ricerca – 56124 Pisa – Italy
carol.peters@isti.cnr.it

Abstract. We describe the objectives and organization of the CLEF 2006 ad hoc track and discuss the main characteristics of the tasks offered to test monolingual, bilingual, and multilingual textual document retrieval systems. The track was divided into two streams. The main stream offered mono- and bilingual tasks using the same collections as CLEF 2005: Bulgarian, English, French, Hungarian and Portuguese. The second stream, designed for more experienced participants, offered the so-called "robust task" which used test collections from previous years in six languages (Dutch, English, French, German, Italian and Spanish) with the objective of privileging experiments which achieve good stable performance over all queries rather than high average performance. The document collections used were taken from the CLEF multilingual comparable corpus of news documents. The performance achieved for each task is presented and a statistical analysis of results is given.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 [**Systems and Software**]: Performance evaluation.

General Terms

Experimentation, Performance, Measurement, Algorithms.

Additional Keywords and Phrases

Multilingual Information Access, Cross-Language Information Retrieval

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. The CLEF 2006 ad hoc track was structured in two streams. The main

stream offered monolingual tasks (querying and finding documents in one language) and bilingual tasks (querying in one language and finding documents in another language) using the same collections as CLEF 2005. The second stream, designed for more experienced participants, was the "robust task", aimed at finding documents for very difficult queries. It used test collections developed in previous years.

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 **Robust** task offered monolingual, bilingual and multilingual tasks using the test collections built over three years: CLEF 2001 - 2003, for six languages: Dutch, English, French, German, Italian and Spanish. Using topics from three years meant that more extensive experiments and a better analysis of the results were possible. The aim of this task was to study and achieve good performance on queries that had proved difficult in the past rather than obtain a high average performance when calculated over all queries.

In this paper we describe the track setup, the evaluation methodology and the participation in the different tasks (Section 2), present the main characteristics of the experiments and show the results (Sections 3 - 5). Statistical testing is discussed in Section 6 and 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 the Ad Hoc section of the Working Notes.

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 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 coordina-

Table 1. Test collections for the main stream Ad Hoc tasks.

Language	Collections
Bulgarian	Sega 2002, Standart 2002
English	LA Times 94, Glasgow Herald 95
French	ATS (SDA) 94/95, Le Monde 94/95
Hungarian	Magyar Hirlap 2002
Portuguese	Público 94/95; Folha 94/95

tion 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 Collections

Different test collections were used in the ad hoc task this year. The main (i.e. non-robust) monolingual and bilingual tasks used the same document collections as in Ad Hoc last year but new topics were created and new relevance assessments made. As has already been stated, the test collection used for the robust task was derived from the test collections previously developed at CLEF. No new relevance assessments were performed for this task.

Documents. The document collections used for the CLEF 2006 ad hoc tasks are part of the CLEF multilingual corpus of newspaper and news agency documents described in the Introduction to these Proceedings.

In the main stream 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 means 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 flexible to handle such variants. The Bulgarian and Hungarian collections used in these tasks were new in CLEF 2005 and consist of national newspapers for the year 2002[1]. This has meant using collections of different time periods for the ad-hoc mono- and bilingual tasks. This had important consequences on topic creation. Table 1 summarizes the collections used for each language.

The robust task used test collections containing data in six languages (Dutch, English, German, French, Italian and Spanish) used at CLEF 2001, CLEF 2002 and CLEF 2003. There are approximately 1.35 million documents and 3.6 gigabytes of text in the CLEF 2006 "robust" collection. Table 2 summarizes the collections used for each language.

---

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

Table 2. Test collections for the Robust task.

Language	Collections
English	LA Times 94, Glasgow Herald 95
French	ATS (SDA) 94/95, Le Monde 94
Italian	La Stampa 94, AGZ (SDA) 94/95
Dutch	NRC Handelsblad 94/95, Algemeen Dagblad 94/95
German	Frankfurter Rundschau 94/95, Spiegel 94/95, SDA 94
Spanish	EFE 94/95

Topics Topics in the CLEF ad hoc track 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 2006 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 were 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, in CLEF 2005 the assessors found that the fact that the collections used in the CLEF 2006 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 all tend to be of broad coverage. In fact, it was difficult to construct topics that would find a limited number of relevant documents in each collection, and consequently a - probably excessive - number of topics used for the 2005 mono- and bilingual tasks have a very large number of relevant documents.

For this reason, we decided to create separate topic sets for the two different time-periods for the CLEF 2006 ad hoc mono- and bilingual tasks. We thus created two overlapping topic sets, with a common set of time independent topics and sets of time-specific topics. 25 topics were common to both sets while 25 topics were collection-specific, as follows:

- Topics C301 - C325 were used for all target collections

- Topics C326 - C350 were created specifically for the English, French and Portuguese collections (1994/1995)

- Topics C351 - C375 were created specifically for the Bulgarian and Hungarian collections (2002).

This meant that a total of 75 topics were prepared in many different languages (European and non-European): Bulgarian, English, French, German, Hungarian, Italian, Portuguese, and Spanish plus Amharic, Chinese, Hindi, Indonesian,

Oromo and Telugu. Participants had to select the necessary topic set according to the target collection to be used.

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

```
<top> <num> C302 </num>
<EN-title> Consumer Boycotts </EN-title> <
EN-desc> Find documents that describe or discuss the impact of consumer
boycotts. </EN-desc>
<EN-narr> Relevant documents will report discussions or points of view on
the efficacy of consumer boycotts. The moral issues involved in such
boycotts are also of relevance. Only consumer boycotts are relevant,
political boycotts must be ignored. </EN-narr> </top>
```

For the robust task, the topic sets used in CLEF 2001, CLEF 2002 and CLEF 2003 were used for evaluation. A total of 160 topics were collected and split into two sets: 60 topics used to train the system, and 100 topics used for the evaluation. Topics were available in the languages of the target collections: English, German, French, Spanish, Italian, Dutch.

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 12 runs each for the mono- and bilingual tasks was allowed (no more than 4 runs for any one language combination - we try to encourage diversity). We accepted a maximum of 4 runs per group and topic language for the multilingual robust task. For bi- and mono-lingual robust tasks, 4 runs were allowed per language or language pair.

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 [1] 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 2006 for the runs submitted for the main stream mono- and bilingual tasks and the relevance assessments were performed by native speakers. Instead, the robust tasks used the original pools and relevance assessments from CLEF 2003.

The individual results for all official ad hoc experiments in CLEF 2006 are given in the Appendix at the end of the on-line Working Notes prepared for the Workshop [2].

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 [3]. For the robust task, we used different measures, see below Section 5.

2.5 Participants and Experiments

As shown in Table 3, a total of 25 groups from 15 different countries submitted results for one or more of the ad hoc tasks - a slight increase on the 23 participants of last year. Table 4 provides a breakdown of the number of participants by country.

A total of 296 experiments were submitted with an increase of 16% on the 254 experiments of 2005. On the other hand, the average number of submitted runs per participant is nearly the same: from 11 runs/participant of 2005 to 11.7 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 (172 out of 296, 58.11%) used this combination of topic fields, 78 (26.35%) used all fields, 41 (13.85%) used the title field, and only 5 (1.69%) used the description field. The majority of experiments were conducted using automatic query construction (287 out of 296, 96.96%) and only in a small fraction

Table 3. CLEF 2006 ad hoc participants – new groups are indicated by *

Participant	Institution	Country
alicante	U. Alicante	Spain
celi	CELI, Torino *	Italy
colesir	U.Coruna and U.Sunderland	Spain
daedalus	Daedalus Consortium	Spain
dcu	Dublin City U.	Ireland
depok	U.Indonesia	Indonesia
dsv	U.Stockholm	Sweden
erss-toulouse	U.Toulouse/CNRS *	France
hildesheim	U.Hildesheim	Germany
hummingbird	Hummingbird Core Technology Group	Canada
indianstat	Indian Statistical Institute *	India
jaen	U.Jaen	Spain
ltrc	Int. Inst. IT *	India
mokk	Budapest U.Tech and Economics	Hungary
nilc-usp	U.Sao Paulo - Comp.Ling. *	Brazil
pucrs	U.Catolica Rio Grande do Sul *	Brazil
queenmary	Queen Mary, U.London *	United Kingdom
reina	U.Salamanca	Spain
rim	EMSE - Ecole Sup. des Mines	France
rsi-jhu	Johns Hopkins U. - APL	United States
saocarlos	U.Fed Sao Carlos - Comp.Sci *	Brazil
u.buffalo	SUNY at Buffalo	United States
ufrgs-usp	U.Sao Paulo and U.Fed. Rio Grande do Sul *	Brazil
unine	U.Neuchatel - Informatics	Switzerland
xldb	U.Lisbon - Informatics	Portugal

Table 4. CLEF 2006 ad hoc participants by country.

Country	# Participants
Brazil	4
Canada	1
France	2
Germany	1
Hungary	1
India	2
Indonesia	1
Ireland	1
Italy	1
Portugal	1
Spain	5
Sweden	1
Switzerland	1
United Kingdom	1
United States	2
Total	**25**

of the experiments (9 out 296, 3.04%) have queries been manually constructed from topics. A breakdown into the separate tasks is shown in Table 5(a).

Fourteen different topic languages were used in the ad hoc experiments. As always, the most popular language for queries was English, with French second. The number of runs per topic language is shown in Table 5(b).

3 Main Stream Monolingual Experiments

Monolingual retrieval was offered for Bulgarian, French, Hungarian, and Portuguese. As can be seen from Table 5(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 6 groups out of 16 (37.5%) submitted monolingual runs only (down from ten groups last year), and 5 of these groups were first time participants in CLEF. This year, most of the groups submitting monolingual runs were doing this as part of their bilingual or multilingual system testing activity. Details on the different approaches used can be found in the papers in this section of the working notes. There was a lot of detailed work with Portuguese language processing; not surprising as we had four new groups from Brazil in Ad Hoc this year. As usual, there was a lot of work on the development of stemmers and morphological analysers ([4], for instance, applies a very deep morphological analysis for Hungarian) and comparisons of the pros and cons of so-called "light" and "heavy" stemming approaches (e.g. [5]). In contrast to previous years, we note that a number of groups experimented with NLP techniques (see, for example, papers by [6], and [7]).

3.1 Results

Table 6 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; and the performance difference between the first and the last participant. Table 6 regards runs using title + description fields only (the mandatory run).

Figures from 1 to 4 compare the performances of the top participants of the Monolingual tasks.

4 Main Stream 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 target language restricted to newcomers in a CLEF cross-language task. This year, in this subtask, we focussed in particular on non-European topic languages and in particular languages for which there are still few processing tools or resources were in existence. We thus offered two Ethiopian languages: Amharic and Oromo; two Indian languages: Hindi and Telugu; and Indonesian. Although, as was to

Table 5. Breakdown of experiments into tracks and topic languages.

(a) Number of experiments per track, participant.

Track	# Part.	# Runs
Monolingual-BG	4	11
Monolingual-FR	8	27
Monolingual-HU	6	17
Monolingual-PT	12	37
Bilingual-X2BG	1	2
Bilingual-X2EN	5	33
Bilingual-X2FR	4	12
Bilingual-X2HU	1	2
Bilingual-X2PT	6	22
Robust-Mono-DE	3	7
Robust-Mono-EN	6	13
Robust-Mono-ES	5	11
Robust-Mono-FR	7	18
Robust-Mono-IT	5	11
Robust-Mono-NL	3	7
Robust-Bili-X2DE	2	5
Robust-Bili-X2ES	3	8
Robust-Bili-X2NL	1	4
Robust-Multi	4	10
Robust-Training-Mono-DE	2	3
Robust-Training-Mono-EN	4	7
Robust-Training-Mono-ES	3	5
Robust-Training-Mono-FR	5	10
Robust-Training-Mono-IT	3	5
Robust-Training-Mono-NL	2	3
Robust-Training-Bili-X2DE	1	1
Robust-Training-Bili-X2ES	1	2
Robust-Training-Multi	2	3
Total		**296**

(b) List of experiments by topic language.

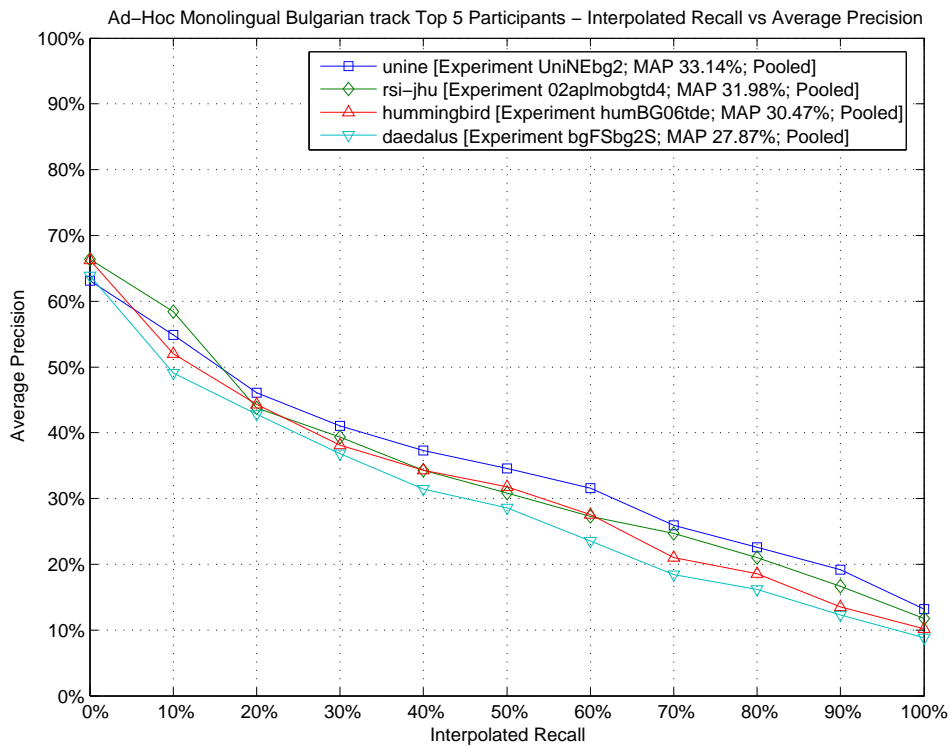| Topic Lang.	# Runs
English	65
French	60
Italian	38
Portuguese	37
Spanish	25
Hungarian	17
German	12
Bulgarian	11
Indonesian	10
Dutch	10
Amharic	4
Oromo	3
Hindi	2
Telugu	2
Total	**296**

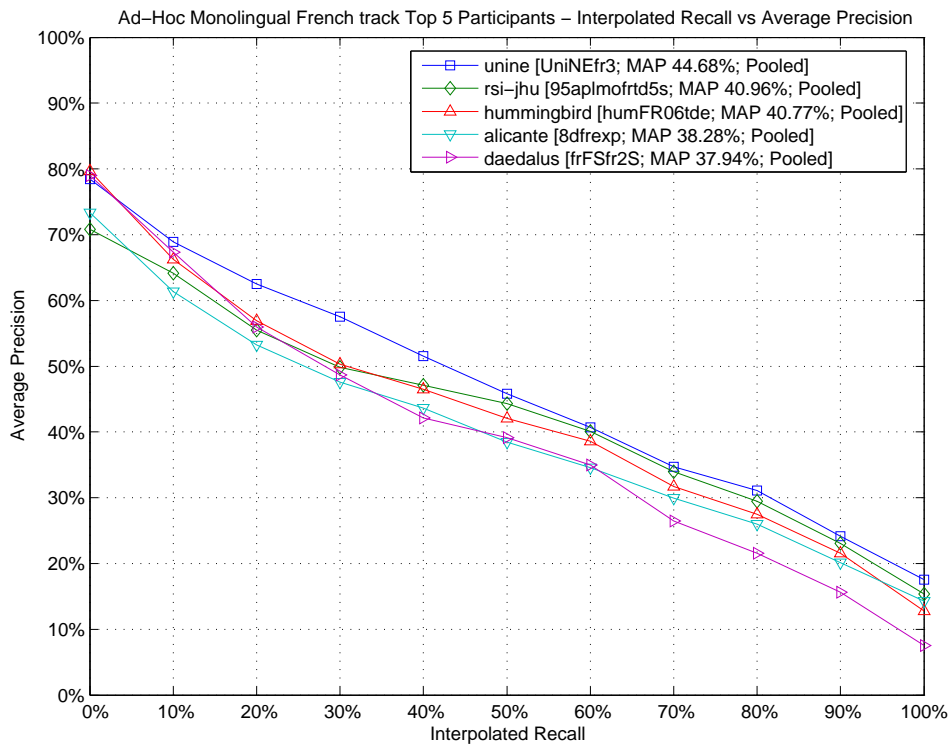**Fig. 1.** Monolingual Bulgarian

Fig. 2. Monolingual French

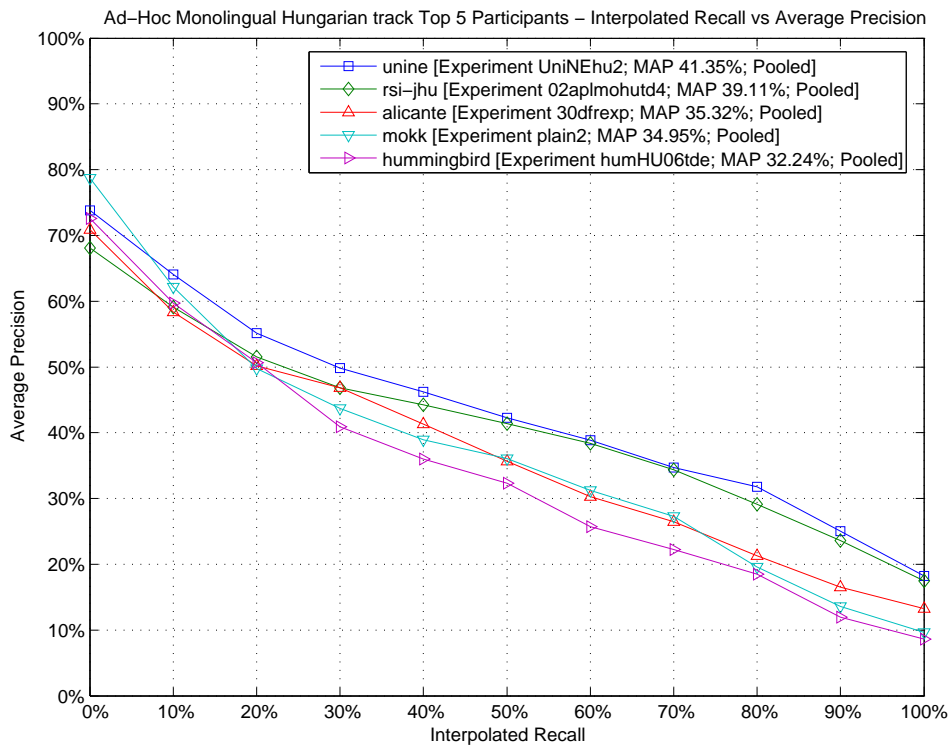**Fig. 3.** Monolingual Hungarian
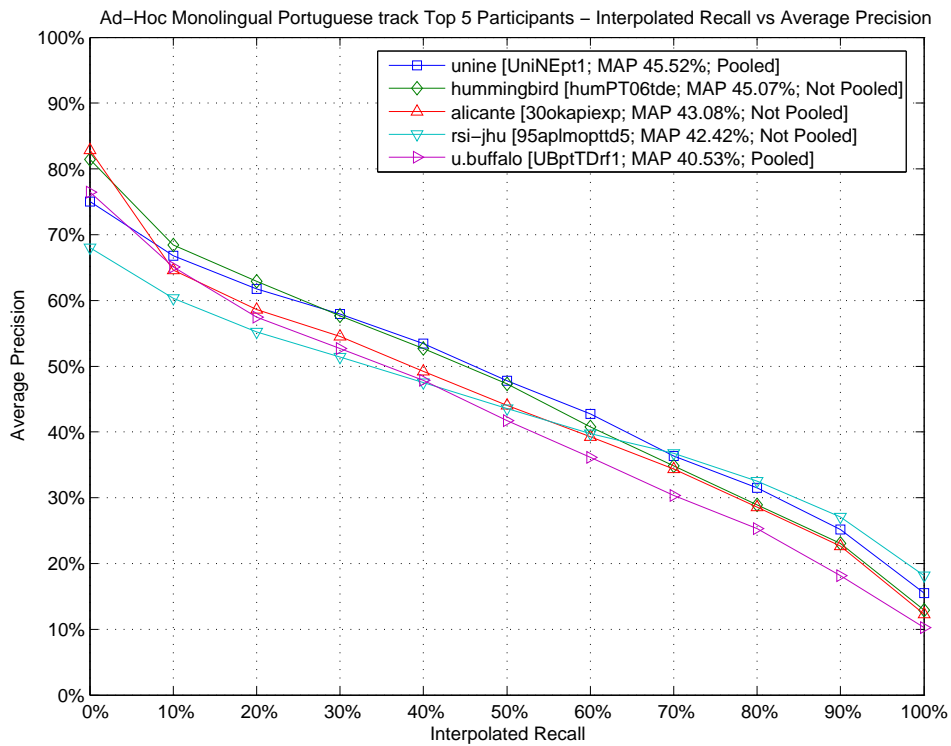
Fig. 4. Monolingual Portuguese

Table 6. Best entries for the monolingual track.

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Bulgarian	unine	rsi-jhu	hummingbird	daedalus		1st vs 4th
MAP	33.14%	31.98%	30.47%	27.87%		20.90%
Run	UniNEbg2	02aplmobgtd4	humBG06tde	bgFSbg2S		
French	unine	rsi-jhu	hummingbird	alicante	daedalus	1st vs 5th
MAP	44.68%	40.96%	40.77%	38.28%	37.94%	17.76%
Run	UniNEfr3	95aplmofrtd5s1	humFR06tde	8dfrexp	frFSfr2S	
Hungarian	unine	rsi-jhu	alicante	mokk	hummingbird	1st vs 5th
MAP	41.35%	39.11%	35.32%	34.95%	32.24%	28.26%
Run	UniNEhu2	02aplmohutd4	30dfrexp	plain2	humHU06tde	
Portuguese	unine	hummingbird	alicante	rsi-jhu	u.buffalo	1st vs 5th
MAP	45.52%	45.07%	43.08%	42.42%	40.53%	12.31%
Run	UniNEpt1	humPT06tde	30okapiexp	95aplmopttd5	UBptTDrf1	

be expected, the results are not particularly good, we feel that experiments of this type with lesser-studied languages are very important (see papers by [8], [9], [10])

4.1 Results

Table 7 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 group is given (in terms of average precision. Again both pooled and not 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 to evaluate performance is to compare results against monolingual baselines. For the best bilingual systems, we have the following results for CLEF 2006:

- X → BG: 52.49% of best monolingual Bulgarian IR system;
- X → FR: 93.82% of best monolingual French IR system;
- X → HU: 53.13% of best monolingual Hungarian IR system.
- X → PT: 90.91% of best monolingual Portuguese IR system;

We can compare these to those for CLEF 2005:

- X → BG: 85% of best monolingual Bulgarian IR system;
- X → FR: 85% of best monolingual French IR system;
- X → HU: 73% of best monolingual Hungarian IR system.
- X → PT: 88% of best monolingual Portuguese IR system;

While these results are very good for the well-established-in-CLEF languages, and can be read as state-of-the-art for this kind of retrieval system, at a first glance they appear very disappointing for Bulgarian and Hungarian. However,

Table 7. Best entries for the bilingual task.

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Bulgarian	daedalus					
MAP	17.39%					
Run	bgFSbgWen2S					
French	unine	queenmary	rsi-jhu	daedalus		1st vs 4th
MAP	41.92%	33.96%	33.60%	33.20%		26.27%
Run	UniNEBifr1	QMUL06e2f10b	aplbienfrd	frFSfrSen2S		
Hungarian	daedalus					
MAP	21.97%					
Run	huFShuMen2S					
Portuguese	unine	rsi-jhu	queenmary	u.buffalo	daedalus	1st vs 5th
MAP	41.38%	35.49%	35.26%	29.08%	26.50%	55.85%
Run	UniNEBipt2	aplbiesptd	QMUL06e2p10b	UBen2ptTDrf2	ptFSptSen2S	
English	rsi-jhu	depok	ltrc	celi	dsv	1st vs 5th
MAP	32.57%	26.71%	25.04%	23.97%	22.78%	42.98%
Run	aplbiinen5	UI_td_mt	OMTD	CELItitleNOEXPANSION	DsvAmhEngFullNofuzz	

we have to point out that, unfortunately, this year only one group submitted cross-language runs for Bulgarian and Hungarian and thus it does not make much sense to draw any conclusions from these, apparently poor, results for these languages. It is interesting to note that when *Cross Language Information Retrieval (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 from 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.

5 Robust Experiments

The robust task was organized for the first time at CLEF 2006. The evaluation of robustness emphasizes stable performance over all topics instead of high average performance [11]. The perspective of each individual user of an information retrieval system is different from the perspective taken by an evaluation initiative. The user will be disappointed by systems which deliver poor results for some topics whereas an evaluation initiative rewards systems which deliver good average results. A system delivering poor results for hard topics is likely to be considered of low quality by a user although it may reach high average results.
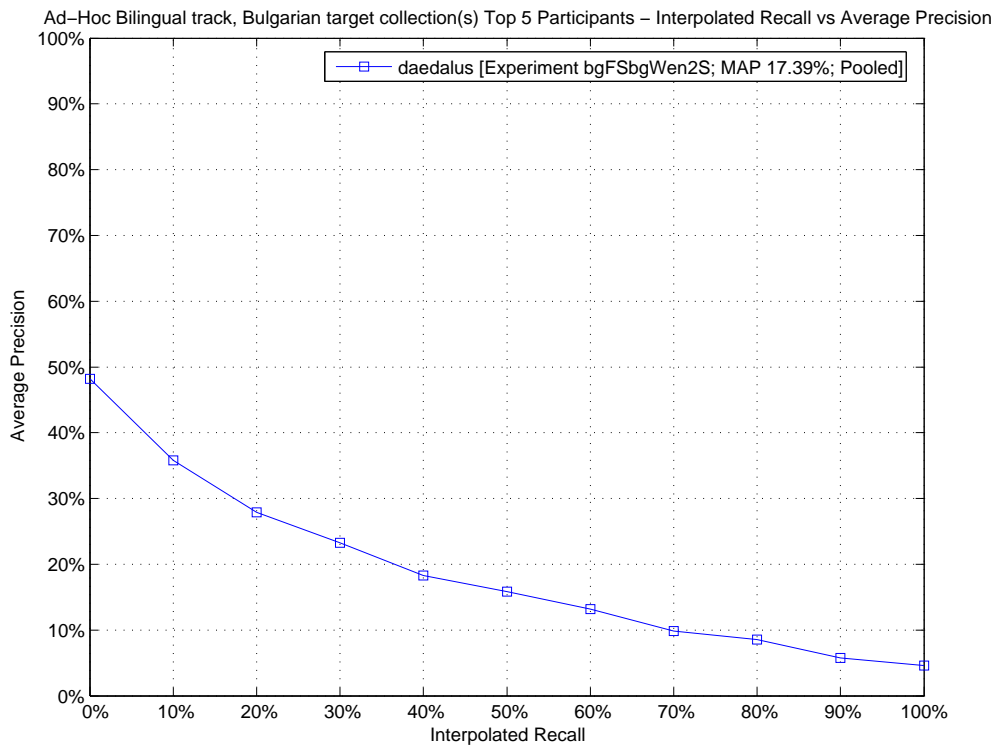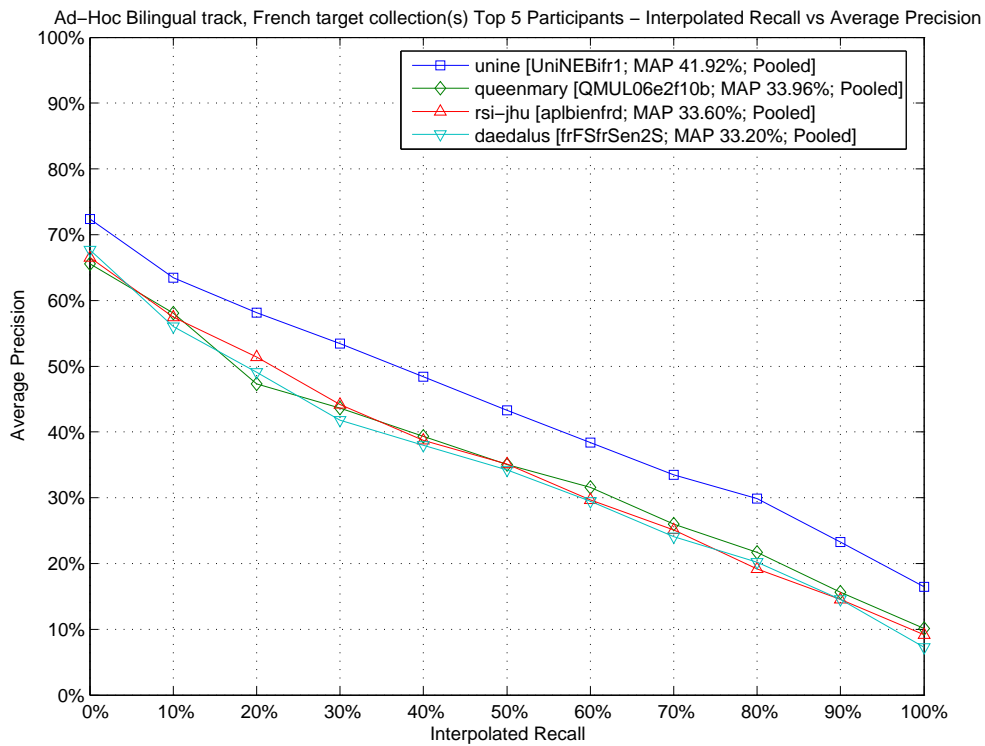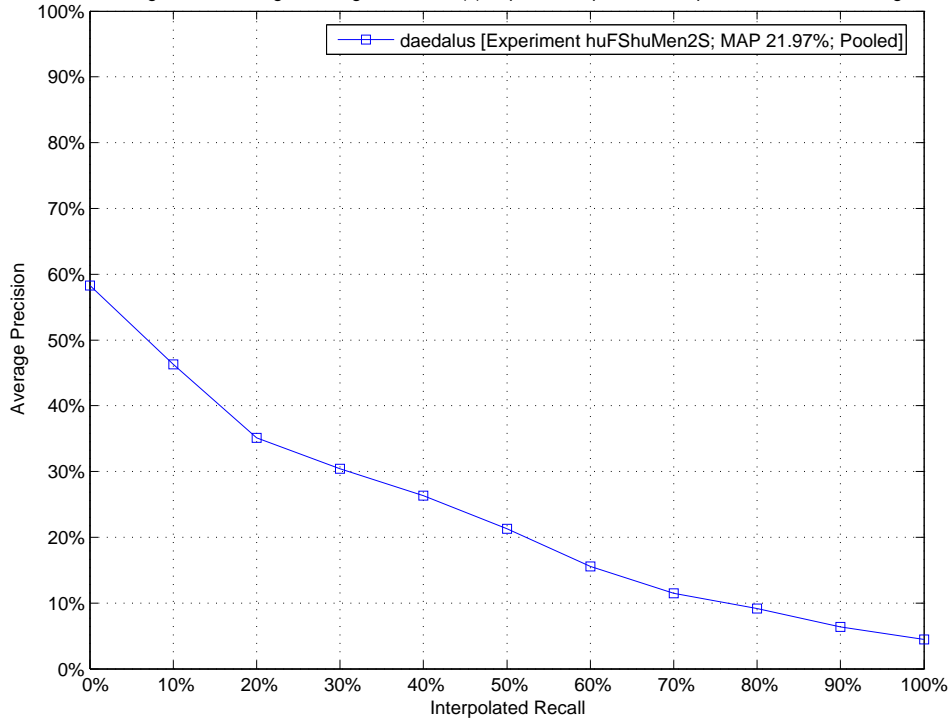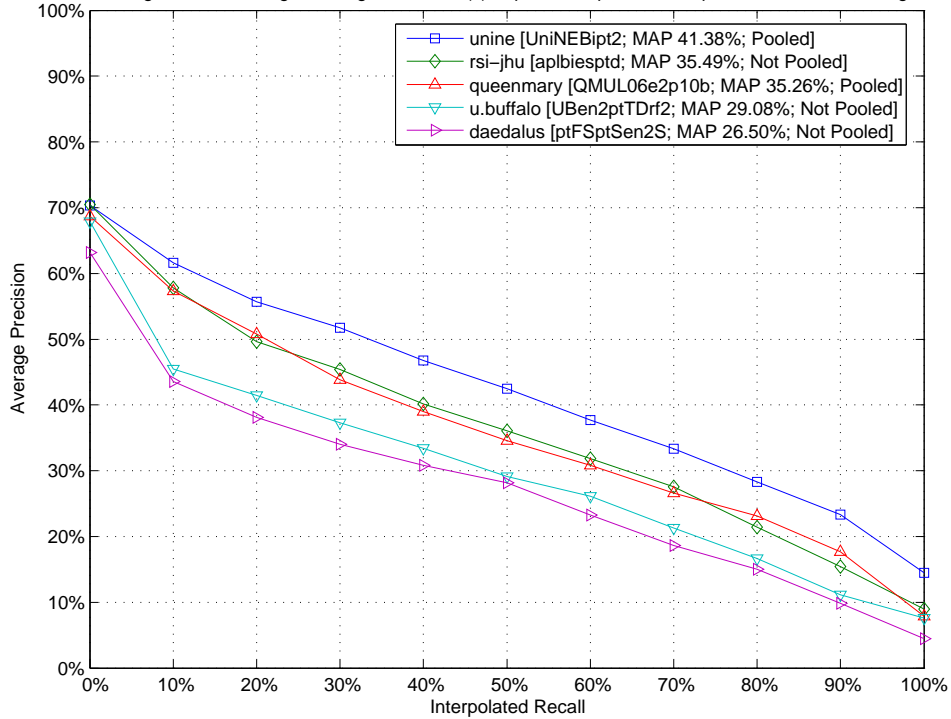
Fig. 5. Bilingual Bulgarian

Fig. 6. Bilingual French

Fig. 7. Bilingual Hungarian

Fig. 8. Bilingual Portuguese
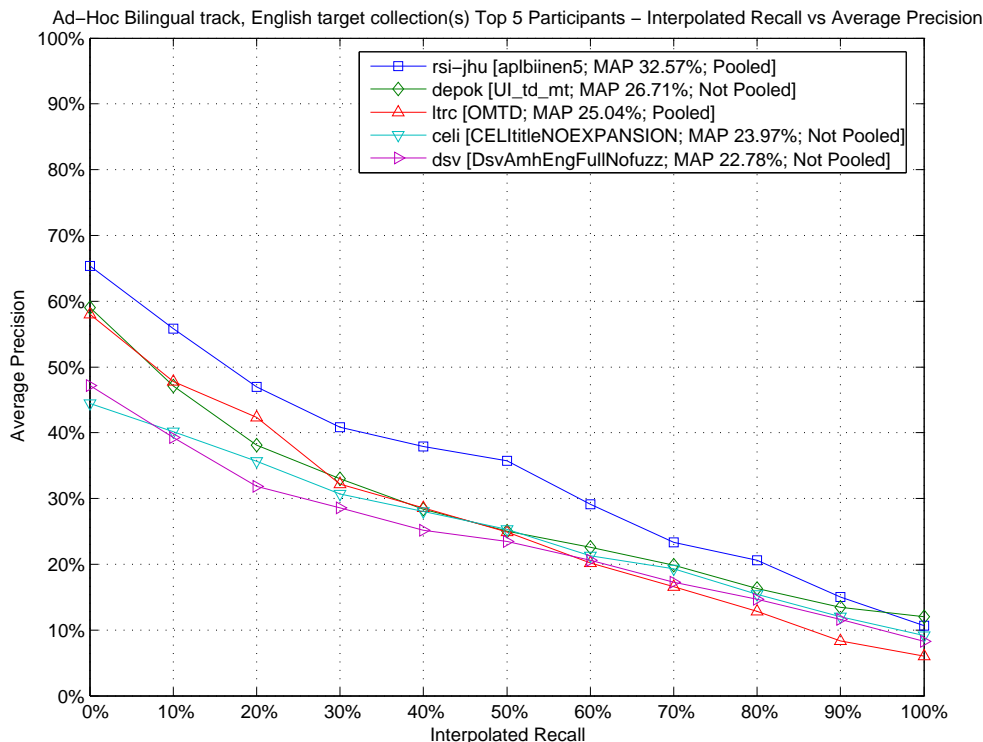
Fig. 9. Bilingual English

The robust task has been inspired by the robust track at TREC where it ran at TREC 2003, 2004 and 2005. A robust evaluation stresses performance for weak topics. This can be done by using the *Geometric Average Precision (GMAP)* as a main indicator for performance instead of the *Mean Average Precision (MAP)* of all topics. Geometric average has proven to be a stable measure for robustness at TREC [11]. The robust task at CLEF 2006 is concerned with the multilingual aspects of robustness. It is essentially an ad-hoc task which offers mono-lingual and cross-lingual sub tasks.

During CLEF 2001, CLEF 2002 and CLEF 2003 a set of 160 topics (Topics #41 - #200) was developed for these collections and relevance assessments were made. No additional relevance judgements were made this year for the robust task. However, the data collection was not completely constant over all three CLEF campaigns which led to an inconsistency between relevance judgements and documents. The SDA 95 collection has no relevance judgements for most topics (#41 - #140). This inconsistency was accepted in order to increase the size of the collection. One participant reported that exploiting the knowledge would have resulted in an increase of approximately 10% in MAP [12]. However, participants were not allowed to use this knowledge. The results of the original

submissions for the data sets were analyzed in order to identify the most difficult topics. This turned out to be an impossible task. The difficulty of a topic varies greatly among languages, target collections and tasks. This confirms the finding of the TREC 2005 robust task where the topic difficulty differed greatly even for two different English collections. It was found that topics are not inherently difficult but only in combination with a specific collection [13]. Topic difficulty is usually defined by low MAP values for a topic. We also considered a low number of relevant documents and high variation between systems as indicators for difficulty. Consequently, the topic set for the robust task at CLEF 2006 was arbitrarily split into two sets. Participants were allowed to use the available relevance assessments for the set of 60 training topics. The remaining 100 topics formed the test set for which results are reported. The participants were encouraged to submit results for training topics as well. These runs will be used to further analyze topic difficulty. The robust task received a total of 133 runs from eight groups listed in Table 5(a).

Most popular among the participants were the mono-lingual French and English tasks. For the multi-lingual task, four groups submitted ten runs. The bilingual tasks received fewer runs. A run using title and description was mandatory for each group. Participants were encouraged to run their systems with the same setup for all robust tasks in which they participated (except for language specific resources). This way, the robustness of a system across languages could be explored.

Effectiveness scores for the submissions were calculated with the GMAP which is calculated as the n-th root of a product of n values. GMAP was computed using the version 8.0 of `trec_eval`[2] program. In order to avoid undefined results, all precision score lower than 0.00001 are set to 0.00001.

5.1 Robust Monolingual Results

Table 8 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 group is given (in terms of average precision).

Figures from 10 to 15 compare the performances of the top participants of the Robust Monolingual.

5.2 Robust Bilingual Results

Table 9 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 group is given (in terms of average precision).

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2006:

- X → DE: 60.37% of best monolingual German IR system;

---

[2] `http://trec.nist.gov/trec_eval/trec_eval.8.0.tar.gz`

Table 8. Best entries for the robust monolingual task.

| Track	Participant Rank					Diff.
	1st	**2nd**	**3rd**	**4th**	**5th**	**Diff.**
Dutch	hummingbird	daedalus	colesir			1st vs 3rd
MAP	51.06%	42.39%	41.60%			22.74%
GMAP	25.76%	17.57%	16.40%			57.13%
Run	humNL06Rtde	nlFSnlR2S	CoLesIRnlTst			
English	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	47.63%	43.66%	43.48%	39.69%	37.64%	26.54%
GMAP	11.69%	10.53%	10.11%	8.93%	8.41%	39.00%
Run	humEN06Rtde	reinaENtdtest	dcudesceng12075	enFSenR2S	CoLesIRenTst	
French	unine	hummingbird	reina	dcu	colesir	1st vs 5th
MAP	47.57%	45.43%	44.58%	41.08%	39.51%	20.40%
GMAP	15.02%	14.90%	14.32%	12.00%	11.91%	26.11%
Run	UniNEfrr1	humFR06Rtde	reinaFRtdtest	dcudescfr12075	CoLesIRfrTst	
German	hummingbird	colesir	daedalus			1st vs 3rd
MAP	48.30%	37.21%	34.06%			41.81%
GMAP	22.53%	14.80%	10.61%			112.35%
Run	humDE06Rtde	CoLesIRdeTst	deFSdeR2S			
Italian	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	41.94%	38.45%	37.73%	35.11%	32.23%	30.13%
GMAP	11.47%	10.55%	9.19%	10.50%	8.23%	39.37%
Run	humIT06Rtde	reinaITtdtest	dcudescit1005	itFSitR2S	CoLesIRitTst	
Spanish	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	45.66%	44.01%	42.14%	40.40%	40.17%	13.67%
GMAP	23.61%	22.65%	21.32%	19.64%	18.84%	25.32%
Run	humES06Rtde	reinaEStdtest	dcudescsp12075	esFSesR2S	CoLesIResTst	
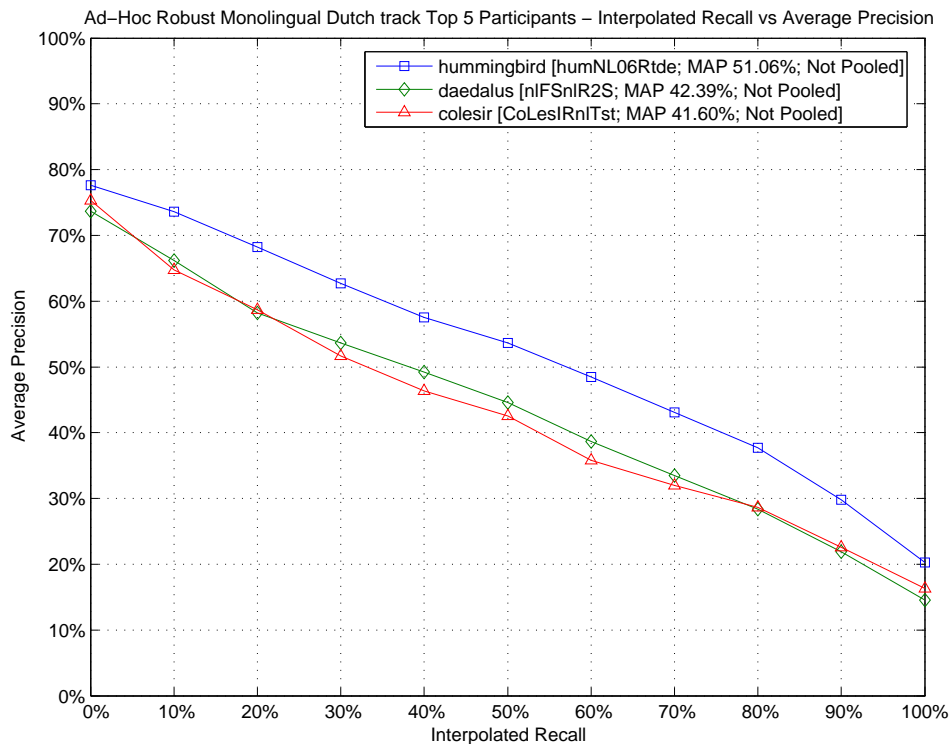
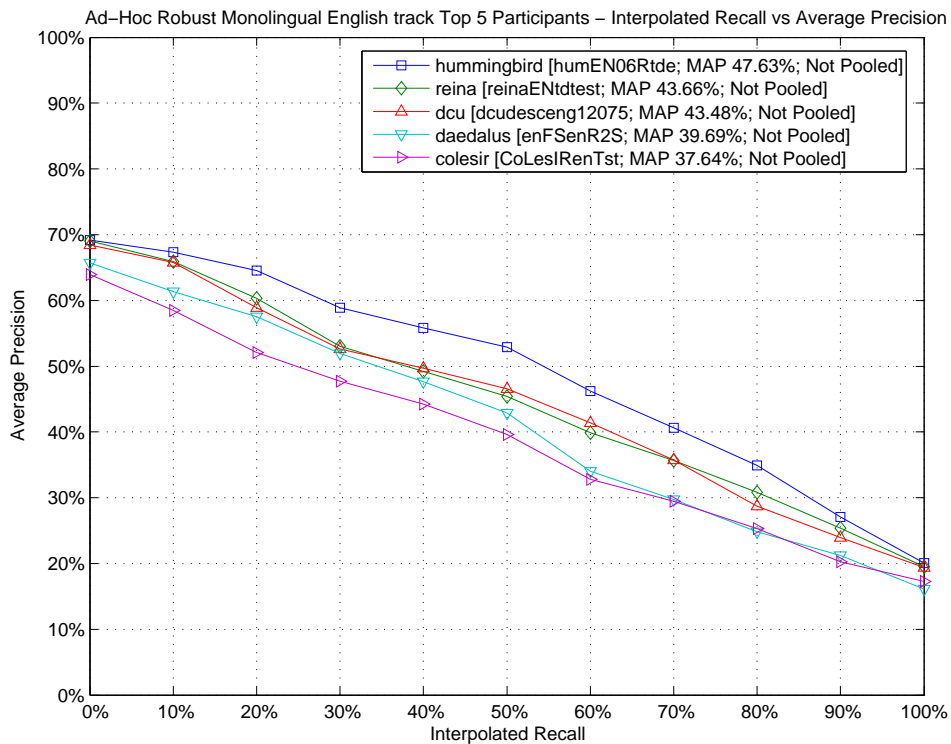**Fig. 10.** Robust Monolingual Dutch.
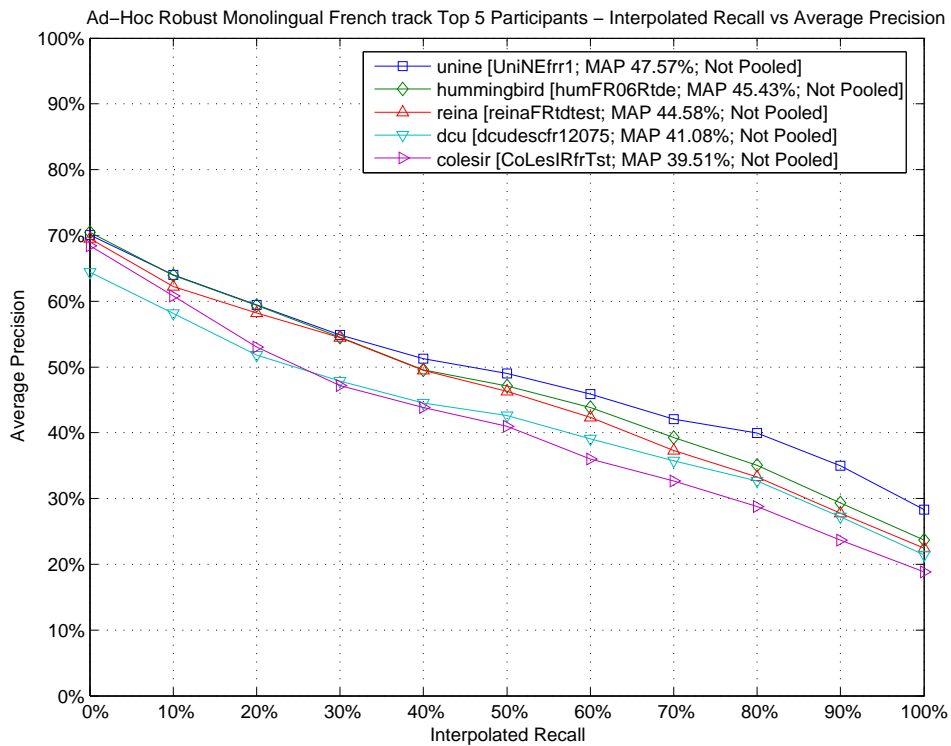

Fig. 11. Robust Monolingual English.

Fig. 12. Robust Monolingual French.

Fig. 13. Robust Monolingual German.

Ad–Hoc Robust Monolingual Italian track Top 5 Participants – Interpolated Recall vs Average Precision

Legend:
- hummingbird [humIT06Rtde; MAP 41.94%; Not Pooled]
- reina [reinaITtdtest; MAP 38.45%; Not Pooled]
- dcu [dcudescit1005; MAP 37.73%; Not Pooled]
- daedalus [itFSitR2S; MAP 35.11%; Not Pooled]
- colesir [CoLesIRitTst; MAP 32.23%; Not Pooled]

Fig. 14. Robust Monolingual Italian.

Ad–Hoc Robust Monolingual Spanish track Top 5 Participants – Interpolated Recall vs Average Precision

Legend:
- hummingbird [humES06Rtde; MAP 45.66%; Not Pooled]
- reina [reinaEStdtest; MAP 44.01%; Not Pooled]
- dcu [dcudescsp12075; MAP 42.14%; Not Pooled]
- daedalus [esFSesR2S; MAP 40.40%; Not Pooled]
- colesir [CoLesIRResTst; MAP 40.17%; Not Pooled]

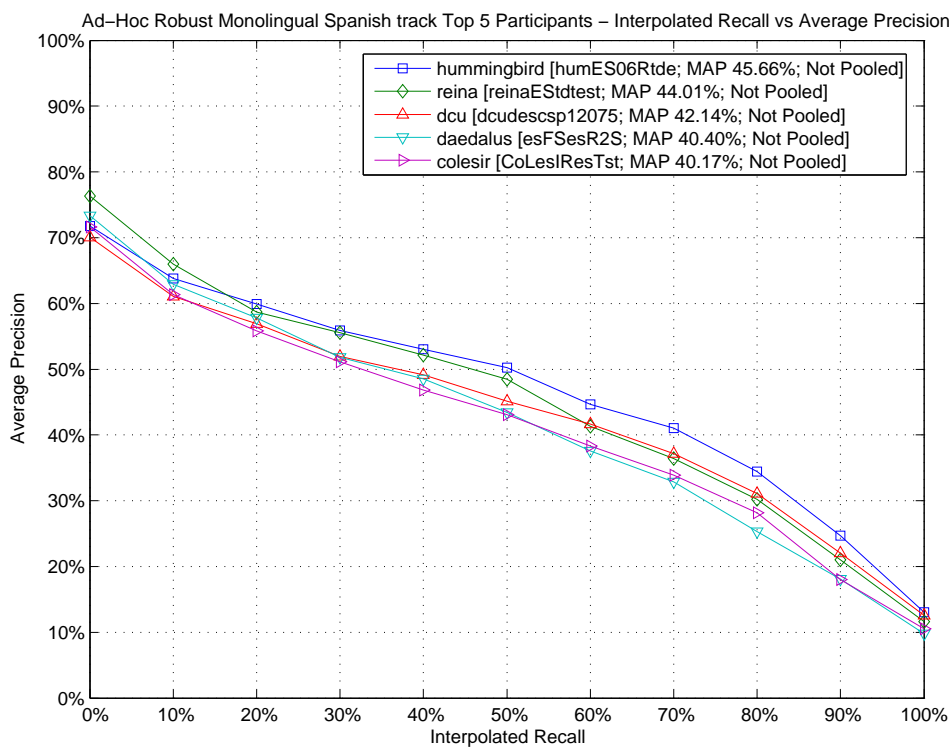**Fig. 15.** Robust Monolingual Spanish.

Table 9. Best entries for the robust bilingual task.

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Dutch	daedalus					
MAP	35.37%					
GMAP	9.75%					
Run	nlFSnlRLfr2S					
German	daedalus	colesir				1st vs 2nd
MAP	29.16%	25.24%				15.53%
GMAP	5.18%	4.31%				20.19%
Run	deFSdeRSen2S	CoLesIRendeTst				
Spanish	reina	dcu	daedalus			1st vs 3rd
MAP	36.93%	33.22%	26.89%			37.34%
GMAP	13.42%	10.44%	6.19%			116.80%
Run	reinaIT2EStdtest	dcuitqydescsp12075	esFSesRLit2S			

Table 10. Best entries for the robust multilingual task.

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Multilingual	jaen	daedalus	colesir	reina		1st vs 4th
MAP	27.85%	22.67%	22.63%	19.96%		39.53%
GMAP	15.69%	11.04%	11.24%	13.25%		18.42%
Run	ujamlrsv2	mlRSFSen2S	CoLesIRmultTst	reinaES2mtdtest		

- X → ES: 80.88% of best monolingual Spanish IR system;
- X → NL: 69.27% of best monolingual Dutch IR system.

Figures from 16 to 18 compare the performances of the top participants of the Robust Bilingual tasks.

### 5.3 Robust Multilingual Results

Table 10 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 group is given (in terms of average precision).

Figure 19 compares the performances of the top participants of the Robust Multilingual task.

### 5.4 Comments on Robust Cross Language Experiments

Some participants relied on the high correlation between the measure and optimized their systems as in previous campaigns. However, several groups worked specifically at optimizing for robustness. The SINAI system took an approach which has proved successful at the TREC robust task, expansion with terms gathered from a web search engine [14]. The REINA system from the University of Salamanca used a heuristic to determine hard topics during training.

Ad–Hoc Robust Bilingual track, Dutch target collection(s) Top 5 Participants – Interpolated Recall vs Average Precision
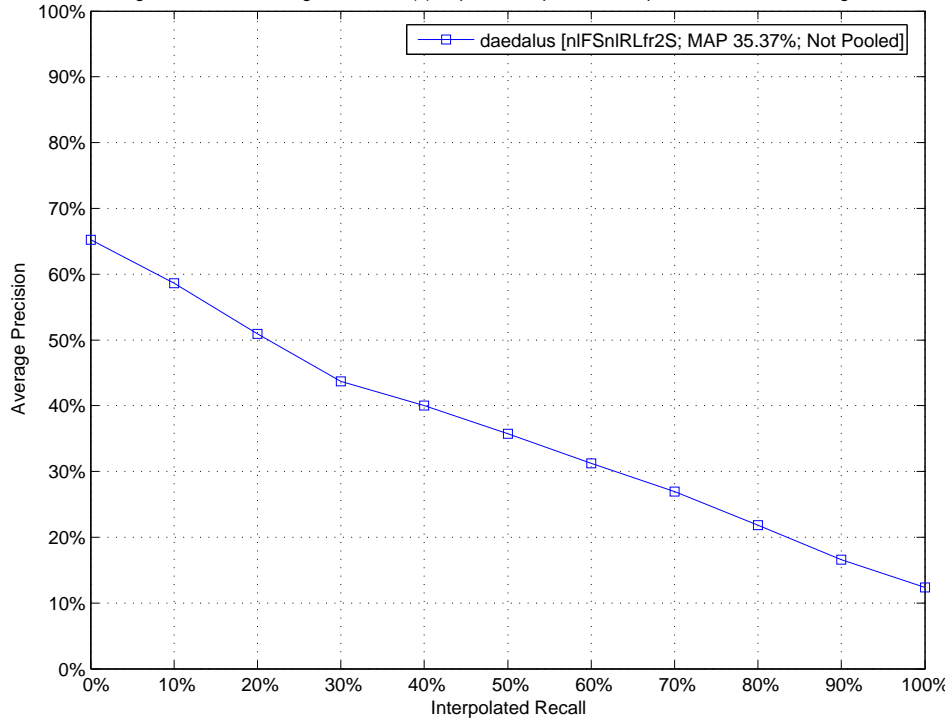
Fig. 16. Robust Bilingual Dutch

Ad–Hoc Robust Bilingual track, German target collection(s) Top 5 Participants – Interpolated Recall vs Average Precision
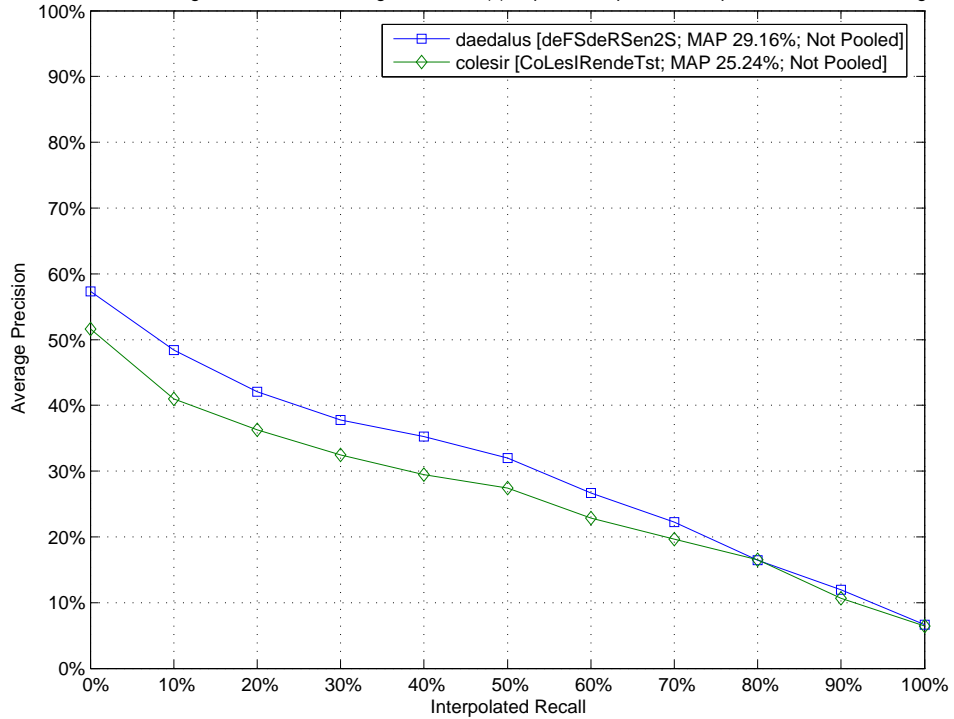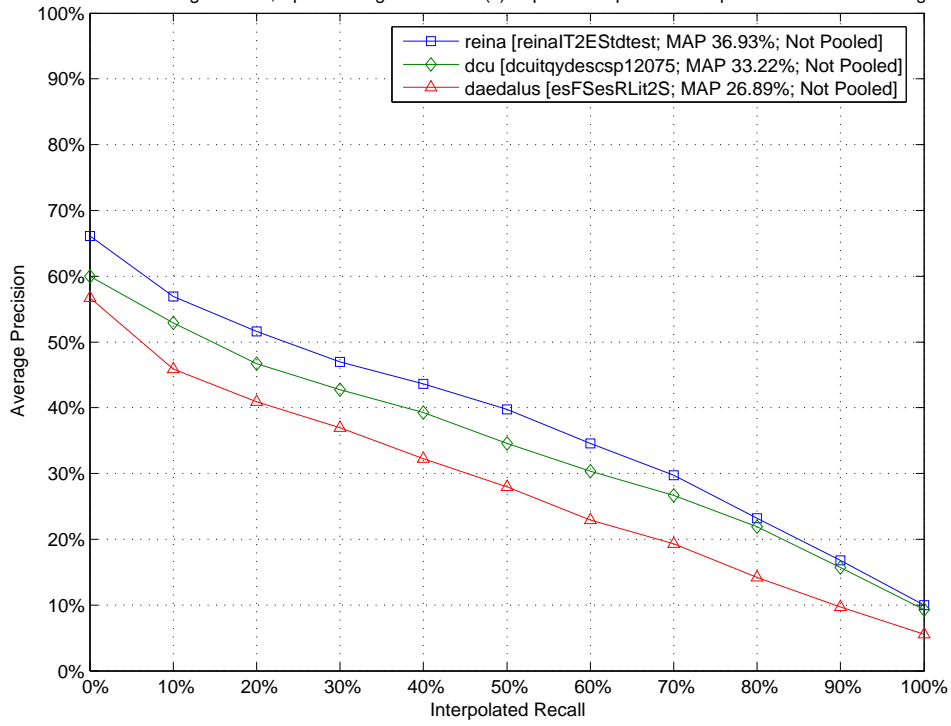

Fig. 17. Robust Bilingual German

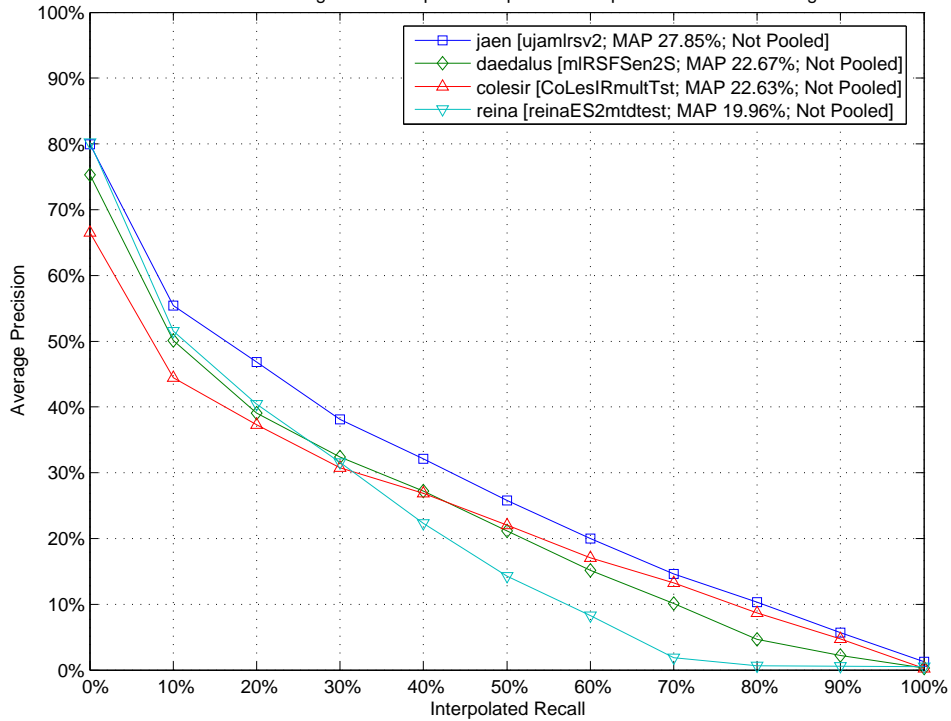**Fig. 18.** Robust Bilingual Spanish

Fig. 19. Robust Multilingual.

Subsequently, different expansion techniques were applied [15]. Hummingbird experimented with other evaluation measures than those used in the track [16]. The MIRACLE system tried to find a fusion scheme which had a positive effect on the robust measure [17].

6   Statistical Testing

When the goal is to validate how well results can be expected to hold beyond a 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 [18].

We used the MATLAB Statistics Toolbox, 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 [18] 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 [19] and the Jarque-Bera test [20]. 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).

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 [21] recommends for use with precision/recall measures.

Table 11 shows the results of both the Lilliefors and Jarque-Bera tests before and after applying the Tague-Sutcliffe transformation. After the transformation the analysis of the normality of samples distribution improves significantly, with some exceptions. 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].

In the following sections, two different graphs are presented to summarize the results of this test. 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. Both for the ad-hoc and robust tasks, only runs where significant differences exist are shown; the remainder of the graphs can be found in the Appendices [2].

Table 11. Lilliefors (LF) and Jarque-Bera (JB) test for each Ad-Hoc track with and without Tague-Sutcliffe (TS) arcsin transformation. Each entry is 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%.

Track	LF	LF & TS	JB	JB & TS
Monolingual Bulgarian	1	6	0	4
Monolingual French	12	25	26	26
Monolingual Hungarian	5	11	8	9
Monolingual Portuguese	13	34	35	37
Bilingual English	0	9	2	2
Bilingual Bulgarian	0	2	0	2
Bilingual French	8	12	12	12
Bilingual Hungarian	0	1	0	0
Bilingual Portuguese	4	12	15	19
Robust Monolingual German	0	5	0	7
Robust Monolingual English	3	9	4	11
Robust Monolingual Spanish	1	9	0	11
Robust Monolingual French	4	3	2	15
Robust Monolingual Italian	6	11	8	10
Robust Monolingual Dutch	0	7	0	7
Robust Bilingual German	0	0	0	4
Robust Bilingual Spanish	0	5	0	4
Robust Bilingual Dutch	0	3	0	4
Robust Multilingual	0	5	0	6

The first graph shows participants' runs (y axis) and performance obtained (x axis). The circle indicates the average performance (in terms of Precision) while the segment shows the interval in which the difference in performance is not statistically significant.

The second graph 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 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 runs that score better than the top entry of that group perform significantly better.

7   Conclusions

We have reported the results of the ad hoc cross-language textual document retrieval track at CLEF 2006. This track is considered to be central to CLEF as
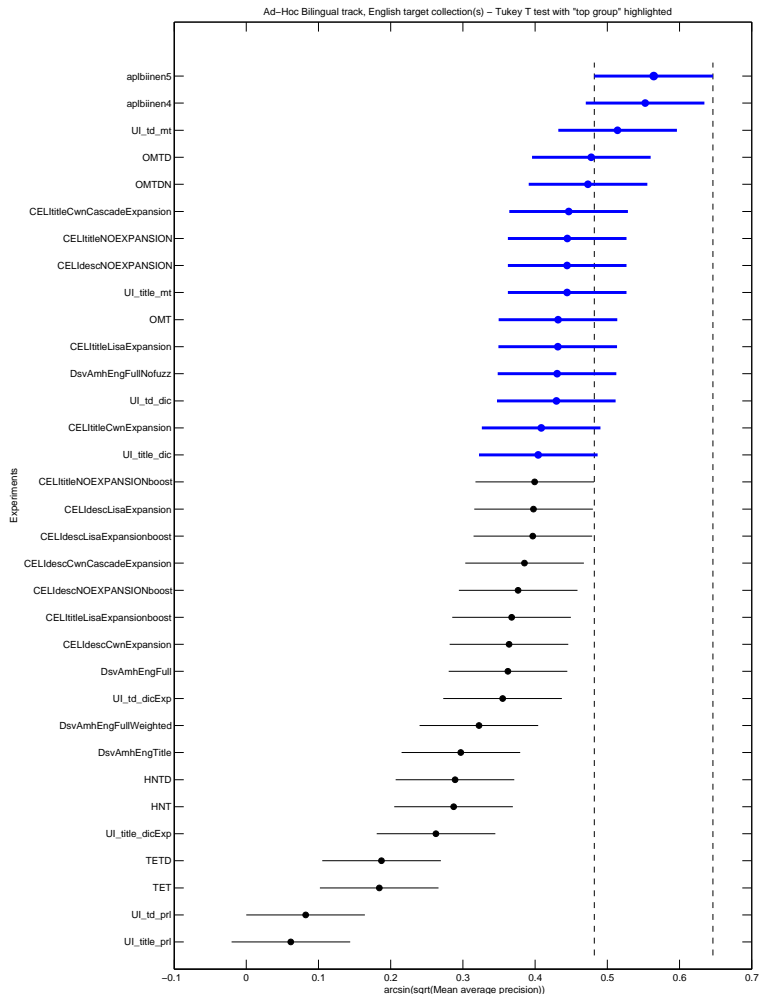
Ad–Hoc Monolingual French track – Tukey T test with "top group" highlighted
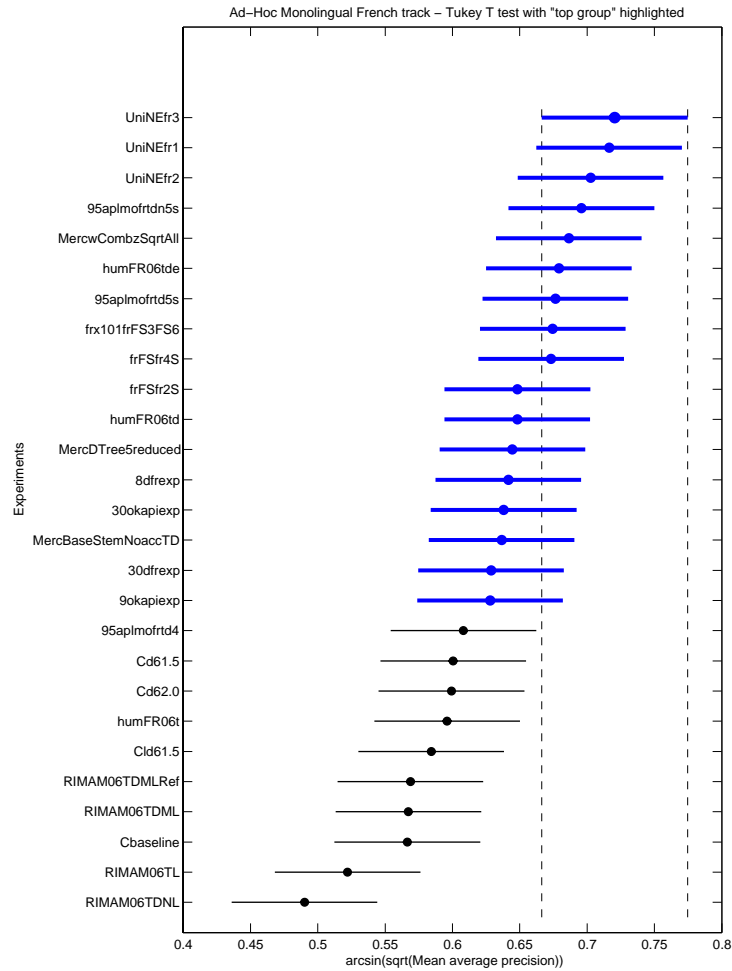
Run ID	Groups						
UniNEfr3	X						
UniNEfr1	X						
UniNEfr2	X	X					
95aplmofrtdn5s	X	X					
MercwCombzSqrtAll	X	X	X				
humFR06tde	X	X	X				
95aplmofrtd5s	X	X	X	X			
frx101frFS3FS6	X	X	X	X	X		
frFSfr4S	X	X	X	X	X		
frFSfr2S	X	X	X	X	X		
humFR06td	X	X	X	X	X		
MercDTree5reduced	X	X	X	X	X		
8dfrexp	X	X	X	X	X		
30okapiexp	X	X	X	X	X		
MercBaseStemNoaccTD	X	X	X	X	X		
30dfrexp	X	X	X	X	X	X	
9okapiexp	X	X	X	X	X	X	
95aplmofrtd4		X	X	X	X	X	
Cd61.5		X	X	X	X	X	
Cd62.0		X	X	X	X	X	
humFR06t		X	X	X	X	X	X
Cld61.5		X	X	X	X	X	X
RIMAM06TDMLRef			X	X	X	X	X
RIMAM06TDML				X	X	X	X
Cbaseline				X	X	X	X
RIMAM06TL						X	X
RIMAM06TDNL							X

Fig. 20. **Ad-Hoc Monolingual French**. Experiments grouped according to the Tukey T Test.

Ad–Hoc Monolingual Hungarian track – Tukey T test with "top group" highlighted
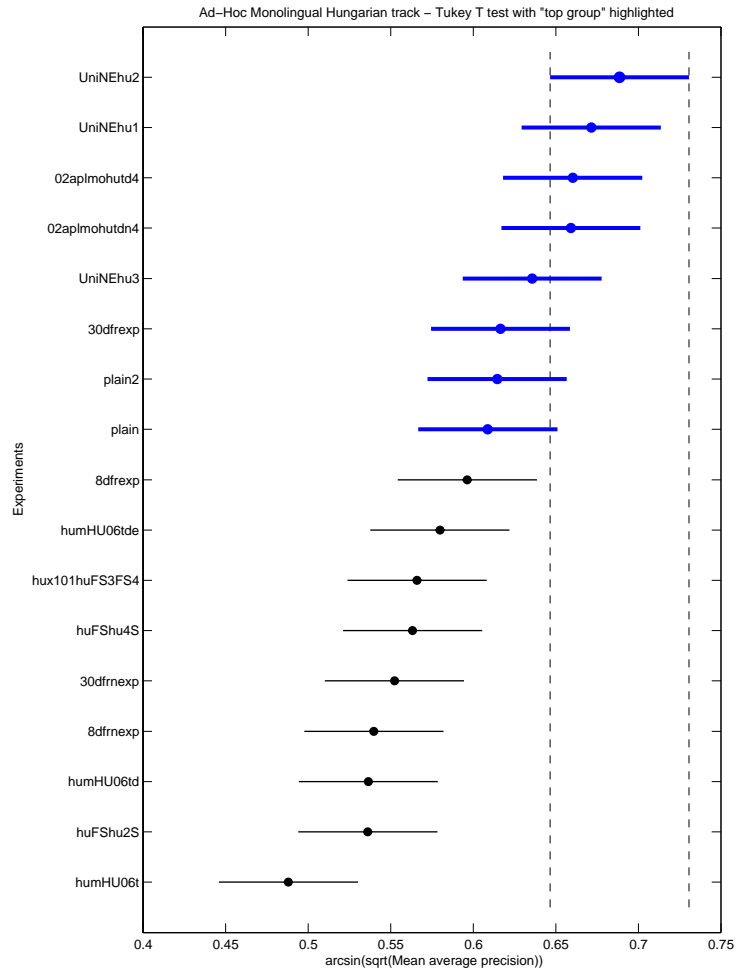
Run ID	Groups					
UniNEhu2	X					
UniNEhu1	X	X				
02aplmohutd4	X	X	X			
02aplmohutdn4	X	X	X			
UniNEhu3	X	X	X	X		
30dfrexp	X	X	X	X	X	
plain2	X	X	X	X	X	
plain	X	X	X	X	X	
8dfrexp		X	X	X	X	
humHU06tde			X	X	X	
hux101huFS3FS4				X	X	X
huFShu4S				X	X	X
30dfrnexp				X	X	X
8dfrnexp					X	X
humHU06td					X	X
huFShu2S					X	X
humHU06t						X

Fig. 21. **Ad-Hoc Monolingual Hungarian**. Experiments grouped according to the Tukey T Test.

Ad–Hoc Bilingual track, English target collection(s) – Tukey T test with "top group" highlighted

Run ID	Groups							
aplbiinen5	X							
aplbiinen4	X	X						
UI_td_mt	X	X	X					
OMTD	X	X	X	X				
OMTDN	X	X	X	X				
CELItitleCwnCascadeExpansion	X	X	X	X	X			
CELItitleNOEXPANSION	X	X	X	X	X			
CELIdescNOEXPANSION	X	X	X	X	X			
UI_title_mt	X	X	X	X	X			
OMT	X	X	X	X	X			
CELItitleLisaExpansion	X	X	X	X	X			
DsvAmhEngFullNofuzz	X	X	X	X	X			
UI_td_dic	X	X	X	X	X			
CELItitleCwnExpansion	X	X	X	X	X	X		
UI_title_dic	X	X	X	X	X	X		
CELItitleNOEXPANSIONboost		X	X	X	X	X		
CELIdescLisaExpansion		X	X	X	X	X		
CELIdescLisaExpansionboost		X	X	X	X	X		
CELIdescCwnCascadeExpansion		X	X	X	X	X		
CELIdescNOEXPANSIONboost		X	X	X	X	X		
CELItitleLisaExpansionboost		X	X	X	X	X		
CELIdescCwnExpansion		X	X	X	X	X		
DsvAmhEngFull		X	X	X	X	X		
UI_td_dicExp		X	X	X	X	X		
DsvAmhEngFullWeighted			X	X	X	X		
DsvAmhEngTitle			X	X	X	X		
HNTD			X	X	X	X		
HNT			X	X	X	X		
UI_title_dicExp						X	X	
TETD							X	X
TET							X	X
UI_td_prl								X
UI_title_prl								X

Fig. 22. Ad-Hoc Bilingual English. Experiments grouped according to the Tukey T Test.

Ad–Hoc Bilingual track, Portuguese target collection(s) – Tukey T test with "top group" highlighted
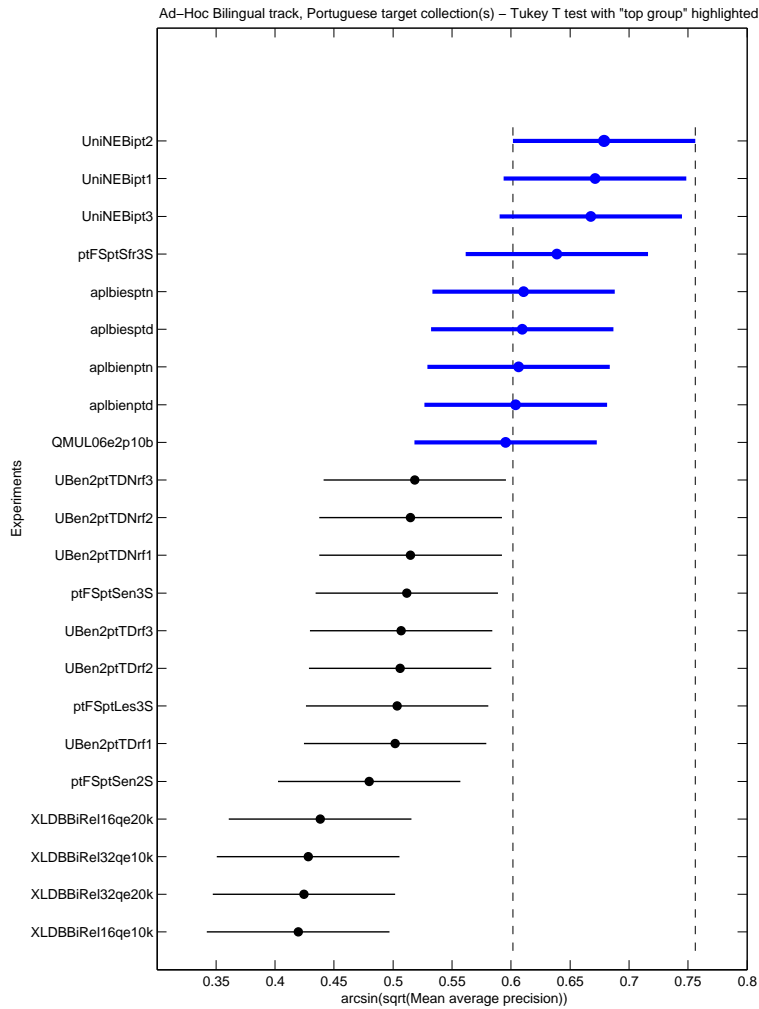
Run ID	Groups					
UniNEBipt2	X					
UniNEBipt1	X	X				
UniNEBipt3	X	X	X			
ptFSptSfr3S	X	X	X	X		
aplbiesptn	X	X	X	X	X	
aplbiesptd	X	X	X	X	X	
aplbienptn	X	X	X	X	X	
aplbienptd	X	X	X	X	X	
QMUL06e2p10b	X	X	X	X	X	
UBen2ptTDNrf3		X	X	X	X	X
UBen2ptTDNrf2			X	X	X	X
UBen2ptTDNrf1			X	X	X	X
ptFSptSen3S				X	X	X
UBen2ptTDrf3				X	X	X
UBen2ptTDrf2				X	X	X
ptFSptLes3S				X	X	X
UBen2ptTDrf1				X	X	X
ptFSptSen2S					X	X
XLDBBiRel16qe20k						X
XLDBBiRel32qe10k						X
XLDBBiRel32qe20k						X
XLDBBiRel16qe10k						X

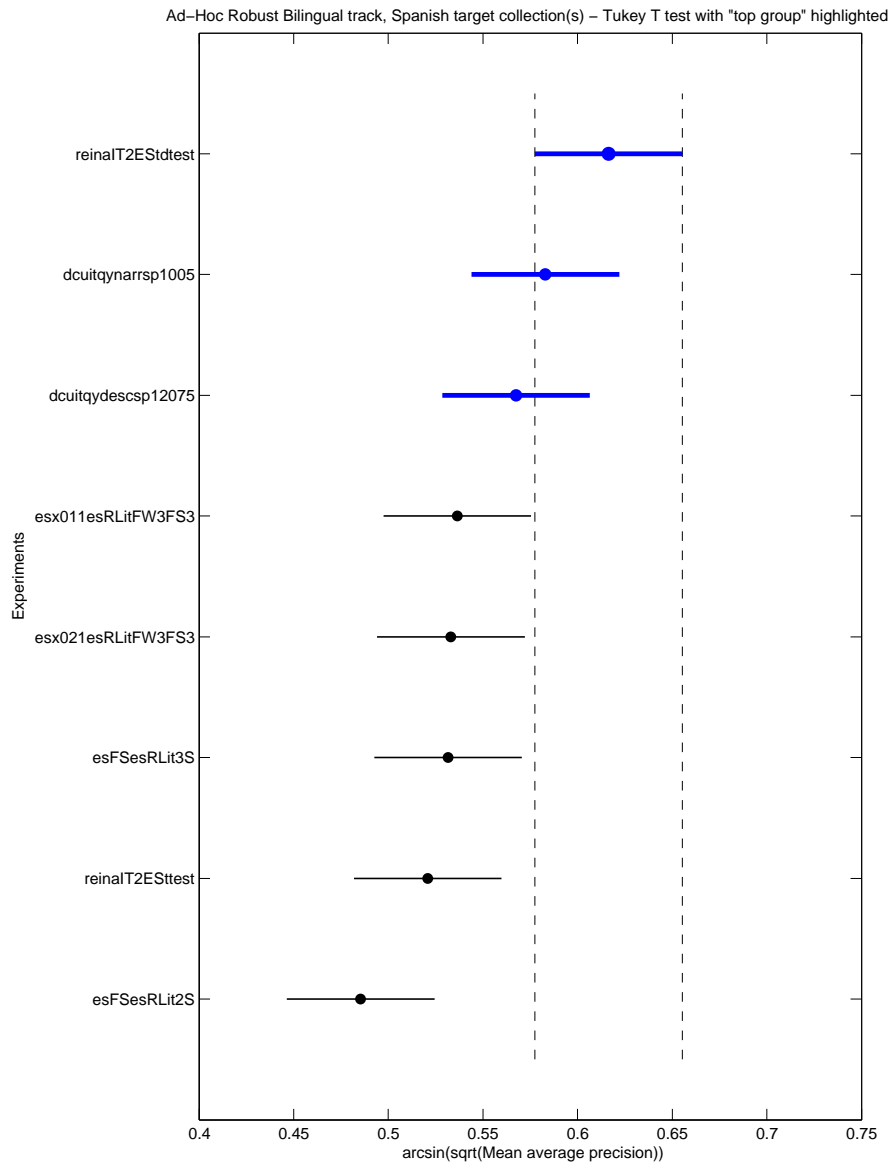**Fig. 23. Ad-Hoc Bilingual Portuguese**. Experiments grouped according to the Tukey T Test.
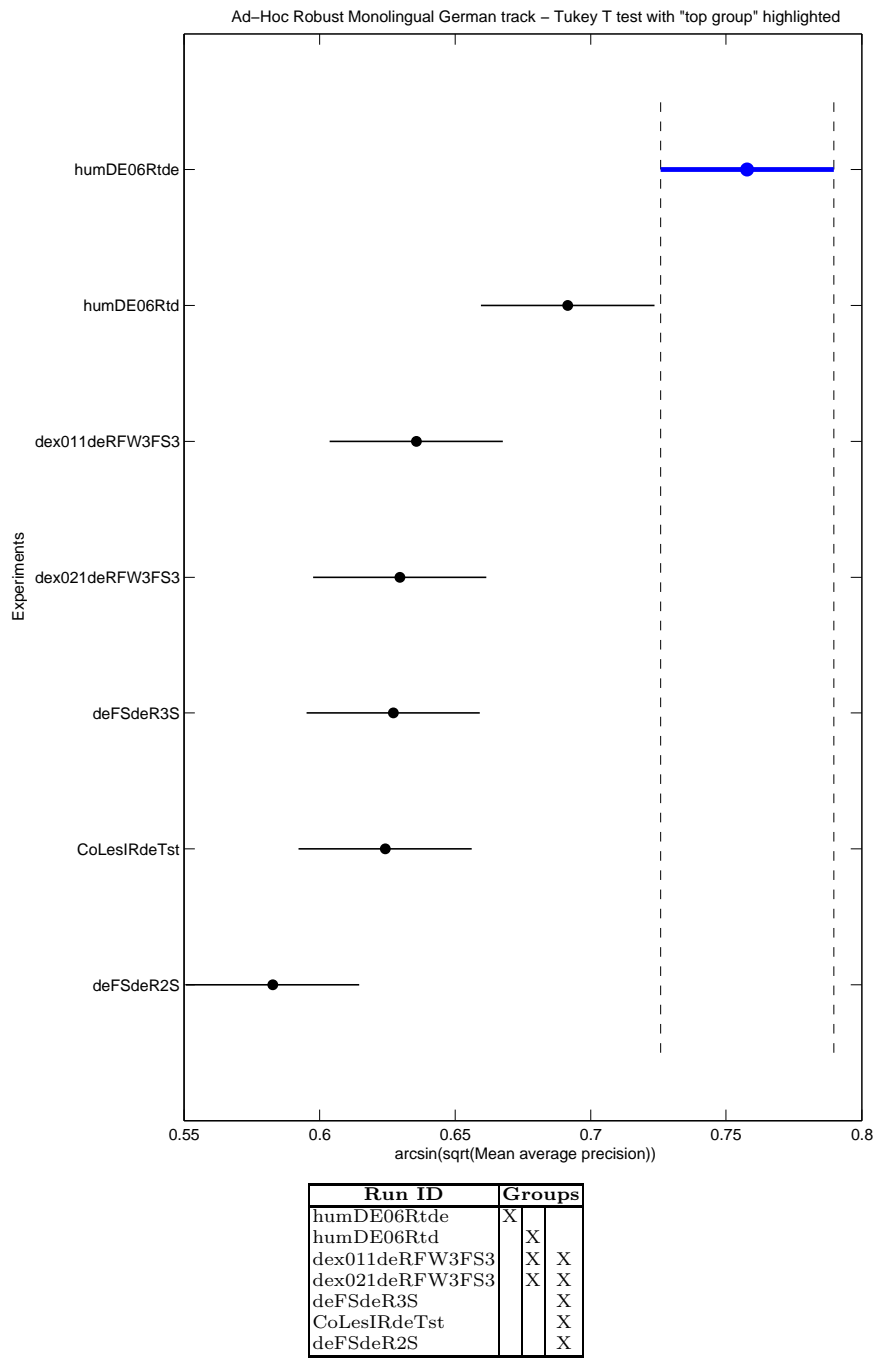
Fig. 24. **Robust Monolingual German**. Experiments grouped according to the Tukey T Test.
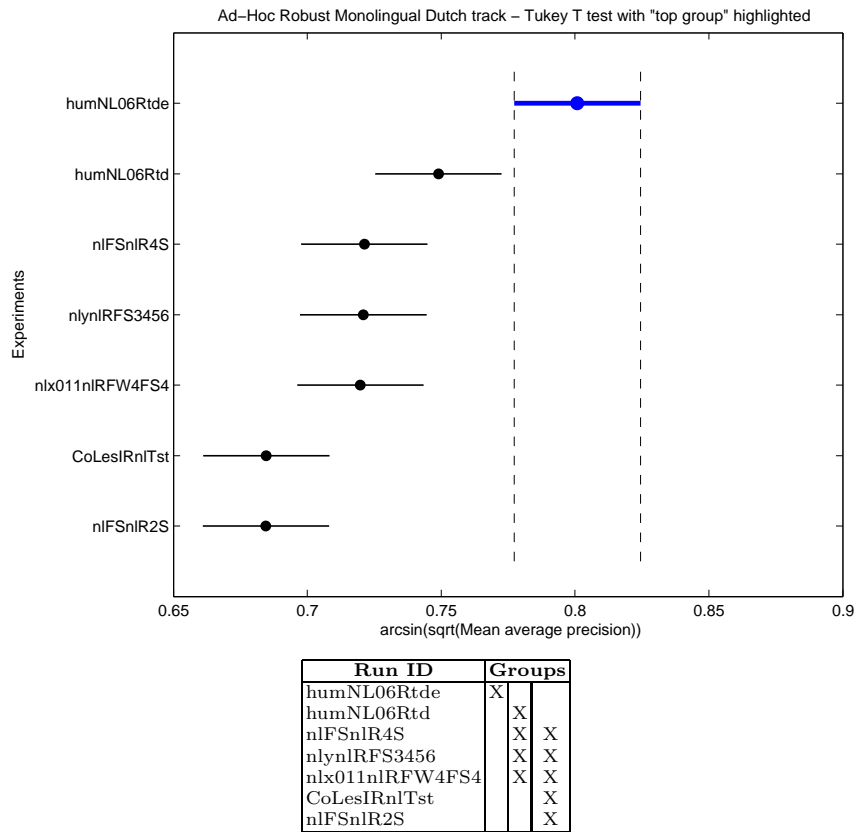
Fig. 25. Robust Monolingual Dutch. Experiments grouped according to the Tukey T Test.

Fig. 26. **Robust Bilingual Spanish**. Experiments grouped according to the Tukey T Test.
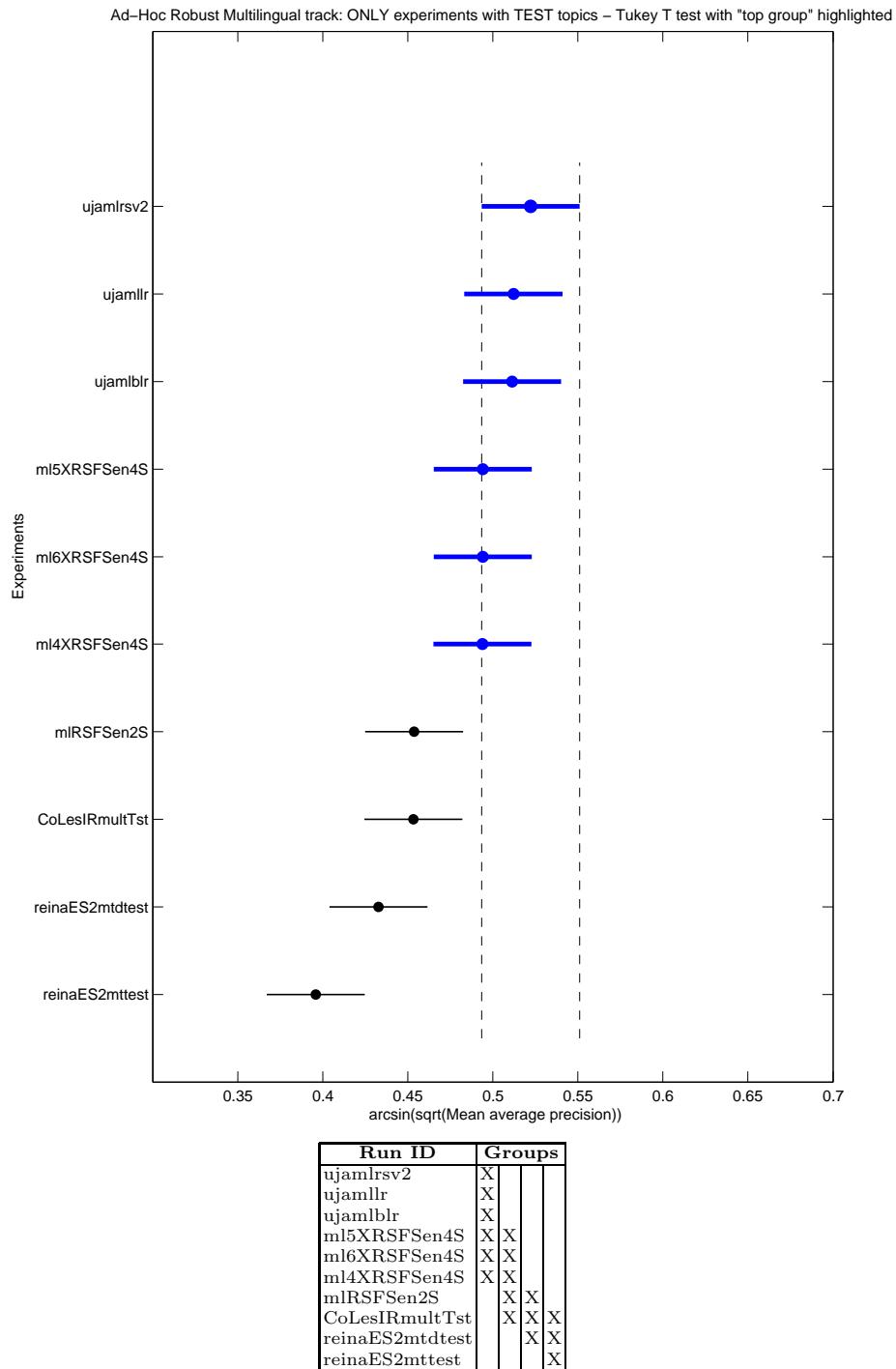
Fig. 27. Robust Multilingual. Experiments grouped according to the Tukey T Test.

for many groups it is the first track in which they participate and provides them with 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 examining the impact of "hard" topics on performance in the "robust" task.

Thus, although the ad hoc track in CLEF 2006 offered the same target languages for the main mono- and bilingual tasks as in 2005, it also had two new focuses. Groups were encouraged to use non-European languages as topic languages in the bilingual task. We were particularly interested in languages for which few processing tools were readily available, such as Amharic, Oromo and Telugu. In addition, we set up the "robust task" with the objective of providing the more expert groups with the chance to do in-depth failure analysis.

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. The test collections for CLEF 2000 - CLEF 2003 are about to be made publicly available on the *Evaluations and Language resources Distribution Agency (ELDA)* catalog[3].

References

1. 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
2. Di Nunzio, G.M., Ferro, N.: Appendix A. Results of the Core Tracks. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)
3. Braschler, M., Peters, C.: CLEF 2003 Methodology and Metrics. 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) 7–20
4. Halácsy, P.: Benefits of Deep NLP-based Lemmatization for Information Retrieval. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)
5. Moreira Orengo, V.: A Study on the use of Stemming for Monolingual Ad-Hoc Portuguese. Information Retrieval (2006)

---

[3] http://www.elda.org/

6. Azevedo Arcoverde, J.M., das Gracas Volpe Nunes, M., Scardua, W.: Using Noun Phrases for Local Analysis in Automatic Query Expansion. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

7. Gonzalez, M., de Lima, V.L.S.: The PUCRS-PLN Group participation at CLEF 2006. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

8. Tune, K.K., Varma, V.: Oromo-English Information Retrieval Experiments at CLEF 2006. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

9. Pingali, P., Varma, V.: Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

10. Hayurani, H., Sari, S., Adriani, M.: Evaluating Language Resources for English-Indonesian CLIR. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

11. Voorhees, E.M.: The TREC Robust Retrieval Track. SIGIR Forum **39** (2005) 11–20

12. Savoy, J., Abdou, S.: UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

13. Voorhees, E.M.: Overview of the TREC 2005 Robust Retrieval Track. In Voorhees, E.M., Buckland, L.P., eds.: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005), `http://trec.nist.gov/pubs/trec14/t14_proceedings.html` [last visited 2006, August 4] (2005)

14. Martinez-Santiago, F., Montejo-Ráez, A., Garcia-Cumbreras, M., Ureña-Lopez, A.: SINAI at CLEF 2006 Ad-hoc Robust Multilingual Track: Query Expansion using the Google Search Engine. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

15. Zazo, A., Figuerola, C., Berrocal, J.: REINA at CLEF 2006 Robust Task: Local Query Expansion Using Term Windows for Robust Retrieval. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

16. Tomlinson, S.: Comparing the Robustness of Expansion Techniques and Retrieval Measures. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

17. Goni-Menoyo, J., Gonzalez-Cristobal, J., Vilena-Román, J.: Report of the MIRA-CLE teach for the Ad-hoc track in CLEF 2006. In Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop, Published Online (2006)

18. 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

19. Conover, W.J.: Practical Nonparametric Statistics. 1st edn. John Wiley and Sons, New York, USA (1971)

20. 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)

21. 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