CLEF 2007: 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, Pisa, Italy
carol.peters@isti.cnr.it

Abstract. We describe the objectives and organization of the CLEF 2007 ad hoc track and discuss the main characteristics of the tasks offered to test monolingual and cross-language textual document retrieval systems. The track was divided into two streams. The main stream offered mono- and bilingual tasks on target collections for central European languages (Bulgarian, Czech and Hungarian). Similarly to last year, a bilingual task encouraging system testing with non-European languages against English documents was also offered; this year, particular attention was given to Indian languages. The second stream, designed for more experienced participants, offered mono- and bilingual "robust" tasks with the objective of privileging experiments which achieve good stable performance over all queries rather than high average performance. These experiments re-used CLEF test collections from previous years in three languages (English, French, and Portuguese). 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. Similarly to last year, the CLEF 2007 ad hoc track was structured in

two streams. The main stream offered mono- and bilingual retrieval tasks on target collections for central European languages plus a bilingual task encouraging system testing with non-European languages against English documents. 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, Czech and Hungarian target collections. Additionally, a bilingual task was offered to test querying with non-European language queries against an English target collection. As a result of requests from a number of Indian research institutes, a special sub-task for Indian languages was offered with topics in Bengali, Hindi, Marathi, Tamil and Telugu. 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 proposed mono- and bilingual experiments using the test collections built over the last six CLEF campaigns. Collections and topics in English, Portuguese and French were used. The goal of the robust analysis is to improve the user experience with a retrieval system. Poor performing topics are more serious for the user than performance losses in the middle and upper interval. The robust task gives preference to systems which achieve a minimal level for all topics. The measure used to assure this, is the geometric mean over all topics. The robust task intends to evaluate stable performance over all topics instead of high average performance.

This was the first year since CLEF began that we have not offered a **Multilingual** ad hoc task (ie searching a target collection in multiple languages).

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 these 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,

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

Language	Collections
Bulgarian	Sega 2002, Standart 2002, Novinar 2002
Czech	Mlada fronta DNES 2002, Lidové Noviny 2002
English	LA Times 2002
Hungarian	Magyar Hirlap 2002

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
Portuguese	Publico 94/95, Folha de Sao Paulo 94/95

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 Collections

Different test collections were used in the ad hoc task this year. The main stream used national newspaper documents from 2002 as the target collections, creating sets of new topics and making new relevance assessments. The robust task reused existing CLEF test collections and did not create any new topics or make any fresh relevance assessments.

Documents. The document collections used for the CLEF 2007 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, Bulgarian, Czech, Hungarian and English national newspapers for 2002 were used. Much of this data represented new additions to the CLEF multilingual comparable text corpora: Czech is a totally new language in the ad hoc track although it was introduced into the speech retrieval track last year; the Bulgarian collection was expanded with the addition of another national newspaper, and in order to have comparable data for English, we acquired a new American-English collection: Los Angeles Times 2002. Table 1 summarizes the collections used for each language.

The robust task used test collections containing news documents for the period 1994-1995 in three languages (English, French, and Portuguese) used in CLEF 2000 through CLEF 2006. Table 2 summarizes the collections used for each language.

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 2007 ad hoc mono- and bilingual tasks. All topic sets were created by native speakers. 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. In CLEF 2006 we deviated from this rule as we were using document collections from two distinct periods (1994/5 and 2002) and created partially separate (but overlapping) sets with a common set of time-independent topics and separate sets of time-specific topics. As we had expected this really complicated our lives as we had to build more topics and had to specify very carefully which topic sets were to be used against which document collections[1]. We determined not to repeat this experience this year and thus only used collections from the same time period.

We created topics in both European and non-European languages. European language topics were offered for Bulgarian, Czech, English, French, Hungarian, Italian and Spanish. The non-European languages were prepared according to demand from participants. This year we had Amharic, Chinese, Indonesian, Oromo plus the group of Indian languages: Bengali, Hindi, Marathi, Tamil and Telugu.

The provision of topics in unfamiliar scripts did lead to some problems. These were not caused by encoding issues (all CLEF data is encoded using UTF-8) but rather by errors in the topic sets which were very difficult for us to spot. Although most such problems were quickly noted and corrected, and the participants were informed so that they all used the right set, one did escape our notice: the title of Topic 430 in the Czech set was corrupted and systems using Czech thus did not do well with this topic. It should be remembered, however, that an error is one topic does not really impact significantly on the comparative results of the systems. The topic will, however, be corrected for future use.

This year topics have been identified by means of a Digital Object Identifier (DOI)[2] of the experiment [1] which allows us to reference and cite them. Below we give an example of the English version of a typical CLEF 2007 topic:

```
<top lang="en">
<num>10.2452/401-AH</num>
<title>Euro Inflation</title>
<desc>Find documents about rises in prices after the introduction of the
Euro.</desc>
```

[1] This is something that anyone reusing the CLEF 2006 ad hoc test collection needs to be very careful about.

[2] http://www.doi.org/

```
<narr>Any document is relevant that provides information on the rise of
prices in any country that introduced the common European
currency.</narr>
</top>
```

For the robust task, the topic sets from CLEF 2001 to 2006 in English, French and Portuguese were used. For English and French, which have been part of CLEF for more time, training topics were offered and a set of 100 topics were used for testing. For Portuguese, no training topics were possible and a set of 150 test topics was used.

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 is 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` 8.0) and available from the TREC website[3].

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). For bi- and monolingual 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 selected runs according to a set of predefined criteria. Traditionally, the top 100 ranked documents from each of the runs selected are included in the pool; in such a case we

[3] `http://trec.nist.gov/trec_eval/`

say that the pool is of depth 100. This pool is then used for subsequent relevance judgments. After calculating the effectiveness measures, the results are analyzed and run statistics produced and distributed.

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. New pools were formed in CLEF 2007 for the runs submitted for the main stream mono- and bilingual tasks. Instead, the robust tasks used the original pools and relevance assessments from previous CLEF campaigns.

The main criteria used when constructing these pools were:

- favour diversity among approaches adopted by participants, according to the descriptions of the experiments provided by the participants;
- choose at least one experiment for each participant in each task, chosen among the experiments with highest priority as indicated by the participant;
- add mandatory title+description experiments, even though they do not have high priority;
- add manual experiments, when provided;
- for bilingual tasks, ensure that each source topic language is represented.

One important limitation when forming the pools is the number of documents to be assessed. We estimate that assessors can judge from 60 to 100 documents per hour, providing binary judgments: relevant / not relevant. This is actually an optimistic estimate and shows what a time-consuming and resource expensive task human relevance assessment is. This limitation impacts strongly on the application of the criteria above - and implies that we are obliged to be flexible in the number of documents judged per selected run for individual pools.

This meant that this year, in order to create pools of more-or-less equivalent size (approx. 20,000 documents), the depth of the Bulgarian, Czech and Hungarian pools varied: 60 for Czech and 80 for Bulgarian and Hungarian, rather than the depth of 100 originally used to judge TREC ad hoc experiments[4]. In his paper in these working notes, Tomlinson [3] makes some interesting observations in this respect. He claims that on average, the percentage of relevant items assessed was less than 60% for Czech, 70% for Bulgarian and 85% for Hungarian. However, as Tomlinson also points out, it has already been shown that test collections created in this way do normally provide reliable results, even if not all relevant documents are included in the pool.

When building the pool for English, in order to respect the above criteria and also to obtain a pool depth of 60, we had to include more than 25,000 documents. Even so, as can be seen from Table 3, it was impossible to include very many runs - just one monolingual and one bilingual run for each set of experiments. We will certainly be performing some post-workshop stability tests on these pools.

The box plot of Figure 1 compares the distributions of the relevant documents across the topics of each pool for the different ad hoc pools; the boxes

[4] Tests made on NTCIR pools in previous years have suggested that a depth of 60 in normally adequate to create stable pools, presuming that a sufficient number of runs from different systems have been included

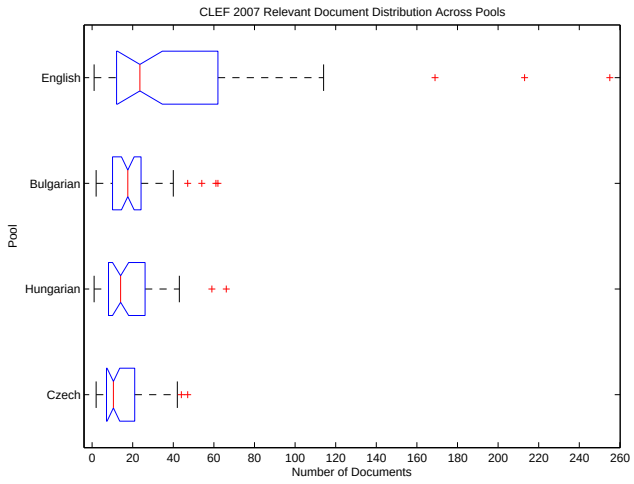


Fig. 1. Distribution of the relevant documents across the pools.

are ordered by decreasing mean number of relevant documents per topic. As can be noted, Bulgarian, Czech, and Hungarian distributions appear similar, even though the Czech and Hungarian ones are slightly more asymmetric towards topics with a greater number of relevant documents. On the other hand, the English distribution presents a greater number of relevant documents per topic, with respect to the other distributions, and is quite asymmetric towards topics with a greater number of relevant documents. All the distributions show some upper outliers, i.e. topics with a great number of relevant document with respect to the behaviour of the other topics in the distribution. These outliers are probably due to the fact that CLEF topics have to be able to retrieve relevant documents in all the collections; therefore, they may be considerably broader in one collection compared with others depending on the contents of the separate datasets. Thus, typically, each pool will have a different set of outliers.

Table 3 reports summary information on the 2007 ad hoc pools used to calculate the results for the main monolingual and bilingual experiments. In particular, for each pool, we show the number of topics, the number of runs submitted, the number of runs included in the pool, the number of documents in the pool (relevant and non-relevant), and the number of assessors.

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 participants and the relevance assessments. Popular measures usually adopted for exercises of this type are Recall and Precision. Details on how they are

Table 3. Summary information about CLEF 2007 pools.

Bulgarian Pool (**DOI** 10.2454/AH-BULGARIAN-CLEF2007)	
Pool size	19,441 pooled documents – 18,429 not relevant documents – 1,012 relevant documents 50 topics
Pooled Experiments	13 out of 18 submitted experiments – monolingual: 11 out of 16 submitted experiments – bilingual: 2 out of 2 submitted experiments
Assessors	4 assessors
Czech Pool (**DOI** 10.2454/AH-CZECH-CLEF2007)	
Pool size	20,607 pooled documents – 19,485 not relevant documents – 762 relevant documents 50 topics
Pooled Experiments	19 out of 29 submitted experiments – monolingual: 17 out of 27 submitted experiments – bilingual: 2 out of 2 submitted experiments
Assessors	4 assessors
English Pool (**DOI** 10.2454/AH-ENGLISH-CLEF2007)	
Pool size	24,855 pooled documents – 22,608 not relevant documents – 2,247 relevant documents 50 topics
Pooled Experiments	20 out of 104 submitted experiments – monolingual: 10 out of 31 submitted experiments – bilingual: 10 out of 73 submitted experiments
Assessors	5 assessors
Hungarian Pool (**DOI** 10.2454/AH-HUNGARIAN-CLEF2007)	
Pool size	18,704 pooled documents – 17,793 not relevant documents – 911 relevant documents 50 topics
Pooled Experiments	14 out of 21 submitted experiments – monolingual: 12 out of 19 submitted experiments – bilingual: 2 out of 2 submitted experiments
Assessors	6 assessors

calculated for CLEF are given in [4]. For the robust task, we used different measures, see below Section 5.

The individual results for all official ad hoc experiments in CLEF 2007 are given in the Appendix at the end of these Working Notes [5,6].

2.5 Participants and Experiments

As shown in Table 4, a total of 22 groups from 12 different countries submitted results for one or more of the ad hoc tasks - a slight decrease on the 25 participants of last year. Table 5 provides a breakdown of the number of participants by country.

A total of 235 runs were submitted with a decrease of about 20% on the 296 runs of 2006. The average number of submitted runs per participant also slightly decreased: from 11.7 runs/participant of 2006 to 10.6 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 (138 out of 235, 58.72%) used this combination of topic fields, 50 (21.28%) used all fields, 46 (19.57%) used the title field, and only 1 (0.43%) used the description field. The majority of experiments were conducted using automatic query construction (230 out of 235, 97.87%) and only in a small fraction of the experiments (5 out 237, 2.13%) were queries been manually constructed from topics. A breakdown into the separate tasks is shown in Table 6(a).

Fourteen different topic languages were used in the ad hoc experiments. As always, the most popular language for queries was English, with Hungarian second. The number of runs per topic language is shown in Table 6(b).

3 Main Stream Monolingual Experiments

Monolingual retrieval focused on central-European languages this year, with tasks offered for Bulgarian, Czech and Hungarian. Eight groups presented results for 1 or more of these languages. We also requested participants in the bilingual-to-English task to submit one English monolingual run, but only in order to provide a baseline for their bilingual experiments and in order to strengthen the English pool for relevance assessment[5].

Five of the participating groups submitted runs for all three languages. One group was unable to complete its Bulgarian experiments, submitting results for just the other two languages. The two groups from the Czech Republic only submitted runs for Czech. From the graphs and from 7, it can be seen that the best performing groups were more-or-less the same for each language and that the results did not greatly differ. It should be noted that these are all veteran participants with much experience at CLEF.

[5] Ten groups submitted runs for monolingual English. We have included a graph showing the top 5 results but it must be remembered that the systems submitting these were actually focusing on the bilingual part of the task.

Table 4. CLEF 2007 ad hoc participants

Participant	Institution	Country
alicante	U.Alicante - Languages&CS	Spain
bohemia	U.W.Bohemia	Czech Republic
bombay-ltrc	Indian Inst. Tech.	India
budapest-acad	Informatics Lab	Hungary
colesun	COLESIR & U.Sunderland	Spain
daedalus	Daedalus & Spanish Univ. Consortium	Spain
depok	U.Indonesia	Indonesia
hildesheim	U.Hildesheim	Germany
hyderabad	International Institute of Information Technology (IIIT)	India
isi	Indian Statistical Institute	India
jadavpur	Jadavpur University	India
jaen	U.Jaen-Intell.Systems	Spain
jhu-apl	Johns Hopkins University Applied Physics Lab	United States
kharagpur	IIT-Kharagpur-CS	India
msindia	Microsoft India	India
nottingham	U.Nottingham	United Kingdom
opentext	Open Text Corporation	Canada
prague	Charles U., Prague	Czech Republic
reina	U.Salamanca	Spain
stockholm	U. Stockholm	Sweden
unine	U.Neuchatel-Informatics	Switzerland
xldb	U.Lisbon	Portugal

Table 5. CLEF 2007 ad hoc participants by country.

Country	# Participants
Canada	1
Cezch Republic	2
Germany	1
Hungary	1
India	6
Indonesia	1
Portugal	1
Spain	5
Sweden	1
Switzerland	1
United Kingdom	1
United States	1
Total	**22**

Table 6. Breakdown of experiments into tracks and topic languages.

(a) Number of experiments per track, participant.

Track	# Part.	# Runs
Monolingual-BG	5	16
Monolingual-CS	8	27
Monolingual-EN	10	31
Monolingual-HU	6	19
Bilingual-X2BG	1	2
Bilingual-X2CS	1	2
Bilingual-X2EN	10	73
Bilingual-X2HU	1	2
Robust-Mono-EN	3	11
Robust-Mono-FR	5	12
Robust-Mono-PT	4	11
Robust-Bili-X2FR	3	9
Robust-Training-Mono-EN	2	6
Robust-Training-Mono-FR	2	6
Robust-Training-Bili-X2FR	2	8
Total		**235**

(b) List of experiments by topic language.

Topic Lang.	# Runs
English	73
Hungarian	33
Czech	26
Bulgarian	16
Indonesian	16
French	14
Hindi	13
Chinese	12
Portuguese	11
Amharic	9
Bengali	4
Oromo	4
Marathi	2
Telugu	2
Total	**235**

As usual in the CLEF monolingual task, the main emphasis in the experiments was on stemming and morphological analysis. The group from University of Neuchatel, which had the best overall performances for all languages, focused very much on stemming strategies, testing both light and aggressive stemmers for the Slavic languages (Bulgarian and Czech). For Hungarian they worked on decompounding. This group also compared performances obtained using word-based and 4-gram indexing strategies [7]. Another of the best performers, JHU-APL, normally uses an n-gram approach. Unfortunately, we have not received a paper yet from this group so cannot comment on their performance. The other group with very good performance for all languages was Opentext. This group also compared 4-gram results against results using stemming for all three languages. They found that while there could be large impacts on individual topics, there was little overall difference in average performance. Their experiments also confirmed past findings that indicate that blind relevance feedback can be detrimental to results, depending on the evaluation measures used [3]. The results of the statistical tests given towards the end of this paper show that the best results of these three groups did not differ significantly.

The group from Alicante also achieved good results testing query expansion techniques [8], while the group from Kolkata compared a statistical stemmer against a rule-based stemmer for both Czech and Hungarian [9]. Czech is a morphologically complex language and the two Czech only groups both used approaches involving morphological analysis and lemmatization [10], [11].

Table 7. Best entries for the monolingual track.

Track	Rank	Participant	Experiment DOI	MAP
Bulgarian	**1st**	unine	10.2415/AH-MONO-BG-CLEF2007.UNINE.UNINEBG4	44.22%
	2nd	jhu-apl	10.2415/AH-MONO-BG-CLEF2007.JHU-APL.APLMOBGTD4	36.57%
	3rd	opentext	10.2415/AH-MONO-BG-CLEF2007.OPENTEXT.OTBG07TDE	35.02%
	4th	alicante	10.2415/AH-MONO-BG-CLEF2007.ALICANTE.IRNBUEXP2N	29.81%
	5th	daedalus	10.2415/AH-MONO-BG-CLEF2007.DAEDALUS.BGFSBG2S	27.19%
	Difference			62.33%
Czech	**1st**	unine	10.2415/AH-MONO-CS-CLEF2007.UNINE.UNINECZ4	42.42%
	2nd	jhu-apl	10.2415/AH-MONO-CS-CLEF2007.JHU-APL.APLMOCSTD4	35.86%
	3rd	opentext	10.2415/AH-MONO-CS-CLEF2007.OPENTEXT.OTCS07TDE	34.84%
	4th	prague	10.2415/AH-MONO-CS-CLEF2007.PRAGUE.PRAGUE01	34.19%
	5th	daedalus	10.2415/AH-MONO-CS-CLEF2007.DAEDALUS.CSFSCS2S	32.03%
	Difference			32.44%
Hungarian	**1st**	unine	10.2415/AH-MONO-HU-CLEF2007.UNINE.UNINEHU4	47.73%
	2nd	opentext	10.2415/AH-MONO-HU-CLEF2007.OPENTEXT.OTHU07TDE	43.34%
	3rd	alicante	10.2415/AH-MONO-HU-CLEF2007.ALICANTE.IRNHUEXP2N	40.09%
	4th	jhu-apl	10.2415/AH-MONO-HU-CLEF2007.JHU-APL.APLMOHUTD5	39.91%
	5th	daedalus	10.2415/AH-MONO-HU-CLEF2007.DAEDALUS.HUFSHU2S	34.99%
	Difference			36.41%
English (only for Bilingual X2EN participants)	**1st**	bombay-ltrc	10.2415/AH-MONO-EN-CLEF2007.BOMBAY-LTRC.IITB_MONO_TITLE_DESC	44.02%
	2nd	jhu-apl	10.2415/AH-MONO-EN-CLEF2007.JHU-APL.APLMOENTD5	43.42%
	3rd	nottingham	10.2415/AH-MONO-EN-CLEF2007.NOTTINGHAM.MONOT	42.74%
	4th	depok	10.2415/AH-MONO-EN-CLEF2007.DEPOK.UIQTDMONO	40.57%
	5th	hyderabad	10.2415/AH-MONO-EN-CLEF2007.HYDERABAD.ENTD_DMENG07	40.16%
	Difference			9.61%

3.1 Results

Table 7 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 experiment; the DOI of the experiment; and the performance difference between the first and the last participant.

Figures 2 to 5 compare the performances of the top participants of the Monolingual tasks.

4 Main Stream Bilingual Experiments

The bilingual task was structured in three tasks (X → BG, CS, or HU target collection) plus a task for non-European topic languages against an English target collection. A special sub-task testing Indian languages against the English collection was also organised in response to requests from a number of research groups working in India. For the bilingual to English task, participating groups also had to submit an English monolingual run, to be used both as baseline and also to reinforce the English pool. All groups participating in the Indian

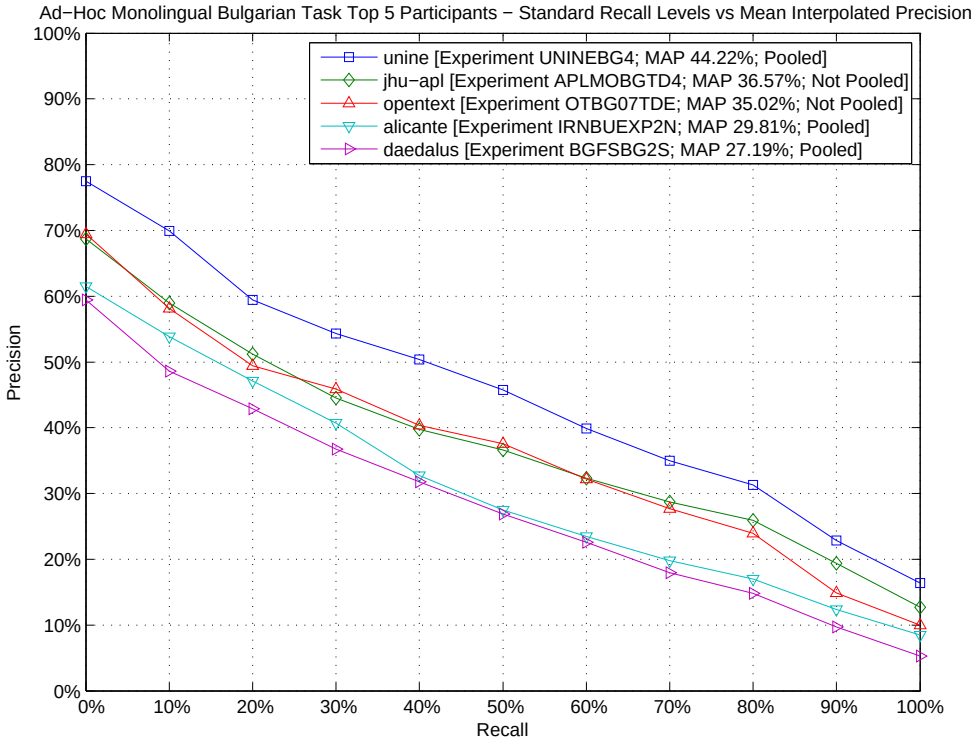


Fig. 2. Monolingual Bulgarian

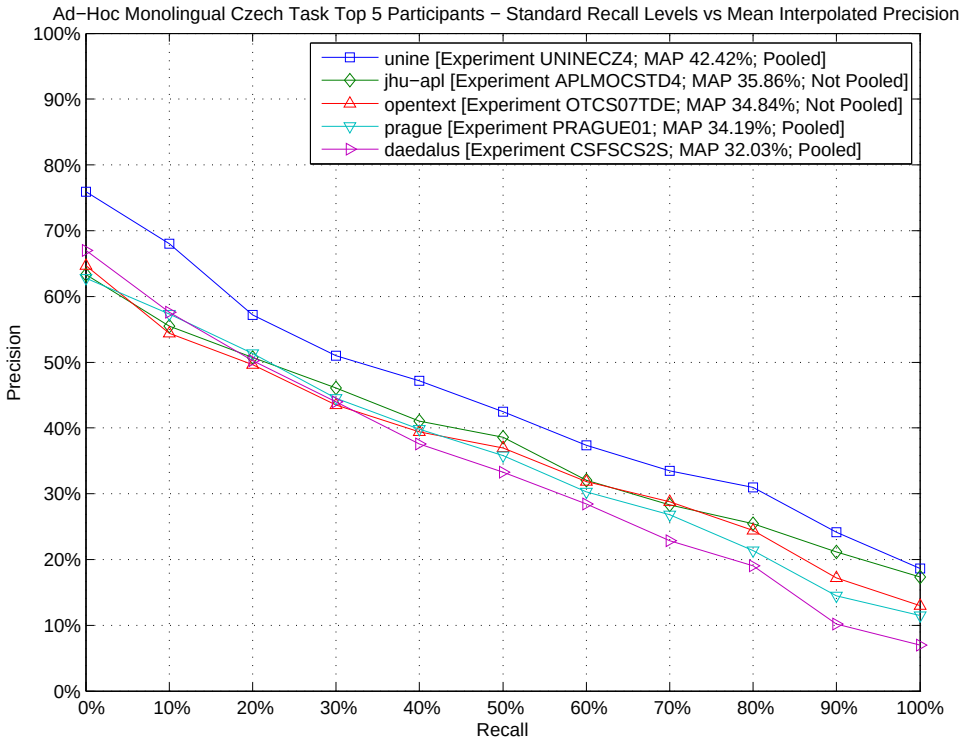


Fig. 3. Monolingual Czech

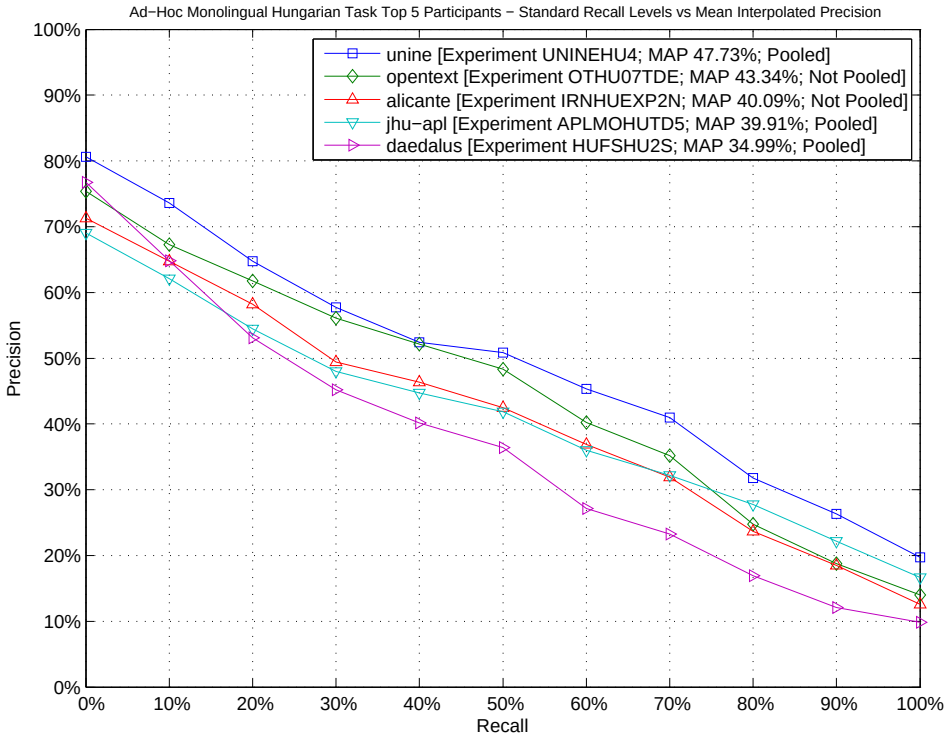


Fig. 4. Monolingual Hungarian

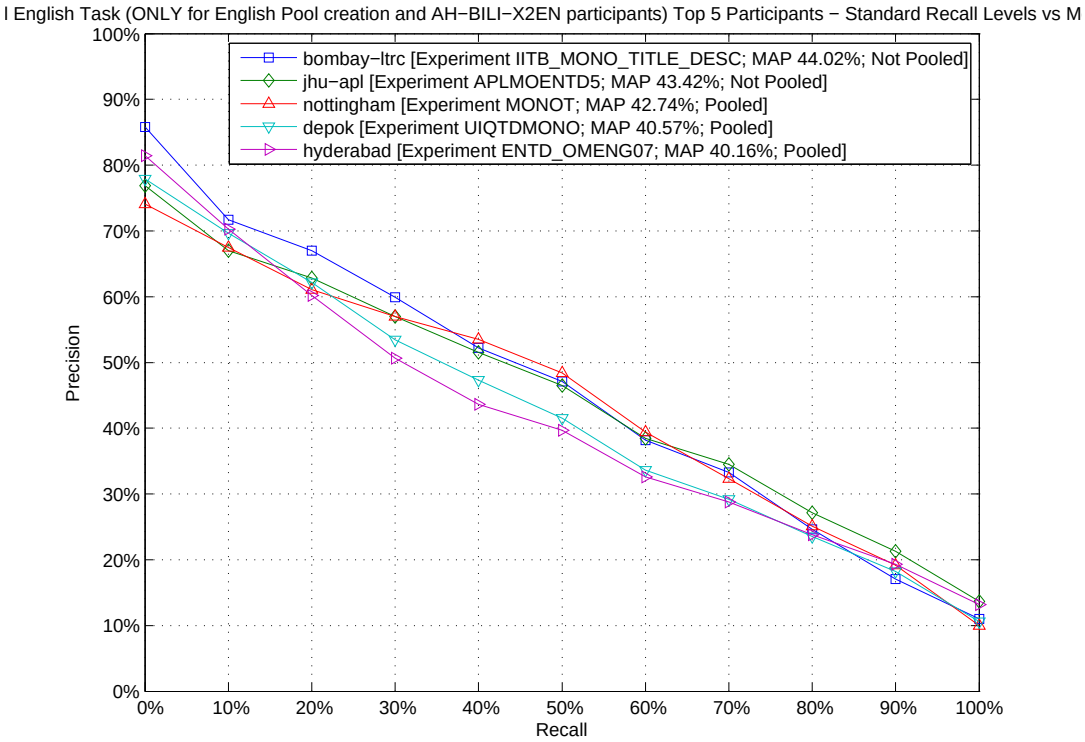


Fig. 5. Monolingual English

languages sub-task also had to submit at least one run in Hindi (mandatory) plus runs in other Indian languages (optional).

We were disappointed to only receive runs from one participant for the X → BG, CS, or HU tasks. Furthermore, the results were quite poor; as this group normally achieves very good performance, we suspect that these runs were probably corrupted in some way. For this reason, we have decided to disregard them as being of little significance. Therefore, in the rest of this section, we only comment briefly on the X → EN results.

We received runs using the following topic languages: Amharic, Chinese, Indonesian and Oromo plus, for the Indian sub-task, Bengali, Hindi, Marathi and Telugu[6].

For many of these languages few processing tools or resources are available. It is thus very interesting to see what measures the participants adopted to overcome this problem. Unfortunately, there has not yet been time to read the submitted reports from each group and, here below, we give just a first cursory glance at some of the approaches and techniques adopted. We will provide a more in-depth analysis at the workshop.

The top performance in the bilingual task was obtained by an Indonesian group; they compared different translation techniques: machine translation using Internet resources, transitive translation using bilingual dictionaries and French and German as pivot languages, and lexicons derived from parallel corpus created by translating all the CLEF English documents into Indonesian using a commercial MT system. They found that they obtained best results using the MT system together with query expansion [12].

The second placed group used Chinese for their queries and a dictionary based translation technique. The experiments of this group concentrated on developing new strategies to address two well-known CLIR problems: translation ambiguity, and coverage of the lexicon [13]. The work by [14] which used Amharic as the topic language also paid attention to the problems of sense disambiguation and out-of-vocabulary terms.

The third performing group also used Indonesian as the topic language; unfortunately we have not received a paper from them so far so cannot comment on their approach. An interesting paper, although slightly out of the task as the topic language used was Hungarian was [15]. This group used a machine readable dictionary approach but also applied Wikipedia data to eliminate unlikely translations according to the conceptual context. The group testing Oromo used linguistic and lexical resources developed at their institute; they adopted a bilingual dictionary approach and also tested the impact of a light stemmer for Afaan Oromo on their performance with positive results [16].

The groups using Indian topic languages tested different approaches. The group from Kolkata submitted runs for Bengali, Hindi and Telugu to English using a bilingual dictionary lookup approach [17]. They had the best performance using Telugu probably because they carried out some manual tasks during indexing. A group from Bangalore tested a statistical MT system trained on

[6] Although topics had also been requested in Tamil, in the end they were not used.

Table 8. Best entries for the bilingual task.

Track	Rank	Part.	Lang.	Experiment DOI	MAP
Bulgarian	**1st**	jhu-apl	en	10.2415/AH-BILI-X2BG-CLEF2007.JHU-APL.APLBIENBGTD4	7.33%
	Difference				
Czech	**1st**	jhu-apl	en	10.2415/AH-BILI-X2CS-CLEF2007.JHU-APL.APLBIENCSTD4	21.43%
	Difference				
English	**1st**	depok	id	10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDTOGGLEFB10D10T	38.78%
	2nd	nottingham	zh	10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAWOTD	34.56%
	3rd	jhu-apl	id	10.2415/AH-BILI-X2EN-CLEF2007.JHU-APL.APLBIIDENTDS	33.24%
	4th	hyderabad	om	10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.OMTDO7	29.91%
	5th	bombay-ltrc	hi	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_DICE	29.52%
	Difference				31.37%
Hungarian	**1st**	jhu-apl	en	10.2415/AH-BILI-X2HU-CLEF2007.JHU-APL.APLBIENHUTD5	29.63%
	Difference				

parallel aligned sentences and a language modelling based retrieval algorithm for a Hindi to English system [18]. The group from Bombay had the best overall performances; they used bilingual dictionaries for both Hindi and Marathi to English and applied term-to-term cooccurrence statistics for sense disambiguation [19]. The Hyderabad group also used bilingual lexicons for query translation from Hindi and Telugu to English together with a variant of the TFIDF algorithm and a hybrid boolean formulation for the queries to improve ranking [20]. Interesting work was done by the group from Kharagpur which submitted runs for Hindi and Bengali. They attempted to overcome the lack of resources for Bengali by using phoneme-based transliterations to generate equivalent English queries from Hindi and Bengali topics [21].

4.1 Results

Table 8 shows the best results for this task. 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.

Figure 6 compares the performances of the top participants of the Bilingual English[7].

For bilingual retrieval evaluation, a common method to evaluate performance is to compare results against monolingual baselines. This year we can only comment on the results for the bilingual to English tasks. The best results were obtained by a system using Indonesian as a topic language. This group achieved 88.10% of the best monolingual English IR system. This is a good result considering that Indonesian is not a language for which a lot of resources and machine-readable dictionaries are available. It is very close to the best results obtained

[7] Since for the other bilingual tasks only one participant submitted experiments, only the graphs for bilingual English are reported

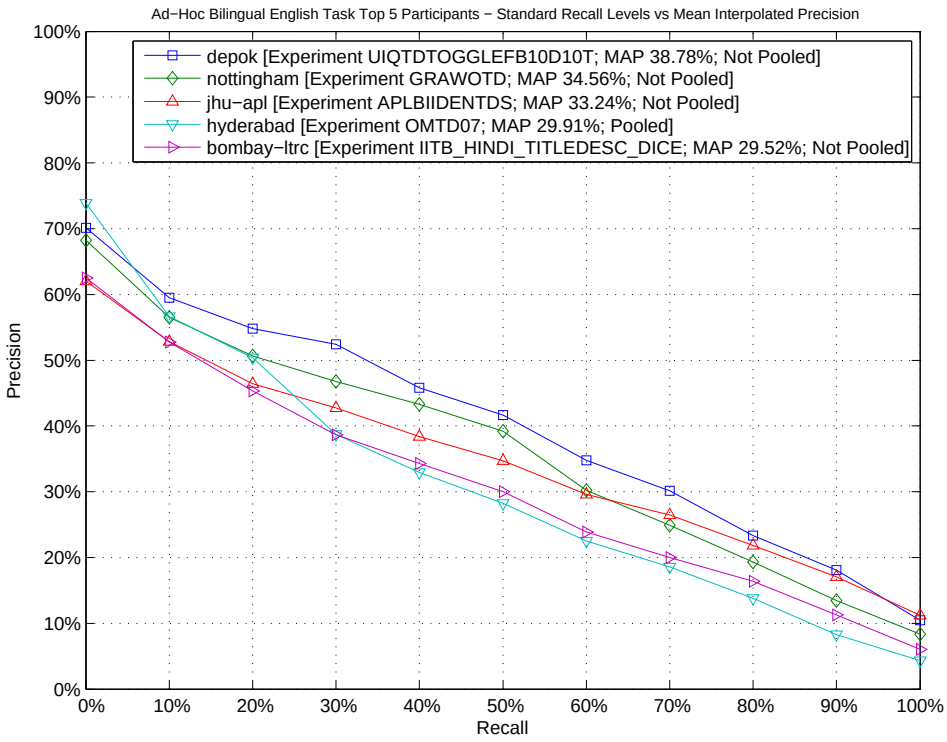


Fig. 6. Bilingual English

last year for two well-established CLEF languages: French and Portuguese, when the equivalent figures were 93.82% and 90.91%, respectively.

4.2 Indian To English Subtask Results

Table 9 shows the best results for the Indian sub-task. The performance difference between the best and the last (up to 6) placed group is given (in terms of average precision). The first set of rows regard experiments for the mandatory topic language: Hindi; the second set of rows report experiments where the source language is one of other Indian languages.

It is interesting to note that in both sets of experiments, the best performing participant is the same. In the second set, we can note that for three (Hindi, Marathi, and Telegu) out of the four Indian languages used the performances of the top groups are quite similar.

The best performance for the Indian sub-task is 76.12% of the best bilingual English system (achieved by veteran CLEF participants) and 67.06% of the monolingual baseline, which is quite encouraging for a new task with languages where encoding issues and linguistic resources make the task difficult. This is in

Table 9. Best entries for the bilingual Indian subtask.

Track	Rank	Part.	Lang.	Experiment DOI	MAP
Hindi to English	**1st**	bombay-ltrc	hi	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_DICE	29.52%
	2nd	msindia	hi	10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.2007_RBLM_ALL_CROSS_1000_POSSCORES	21.80%
	3rd	hyderabad	hi	10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.HITD	15.60%
	4th	jadavpur	hi	10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.AHBILIHI2ENR1	10.86%
	5th	kharagpur	hi	10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.HINDITITLE	4.77%
	6th				
	Difference				518.87%
Bengali/ Hindi/ Marathi/ Telugu to English	**1st**	bombay-ltrc	hi	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_DICE	29.52%
	2nd	msindia	hi	10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.2007_RBLM_ALL_CROSS_1000_POSSCORES	21.80%
	3rd	bombay-ltrc	mr	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_MAR_TITLE_DICE	21.63%
	4th	hyderabad	te	10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.TETD	21.55%
	5th	jadavpur	te	10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.AHBILITE2ENR1	11.28%
	6th	kharagpur	bn	10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.BENGALITITLEDESC	7.25%
	Difference				307.17%

fact comparable with the performances of some newly introduced European languages. For example, we can compare them to those for Bulgarian and Hungarian in CLEF 2006:

- X → BG: 52.49% of best monolingual Bulgarian IR system;
- X → HU: 53.13% of best monolingual Hungarian IR system.

5 Robust Experiments

The robust task ran for the second time at CLEF 2007. It is an ad-hoc retrieval task based on data of previous CLEF campaigns. The evaluation approach is modified and a different perspective is taken. The robust task emphasizes the difficult topics by a non-linear integration of the results of individual topics into one result for a system [22,23]. By doing this, the evaluation results are interpreted in a more user oriented manner. Failures and very low results for some topics hurt the user experience with a retrieval system. Consequently, any system should try to avoid these failures. This has turned out to be a hard task [24]. Robustness is a key issue for the transfer of research into applications. The robust task rewards systems which achieve a minimal performance level for all topics.

In order to do this, the robust task uses the geometric mean of the average precision for all topics (GMAP) instead of the mean average of all topics (MAP). This measure has also been used at a roust track at the Text Retrieval Conference (TREC) where robustness was explored for monolingual English retrieval [23]. At CLEF, robustness is evaluated for monolingual and bilingual retrieval for several European languages.

The robust task at CLEF exploits data created for previous CLEF editions. Therefore, a larger data set can be used for the evaluation. A larger number of

Table 10. Data for the Robust Task 2007.

Language	Target Collection	Training Topic DOIs	Test Topic DOIs
English	LA Times 1994	10.2452/41-AH–10.2452/200-AH	10.2452/251-AH–10.2452/350-AH
French	Le Monde 1994 SDA 1994	10.2452/41-AH–10.2452/200-AH	10.2452/251-AH–10.2452/350-AH
Portuguese	Público 1995	–	10.2452/201-AH–10.2452/350-AH

topics allows a more reliable evaluation [25]. A secondary goal of the robust task is the definition of larger data sets for retrieval evaluation.

As described above, the CLEF2007 robust task offered three languages often used in previous CLEF campaigns: English, French and Portuguese. The data used has been developed during CLEF 2001 through 2006. Generally, the topics from CLEF 2001 until CLEF 2003 were training topics whereas the topics developed between 2004 and 2006 were the test topics on which the main evaluation measures are given.

Thus, the data used in the robust task in 2007 is different from the set defined for the roust task at CLEF 2006. The documents which need to be searched are articles from major newspapers and news providers in the three languages. Not all collections had been offered consistently for all CLEF campaigns, therefore, not all collections were integrated into the robust task. Most data from 1995 was omitted in order to provide a homogeneous collection. However, for Portuguese, for which no training data was available, only data from 1995 was used. Table 10 shows the data for the robust task.

The robust task attracted 63 runs submitted by 7 groups (CLEF 2006: 133 runs from 8 groups). Effectiveness scores were calculated with the version 8.0 of the program which provides the *Mean Average Precision (MAP)*, while the *Geometric Average Precision (GMAP)* was calculated using DIRECT version 2.0.

5.1 Robust Monolingual Results

Table 11 shows the best results for this task. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

The results cannot be compared to the results of the CLEF 2005 and CLEF 2006 campaign in which the same topics were used because a smaller collection had to be searched.

Figures from 7 to 9 compare the performances of the top participants of the Robust Monolingual.

5.2 Robust Bilingual Results

Table 12 shows the best results for this task. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision). All the experiments where from English to French.

Table 11. Best entries for the robust monolingual task.

Track	Rank	Participant	Experiment DOI	MAP	GMAP
English	**1st**	reina	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.REINA.REINAENTDNT	38.97%	18.50%
	2nd	daedalus	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.DAEDALUS.ENFSEN22S	37.78%	17.72%
	3rd	hildesheim	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.HILDESHEIM.HIMOENBRFNE	5.88%	0.32%
	4th				
	5th				
	Difference			562.76%	5,681.25%
French	**1st**	unine	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.UNINE.UNINEFR1	42.13%	14.24%
	2nd	reina	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.REINA.REINAFRTDET	38.04%	12.17%
	3rd	jaen	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.JAEN.UJARTFR1	34.76%	10.69%
	4th	daedalus	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.DAEDALUS.FRFSFR22S	29.91%	7.43%
	5th	hildesheim	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.HILDESHEIM.HIMOFRBRF2	27.31%	5.47%
	Difference			54.27%	160.33%
Portuguese	**1st**	reina	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.REINA.REINAPTTDNT	41.40%	12.87%
	2nd	jaen	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.JAEN.UJARTPT1	24.74%	0.58%
	3rd	daedalus	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.DAEDALUS.PTFSPT2S	23.75%	0.50%
	4th	xldb	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.XLDB.XLDBROB16_10	1.21%	0.07%
	5th				
	Difference			3,321,49%	18,285.71%

Table 12. Best entries for the robust bilingual task.

Track	Rank	Participant	Experiment DOI	MAP	GMAP
French	**1st**	reina	10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.REINA.REINAE2FTDNT	35.83%	12.28%
	2nd	unine	10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.UNINE.UNINEBILFR1	33.50%	5.01%
	3rd	colesun	10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.COLESUN.EN2FRTST4GRINTLOGLU001	22.87%	3.57%
	4th				
	5th				
	Difference			54.27%	243.98%

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2007:

– X → FR: 85.05% of best monolingual French IR system;

Figure 10 compares the performances of the top participants of the Robust Bilingual task.

5.3 Approaches Applied to Robust Retrieval

The REINA system applied different measures of robustness during the training phase in order to optimize the performance. A local query expansion technique added terms. The CoLesIR system experimented with n-gram based translation for bi-lingual retrieval which requires no languages specific components. SINAI tried to increase the robustness of the results by expanding the query with an external knowledge source. This is a typical approach in order to obtain additional

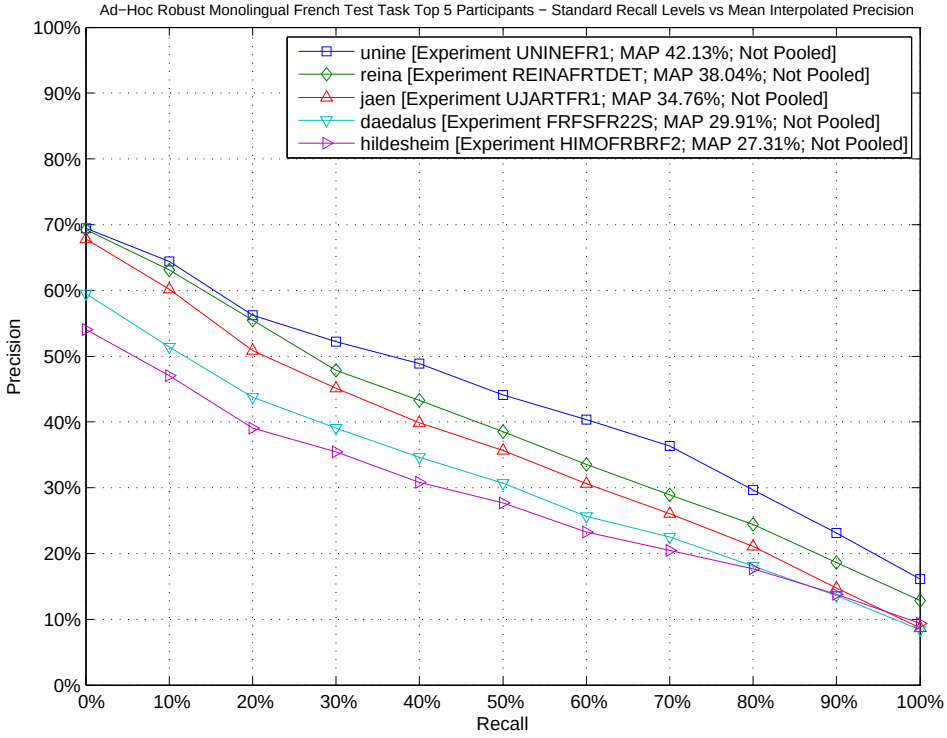


Fig. 7. Robust Monolingual French.

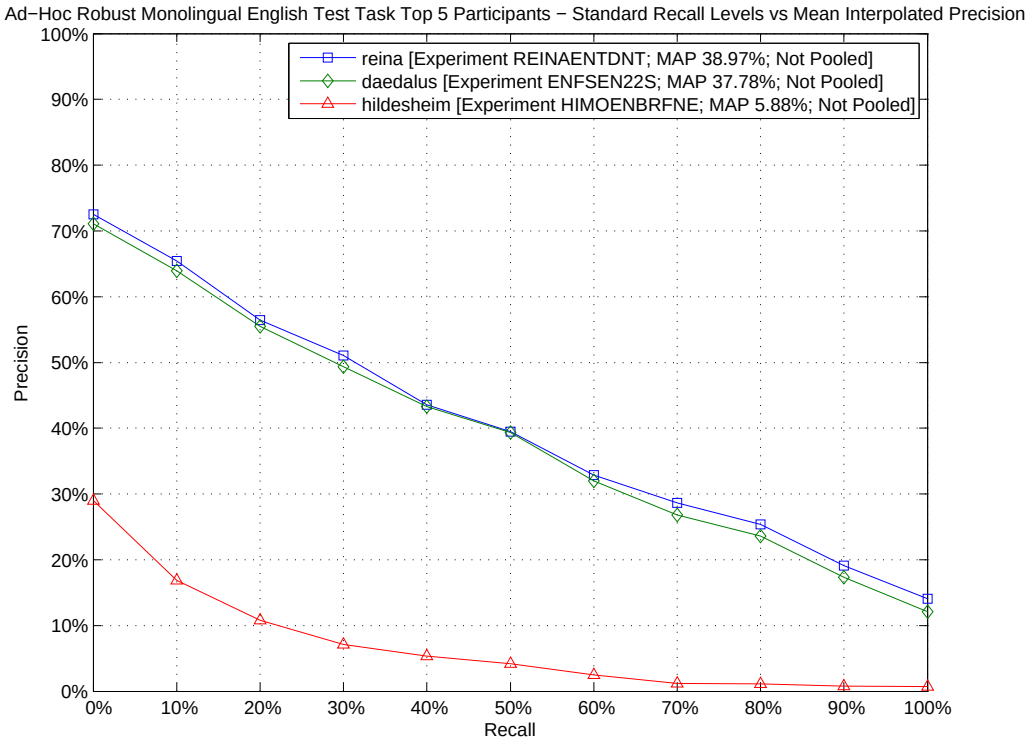


Fig. 8. Robust Monolingual English.

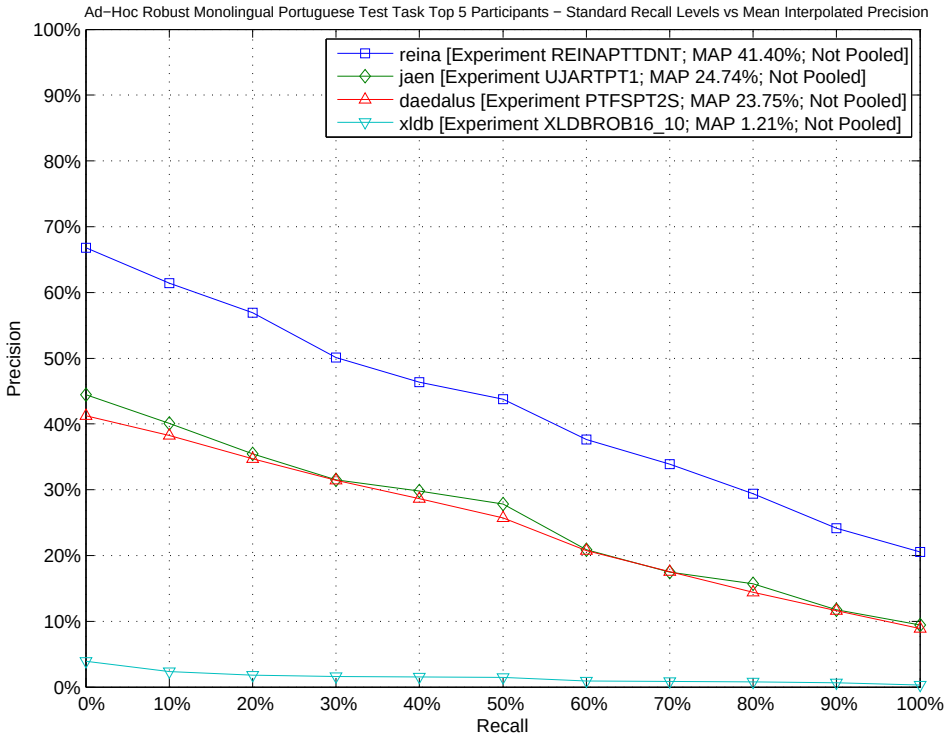


Fig. 9. Robust Monolingual Portuguese.

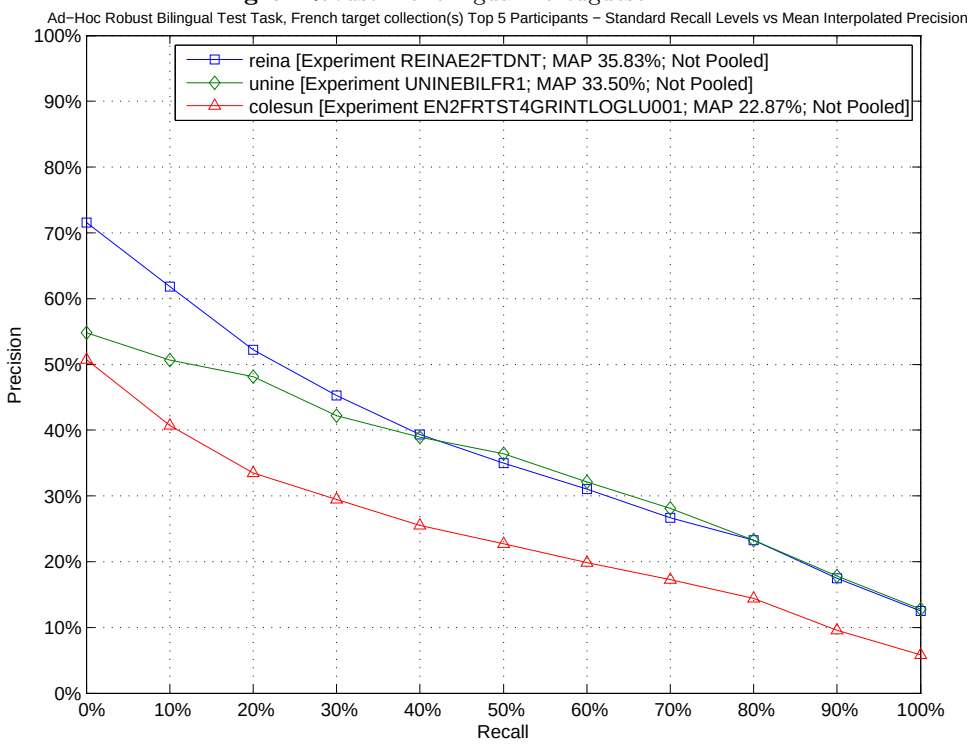


Fig. 10. Robust Bilingual French

query terms and avoid zero hits in case of out of vocabulary problems. Contrary to standard query expansion techniques, the new terms form a second query and results of both initial and second query are integrated under a logistic fusion strategy. The Daedalus group submitted experiments wit the Miracle system. BM25 weighting without blind relevance feedback was applied. For descriptions of all the robust experiments, see the Robust section in these Working Notes.

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 [26].

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 [26] 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 [27] and the Jarque-Bera test [28]. 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 [29] recommends for use with precision/recall measures.

Table 13 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

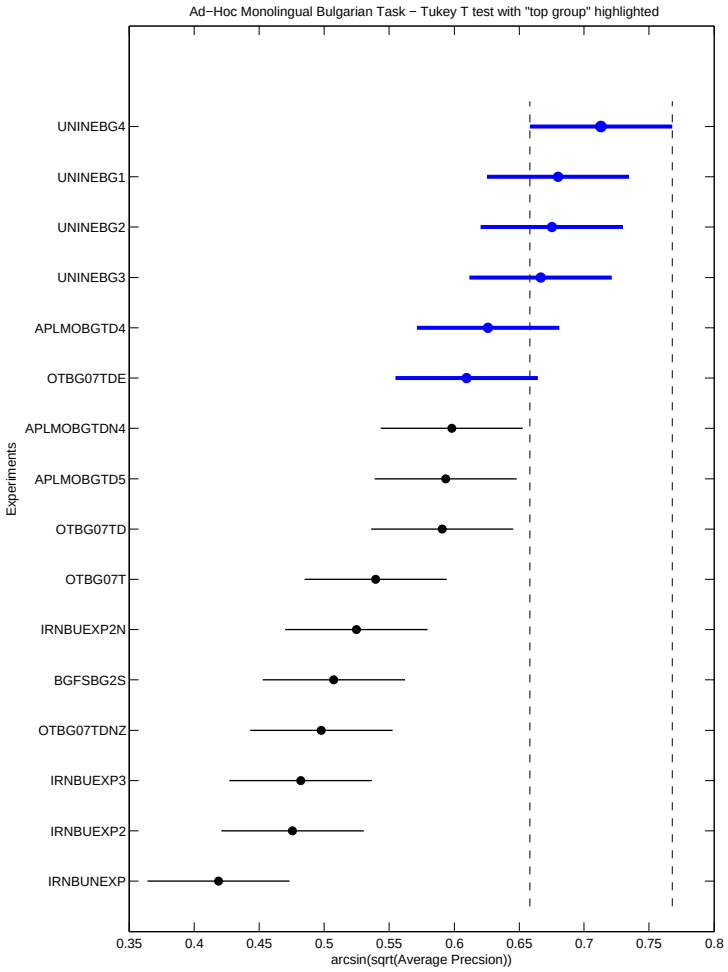
Table 13. Lilliefors (LF) and Jarque-Bera (JB) test for each Ad-Hoc track with and without Tague-Sutcliffe (TS) arcsin transformation. Each entry is the number of experiments whose performance distribution can be considered drawn from a Gaussian distribution, with respect to the total number of experiment of the track. The value of alpha for this test was set to 5%.

Track	LF	LF & TS	JB	JB & TS
Monolingual Bulgarian	8	15	12	16
Monolingual Czech	4	22	14	21
Monolingual English	22	24	28	27
Monolingual Hungarian	6	16	15	17
Bilingual Bulgarian	0	0	0	0
Bilingual Czech	0	0	0	0
Bilingual English	6	35	28	41
Bilingual Hungarian	0	1	1	2
Robust Monolingual English	3	6	1	4
Robust Monolingual French	2	9	5	9
Robust Monolingual Portuguese	0	3	0	2
Robust Bilingual French	0	5	1	6

and robust tasks, only runs where significant differences exist are shown; the remainder of the graphs can be found in the Appendices [5,6].

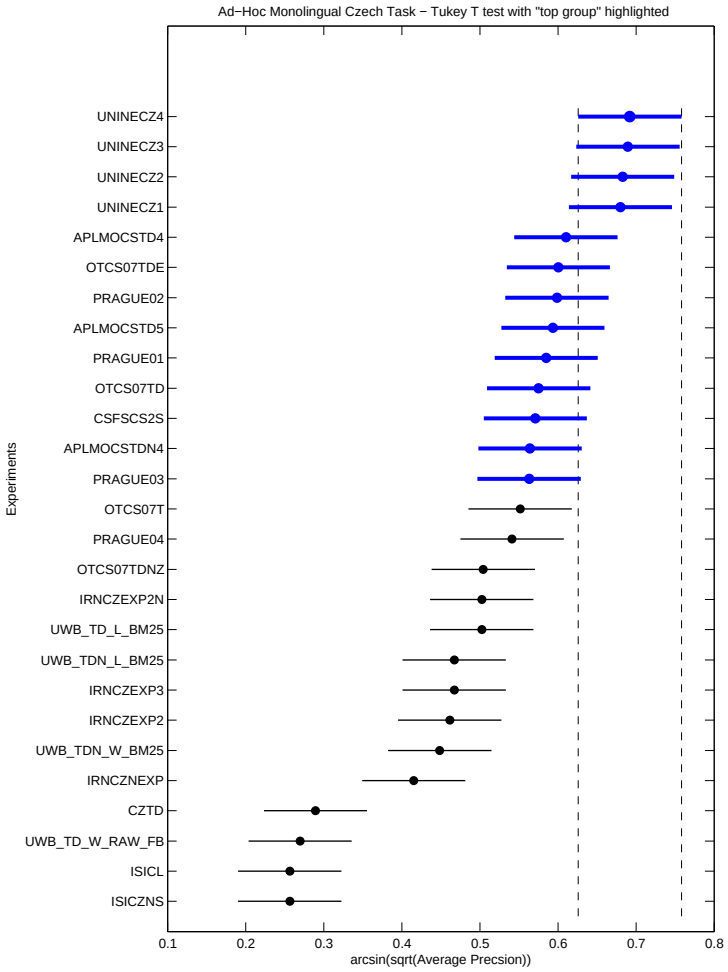
The first graph shows participants' runs (y axis) and performance obtained (x axis). The circle indicates the average performance (in terms of Precision) while the segment shows the interval in which the difference in performance is not statistically significant.

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.



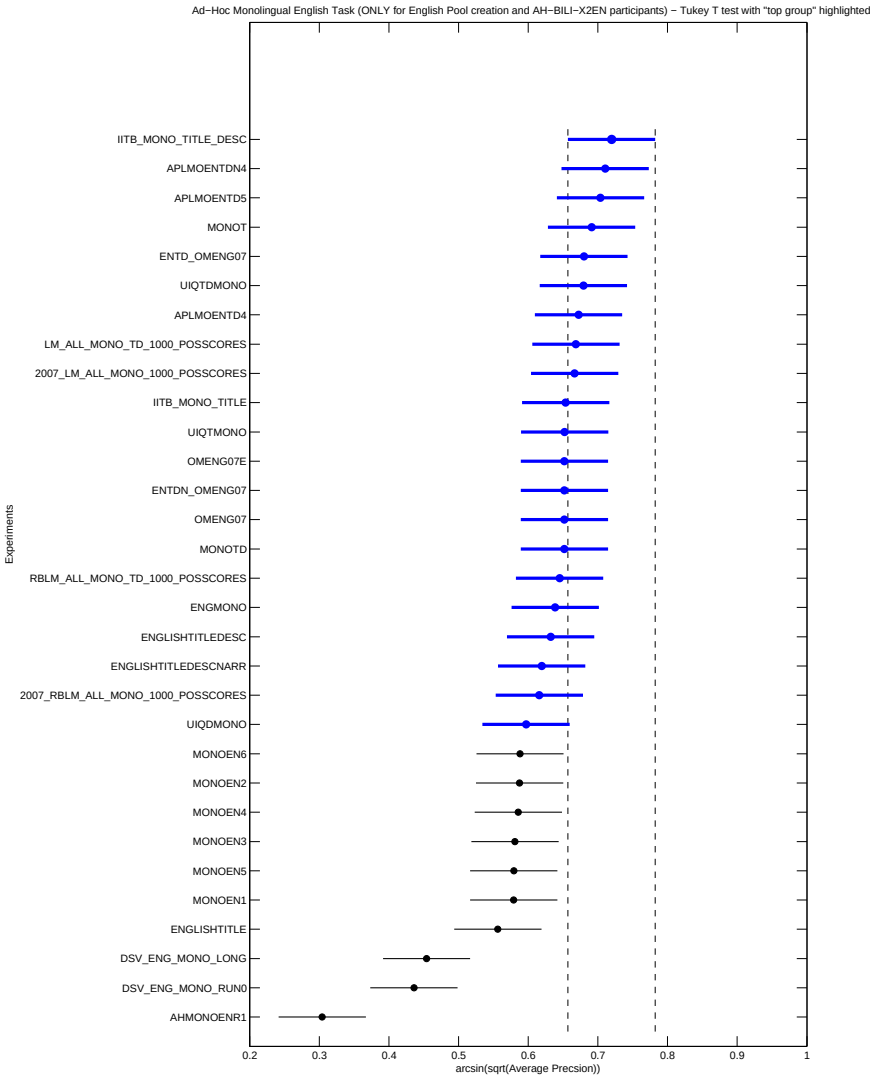
Experiment DOI	Groups							
10.2415/AH-MONO-BG-CLEF2007.UNINE.UNINEBG4	X							
10.2415/AH-MONO-BG-CLEF2007.UNINE.UNINEBG1	X	X						
10.2415/AH-MONO-BG-CLEF2007.UNINE.UNINEBG2	X	X						
10.2415/AH-MONO-BG-CLEF2007.UNINE.UNINEBG3	X	X						
10.2415/AH-MONO-BG-CLEF2007.JHU-APL.APLMOBGTD4	X	X	X					
10.2415/AH-MONO-BG-CLEF2007.OPENTEXT.OTBG07TDE	X	X	X	X				
10.2415/AH-MONO-BG-CLEF2007.JHU-APL.APLMOBGTDN4		X	X	X	X			
10.2415/AH-MONO-BG-CLEF2007.JHU-APL.APLMOBGTD5		X	X	X	X			
10.2415/AH-MONO-BG-CLEF2007.OPENTEXT.OTBG07TD		X	X	X	X	X		
10.2415/AH-MONO-BG-CLEF2007.OPENTEXT.OTBG07T			X	X	X	X	X	
10.2415/AH-MONO-BG-CLEF2007.ALICANTE.IRNBUEXP2N			X	X	X	X	X	X
10.2415/AH-MONO-BG-CLEF2007.DAEDALUS.BGFSBG2S				X	X	X	X	X
10.2415/AH-MONO-BG-CLEF2007.OPENTEXT.OTBG07TDNZ					X	X	X	X
10.2415/AH-MONO-BG-CLEF2007.ALICANTE.IRNBUEXP3						X	X	X
10.2415/AH-MONO-BG-CLEF2007.ALICANTE.IRNBUEXP2							X	X
10.2415/AH-MONO-BG-CLEF2007.ALICANTE.IRNBUNEXP								X

Fig. 11. Ad-Hoc Monolingual Bulgarian. Experiments grouped according to the Tukey T Test (DOI 10.2455/TUKEY_T_TEST.9633F4C12FC97855A347B6AF7DD9B7A5).



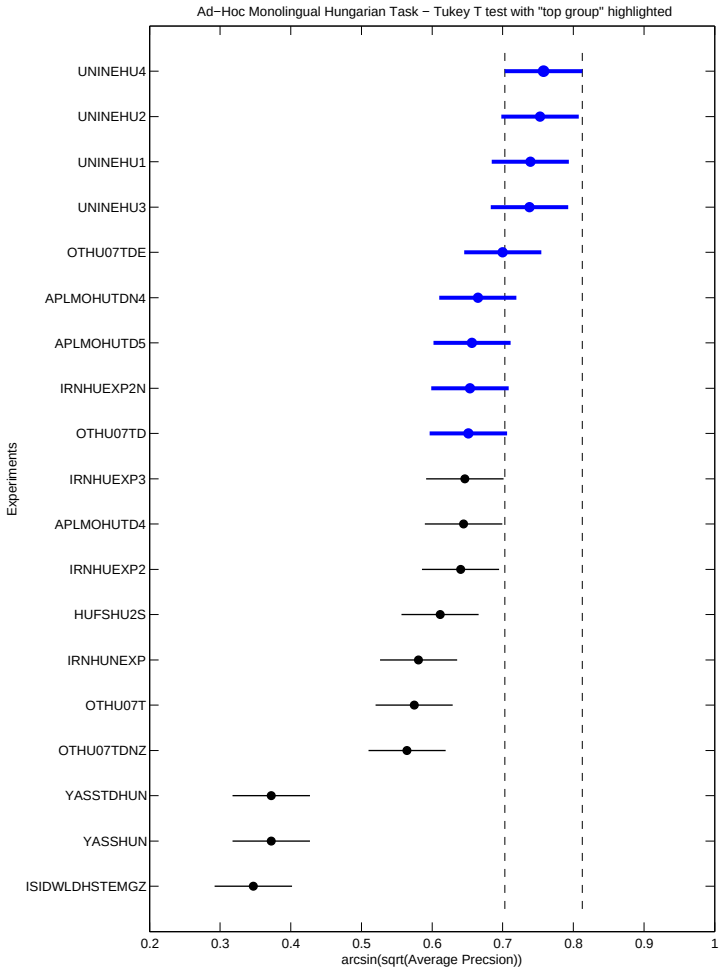
Experiment DOI	Groups							
10.2415/AH-MONO-CS-CLEF2007.UNINE.UNINECZ4	X							
10.2415/AH-MONO-CS-CLEF2007.UNINE.UNINECZ3	X							
10.2415/AH-MONO-CS-CLEF2007.UNINE.UNINECZ2	X	X						
10.2415/AH-MONO-CS-CLEF2007.UNINE.UNINECZ1	X	X						
10.2415/AH-MONO-CS-CLEF2007.JHU-APL.APLMOCSTD4	X	X	X					
10.2415/AH-MONO-CS-CLEF2007.OPENTEXT.OTCS07TDE	X	X	X					
10.2415/AH-MONO-CS-CLEF2007.PRAGUE.PRAGUE02	X	X	X	X				
10.2415/AH-MONO-CS-CLEF2007.JHU-APL.APLMOCSTD5	X	X	X	X	X			
10.2415/AH-MONO-CS-CLEF2007.PRAGUE.PRAGUE01	X	X	X	X	X			
10.2415/AH-MONO-CS-CLEF2007.OPENTEXT.OTCS07TD	X	X	X	X	X	X		
10.2415/AH-MONO-CS-CLEF2007.DAEDALUS.CSFSCS2S	X	X	X	X	X	X		
10.2415/AH-MONO-CS-CLEF2007.JHU-APL.APLMOCSTDN4	X	X	X	X	X	X		
10.2415/AH-MONO-CS-CLEF2007.PRAGUE.PRAGUE03	X	X	X	X	X	X		
10.2415/AH-MONO-CS-CLEF2007.OPENTEXT.OTCS07T		X	X	X	X	X		
10.2415/AH-MONO-CS-CLEF2007.PRAGUE.PRAGUE04			X	X	X	X	X	
10.2415/AH-MONO-CS-CLEF2007.OPENTEXT.OTCS07TDNZ			X	X	X	X	X	
10.2415/AH-MONO-CS-CLEF2007.ALICANTE.IRNCZEXP2N			X	X	X	X	X	
10.2415/AH-MONO-CS-CLEF2007.BOHEMIA.UWB_TD_L_BM25			X	X	X	X	X	
10.2415/AH-MONO-CS-CLEF2007.BOHEMIA.UWB_TDN_L_BM25				X	X	X	X	
10.2415/AH-MONO-CS-CLEF2007.ALICANTE.IRNCZEXP3				X	X	X	X	
10.2415/AH-MONO-CS-CLEF2007.ALICANTE.IRNCZEXP2					X	X	X	
10.2415/AH-MONO-CS-CLEF2007.BOHEMIA.UWB_TDN_W_BM25						X	X	
10.2415/AH-MONO-CS-CLEF2007.ALICANTE.IRNCZNEXP							X	X
10.2415/AH-MONO-CS-CLEF2007.ISI.CZTD								X X
10.2415/AH-MONO-CS-CLEF2007.BOHEMIA.UWB_TD_W_RAW_FB								X
10.2415/AH-MONO-CS-CLEF2007.ISI.ISICL								X
10.2415/AH-MONO-CS-CLEF2007.ISI.ISICZNS								X

Fig. 12. Ad-Hoc Monolingual Czech. Experiments grouped according to the Tukey T Test (DOI 10.2455/TUKEY_T_TEST.2654ECBA17A46006F564F803675F163C).



Experiment DOI	Groups					
10.2415/AH-MONO-EN-CLEF2007.BOMBAY-LTRC.IITB_MONO_TITLE_DESC	X					
10.2415/AH-MONO-EN-CLEF2007.JHU-APL.APLMOENTDN4	X	X				
10.2415/AH-MONO-EN-CLEF2007.JHU-APL.APLMOENTD5	X	X	X			
10.2415/AH-MONO-EN-CLEF2007.NOTTINGHAM.MONOT	X	X	X			
10.2415/AH-MONO-EN-CLEF2007.HYDERABAD.ENTD_OMENG07	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.DEPOK.UIQTDMONO	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.JHU-APL.APLMOENTD4	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.MSINDIA.LM_ALL_MONO_TD_1000_POSSCORES	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.MSINDIA.2007_LM_ALL_MONO_1000_POSSCORES	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.BOMBAY-LTRC.IITB_MONO_TITLE	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.DEPOK.UIQTMONO	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.HYDERABAD.OMENG07E	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.HYDERABAD.ENTDN_OMENG07	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.HYDERABAD.OMENG07	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.NOTTINGHAM.MONOTD	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.MSINDIA.RBLM_ALL_MONO_TD_1000_POSSCORES	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.HYDERABAD.ENGMONO	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.KHARAGPUR.ENGLISHTITLEDESC	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.KHARAGPUR.ENGLISHTITLEDESCNARR	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.MSINDIA.2007_RBLM_ALL_MONO_1000_POSSCORES	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.DEPOK.UIQDMONO	X	X	X	X		
10.2415/AH-MONO-EN-CLEF2007.BUDAPEST-ACAD.MONOEN6		X	X	X		
10.2415/AH-MONO-EN-CLEF2007.BUDAPEST-ACAD.MONOEN2		X	X	X		
10.2415/AH-MONO-EN-CLEF2007.BUDAPEST-ACAD.MONOEN4		X	X	X		
10.2415/AH-MONO-EN-CLEF2007.BUDAPEST-ACAD.MONOEN3			X	X		
10.2415/AH-MONO-EN-CLEF2007.BUDAPEST-ACAD.MONOEN5			X	X	X	
10.2415/AH-MONO-EN-CLEF2007.BUDAPEST-ACAD.MONOEN1			X	X	X	
10.2415/AH-MONO-EN-CLEF2007.KHARAGPUR.ENGLISHTITLE				X	X	
10.2415/AH-MONO-EN-CLEF2007.STOCKHOLM.DSV_ENG_MONO_LONG					X	
10.2415/AH-MONO-EN-CLEF2007.KHARAGPUR.ENGLISHTITLE						X
10.2415/AH-MONO-EN-CLEF2007.STOCKHOLM.DSV_ENG_MONO_LONG						X
10.2415/AH-MONO-EN-CLEF2007.STOCKHOLM.DSV_ENG_MONO_RUN0						X

Fig. 13. Ad-Hoc Monolingual English. Experiments grouped according to the Tukey T Test (DOI 10.2455/TUKEY_T_TEST.430FD660255B5646DE62C1BFFFB0E298).



Experiment DOI	Groups					
10.2415/AH-MONO-HU-CLEF2007.UNINE.UNINEHU4	X					
10.2415/AH-MONO-HU-CLEF2007.UNINE.UNINEHU2	X	X				
10.2415/AH-MONO-HU-CLEF2007.UNINE.UNINEHU1	X	X	X			
10.2415/AH-MONO-HU-CLEF2007.UNINE.UNINEHU3	X	X	X			
10.2415/AH-MONO-HU-CLEF2007.OPENTEXT.OTHU07TDE	X	X	X	X		
10.2415/AH-MONO-HU-CLEF2007.JHU-APL.APLMOHUTDN4	X	X	X	X	X	
10.2415/AH-MONO-HU-CLEF2007.JHU-APL.APLMOHUTD5	X	X	X	X	X	
10.2415/AH-MONO-HU-CLEF2007.ALICANTE.IRNHUEXP2N	X	X	X	X	X	
10.2415/AH-MONO-HU-CLEF2007.OPENTEXT.OTHU07TD	X	X	X	X	X	
10.2415/AH-MONO-HU-CLEF2007.ALICANTE.IRNHUEXP3		X	X	X	X	
10.2415/AH-MONO-HU-CLEF2007.JHU-APL.APLMOHUTD4		X	X	X	X	
10.2415/AH-MONO-HU-CLEF2007.ALICANTE.IRNHUEXP2			X	X	X	
10.2415/AH-MONO-HU-CLEF2007.DAEDALUS.HUFSHU2S				X	X	
10.2415/AH-MONO-HU-CLEF2007.ALICANTE.IRNHUNEXP					X	
10.2415/AH-MONO-HU-CLEF2007.OPENTEXT.OTHU07T					X	
10.2415/AH-MONO-HU-CLEF2007.OPENTEXT.OTHU07TDNZ					X	
10.2415/AH-MONO-HU-CLEF2007.ISI.YASSTDHUN						X
10.2415/AH-MONO-HU-CLEF2007.ISI.YASSHUN						X
10.2415/AH-MONO-HU-CLEF2007.ISI.ISIDWLDHSTEMGZ						X

Fig. 14. Ad-Hoc Monolingual Hungarian. Experiments grouped according to the Tukey T Test (DOI 10.2455/TUKEY_T_TEST.D46DC11E6C986891646E633B0E27104C).

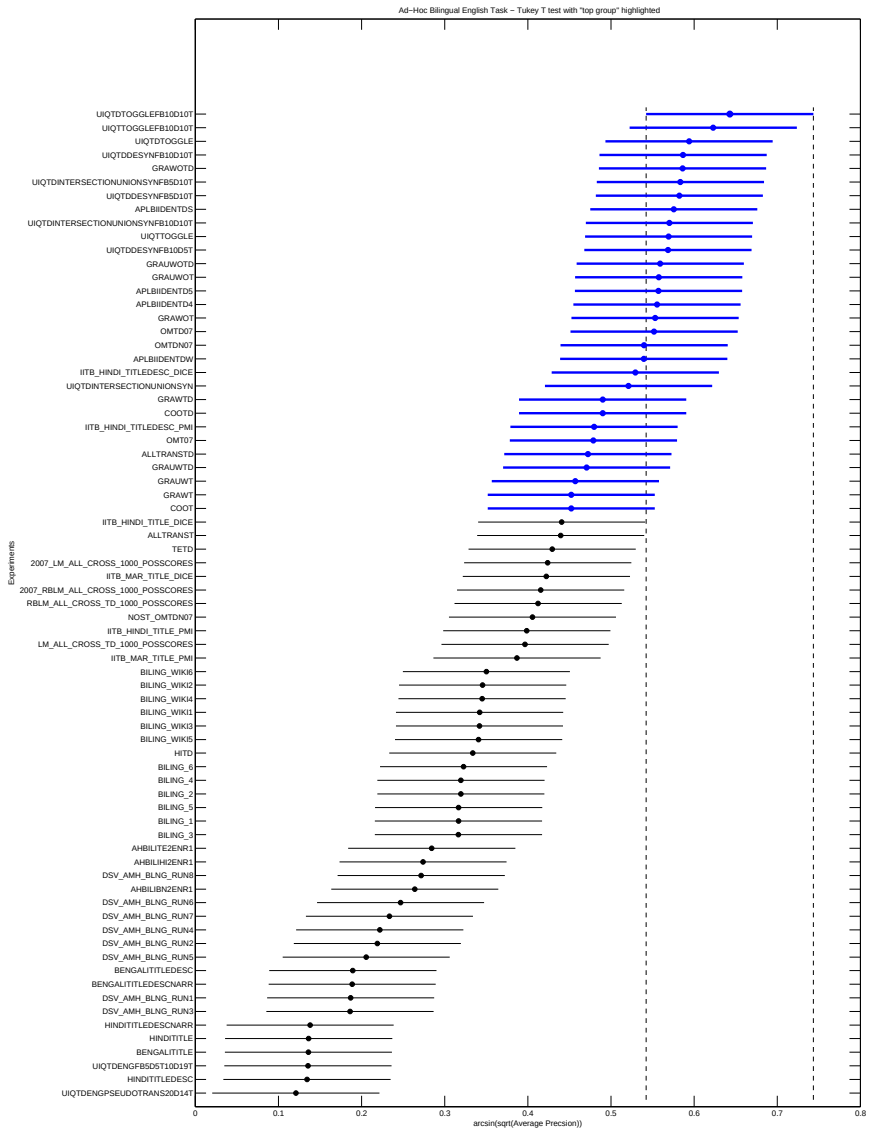


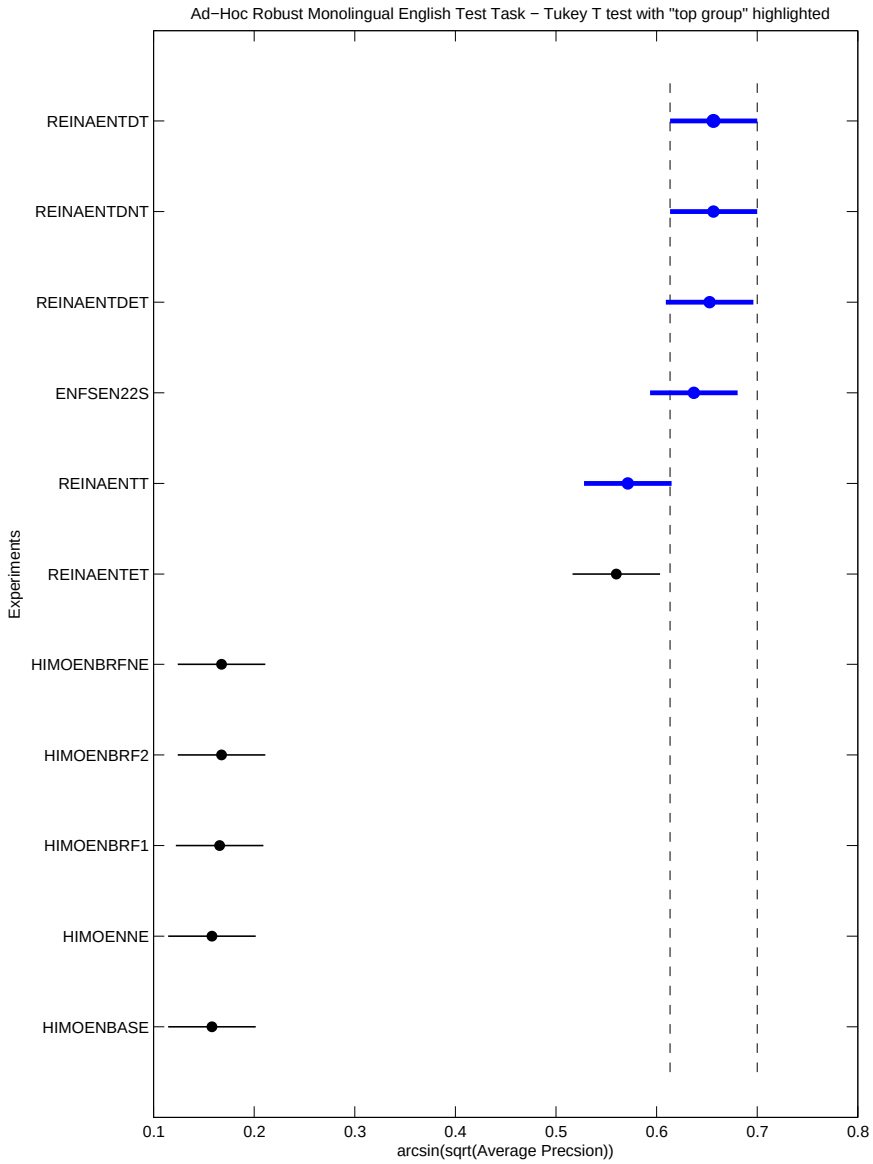
Fig. 15. Ad-Hoc Bilingual English. The figure shows the Tukey T Test (DOI 10.2455/TUKEY_T_TEST.4E800CB597F00B0A2CEF6D50479867EF).

Table 14. Ad-Hoc Bilingual English. The table shows the first ten groups of the Tukey T Test.

Run ID	Groups									
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDTOGGLEFB10D10T	X									
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTTOGGLEFB10D10T	X	X								
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDTOGGLE	X	X	X							
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDDESYNFB10D10T	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAWOTD	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDINTERSECTIONUNIONSYNFB5D10T	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDDESYNFB5D10T	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.JHU-APL.APLBIIDENTDS	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDINTERSECTIONUNIONSYNFB10D10T	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTTOGGLE	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDDESYNFB10D5T	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAUWOTD	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAUWOT	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.JHU-APL.APLBIIDENTD5	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.JHU-APL.APLBIIDENTD4	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAWOT	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.OMTD07	X	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.OMTDN07	X	X	X	X	X					
10.2415/AH-BILI-X2EN-CLEF2007.JHU-APL.APLBIIDENTDW	X	X	X	X	X					
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_DICE	X	X	X	X	X	X				
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDINTERSECTIONUNIONSYN	X	X	X	X	X	X	X			
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAWTD	X	X	X	X	X	X	X	X		
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.COOTD	X	X	X	X	X	X	X	X		
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_PMI	X	X	X	X	X	X	X	X	X	
10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.OMT07	X	X	X	X	X	X	X	X	X	
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.ALLTRANSTD	X	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAUWTD	X	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAUWT	X	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAWT	X	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.COOT	X	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLE_DICE		X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.ALLTRANST		X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.TETD		X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.2007_LM_ALL_CROSS_1000_POSSCORES		X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_MAR_TITLE_DICE		X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.2007_RBLM_ALL_CROSS_1000_POSSCORES			X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.RBLM_ALL_CROSS_TD_1000_POSSCORES			X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.NOST_OMTDN07			X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLE_PMI			X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.LM_ALL_CROSS_TD_1000_POSSCORES			X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_MAR_TITLE_PMI				X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI6					X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI2					X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI4					X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI1					X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI3					X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI5					X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.HITD						X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_6							X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_4								X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_2									X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_5								X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_1								X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_3								X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.AHBILITE2ENR1									X	X
10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.AHBILIHI2ENR1										X
10.2415/AH-BILI-X2EN-CLEF2007.STOCKHOLM.DSV_AMH_BLNG_RUN8										X

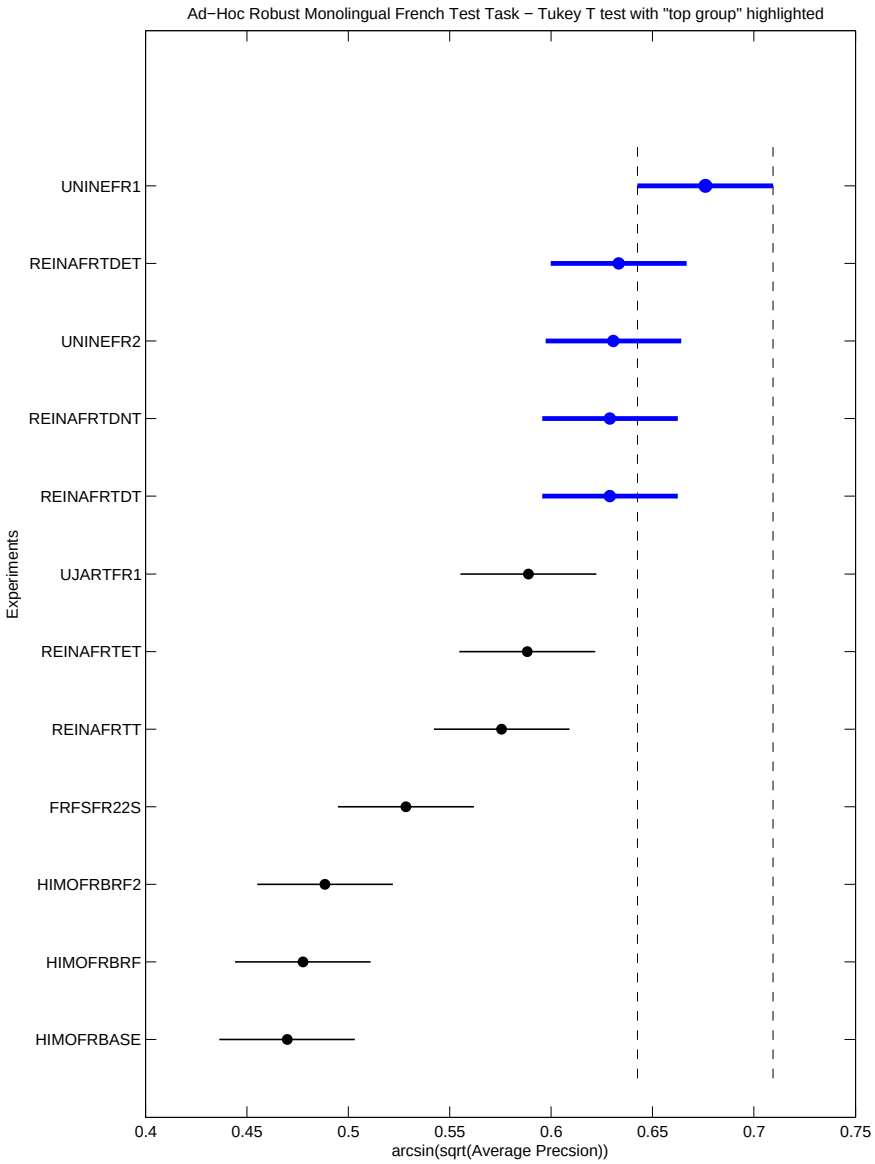
Table 15. Ad-Hoc Bilingual English. The table shows the last nine groups of the Tukey T Test.

Run ID	Groups								
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAUWT	X								
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAWT	X								
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.COOT	X								
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLE_DICE	X	X							
10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.ALLTRANST	X	X							
10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.TETD	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.2007_LM_ALL_CROSS_1000_POSSCORES	X	X	X						
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_MAR_TITLE_DICE	X	X	X	X					
10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.2007_RBLM_ALL_CROSS_1000_POSSCORES	X	X	X	X	X				
10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.RBLM_ALL_CROSS_TD_1000_POSSCORES	X	X	X	X	X				
10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.NOST_OMTDN07	X	X	X	X	X	X			
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLE_PMI	X	X	X	X	X	X			
10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.LM_ALL_CROSS_TD_1000_POSSCORES	X	X	X	X	X	X			
10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_MAR_TITLE_PMI	X	X	X	X	X	X	X		
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI6	X	X	X	X	X	X	X		
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI2	X	X	X	X	X	X	X		
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI4	X	X	X	X	X	X	X		
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI1	X	X	X	X	X	X	X		
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI3	X	X	X	X	X	X	X		
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_WIKI5	X	X	X	X	X	X	X		
10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.HITD	X	X	X	X	X	X	X	X	
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_6	X	X	X	X	X	X	X	X	
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_4	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_2	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_5	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_1	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.BUDAPEST-ACAD.BILING_3	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.AHBILITE2ENR1	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.AHBILIHI2ENR1	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.STOCKHOLM.DSV_AMH_BLNG_RUN8	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.AHBILIBN2ENR1	X	X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.STOCKHOLM.DSV_AMH_BLNG_RUN6		X	X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.STOCKHOLM.DSV_AMH_BLNG_RUN7			X	X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.STOCKHOLM.DSV_AMH_BLNG_RUN4				X	X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.STOCKHOLM.DSV_AMH_BLNG_RUN2					X	X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.STOCKHOLM.DSV_AMH_BLNG_RUN5						X	X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.BENGALITITLEDESC							X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.BENGALITITLEDESCNARR							X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.STOCKHOLM.DSV_AMH_BLNG_RUN1							X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.STOCKHOLM.DSV_AMH_BLNG_RUN3							X	X	X
10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.HINDITITLEDESCNARR								X	X
10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.HINDITITLE								X	X
10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.BENGALITITLE								X	X
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDENGFB5D5T10D19T								X	X
10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.HINDITITLEDESC								X	X
10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDENGPSEUDOTRANS20D14T									X



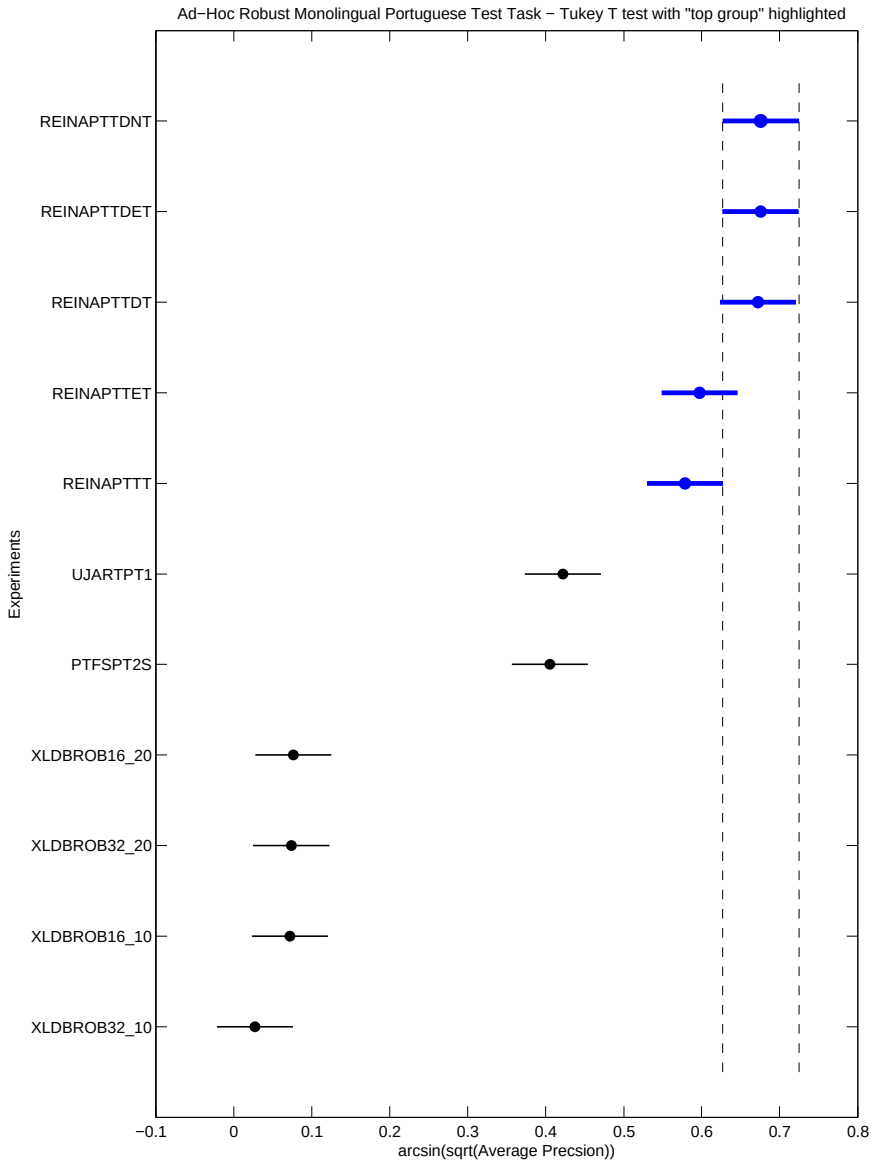
Experiment DOI	Groups		
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.REINA.REINAENTDT	X		
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.REINA.REINAENTDNT	X		
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.REINA.REINAENTDET	X		
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.DAEDALUS.ENFSEN22S	X	X	
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.REINA.REINAENTT	X	X	
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.REINA.REINAENTET		X	
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.HILDESHEIM.HIMOENBRFNE			X
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.HILDESHEIM.HIMOENBRF2			X
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.HILDESHEIM.HIMOENBRF1			X
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.HILDESHEIM.HIMOENNE			X
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.HILDESHEIM.HIMOENBASE			X

Fig. 16. Robust Monolingual English. Experiments grouped according to the Tukey T Test (DOI 10.2455/TUKEY_T_TEST.90F31E2BBDA52E383201421CD2207C37).



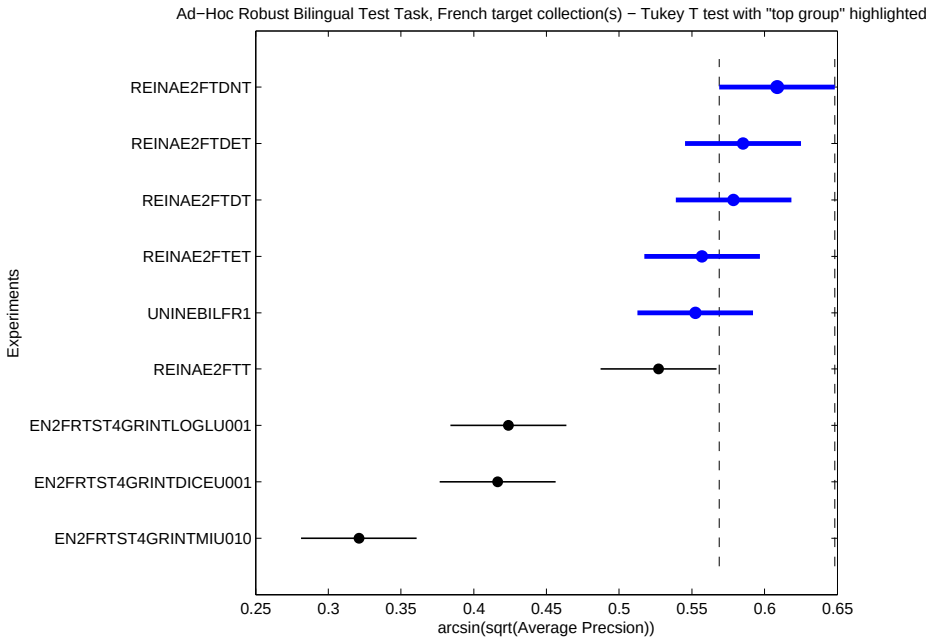
Experiment DOI	Groups			
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.UNINE.UNINEFR1	X			
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.REINA.REINAFRTDET	X	X		
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.UNINE.UNINEFR2	X	X		
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.REINA.REINAFRTDNT	X	X		
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.REINA.REINAFRTDT	X	X		
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.JAEN.UJARTFR1		X	X	
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.REINA.REINAFRTET		X	X	
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.REINA.REINAFRTT		X	X	
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.DAEDALUS.FRFSFR22S			X	X
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.HILDESHEIM.HIMOFRBRF2				X
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.HILDESHEIM.HIMOFRBRF				X
10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.HILDESHEIM.HIMOFRBASE				X

Fig. 17. Robust Monolingual French. Experiments grouped according to the Tukey T Test (DOI 10.2455/TUKEY_T_TEST.66736BFA0C417F72BA727E3EF6324986).



Experiment DOI	Groups		
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.REINA.REINAPTTDNT	X		
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.REINA.REINAPTTDET	X		
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.REINA.REINAPTTDT	X		
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.REINA.REINAPTTET	X		
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.REINA.REINAPTTT	X		
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.JAEN.UJARTPT1		X	
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.DAEDALUS.PTFSPT2S		X	
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.XLDB.XLDBROB16_20			X
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.XLDB.XLDBROB32_20			X
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.XLDB.XLDBROB16_10			X
10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.XLDB.XLDBROB32_10			X

Fig. 18. Robust Monolingual Portuguese. Experiments grouped according to the Tukey T Test (DOI 10.2455/TUKEY_T_TEST.CBF851699C4817B013FA33838640008A).



Experiment DOI	Groups		
10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.REINA.REINAE2FTDNT	X		
10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.REINA.REINAE2FTDET	X	X	
10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.REINA.REINAE2FTDT	X	X	
10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.REINA.REINAE2FTET	X	X	
10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.UNINE.UNINEBILFR1	X	X	
10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.REINA.REINAE2FTT		X	
10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.COLESUN.EN2FRTST4GRINTLOGLU001			X
10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.COLESUN.EN2FRTST4GRINTDICEU001			X

Fig. 19. Robust Bilingual French. Experiments grouped according to the Tukey T Test (DOI 10.2455/TUKEY_T_TEST.D284127041AE7A69919B3C09EBA3769F).

7 Conclusions

We have reported the results of the ad hoc cross-language textual document retrieval track at CLEF 2007. 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 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. This year, the monolingual task focused on central European languages while the bilingual task included an activity for groups that wanted to use non-European topic languages and languages with few processing tools and resources. Each year, we also include a task aimed at examining particular aspects of cross-language text retrieval. Again this year, the focus was examining the impact of "hard" topics on performance in the "robust" task.

The paper also describes in some detail the creation of the pools used for relevance assessment this year. We still have to do stability tests on these pools; the results will be published in the CLEF 2007 post-workshop Proceedings.

Although there was quite a good participation in the monolingual Bulgarian, Czech and Hungarian tasks and the experiments report some interesting work on stemming and morphological analysis, we were very disappointed by the lack of participation in bilingual tasks for these languages. On the other hand, the interest in the task for non-European topic languages was encouraging and the results reported can be considered positively. We are currently undecided about the future of the main mono- and cross-language tasks in the ad hoc track; this will be a topic for discussion at the breakout session during the workshop.

The robust task has analyzed the performance of systems for older CLEF data under a new perspective. A larger data set which allows a more reliable comparative analysis of systems was assembled. Systems needed to avoid low performing topics. Their success was measured with the geometric mean (GMAP) which introduces a bias on poor performing topics. Results for the robust task for mono-lingual retrieval or English, French and Portuguese as well as for bilingual retrieval from English to French are reported. Robustness can also be interpreted as the fitness of a system under a variety of conditions. The definition on what robust retrieval means has to continue. All participants in CLEF 2007 are invited to engage in the discussion of the future of the robust task.

The test collections for CLEF 2000 - CLEF 2003 are now publicly available on the *Evaluations and Language resources Distribution Agency (ELDA)* catalog[8].

8 Acknowledgements

We should like to acknowledge the enormous contribution of the groups responsible for topic creation and relevance assessment. In particulat, we thank the group responsible for the work on Bulgarian led by Kiril Simov and Petya Osenova,

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

the group responsible for Czech led by Pavel Pecina[9], and the group responsible for Hungarian led by Tamás Váradi and Gergely Bottyán. These groups worked very hard under great pressure in order to complete the heavy load of relevance assessments in time.

References

1. Paskin, N., ed.: The DOI Handbook – Edition 4.4.1. International DOI Foundation (IDF). http://dx.doi.org/10.1000/186 [last visited 2007, August 30] (2006)
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. Tomlinson, S.: Sampling Precision to Depth 10000: Evaluation Experiments at CLEF 2007. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
4. 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
5. Di Nunzio, G.M., Ferro, N.: Appendix A: Results of the Core Tracks – Ad-hoc Bilingual and Monolingual Tasks. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
6. Di Nunzio, G.M., Ferro, N.: Appendix B: Results of the Core Tracks – Ad-hoc Robust Bilingual and Monolingual Tasks. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
7. Dolamic, L., Savoy, J.: Stemming Approaches for East European Languages. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
8. Noguera, E., Llopis, F.: Applying Query Expansion techniques to Ad Hoc Monolingual tasks with the IR-n system. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
9. Majumder, P., Mitra, M., Pal, D.: Hungarian and Czech Stemming using YASS. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
10. Češka, P., Pecina, P.: Charles University at CLEF 2007 Ad-Hoc Track. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
11. Ircing, P., Müuller, L.: Czech Monolingual Information Retrieval Using Off-The-Shelf Components – the University of West Bohemia at CLEF 2007 Ad-Hoc track. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)

[9] Ministry of Education of the Czech Republic, project MSM 0021620838

12. Hayurani, H., Sari, S., Adriani, M.: Evaluating Language Resources for CLEF 2007. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
13. Zhou, D., Truran, M., Brailsford, T.: Ambiguity and Unknown Term Translation in CLIR. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
14. Argaw, A.A.: Amharic-English Information Retrieval with Pseudo Relevance Feedback. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
15. Schönhofen, P., Benczúr, A., Bíró, I., Csalogány, K.: Performing Cross-Language Retrieval with Wikipedia. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
16. Tune, K.K., Varma, V.: Oromo-English Information Retrieval Experiments at CLEF 2007. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
17. Bandyopadhyay, S., Mondal, T., Naskar, S.K., Ekbal, A., Haque, R., Godavarthy, S.R.: Bengali, Hindi and Telugu to English Ad-hoc Bilingual task at CLEF 2007. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
18. Jagarlamudi, J., Kumaran, A.: Cross-Lingual Information Retrieval System for Indian Languages. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
19. Chinnakotla, M.K., Ranadive, S., Bhattacharyya, P., Damani, O.P.: Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
20. Pingali, P., Varma, V.: IIIT Hyderabad at CLEF 2007 – Adhoc Indian Language CLIR task. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
21. Mandal, D., Dandapat, S., Gupta, M., Banerjee, P., Sarkar, S.: Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources. In Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2007 Workshop, http://www.clef-campaign.org/ [last visited 2007, September 5] (2007)
22. Robertson, S.: On GMAP: and Other Transformations. In Yu, P.S., Tsotras, V., Fox, E.A., Liu, C.B., eds.: Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006), ACM Press, New York, USA (2006) 78–83
23. Voorhees, E.M.: The TREC Robust Retrieval Track. SIGIR Forum **39** (2005) 11–20
24. Savoy, J.: Why do Successful Search Systems Fail for Some Topics. In Cho, Y., Wan Koo, Y., Wainwright, R.L., Haddad, H.M., Shin, S.Y., eds.: Proc. 2007 ACM Symposium on Applied Computing (SAC 2007). ACM Press, New York, USA (2007) 872–877
25. Sanderson, M., Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. 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) 162–169

26. 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
27. Conover, W.J.: Practical Nonparametric Statistics. 1st edn. John Wiley and Sons, New York, USA (1971)
28. 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)
29. 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