# GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview

Thomas Mandl[1], Paula Carvalho[2], Fredric Gey[3],
Ray Larson[3], Diana Santos[4], Christa Womser-Hacker[1]

[1] Information Science, University of Hildesheim, GERMANY
mandl@uni-hildesheim.de, womser@uni-hildesheim.de
[3] University of California, Berkeley, CA, USA
gey@berkeley.edu , ray@sims.berkeley.edu
[2] University of Lisbon, DI, LasiGE, XLDB
Linguateca, PORTUGAL
pqfcarvalho@gmail.com
[4] Linguateca, SINTEF ICT, NORWAY
Diana.Santos@sintef.no

WITH

Giorgio Di Nunzio[5], Nicola Ferro[5]
[5] Department of Information Engineering, University of Padua, Italy
{dinunzio|ferro}@dei.unipd.it

**Abstract**

GeoCLEF is an evaluation initiative for testing queries with a geographic specification in large set of text documents. GeoCLEF ran a regular track for the third time within the Cross Language Evaluation Forum (CLEF) 2008. The purpose of GeoCLEF is to test and evaluate cross-language geographic information retrieval (GIR). GeoCLEF 2008 consisted of two sub tasks. A search task ran for the third time and a Wikipedia pilot task (GikiP) was organized for the first time. For the GeoCLEF 2008 search task, twenty-five search topics were defined by the organizing groups for searching English, German and Portuguese document collections. Topics were developed also for English, German and Portuguese. Many topics were geographically challenging. Eleven groups submitted 131 runs. The groups used a variety of approaches, including sample documents, named entity extraction and ontology based retrieval. The evaluation methodology and results are presented in the paper.

## 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

## General Terms

Measurement, Performance, Experimentation

## Keywords

Multilingual Information Retrieval, Geographic Information Retrieval, Evaluation Benchmarks, Retrieval Experiments.

# 1 Introduction

The Internet has brought a large number of geographic services which range from map services to route planning and hotel reservation systems. Many queries for search engines are of geographic nature. The development and the evaluation of information retrieval systems which allow and optimize the geographically oriented access to information is of high practical relevance.

Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Many documents contain some kind of spatial reference which may be important for IR. For example, to retrieve, rank and visualize search results based on a spatial dimension (e.g. "find me news stories about riots near Paris and their consequences").

GeoCLEF is the first track at an evaluation campaign dedicated to evaluating geographic information retrieval systems. The aim of GeoCLEF is to provide the necessary framework in which to evaluate GIR systems for search tasks involving both spatial and multilingual aspects. Participants are offered a TREC style ad hoc retrieval task based on existing CLEF newspaper collections. GeoCLEF 2005 was run as a pilot track in 2005 and in since 2006, GeoCLEELF was has been a regular CLEF track. GeoCLEF has continued to evaluate retrieval of documents with an emphasis on geographic information retrieval from text. As such, searches with a geographic condition require the combination of spatial information and content based relevance into one result. Often, spatial reasoning is necessary to solve the search tasks.

Eleven research groups (13 in 2007) from a variety of backgrounds and nationalities submitted 131 runs (108 in 2007) to GeoCLEF.

For 2008, Portuguese, German and English were available as document and topic languages. As in previous years, there were two Geographic Information Retrieval tasks: monolingual (English to English, German to German and Portuguese to Portuguese) and bilingual (language X to language Y, where X or Y was one of English, German, or Portuguese).

GeoCLEF developed a standard evaluation collection which supports long-term research. Altogether, 100 topics including relevance assessments have been developed over the four last years (one pilot run and the three regular tracks). Another set of 26 CLEF ad-hoc topics with spatial restrictions has been identified and can also be used as a benchmark. It is intended to make the topics and the relevance judgment files publicly available on the GeoCLEF website[1].

**Table 1.** GeoCLEF test collection – collection and topic languages

| GeoCLEF Year | Collection Languages | Topic Languages |
|---|---|---|
| 2005 (pilot) | English, German | English, German |
| 2006 | English, German, Portuguese, Spanish | English, German, Portuguese, Spanish, Japanese |
| 2007 | English, German, Portuguese | English, German, Portuguese |
| 2008 | English, German, Portuguese | English, German, Portuguese |

Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Many documents contain some kind of spatial reference which may be important for IR. For example, to retrieve, rank and visualize search results based on a spatial dimension (e.g. "find me news stories about riots near Paris and their consequences").

Many challenges of geographic IR involve geographical references which are often vague, ambiguous multi-lingually challenging [2, 3, 9]. Ambiguity e.g. can occur as homonymy between a geographical and a non-geographical term (*Santos* as a Brazilian city and *santos* as a Brazilian and Spanish word for saints) or between several places (e.g. *Neustadt*, *Albertville*) and even as combination of the two (e.g. *Halle* is the name of two mid-sized German cities and a word for hall or gym). Some references to places are multi word groups and need to be recognized (e.g. *Rio Grande do Sul*, *Newcastle upon Tyne*). Multi-lingual retrieval requires systems to match references to a place from one language to another. Sometimes these may differ (e.g. *Athens*, *Athen*, *Atenas*, *Atina*) and sometimes not. Often spatial reasoning is necessary to solve information needs (e.g. demonstrations in cities in *Northern Germany*).

The GeoCLEF track comprises two sub tasks. The main search task with newspaper collections is described in the following sections. The GikiP task[2] evaluates searches for Wikipedia entries which require some geographical processing. It is describes in a separate overview paper [11].

[1] http://www.uni-hildesheim.de/geoclef/

[2] http://www.linguateca.pt/GikiP/

# 2 GeoCLEF 2008 Search Task

The geographic search task is the main task of GeoCLEF and it is developed like the CLEF ad-hoc task. The following sections describe the test design adopted by GeoCLEF.

## 2.1 Document Collections used in GeoCLEF 20087

The document collections for this year's GeoCLEF experiments areis identical to the ones used for GeoCLEF 20076,. It consists of newspaper and newswire stories from the years 1994 and 1995 used in previous CLEF ad-hoc evaluations [1]. The collections contain stories covering international and national news events, therefore representing a wide variety of geographical regions and places. The English document collection consists of 169,477 documents and was composed of stories from the British newspaper The Glasgow Herald (1995) and the American newspaper The Los Angeles Times (1994). The German document collection consists of 294,809 documents from the German news magazine Der Spiegel (1994/95), the German newspaper Frankfurter Rundschau (1994) and the Swiss newswire agency Schweizer Depeschen Agentur (SDA, 1994/95). For Portuguese, two newspaper collections were utilized: *Público* (106,821 documents) and *Folha de São Paulo* (103,913 documents). Both are major daily newspapers in their countries. The Portuguese collections are also distributed for IR and NLP research by Linguateca as the CHAVE collection[3] [6].

In all three collections, the documents have a common structure: newspaper-specific information like date, (optionally) page, issue, special filing numbers and usually one or more titles, a byline and the actual text. The document collections were not geographically tagged and contained no semantic location-specific information. Although the Portuguese collection can also be found in a fully parsed version, we have no information that this has been used by any participant.

**Table 2.** GeoCLEF 20087 test collection size

| Language | English | German | Portuguese |
|---|---|---|---|
| Number of documents | 169,477 | 294,809 | 210,734 |

## 2.2 Generating Search Topics

A total of 25 topics were generated for this year's GeoCLEF (GC76 - GC100). Topic creation was shared among the Portuguese and the German groups, who created all topics utilizing the DIRECT System provided by the University of Padua. As in the past years, DIRECT included a search utility for the collections. Topics are meant to express a natural information need which a user of the collection might have.

Topic creation was performed in two stages. First, each group devised a set of candidate topics in their own language, whose appropriateness was checked in the text collection available for that language. The first choice was based on monolingual criteria These topic candidates were subsequently checked for relevant documents in the other collections. In many cases, it is difficult to find geographically interesting topics below the granularity of a country. Regional events with much coverage in one country do not often lead to many newspaper articles in other countries. As a consequence, the topics need to be partially modified or refined by relaxing or tightening the content or the geographic focus. Reasons driving this process were the absence of relevant documents in one of the languages, the complexity of topic interpretation and/or even translation into the other languages. For example, a candidate topic on fish on the Iberian Peininsula had relevant matches in the Portuguese collection. But not only did it lack matches in the other language newspapers but some of the species (e.g. *"saramugo*") were very hard to translate into German or English. The fish called *saramugo* seems to a species which lives only in Spanish and Portuguese rivers. Consequently, the spatial parameter (*Iberian Peninsula*) remained, but the subject was replaced by a matter more prone to attention from international mass media, (directly or indirectly) related to political and/or economical issues: in this case the state of agriculture. It should be stressed, however, that, in the majority of cases, the changes were not so extreme, corresponding to a simple topic extension. For instance, the initial candidate topic "Nobel Prize winners in Physics from Northern European countries" was replaced by a more general and simpler one: "Nobel prize winners from Northern European countries". In other cases we replaced the geographic term by other(s) involving a more difficult (but also more interesting) exercise of geographic reasoning and processing, as "Most visited sights in the capital of

[3] http://www.linguateca.pt/CHAVE/

France" gave way to the topic: "Most visited sights in the capital of France and its vicinity", which is definitively much more challenging in the geographic point of view.

The final topic set was agreed upon after intensive discussion. Finally, all missing topics were translated into Portuguese and German and all translations were checked. The following section will discuss the creation of topics with spatial parameters for the track.

One goal of GeoCLEELF is the creation of a geographically challenging topic set. Geographic knowledge should be necessary to successfully retrieve relevant documents for most documents. While many geographic searches may be well served by keyword approaches, others require a profound geographic reasoning. We speculate that most systems and especially keyword based systems might perform better faces with a realistic topic set where these difficulties might occur less frequently.

In order to increase the difficulty of the topic set, several issues were explicitly included in the topics of GeoCLEF 20087. Some of them are the following ones:

- vague geographic regions (Sub-Saharan Africa , Western Europe )
- geographical relations beyond IN (forest fires on Spanish islands)
- granularity below the country level (Industrial or cultural fairs in Lower Saxony)
- terms which do not occur in documents (Portuguese communities in other countries, demonstrations in German cities)

We aimed at creating a set of topics representing different kinds of geographic queries that present different levels of complexity and may require different kinds of approaches to process them adequately and to successfully retrieve relevant documents. Rather than privileging specific geographical places, like a country or city, preference was given to reference to geographical regions, comprehending more than one physical or administrative place. Different kinds of regions were, then, considered, which may correspond, for instance, to a delimited geographical area of a given continent (e.g. Sub-Saharan Africa, Southeast Asia, Northern Africa, Western Europe) or country (e.g. Western USA, Lower Saxony, Spanish islands). Other interesting geo-economic-political terms, such as OCDE countries, were also considered in the topic creation.

The majority of the GeoCLEF 2008 topics specify complex (multiply defined) geographical relations, a property introduced in the GeoCLEF 2007 [8] kept in this evaluation. Such geographical relations, which can be explicitly or implicitly mentioned in the topic, may represent, for instance:

- Proximity (e.g. Most visited sights in the capital of France and its vicinity);
- Inclusion (e.g. Attacks in Japanese subways);
- Exclusion (e.g. Portuguese immigrant communities in the world).

Notice that the example illustrated in (i) also presents a relation of inclusion (established between sights and capital of France, and which is explicitly formalized by the preposition "in"). That relation can also be inferred in the NP Japanese subways occurring in the topic illustrated in (ii), which can be paraphrase by the expression "subways in Japan".

Contrarily to the topic creation performed last year, which contained explicit relations of exclusion (e.g. Europe excluding the Alps), such kind of relations were only implicitly represented in the topics of GeoCLEF 2008, as illustrated in (iii). This topic is indeed interesting because it refers simultaneously to an inclusion (communities from Portugal in the world) and exclusion (the generic geographical term world must be interpreted, in this context, as the entire world excluding Portugal).

We also chose to use vague geographic designations for certain topics, as done in previous GeoCLEF editions. For example, in the topic: Nuclear tests in the South Pacific, the geographical term South Pacific may refer both Australasia ("an an area including Australia, New Zealand, New Guinea and other islands including the eastern part of Indonesia") and Oceania ("a geographical (often geopolitical) region of many countries/territories (mostly islands) in the southern Pacific Ocean"). The interpretation of this geographical term (namely, in what concerns the determination if it refers to a land place or to the ocean) is only possible if we consider the full topic content.

A similar situation is observed in the topic "American troops in the Persian Gulf". In this case, the Persian Gulf does not stand for the gulf itself but to a Southwest Asian region, which is an extension of the Indian Ocean located between Iran and the Arabian Peninsula. Once again, the term disambiguation is only possible if taking into account the rest of the information described in the topic.

Another case of intended vagueness is the topic Environmental pollution in European waters, where the term waters can refer to rivers, lakes or sea

We repeat here that no markup of geographic entities in the topics was provided. Systems were expected to reveal that information be themselves from the topic which resembles a more realistic task. However, it was

difficult to develop topics which fulfilled all criteria. For example, local and regional events which allow queries on a low level of granularity and which require local knowledge often do not result in newspaper articles outside the national press. This impedes made cross-lingual topic development hard.

### 2.3 Format of Topic Description

The format of GeoCLEF 20087 is the same as that of 2006 and 2007. No explicit geographic information was given. Two examples for full topics are shown in table 3.

**Tab. 3:** Topics GC08958 and GC08475

| | |
|---|---|
| <num>10.2452/89-GC</num><br><title>Trade fairs in Lower Saxony </title><br><desc>Documents reporting about industrial or cultural fairs in Lower Saxony. </desc><br><narr>Relevant documents should contain information about trade or industrial fairs which take place in the German federal state of Lower Saxony, i.e. name, type and place of the fair. The capital of Lower Saxony is Hanover. Other cities include Braunschweig, Osnabrück, Oldenburg and Göttingen. </narr><br></top> | <num>10.2452/84-GC</num><br><title>Atentados à bomba na Irlanda do Norte </title><br><desc>Os documentos relevantes mencionem atentados bombistas em localidades da Irlanda do Norte </desc><br><narr>Documentos relevantes devem mencionar atentados à bomba na Irlanda do Norte, indicando a localização do atentado. </narr><br></top> |

As can be seen, after the brief descriptions within the title and description tags, the narrative tag contains detailed description of the geographic detail sought and the relevance criteria. In some topics, lists of relevant geographic names are given.

### 2.4 Approaches to Geographic Information Retrieval

In the last three editions of GeoCLEF, traditional ad-hoc retrieval approaches and specific geographic reasoning systems have been explored in parallel. Good success has often been achieved by ad-hoc techniques without any specific geographic knowledge or processing. These approaches have sometimes been developed as a baseline for more sophisticated systems. It has been observed that some of the traditional techniques may have effects which are beneficial for geographic search tasks. Blind relevance feedback can lead to a geographic term expansion necessary to solve a search problem. For example, a query for *riots in German cities* does not contain the name of any German city. A query including the term *German* may lead to documents which contain the word *German* as well as the names of some cities which can be included in subsequent optimized queries. As a result, geographic term expansion has been achieved without proper geographic knowledge being available for the system. This form of pseudo-geographic processing is obviously not very reliable, but the specific components often have a high error rate as well or do introduce too much noise. In GeoCLEF 2007, some systems tried combinations of the two approaches and the dedicated geographic systems had further matured. This year, new ideas have been introduced and an ontology based approach by the DFKI has been very successful for the most competitive task which is monolingual English. On the other hand, the University of Berkeley implemented a system designed like an ad-hoc system without any geographic elements and achieved excellent results for monolingual and bilingual tasks.

The participants used a wide variety of approaches to the GeoCLEF tasks, ranging from basic IR approaches to deep natural language processing (NLP) processing. The approaches include the use of full documents for ranking the result set, map based techniques and Wikipedia as a knowledge source. For more details we refer the reader to the description of the systems in the papers of the participants in this volume.

### 2.5 Relevance assessment

English assessment was shared by Berkeley and Hildesheim Universities. German assessment was done by the University of Hildesheim and Portuguese assessment by Linguateca. The DIRECT System already used for topic development, was also utilized for assessment. The system provided by the University of Padua allowed the

automatic submission of runs by participating groups and supported assembling the GeoCLEF assessment pools by language.

**Table 4.** GeoCLEF 2008 Size of Pools

| Language | # Documents |
|---|---|
| English | 14.528 |
| German | 15.081 |
| Portuguese | 14.780 |

**Table 5.** GeoCLEF 2008 Relevant Documents per Topic

| Language | Min | Max |
|---|---|---|
| English | 0 | 109 |
| German | 1 | 146 |
| Portuguese | 2 | 158 |

During the assessment process, the assessor tried to find the best collection of keywords – based on the detailed information in the narrative and his/her knowledge of the geographical concepts and subjects involved – and queried the DIRECT system. Some of the issues are reported in the following section for each language.

### 2.5.1 Portuguese Relevance Assessment

There were some topics which caused assessment difficulties: especially when the narrative required specific information. For example, given a sentence: Bonn ... former chancellor Willy Brandt ... Nobel Peace prize winner... Is this enough to infer that Willy Brandt was German? Opinions diverged but a consensus was reached. One topic that caused difficulties was the translation of Portuguese "monumentos" by English "sights" (or vice versa) when Euro Disney was to be assessed. It cannot by any means refer in Portuguese to such, but it can obvioulsy do that in the much more general English expression "sights". Also the German word *Sehenswürdigkeiten* (literally “things worth seeing”) can refer to more than “monuments” (*Monumente* in German). One could also subsume Euro Disney under this German term for “sights”. For the assessment, the more restrictive Portuguese interpretation was adopted for the pools in German and English. Discussing these issues among a geographically highly distributed group under serious time constraints which are set for the assessment is a challenge.

In assessments, topic drifts typically occur. This was also the case in the GeoCLEF 2008 assessment. Is a document about kidnapping of a French aid worker in Kenya relevant for "foreign aid in Sub-Saharan Africa"? On the one hand, his existence is a proof of the existence of foreign aid, but on the other hand, his kidnapping is not foreign aid.

### 2.5.2 German relevance assessment

Often, the assessment provides hints on why systems failed. This was the case for the German topic about fairs in Lower Saxony. The German word for fails (Messe) was matched against similar words which have a different meaning (e.g. angemessen -> appropriate, Messer -> knife). This may be due to inappropriate stemming rules or due to high n-gram similarity.

### 2.5.3 English Assessment

The English document pool also led to borderline cases for the assessors which needed to be discussed intensively. One topic required documents on natural disasters in Western states of the USA. Some documents only reported about the insurance costs caused by natural disaster overall. These were only considered relevant when the mentioned a geographically relevant place (sometimes they mentioned *Los Angeles*) even when they did not mention the disaster explicitly and directly.

# 3 GeoCLEF 2008 Results

The results of the participating groups are reported in the following sections.

## 3.1 Participants and Experiments

As shown in Table 6, a total of 11 groups from seven different countries submitted results for one or more of the GeoCLEF tasks. A total of 131 experiments (runs) were submitted. Five of these groups participated in GeoCLEF for the first time.

**Table 6.** GeoCLEF 2008 participants – new groups are indicated by *

| Participant | Institution | Country |
|---|---|---|
| alivale* | U.Jaén & U.Politecnica Valencia | Spain |
| cheshire | U.C.Berkeley | United States |
| Csusm | Cal. State U.- San Marcos | United States |
| dfki* | German Research Center for AI | Germany |
| Hagen | U.Hagen-Comp.Sci | China |
| icl | Imperial College London | United Kingdom |
| Inaoe* | Lab. Tecnologıas del Lenguaje | Mexico |
| jaen* | U.Jaén | Spain |
| pittsburgh* | U.Chengdu & U.Pittsburgh, | China & United States |
| valencia | U.Politecnica Valencia | Spain |
| xldb | U.Lisbon | Portugal |

Table 7 reports the number of participants by their country of origin.

**Table 7.** GeoCLEF 2008 participants by country

| Country | # Participants |
|---|---|
| China | 1 |
| Germany | 2 |
| Mexico | 1 |
| Portugal | 1 |
| Spain | 3 |
| United Kingdom | 1 |
| United States | 3 |

Table 8 provides a breakdown of the experiments submitted by each participant for each of the offered tasks.

**Table 8.** GeoCLEF 2008 experiments by task

| Participant | Monolingual Tasks | | | Bilingual Tasks | | | TOTAL |
|---|---|---|---|---|---|---|---|
| | DE | EN | PT | X2DE | X2EN | X2PT | |
| alivale* | | 9 | | | | | **9** |
| cheshire | 3 | 3 | 3 | 6 | 6 | 6 | **27** |
| csusm | 1 | 1 | 2 | 1 | 1 | 1 | **7** |
| dfki* | | 5 | | | | | **5** |
| hagen | 5 | | | 10 | | | **15** |
| icl | | 9 | | | | | **9** |
| inaoe* | | 12 | | | | | **12** |
| jaen* | | 7 | | | 6 | | **13** |
| pittsburgh* | | 4 | | | | | **4** |
| valencia | | 6 | | | | | **6** |
| xldb | | 12 | 12 | | | | **24** |
| **TOTAL** | **9** | **68** | **17** | **17** | **13** | **7** | **131** |

Three different topic languages were possible for the GeoCLEF bilingual experiments. Again, the most popular language for queries was English; German took the second place. The number of bilingual runs by topic language is shown in Table 9.

**Table 9.** GeoCLEF 2008 Bilingual experiments by topic language

| Track | Source Language | | | TOTAL |
|---|---|---|---|---|
| | DE | EN | PT | |
| Bilingual X2DE | | 10 | 7 | **17** |
| Bilingual X2EN | 4 | | 3 | **7** |
| Bilingual X2PT | 7 | 6 | | **13** |
| **TOTAL** | **11** | **16** | **10** | **27** |

### 3.2 Monolingual Experiments

Monolingual retrieval was offered for the following target collections: English, German, and Portuguese. Figures 1 to 3 show the interpolated recall vs. average precision for the top participants of the monolingual tasks.

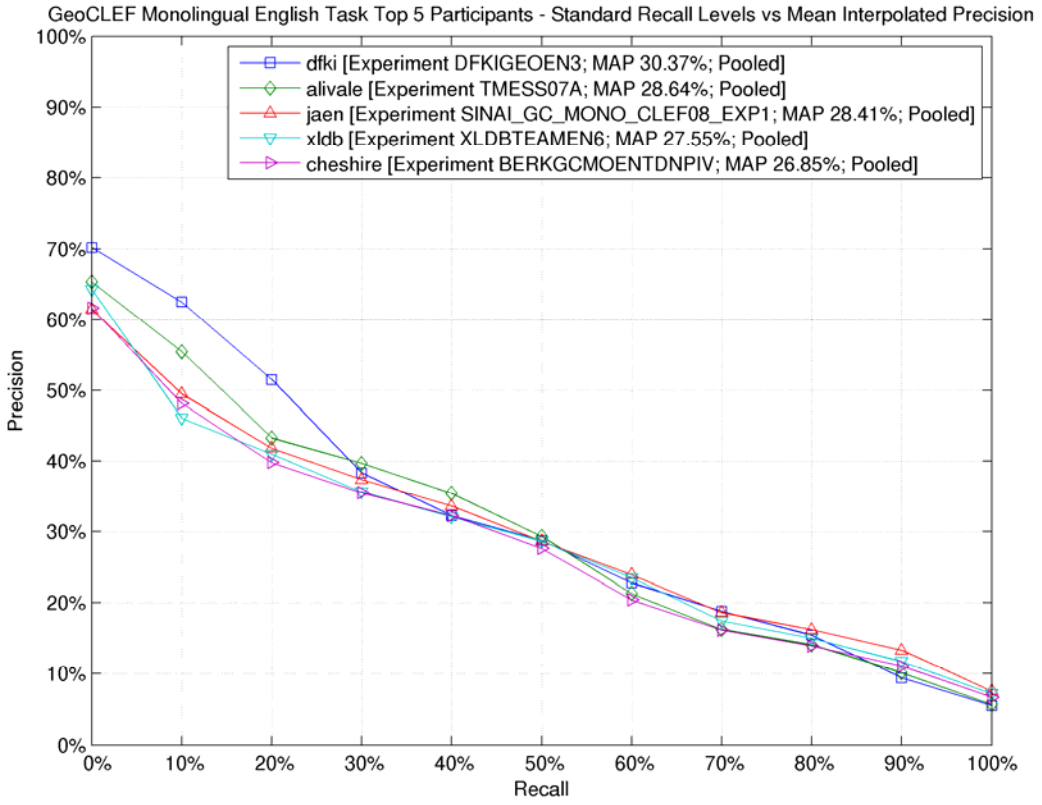



**Fig. 1.** Monolingual English top participants. Interpolated Recall vs. Average Precision.

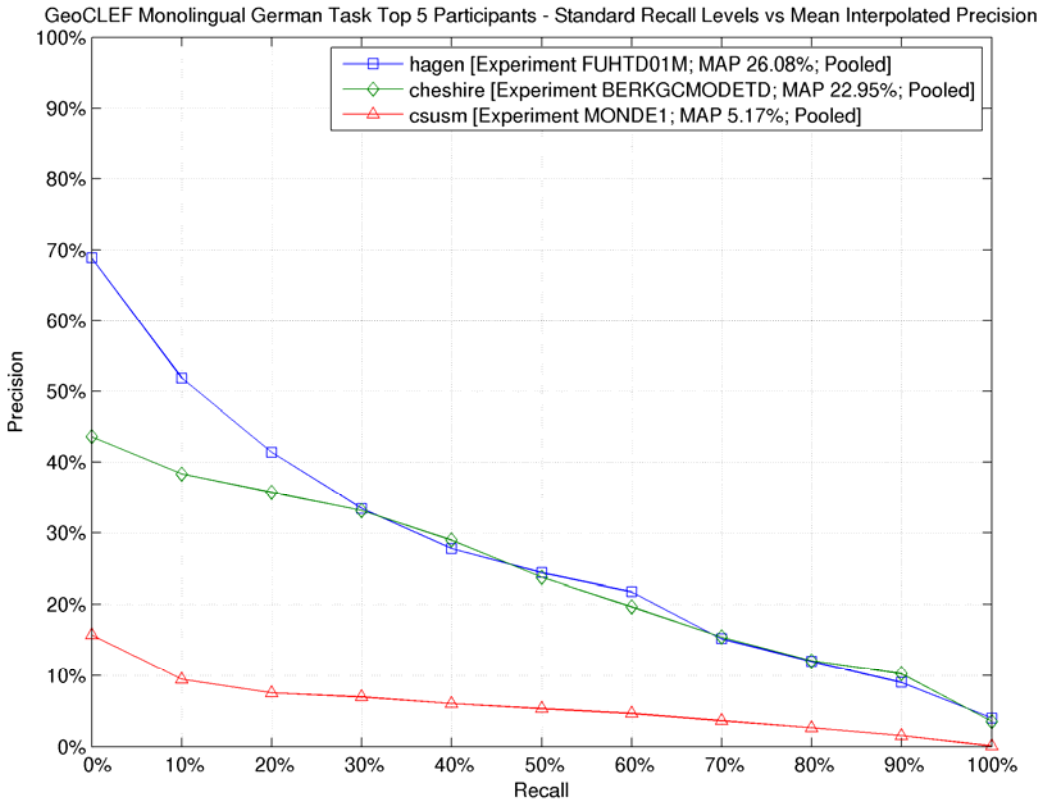



**Fig. 2.** Monolingual German top participants. Interpolated Recall vs. Average Precision.

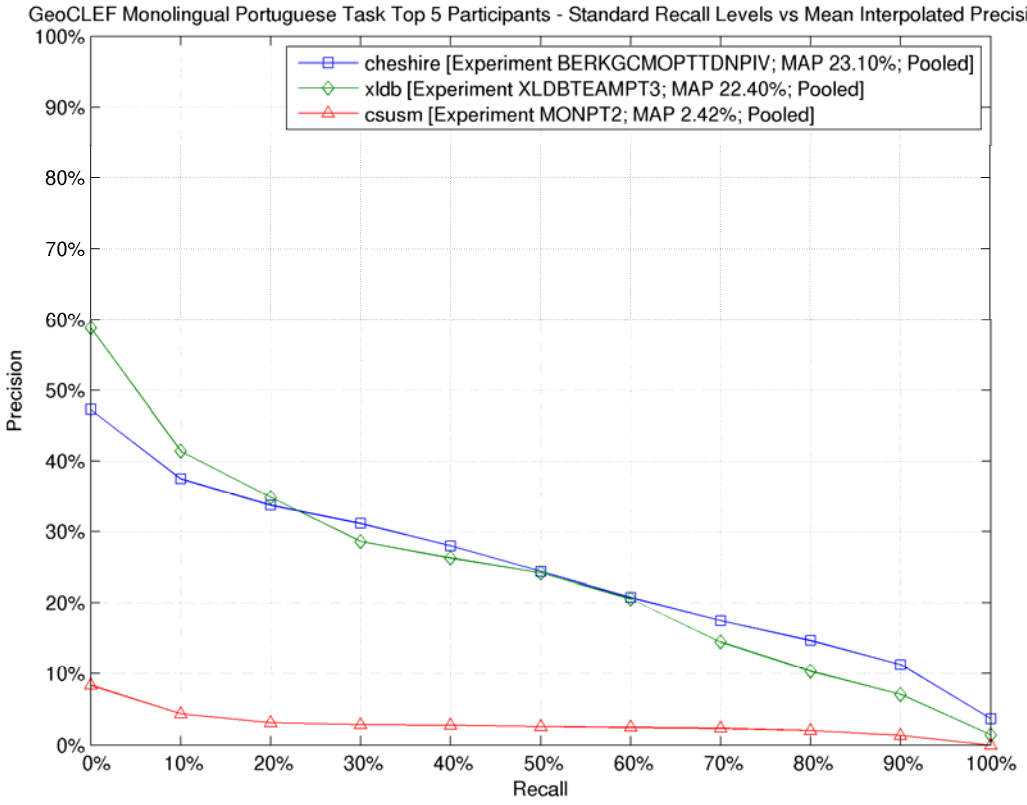



**Fig. 3.** Monolingual Portuguese top participants. Interpolated Recall vs. Average Precision.

### 3.3 Bilingual Experiments

The bilingual task was structured in four subtasks (X → DE, EN or PT target collection). The best system for each of the three bilingual sub-tasks was presented by the University of California at Berkeley who did not use any specific geographic reasoning or knowledge source. Figure 4 to 6 show the interpolated recall vs. average precision graph for the top participants of the different bilingual tasks.

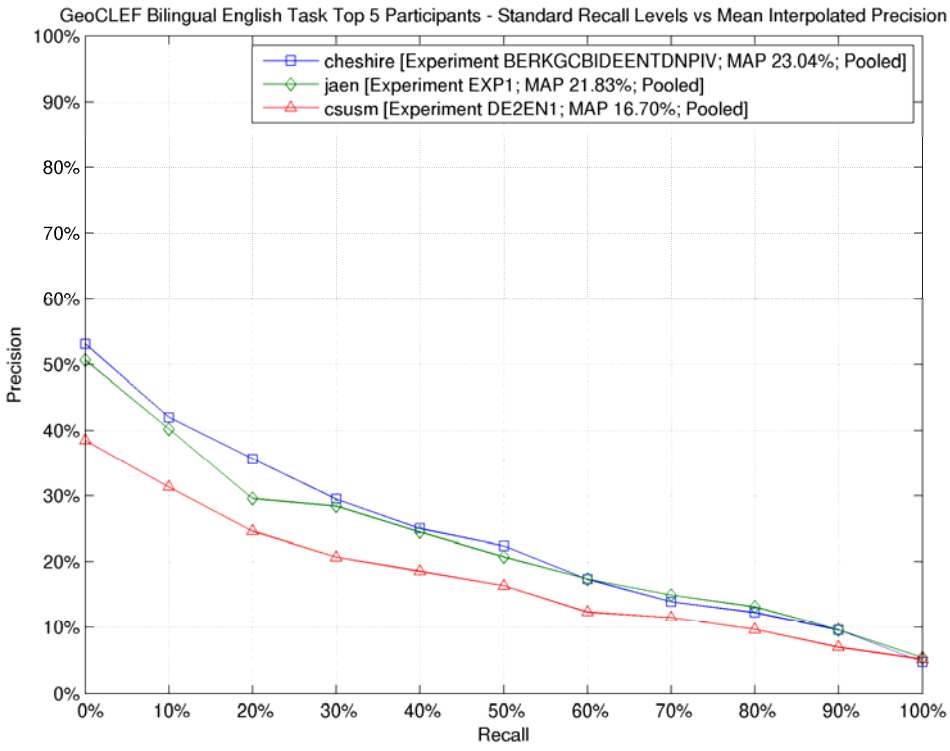



**Fig. 4.** Bilingual English top participants. Interpolated Recall vs Average Precision.

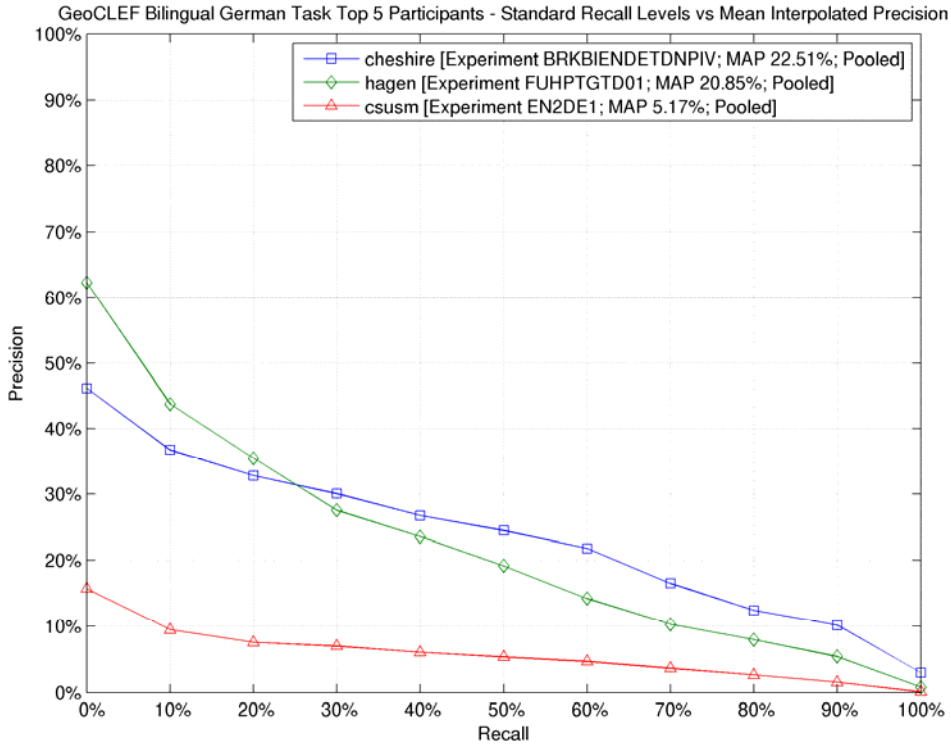



**Fig. 5.** Bilingual German top participants. Interpolated Recall vs Average Precision.

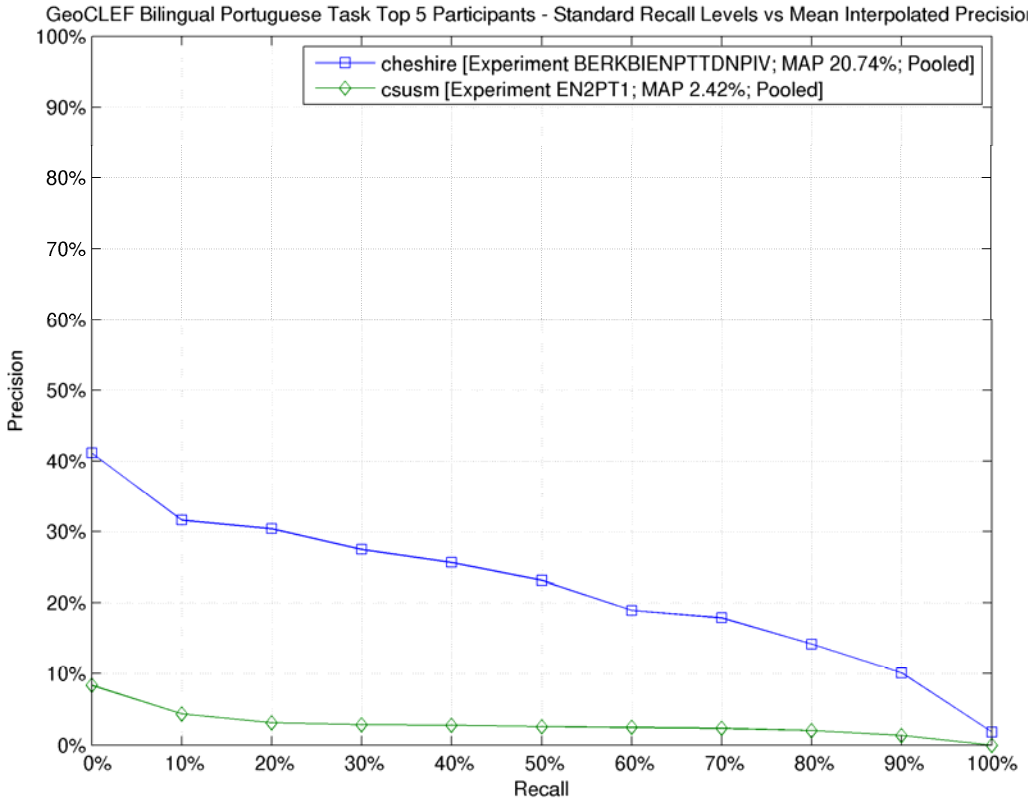



**Fig. 6.** Bilingual Portuguese top participants. Interpolated Recall vs Average Precision.

# 4 Result Analysis

The test collection of GeoCLEF grew of 25 topics each year. This is usually considered the minimal test collection size to produce reliable results. Therefore, statistical testing and further analysis are performed to assess the validity of the results obtained. The range of difficulties in the topics might have led to topics more difficult and more diverse than in traditional ad-hoc evaluations. To gain some insight on this issue, a topic performance analysis was also conducted.

## 4.1 Statistical Testing

Statistical testing for retrieval tests is intended to determine whether the order of the systems which results from the evaluation reliably measures the quality of the systems [2]. In most cases, the statistical analysis gives an conservative estimate of the upper level of significance.

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.

**Table 10.** Lilliefors test for each track with (LL) and without Tague-Sutcliffe arcsin transformation (LL & TS). Jarque-Bera test for each track with (JB) and without Tague-Sutcliffe arcsin transformation (JB & TS).

| Track | LL | LL & TS | JB | JB & TS |
|---|---|---|---|---|
| Monolingual English | 12 | 50 | 0 | 29 |
| Monolingual German | 1 | 6 | 1 | 6 |
| Monolingual Portuguese | 3 | 14 | 3 | 14 |
| Bilingual English | 0 | 2 | 0 | 8 |
| Bilingual German | 2 | 6 | 0 | 5 |
| Bilingual Portuguese | 0 | 3 | 0 | 2 |

Table 10 shows the results of the Lilliefors test before and after applying the Tague-Sutcliffe transformation. The following tables 11 to 16 show the result of the statistical testing.

**Table 11.** Monolingual English: experiment groups according to the Tukey T Test.

| Experiment DOI | Groups |
|---|---|
| 10.2415/GC-MONO-EN-CLEF2008.DFKI.DFKIGEOEN3 | X |
| 10.2415/GC-MONO-EN-CLEF2008.DFKI.DFKIGEOEN2 | X |
| 10.2415/GC-MONO-EN-CLEF2008.DFKI.DFKIGEOEN1 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ALIVALE.TMESS07A | X |
| 10.2415/GC-MONO-EN-CLEF2008.DFKI.DFKIGEOEN4 | X |
| 10.2415/GC-MONO-EN-CLEF2008.JAEN.SINAI_GC_MONO_CLEF08_EXP1 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ICL.IMPCOLLFILTERNODIS | X |
| 10.2415/GC-MONO-EN-CLEF2008.JAEN.EXP4 | X |
| 10.2415/GC-MONO-EN-CLEF2008.CHESHIRE.BERKGCMOENTDNPIV | X |
| 10.2415/GC-MONO-EN-CLEF2008.PITTSBURGH.PITTQP1 | X |
| 10.2415/GC-MONO-EN-CLEF2008.PITTSBURGH.PITTQP2 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN6 | X |
| 10.2415/GC-MONO-EN-CLEF2008.CHESHIRE.BERKGCMOENTD | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN12 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN11 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN5 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN4 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.RBF52 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN10 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN3 | X |
| 10.2415/GC-MONO-EN-CLEF2008.VALENCIA.NLEL0806 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ICL.IMPCOLLFILTERMR | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN8 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN9 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN7 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.RRGEOEXP55 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.RRGEO55 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.RRBF55 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ICL.IMPCOLLFILTERTFIDF | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.RBF55 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ICL.IMPCOLLFILTERNEIG | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.BASELINE1 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN2 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.BASELINE2 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ICL.IMPCOLLTEXTONLY | X |
| 10.2415/GC-MONO-EN-CLEF2008.VALENCIA.NLEL0804 | X |
| 10.2415/GC-MONO-EN-CLEF2008.XLDB.XLDBTEAMEN1 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ALIVALE.TMESS05 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ALIVALE.TMESS04 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.RRGEOEXP52 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.RRBF52 | X |
| 10.2415/GC-MONO-EN-CLEF2008.VALENCIA.NLEL0803 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.RRGEO52 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ALIVALE.TMESS06 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ALIVALE.TMESS02 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ALIVALE.TMESS01 | X |
| 10.2415/GC-MONO-EN-CLEF2008.VALENCIA.NLEL0505 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ALIVALE.TMESS08 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ALIVALE.TMESS03 | X |
| 10.2415/GC-MONO-EN-CLEF2008.JAEN.SINAI_GC_MONO_EN_CLEF08_EXP2 | X |
| 10.2415/GC-MONO-EN-CLEF2008.JAEN.EXP7 | X |
| 10.2415/GC-MONO-EN-CLEF2008.JAEN.SINAI_GC_MONO_EN_CLEF2008_EXP3 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.BASELINE3 | X |
| 10.2415/GC-MONO-EN-CLEF2008.VALENCIA.NLEL0807 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ALIVALE.TMESS09 | X |
| 10.2415/GC-MONO-EN-CLEF2008.VALENCIA.NLEL0802 | X |
| 10.2415/GC-MONO-EN-CLEF2008.ICL.IMPCOLLWEIGHTEDNEIGH | X |
| 10.2415/GC-MONO-EN-CLEF2008.ICL.IMPCOLLWEIGHTEDNODIS | X |
| 10.2415/GC-MONO-EN-CLEF2008.ICL.IMPCOLLWEIGHTEDTFIDF | X |
| 10.2415/GC-MONO-EN-CLEF2008.CHESHIRE.BERKGCMOENTDN | X |
| 10.2415/GC-MONO-EN-CLEF2008.ICL.IMPCOLLWEIGHTEDMR | X |
| 10.2415/GC-MONO-EN-CLEF2008.PITTSBURGH.PITTQI1 | X |
| 10.2415/GC-MONO-EN-CLEF2008.JAEN.EXP5 | X |
| 10.2415/GC-MONO-EN-CLEF2008.INAOE.BASELINE4 | X |
| 10.2415/GC-MONO-EN-CLEF2008.JAEN.EXP6 | X |
| 10.2415/GC-MONO-EN-LEF2008.PITTSBURGH.PITTQI2 | X |
| 10.2415/GC-MONO-EN-CLEF2008.DFKI.DFKIGEOEN5 | X |
| 10.2415/GC-MONO-EN-CLEF2008.CSUSM.MONEN1 | X |

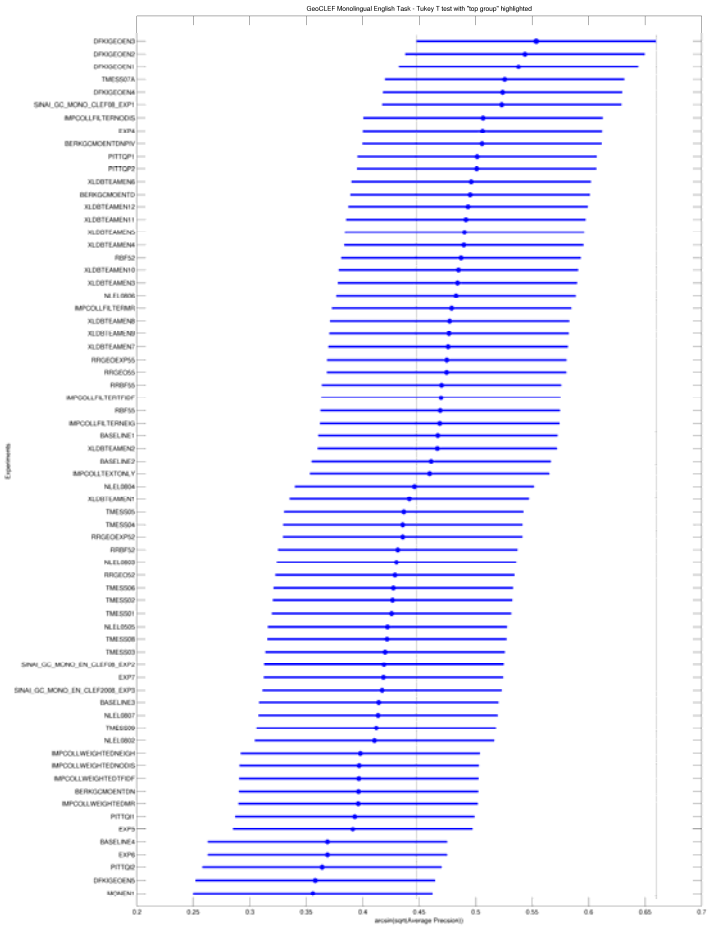


**Table 12.** Monolingual German: experiment groups according to the Tukey T Test.

| Experiment DOI | Groups | | |
|---|---|---|---|
| 10.2415/GC-MONO-DE-CLEF2008.HAGEN.FUHTD01M | X | | |
| 10.2415/GC-MONO-DE-CLEF2008.HAGEN.FUHTD01 | X | X | |
| 10.2415/GC-MONO-DE-CLEF2008.CHESHIRE.BERKGCMODETDNPIV | X | X | |
| 10.2415/GC-MONO-DE-CLEF2008.CHESHIRE.BERKGCMODETD | X | X | |
| 10.2415/GC-MONO-DE-CLEF2008.HAGEN.FUHTD20M | X | X | |
| 10.2415/GC-MONO-DE-CLEF2008.CHESHIRE.BERKGCMODETDN | X | X | |
| 10.2415/GC-MONO-DE-CLEF2008.HAGEN.FUHTD20 | X | X | |
| 10.2415/GC-MONO-DE-CLEF2008.HAGEN.FUHTDN20 | | X | X |
| 10.2415/GC-MONO-DE-CLEF2008.CSUSM.MONDE1 | | | X |

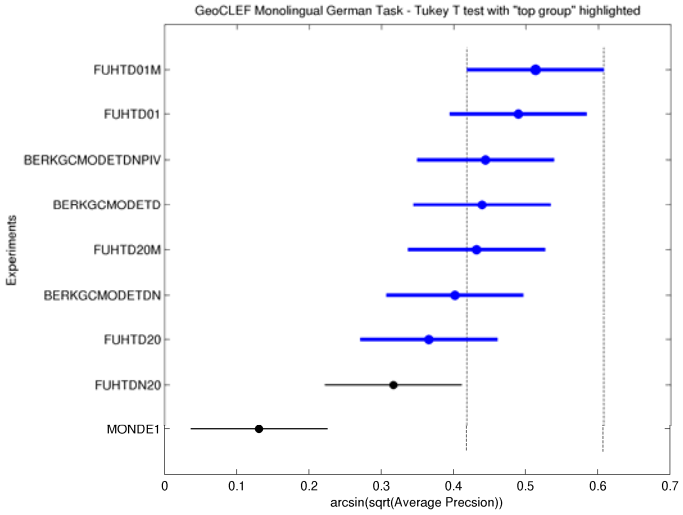


**Table 13.** Monolingual Portuguese: experiment groups according to the Tukey T Test.

| Experiment DOI | Groups | |
|---|---|---|
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT3 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT2 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT1 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.CHESHIRE.BERKGCMOPTTDNPIV | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT11 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT10 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT12 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT4 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT8 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT9 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT5 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT7 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.XLDB.XLDBTEAMPT6 | X | |
| 10.2415/GC-MONO-PT-CLEF2008.CHESHIRE.BERKGCMOPTTD | X | |
| 10.2415/GC-MONO-PT-CLEF2008.CHESHIRE.BERKGCMOPTTDN | X | |
| 10.2415/GC-MONO-PT-CLEF2008.CSUSM.MONPT1 | | X |
| 10.2415/GC-MONO-PT-CLEF2008.CSUSM.MONPT2 | | X |

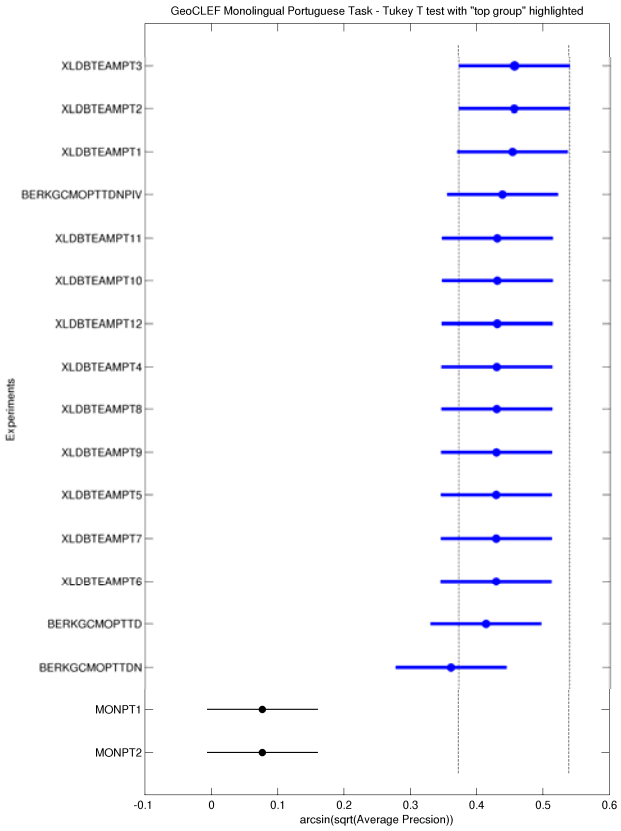


**Table 14.** Bilingual English: experiment groups according to the Tukey T Test.

| Experiment DOI | Groups | | |
|---|---|---|---|
| 10.2415/GC-BILI-X2EN-CLEF2008.CHESHIRE.BERKGCBIDEENTDNPIV | X | | |
| 10.2415/GC-BILI-X2EN-CLEF2008.CHESHIRE.BERKGCBIDEENTD | X | | |
| 10.2415/GC-BILI-X2EN-CLEF2008.CHESHIRE.BERKGCBIPTENTDNPIV | X | | |
| 10.2415/GC-BILI-X2EN-CLEF2008.JAEN.EXP1 | X | | |
| 10.2415/GC-BILI-X2EN-CLEF2008.JAEN.EXP4 | X | | |
| 10.2415/GC-BILI-X2EN-CLEF2008.CHESHIRE.BERKGCBIPTENTD | X | X | |
| 10.2415/GC-BILI-X2EN-CLEF2008.CHESHIRE.BERKGCBIDEENTDN | X | X | X |
| 10.2415/GC-BILI-X2EN-CLEF2008.JAEN.EXP2 | X | X | X |
| 10.2415/GC-BILI-X2EN-CLEF2008.JAEN.EXP5 | X | X | X |
| 10.2415/GC-BILI-X2EN-CLEF2008.CSUSM.DE2EN1 | X | X | X |
| 10.2415/GC-BILI-X2EN-CLEF2008.CHESHIRE.BERKGCBIPTENTDN | X | X | X |
| 10.2415/GC-BILI-X2EN-CLEF2008.JAEN.EXP6 | | X | X |
| 10.2415/GC-BILI-X2EN-CLEF2008.JAEN.EXP3 | | | X |

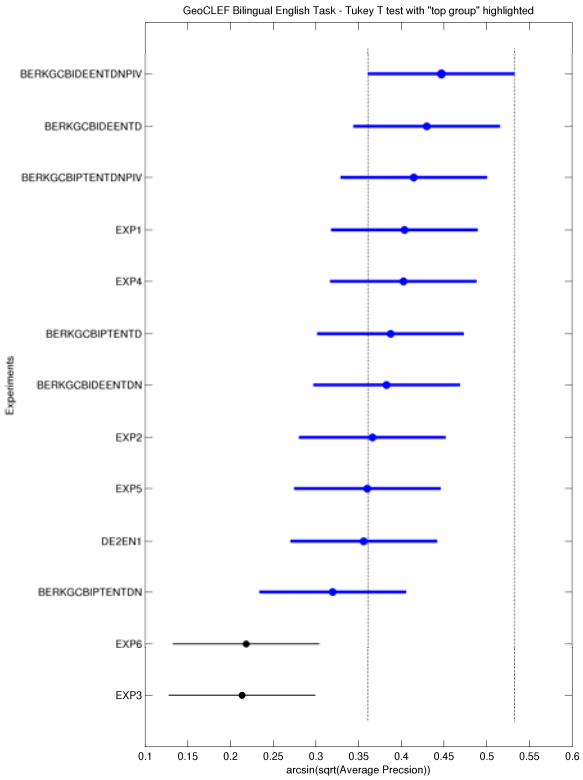


**Table 15.** Bilingual German: experiment groups according to the Tukey T Test.

| Experiment DOI | Groups | |
|---|---|---|
| 10.2415/GC-BILI-X2DE-CLEF2008.CHESHIRE.BRKBIENDETDNPIV | X | |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHPTGTD01 | X | |
| 10.2415/GC-BILI-X2DE-CLEF2008.CHESHIRE.BERKGCBIENDETD | X | |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHPTGTD21 | X | |
| 10.2415/GC-BILI-X2DE-CLEF2008.CHESHIRE.BERKGCBIPTDETDNPIV | X | |
| 10.2415/GC-BILI-X2DE-CLEF2008.CHESHIRE.BERKGCBIPTDETD | X | |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHPTGTD20 | X | |
| 10.2415/GC-BILI-X2DE-CLEF2008.CHESHIRE.BERKGCBIENDETDN | X | |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHPTGTDN21 | X | X |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHPTGTDN20 | X | X |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHENGTDN20 | X | X |
| 10.2415/GC-BILI-X2DE-CLEF2008.CHESHIRE.BERKGCBIPTDETDN | X | X |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHENOTDN20 | X | X |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHENOTD20 | X | X |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHENATD20 | X | X |
| 10.2415/GC-BILI-X2DE-CLEF2008.HAGEN.FUHENATDN20 | X | X |
| 10.2415/GC-BILI-X2DE-CLEF2008.CSUSM.EN2DE1 | | X |

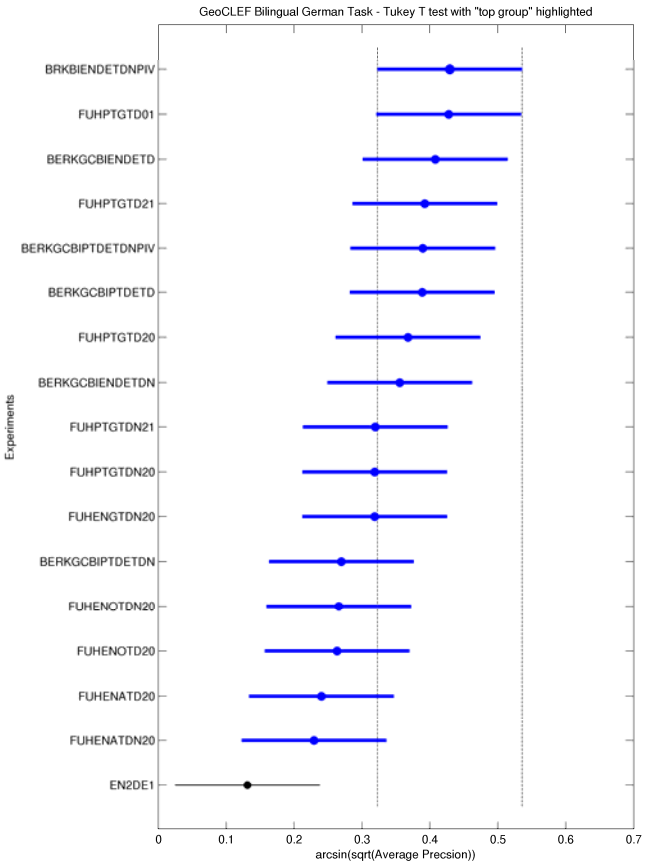
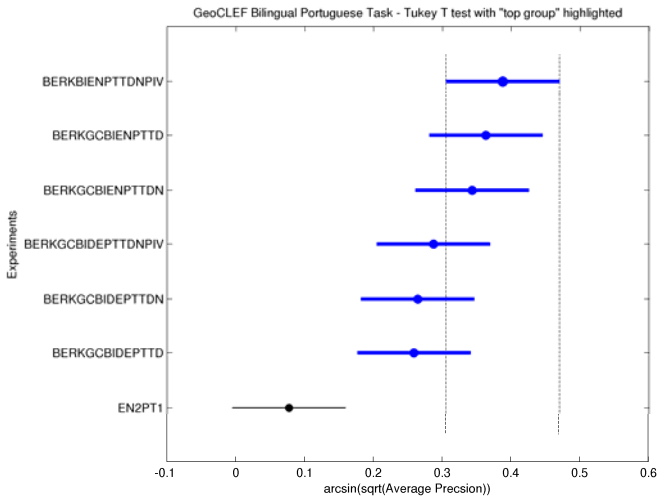


**Table 16.** Bilingual Portuguese: experiment groups according to the Tukey T Test.

| Experiment DOI | Groups | |
|---|---|---|
| 10.2415/GC-BILI-X2PT-CLEF2008.CHESHIRE.BERKBIENPTTDNPIV | X | |
| 10.2415/GC-BILI-X2PT-CLEF2008.CHESHIRE.BERKGCBIENPTTD | X | |
| 10.2415/GC-BILI-X2PT-CLEF2008.CHESHIRE.BERKGCBIENPTTDN | X | |
| 10.2415/GC-BILI-X2PT-CLEF2008.CHESHIRE.BERKGCBIDEPTTDNPIV | X | |
| 10.2415/GC-BILI-X2PT-CLEF2008.CHESHIRE.BERKGCBIDEPTTDN | X | |
| 10.2415/GC-BILI-X2PT-CLEF2008.CHESHIRE.BERKGCBIDEPTTD | X | |
| 10.2415/GC-BILI-X2PT-CLEF2008.CSUSM.EN2PT1 | | X |




# 5 Conclusions and Future Work

GeoCLEF has developed 100 topics and relevance judgments for geographic information retrieval. Another 26 topics with geographic specification were selected out of previous ad-hoc topics from CLEF. This test collection is the first GIR test collection available for the research community and it will be a benchmark for future research.

GIR is receiving increased notice both through the GeoCLEF effort as well as through scientific workshops on the topic. The wide availability of geographic systems on the Internet will further increase the demand for and the interest in geographic information retrieval.

For GeoCLEF 2009, a new GikIP track is again planned.

In addition, a query parsing and classification task is planned for GeoCLEF 2009. Such a task has been part of GeoCLEF 2007 [8] and it requires the participants to identify geographic queries within a large set of queries from a search engine log. All participants of GeoCLEF 2008 are invited to participate in the discussion of the future of GeoCLEF.

## Acknowledgments

The organization of GeoCLEF was mainly volunteer work. Especially the labor intensive relevance assessment required substantial efforts. The English assessment was performed by Fredric Gey and Ray Larson from the University of California at Berkeley and Samaneh Beheshti-Kashi and Wiebke Alscher from the University of Hildesheim. German assessment was carried out by the following group of people from the University of Hildesheim: Lea Drolshagen, Kathrin Stackmann,, Julia Schulz, Nadine Mahrholz, Daniela Wilczek, Ralph Kölle and Thomas Mandl. The Portuguese documents were assessed by Ana Frankenberg, Cláudia Freitas, Cristina Mota, David Cruz, Diana Santos, Hugo Oliveira, Luís Costa, Luís Miguel Cabral, Marcirio Chaves, Paula Carvalho, Paulo Rocha, Pedro Martins, Rosário Silva, Sérgio Matos and Susana Inácio, all of Linguateca (thanks to grant POSI/PLP/43931/2001 from Portuguese FCT, co-financed by POSI). The topics were thoroughly checked by Sven Hartrumpf from the University of Hagen (Germany).

## References

[1] Braschler, Martin; Peters, Carol: Cross-Language Evaluation Forum: Objectives, Results, Achievements. In: *Information Retrieval* vol. 7 (1-2) 2004. pp. 7-31

[2] Buckley, Chris & Voorhees, Ellen: Retrieval System Evaluation. In: *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge & London: MIT Press. 2005. pp. 53-75.

[3] Chaves, Marcirio Silveira; Martins, Bruno & Silva, Mário J.: Challenges and Resources for Evaluating Geographical IR. In: *Proceedings of the 2nd International Workshop on Geographic Information Retrieval, CKIM 2005*. Nov. 2005, Bremen, Germany. pp. 65-69.

[4] Gey, Fredric; Larson, Ray; Sanderson, Mark; Joho, H.; Clough, Paul & Petras, Vivien: GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview. In *6th Workshop of the Cross-Language Evaluation Forum: CLEF 2005*. Springer (Lecture Notes in Computer Science 4022), 2006.

[5] Gey, Fredric; Larson, Ray; Sanderson, Mark; Bishoff, Kerstin; Mandl, Thomas; Womser-Hacker, Christa; Santos, Diana; Rocha, Paulo; Di Nunzio, Giorgio and Ferro, Nicola: GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In *7th Workshop of the Cross-Language Evaluation Forum: CLEF 2006*, Alicante, Spain, Revised Selected Papers. Berlin et al.: Springer (Lecture Notes in Computer Science 4730) 2007. pp. 852-876.

[6] Santos, Diana & Rocha, Paulo: The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck & Bernardo Magnini (eds.), *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*. Berlin/Heidelberg: Springer, Lecture Notes in Computer Science, 2005, pp. 821-832.

[7] Mandl, Thomas; Gey, Fredric; Di Nunzio, Giorgio; Ferro, Nicola; Sanderson, Mark; Santos, Diana & Womser-Hacker, Christa: An evaluation resource for Geographical Information Retrieval. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)* (Marrakech, 28-30 May 2008), European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/summaries/8.html

[8] Mandl, Thomas; Gey, Fredric; Di Nunzio, Giorgio; Ferro, Nicola; Larson, Ray; Sanderson, Mark; Santos, Diana; Womser-Hacker, Christa; Xing, Xie: GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Infor¬ma¬tion Retrie¬val Track Overview. In: Peters, Carol; Jijkoun, Valentin; Mandl, Thomas; Müller, Henning Oard, Doug; Peñas, Anselmo; Petras, Vivien; Santos, Diana (Eds.): *Advances in Multilingual and Multi¬modal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum. CLEF 2007, Budapest, Hungary, Revised Selected Papers*. Berlin et al.: Springer [Lecture Notes in Computer Science 5152] 2008

[9] Santos, Diana & Cardoso, Nuno: Portuguese at CLEF 2005: Reflections and challenges. In *Cross Language Evaluation Forum: LNCS CLEF 2005 Workshop (Vienna, Austria, 21-23 September 2005)*

[10] Santos, Diana & Chaves, Marcirio Silveira;. The place of place in geographical information retrieval. In Chris Jones and Ross Purves, editors, *Workshop on Geographic Information Retrieval (GIR06), SIGIR06, Seattle, 10 August 2006*, pages 5-8, 2006

[11] Santos, Diana & Cardoso, Nuno; Carvalho, Paula; Dornescu, Iustin; Hartrumpf, Sven; Leveling, Johannes & Skalban, Yvonne. Getting geographical answers from Wikipedia: the GikiP pilot at CLEF. *In this volume*.