GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia

Diana Santos, Luís Miguel Cabral

Linguateca, Oslo node, SINTEF ICT, Norway

{Diana.Santos, luis.m.cabral}@sintef.no

Abstract

In this paper we provide a full overview of GikiCLEF, an evaluation contest (track) that was specifically designed to expose and investigate cultural and linguistic issues involved in multimedia collections and searching. In GikiCLEF, 50 topics were developed by a multilingual team with non-English users in mind. Answers should be found in Wikipedia, but not trivially, in the sense that the task should be difficult for human users as well. Crosslinguality was fostered and encouraged by the evaluation measures employed. We present the motivation and the organization process, the management system developed, dubbed SIGA, an overview of participation and results, concluding with what we have learned from the whole initiative.

Keywords

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; H.3.7 Digital Libraries; H.2.3 [Database Managment]: Languages– Query Languages

1 Motivation

It is often stated that multilinguality is just about coping with the same information coded in different ways, and that natural languages are simply a hindrance in our way to to getting at the information (whatever the language). This naive view does not take into consideration that the different members of different language communities also have different views about the information itself. Furthermore, different information is coded in different languages, as is a widely known in disciplines such as linguistics [19, 12], translation studies [16, 2], social studies [6] and usability [14].

Now that everyone is aware of the need to process more than one language, one should be wary of processing the **same** information in all languages, and should instead focus on the ability to look for, and make special use of, **different** information encoded in different languages and cultures. In a similar vein, we believe that systems should cater for different kinds of users and not expect the same user needs overall [8].

GikiCLEF, a follow-up of GikiP [17], was devised on the assumption that not all answers and questions are formulated and answered equally well in any language. Users are different.

Considering that Wikipedia is an information source widely consulted in many languages, GikiCLEF's aim was to foster the development of systems that helped real users. These users can be loosely defined as everyone interested in knowledge already embedded in Wikipedia, but who cannot attain it easily, either for lack of time or ingenuity, or simply for not being able to browse hundreds of pages manually.

1.1 Why Wikipedia

Wikipedia is here being used as the source of freely available multilingual data, semistructured and with some quality control, that is, as an invaluable resource to gather semantic data for natural language processing, as advocated by many as a solution to the knowledge acquisition bottleneck. We are obviously not the first to see Wikipedia in this light, cf. [1, 5, 21, 7, 20]. However, in our present case Wikipedia is rather seen as a user environment which has billions of users and to which – through GikiCLEF – we are contributing to provide a better user experience: one should be able to pose questions to Wikipedia and find a list of articles that provide the answer. Furthermore, in a third way, we are also looking at Wikipedia as providing raw material for an evaluation contest, as has been done by [11] some years ago. It should be however clear that the systems developed for querying Wikipedia in an intelligent way are not necessarily usable only in that context: on the contrary, we expect that the insights and techniques used could be generalized or adapted to all other sources of (multilingual) encyclopedic information as well as other large sized wiki-like sites, and should not be too dependent on particular Wikipedia idyosincrasies.

To our knowledge, Wikipedia snapshots are by far the largest (partially aligned) multilingual corpora that have the highest number of crosslingual links. Most other Web pages have just one or two other languages to which they are linked, as can be appreciated e.g. in [15].[1] But we should hasten to say that we do not believe that the existence of crosslingual links means the existence of independently edited and equally reliable information: in fact, the more paralell the information in two language versions of the same topic, the more probable that one is the translation of the other. Also, we are quite aware that there is a Wikipedia bias in terms of subjects covered, as pointed e.g. by Veale [20]: there is a much higher population of science fiction and comics heros as compared for example with traditional desserts.

This said, and given that Wikipedia is something that evolves daily, it is challenging to process something real (and therefore with inconsistencies and problems), rather than a formal model which is an idealization, and keeps us closed in a lab.

2   Task description

In GikiCLEF, systems need to answer or address geographically challenging topics, on the Wikipedia collections, returning list of answers in the form of Wikipedia document titles.

The "geographical domain" was chosen, not only on internal CLEF grounds (to maintain the tradition started by GeoCLEF and continued with GikiP) or because it is a hot topic nowadays, but because it displays a huge variety in natural language that current gazetteer compilers are often not aware of. We believed it made sense to look at geographically-related queries in order to highlight what is or may be different from language to language, or culture to culture. In fact, in other spheres of thought there have been strong claims for different spatial conceptualizations in languages, see [3, 18], and this is a recurring theme in the GeoCLEF series papers [9, 10, 13] as well.

In practice, a system participating in GikiCLEF receives a set of topics representing valid and realistic user needs, coming from a range of different cultures and languages – in all GikiCLEF languages, namely Bulgarian, Dutch, English, German, Italian, Norwegian – both Bokmål and Nynorsk[2] –, Portuguese, Romanian and Spanish, and its output is a list of answers, in all languages it can find answers.

This kind of output seems to be appropriate, considering that it would be followed by a "output formatter module": For different kinds of human users, and depending on the languages those users could read, different possible output formats would filter the information per language, as well as rank it in order of preference. We are assuming here that people prefer to read answers in their native languages, but that most people are happier with answers (remember, answers are titles of Wikipedia entries) in other languages they also know or even just slightly understand, than with no answers at all.

Since we are aware that not all GikiCLEF participants have the resources and interest to process or give answers in the ten collections, we have added the option of "languages of participation" to the registration

---

[1]The exception is probably the Bible, but it is not so widely accessed as Wikipedia in our days.

[2]Norwegian has two written standards, and Norwegians therefore decided to maintain Wikipedia in two "parallel" versions, so GikiCLEF covers nine languages and ten collections. We have therefore created questions in and/or translated them into both written standards of Norwegian.

process, that is, languages of the users the systems want to please. However, as will be explained presently, systems not tackling all languages will at once have a lower score.

The evaluation measures are then as follows for a given run, and for each language:

- C: number of correct (that is, justified in at least one language) answers

- N: total number of answers provided by the system

- GikiCLEF score per language: C*C/N (so one has a score for de, pt, etc, as $C_{de} * C_{de}/N_{de}$, $C_{pt} * C_{pt}/N_{pt}$, etc.)

The final score of any system is given by the sum of the scores for each individual language. So, the more languages a system returns answers in, the better its scores. Furthermore, a language with no answers for a particular topic (C=0) will not contribute for the relative ordering of the systems.

Note that a score for a particular language is the sum for all topics, not the average of the scores per topic. This is in order not to penalize languages which have no information on a particular topic in their Wikipedia.

3 The organization of GikiCLEF

The Wikipedia collections for all GikiCLEF languages were released on 20 January, 2009, and correspond to the Wikipedia snapshots from June 2008. They were converted to XML with the WikiXML tool created by the University of Amsterdam, which is available from http://ilps.science.uva.nl/WikiXML/. Figure 1 presents their relative sizes. Later on, due to some problems in the conversion, we allowed participants to use the HTML versions as well.
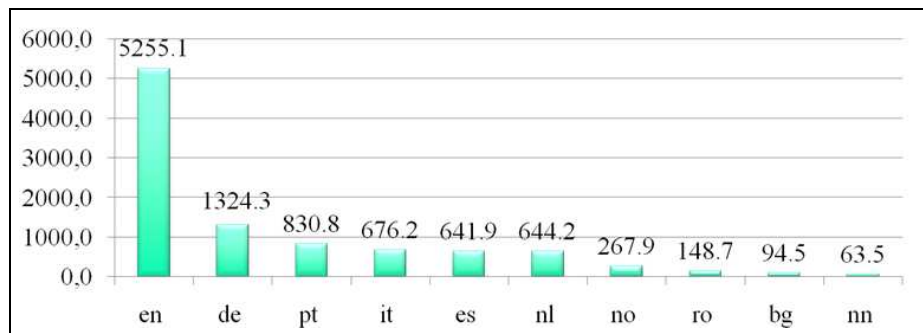

Figure 1: Size of the different GikiCLEF Wikipedia collections

This was the only task performed prior to the development of SIGA[3], which we then developed in order to assist both organizers and participants in the GikiCLEF task. In fact, four distinct roles had to be implemented, with different access modes and privileges: participant, topic manager, assessor, and administrator.

Briefly, the different tasks involved in the several phases of GikiCLEF, in a loose chronological order, were:

Topic management The process of developing topics, finding some answers (pre-determined, and mark if they were self-justified or required further information), translate the wording into other languages and provide a motivation for them (for topic managers);

Participation Fetching the topics, submitting answers and validating them , getting final individual scores (for participants);

---

[3]SIGA stands for *SIstema de Gestão e Avaliação do GIKICLEF*, Portuguese for "Management and Evaluation System of Giki-CLEF". The word *siga* means "Go on!" (imperative of verb *seguir*, "continue").

Answer pool creation The process of merging all answers from all runs, come up with a pool of unique answers to be assessed, and attribute them to different assessors, with some overlap per language (for administrators);

Topic assessment The process of evaluating individual answers (and their justifications) as well as discuss hard cases (for assessors);

Conflict resolution Comparing the assessments done by different assessors and proceed to a final decision (for administrators);

Results computation For each run, propagate the justification to other languages, do another (crosslingual) conflict resolution, obtain individual scores, and provide aggregated results (for administrators).

A system helping during all these phases was necessary since GikiCLEF had a really large (and geographically distributed) organizer committee, and the same was even more true of the assessors and participants masses.

SIGA was developed in MySQL, Perl and PhP and its source code is available, under a Gnu license, from the GikiCLEF site. We will be presenting SIGA along with the description of the process followed in 2009.

4   Topics: their preparation and related issues

The final topics were released 15 May 2009, after a set of 24 example topics, displayed in Table 1, had been made available some months before.

Table 1: Example topics in GikiCLEF.

ID	Topic in English
EX01	Name Portuguese-speaking Nobel prize winners
EX02	List Portuguese Pop/Rock groups created in the 90s.
EX03	Which Brazilian football players play in clubs in the Iberian Pensinsula?
EX04	What capitals of Dutch provinces received their town privileges during the sixteenth century?
EX05	In which places did Italo Calvino live during adulthood?
EX06	Name Mexican poets who published volumes with ballads until 1930.
EX07	Name authors born in Alaska and who wrote fiction about it.
EX08	What Belgians won the Tour de France exactly twice?
EX09	Find Amazon tribes which have no written language
EX10	Find Northern Europe companies which produce nano-electronic components for planes.
EX11	Which MPB musicians are also distinguished Spanish guitar players?
EX12	Which stave churches are still used for religious purposes in Norway?
EX13	Name Brazilian theologists connected with Liberation Theology.
EX14	Amusement parks in East Germany
EX15	Technical universities in Germany with more than 10,000 students
EX16	Carnival strongholds in the Rhineland
EX17	Cathedral libraries in Germany
EX18	Cities having an Olympic stadium
EX19	Cities situated by the former Eifel Aqueduct
EX20	Castles in Upper Bavaria, that are still inhabited
EX21	German zoos with animals from the African savannah
EX22	Cities, that have a womens's football (soccer) team, that won the Women's Football Bundesliga
EX23	Terminal stations in Germany with more than 100,000 passengers a day
EX24	Transalpine UNESCO World Heritage Sites, listed before 2000

Table 2: Final GikiCLEF 2009 topics.

Topic ID	Topic in English
GC-2009-01	List the Italian places which Ernest Hemingway visited during his life.
GC-2009-02	Which countries have the white, green and red colors in their national flag?
GC-2009-03	In which countries outside Bulgaria are there published opinions on Petar Dunov's (Beinsa Duno's) ideas?
GC-2009-04	Name Romanian poets who published volumes with ballads until 1941.
GC-2009-05	Which written fictional works of non-Romanian authors have as subject the Carpathians mountains?
GC-2009-06	Which Dutch violinists held the post of concertmaster at the Royal Concertgebouw Orchestra in the twentieth cent
GC-2009-07	What capitals of Dutch provinces received their town privileges before the fourteenth century?
GC-2009-08	Which authors are born in and write about the Bohemian Forest?
GC-2009-09	Name places where Goethe fell in love.
GC-2009-10	What Flemish towns hosted a restaurant with two or three Michelin stars in 2008?
GC-2009-11	What Belgians won the Ronde van Vlaanderen exactly twice?
GC-2009-12	Present monarchies in Europe headed by a woman.
GC-2009-13	Romantic and realist European novelists of the XIXth century who died of tuberculosis.
GC-2009-14	Name rare diseases with dedicated research centers in Europe.
GC-2009-15	List the basic elements of the cassata.
GC-2009-16	In which European countries is the bidet commonly used?
GC-2009-17	List the 5 Italian regions with a special statute.
GC-2009-18	In which Tuscan provinces is the Chianti produced?
GC-2009-19	Name mountains in Chile with permanent snow.
GC-2009-20	List the name of the sections of the North-Western Alps.
GC-2009-21	List the left side tributaries of the Po river.
GC-2009-22	Which South American national football teams use the yellow color?
GC-2009-23	Name American museums which have any Picasso painting.
GC-2009-24	Which countries have won a futsal European championship celebrated in Spain?
GC-2009-25	Name Spanish drivers who have driven in Minardi.
GC-2009-26	Which Bulgarian fighters were awarded the "Diamond belt"?
GC-2009-27	Which Dutch bands are named after a Bulgarian footballer?
GC-2009-28	Find coastal states with Petrobras refineries.
GC-2009-29	Places above the Arctic circle with a population larger than 100,000 people
GC-2009-30	Which Japanese automakers companies have manufacturing or assembling factories in Europe?
GC-2009-31	Which countries have Italian as official language?
GC-2009-32	Name Romanian writers who were living in USA in 2003.
GC-2009-33	What European Union countries have national parks in the Alps?
GC-2009-34	What eight-thousanders are at least partially in Nepal?
GC-2009-35	Which Romanian mountains are declared biosphere reserves?
GC-2009-36	Name Romanian caves where Paleolithic human fossil remains were found.
GC-2009-37	Which Norwegian musicians were convicted for burning churches?
GC-2009-38	Which Norwegian waterfalls are higher than 200m?
GC-2009-39	National team football players from Scandinavia with sons who have played for English clubs.
GC-2009-40	Which rivers in North Rhine Westphalia are approximately 10km long?
GC-2009-41	Chefs born in Austria who received a Michelin Star.
GC-2009-42	Political parties in the National Council of Austria which have been founded after the end of World War II
GC-2009-43	Austrian ski resorts with a total piste length of at least 100 km
GC-2009-44	Find Austrian grape varieties with a vineyard area below 100 ha.
GC-2009-45	Find Swiss casting show winners.
GC-2009-46	German writers which are Honorary Citizens in Switzerland.
GC-2009-47	Which cities in Germany have more than one university?
GC-2009-48	Which German-speaking movies have been nominated for an Oscar?
GC-2009-49	Formula One drivers who moved to Switzerland.
GC-2009-50	Which Swiss people were Olympic medalists in snowboarding at the Winter Olympic Games in 2006?

Figure 2: SIGA in topic creation mode: editing translations

Before disclosing the final topics in Table 2, let us present the topic creation guidelines and our expectations (not necessarily met by the final set, as will be hinted at in the final discussion):

- One should strive for realistic topics which can be answered in **some** Wikipedia covered by Giki-CLEF, chosen with a conscious cultural bias so that not all Wikipedia would have that information.

- Ideal topics for GikiCLEF may require knowledge of culture to understand the way they should be answered (or better, what it is that is being sought). This requirement entails that translation into other languages may require lengthy explanations. For example, *Spanish guitar* is a technical term in music that is probably not the best way to translate *violão*, the Brazilian (original) term. Also, to render the Norwegian *oppvekstroman* requires the clarification that this is close, but not the same as what, in English, literature experts use the German (!) term *Bildungsroman* to express. Similary, Romanian *balade* is probably a false friend with Spanish *ballada*, and had to be translated by *romance*. Interestingly, this is again a false friend with Portuguese *romance*, denoting what in English is called *novel*.

- Answers to the questions had to be justified in at least one Wikipedia (that is, the string may be found as a entry in all Wikipedias, but the rest of the information has to be found in at least one). So, we are not looking for absolute truth, we are looking for answers which are justified in Wikipedia.

- Questions may include ambiguous concepts or names, especially when translated. In that case, participants were warned that only answers related to the proper disambiguation will be considered correct e.g. *Which countries did Bush visit in the first two years of his mandate?* will not be correctly answered by the singer Kate Bush's travels in whatever mandate she may have (had). Narratives[4] should thus clearly specify and explain the exact user need.

- In case there appear ambiguities in the topic formulation that have not been discussed or clarified in the narrative, and which have more than one interpretation acceptable (with respect to the user model at stake), assessment will accept both. For example, in *Award-winning Romanian actresses in international cinema festivals*, one would have to accept not only those actresses actually receiving prizes, but also those just in the audience or even hosting the event, if that had not been made clear beforehand (in the Further clarification text).

---

[4]In fact, the term *Further clarification* was employed in SIGA instead. Participants did not have access to them during submission, only after their participation.

- Different answers about the same subject are welcome, provided they have support in the material. Examples are "Who is (considered to be) the founder of mathematics?" or "Name the greatest scientific breakthroughs in the XIXth century", which are obviously open to different opinions.[5]

During the topic discussion phase, the topic creation group came up with 75 topics, from which the final 50 were chosen according to the following additional criteria: avoid repetition, avoid quizz-like flavour, avoid hard to interpret topics, and then removing randomly until the number 50 was reached.

As an integral part of topic choice and preparation, SIGA helped the topic managers to look for answers in titles of Wikipedia documents pertaining to the GikiCLEF collection, as illustrated by Figure 3.

We expected that this process of finding candidates by just looking in the titles would be of considerable help for topic managers, who would not need to deal with the large collections in order to list correct answers. However, we did not require that people stored the answers there during topic creation.

Figure 3: SIGA in topic creation mode: finding candidates in the GikiCLEF colection.
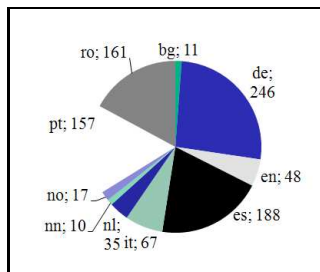

Figure 4: Pre determined answers per language

This was something we provided as a facility in order to avoid, later, much work during assessment. Interestingly, only half of the members of the topic group used this, and also for different topics and for different languages there were different policies. Some people did it for the topics they owned in almost all languages, others did it for all topics only in their language, some did no pre-storing at all, and the majority did just some and in some languages. In Figure 4 one can see the result of this process.

_____

[5]For the record, no topic owner chose to do this kind of opinion questions in GikiCLEF 2009.

5 Expected answers in GikiCLEF

Systems were supposed to deliver as many answers (in as many languages) as possible, but answers had to be justified in at least one language. For an answer to be considered justified, it required simply that a person would accept the answer by reading it (the article) and further documents offered as additional material. Of course this is ultimately subjective, but all evaluation in information retrieval is. In order to ensure a maximum of fairness, guidelines for borderline cases had to be discussed among the assessors and normalized in the end, to the best of our abilities (and to the strain of the assessors, who had often to reassess their answers).

The Wikipedia page about the answer may be its own justification (the simplest case), but we imagined, and catered for, cases where other pages would have to be brought to bear (such as disambiguation or lists, or even images).

An answer without justification was never to be considered right.

Let us provide two examples in more detail:

Question *In which places did Italo Calvino live during adulthood?* would require a system to go to the page(s) devoted to this writer, find that information, and get the places, namely e.g. Turin and Paris. In order to have these accepted as correct answers, the page about Italo Calvino which describes his life, e.g. `http://en.wikipedia.org/wiki/ItaloCalvino`, would have to be included as (further) justification for Paris and Turin.

Once there was a justification (in this example and to make it easier for the present paper, in English – although the most complete is probably in the Italian Wikipedia), any answers like *Turim* in Portuguese or *Parisj* in Dutch would be considered correct: in other words – once justified in a particular language, justified for all languages.

Now to a more complex example, to show how the GikiCLEF format allows arbitrary chains of reasoning – which is not to say that we expected current systems to be able to do it. Take question *Name American cities where people who killed presidents lived for more than one year*. To answer it, in addition to the name of the city, systems would have to find the names of presidents who were killed, and – although in this particular case there is even a category in Wikipedia entitled "United States presidential assassination attempts" – this might require that systems go through all pages concerning presidents and investigate the deaths, their causes and the names of the assassins, then check the assassins' pages, and finally extract the cities where they lived.

In order for an answer to be justified, let us say the answer "Chicago", the justification would have to include the page of the assassin that mentions that place, and the name of the president killed as well if this is not mentioned in the assassin's page. So, in principle at least, one may have to include several pages in order to justify any given answer.

Figure 5: SIGA in assessment mode: Is this answer correct and justified?

6 Asssessment and evaluation

In GikiCLEF, as before in GikiP, only answers / documents of the correct type were considered correct. That is, if the question is about people, an answer of an organization is considered wrong, even if in that document whose title was an organization there is the person one would want as answer.

After pooling all answers returned by the participant systems, they were manually assessed by the assessors' group. SIGA's assessment interface, displayed in Figure 5, allows the assessors to judge the candidate answers, and check the correctness of their justifications.

Prior to this, to ease the assessment task, an automatic process assesses the answer documents that were listed as correct answers during the topic preparation period, as well as eliminates invalid document answers (such as redirects).

Assessment in GikiCLEF proceeds in several phases:

1. All pre-stored correct answers which are self-justified are automatically classified. The ones which require a justification are marked as **Correct**, but are still presented to the assessors for them to assign a **Justified** (or **Not Justified**) verdict.

2. Assessors assess individual answers, assigning either **Incorrect**, **Correct**, or **Unknown**. If it is **Correct**, they have to indicate whether the individual answer they are assessing (which includes the justification chain) is **Justified**, or whether it is **Not Justified**.

3. A process of conflict resolution among different assessments of the very same answer is then run, which allows people to discuss and get aware of complications and/or mistakes or mistaken assumptions. Only after all conflicts are resolved can one proceed to:

4. Evaluate runs, by propagating justification across languages

5. A new process of crosslingual conflict resolution then ensues, with the net result that positively conflicting information for one topic brings about the inhibition of multilingual propagation: for those topics, only monolingually correct and justified answers will be considered correct.

6. Final scores are computed and displayed

It goes without saying that all these phases and checkpoints allowed us to find problems, inconsistencies and even wrongly pre-defined answers in the original topic set.

Figure 6 displays SIGA's assistance in conflict solving. The administrator can choose to send a question to the diverging assessors, or decide herself, if it is a straightforward case.

Figure 6: SIGA in conflict resolution mode: choose between assessments

The final scores are automatically computed after the assessment task and made available to the participants, who are granted access to several scores and the detailed assessment of their answers, as illustrated in Figure 7.

Participant _____ run *3*: Score per language									
Bulgarian (BG)	Dutch (NL)	English (EN)	German (DE)	Italian (IT)	Norwegian Bokmål(NO)	Norwegian nynorsk(NN)	Portuguese (PT)	Romanian (RO)	Spanish (ES)
0.143	0.000	1.778	1.333	1.000	1.800	1.000	1.000	0.000	0.500

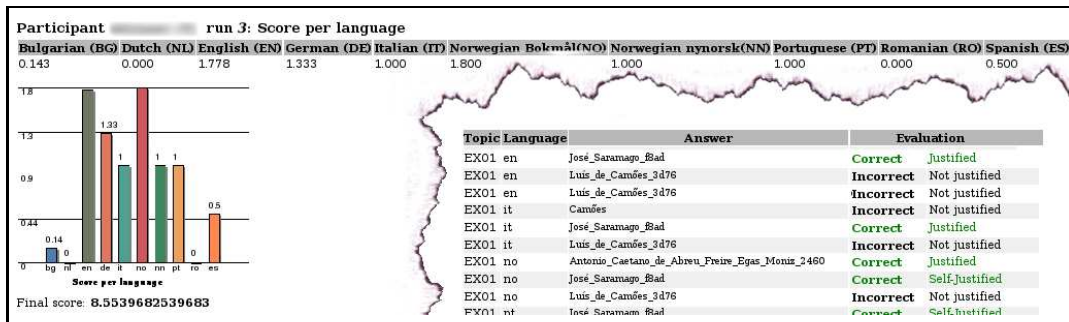| Topic	Language	Answer	Evaluation	
EX01	en	José_Saramago_fBad	Correct	Justified
EX01	en	Luis_de_Camões_3d76	Incorrect	Not justified
EX01	en	Luis_de_Camões_3d76	Incorrect	Not justified
EX01	it	Camões	Incorrect	Not justified
EX01	it	José_Saramago_fBad	Correct	Justified
EX01	it	Luis_de_Camões_3d76	Incorrect	Not justified
EX01	no	Antonio_Caetano_de_Abreu_Freire_Egas_Moniz_2460	Correct	Justified
EX01	no	José_Saramago_fBad	Correct	Self-Justified
EX01	no	Luis_de_Camões_3d76	Incorrect	Not justified
EX01	pt	José_Saramago_fBad	Correct	Self-Justified

Final score: 8.5539682539683

Figure 7: Displaying SIGA results: on the left a graphic with language score, and on the right the assessment of each answer given.

While this seems a complete enough description of the assessment process, one should document that a lot of other more specific decisions and guidelines had to be decided during the process. By writing them down here we intend not only to illustrate the kinds of problems that arise, but also provide an initial set for further initiatives or GikiCLEF editions.

1. If the answer is already contained in the question, it is considered incorrect. For example, *Italy* is not a fair answer to a question starting by *List the Italian places*

2. If there is principled disagreement about vague, complex categories and different people have strong reasons for disagreement, for GikiCLEF we accept the union of all.

3. Speaking/writing poets in other languages than Romanian are Romanian poets? We decided for a yes.

4. Studying in a place, taking a short visit to another place and coming back in love to that place, does it qualify as a place where someone falls in love? Again, yes.

5. If a ciclist won the junior Tour de Flandres and then the adult one, is s/he considered a winner twice? We decided for yes, although this is a recurrent issue in sports questions. Often, without further specification, only the major competition is meant.

6. Very slight differences which very strongly convey the probability of yes are accepted, because we would expect most people (except lawyers and logicians) to accept that:

   - Eight thousanders accept a 50 m deviation (if a mountain is higher than 7950 m)
   - Norwegian musicians convicted for burning (even if the article does not mention they burned churches) must be the ones looked for
   - People wro wrote ballads and published a lot of volumes of poetry is expected to have published volumes with ballads although the article does not say so
   - People who have two residences, one in Switzerland and another somewhere else, can be considered to have moved to Switzerland some time in their lives.

7. If two of three sisters died of tuberculosis and for the third the cause of death is not certain, is a page entitled "the sisters Brontë" correct? We relaxed the strict requirement that writers should be people and not a group of people, because beforehand we did not expect groups of writers to stand as an article. So we accepted it as correct.

8. No longer existing Austrian parties, provided they were founded after the Second World War and had – at some time – people in the National Council of Austria, were considered correct. This brings about the often noted fact that most questions are not independent on time.

9. Finally, what are American museums? This expression should be interpreted according to the natural meaning of the corresponding word (*american, amerikansk, americanos* etc.) in the corresponding language – at least this was how we asked people to translate the question. But apparently Canadian and Brazilian museums do not mention their Picassos in their Wikipedia pages (so, even if correct, these answers will turn out **Not justified**, hence **Incorrect**), and the only hits found corresponded to museums in the USA. In this case, it was probably the organization's fault not to emphasize to the participants that each language topics should be understood and answered in that language, even if the correct interpretation of the terms in the different languages turned out to be different.[6]

We do not want to convey the idea that anything goes, though. In fact there were several other cases which were negatively decided:

1. *Gulag* can be metonymically used for the places where people were imprisoned and most of them were above the Arctic. However, we did not consider it as a valid place.

2. Fictional countries were not considered as correct when asking for fictional works, even if they were created or presented in the scope of a written fictional work.

3. Cases where the expected answer type (EAT) was clearly different from the one returned were considered downright incorrect, notwithstanding our agreement that the answers could be useful. So,

   - Flags were not accepted as answers to questions *Which countries had a flag ...?*
   - Queens were not accepted as answers to questions *Which countries have a queen...?*
   - Countries were not accepted as answers to questions *Which national teams...?*
   - Reserves were not accepted as answers to questions *Which mountains...?*

   This is in line with our belief that assessment would have become a nightmare if any answer, whatever its type, had to be investigated by the assessors to see whether it could be indirectly useful. However, we are also aware that different participants in GikiCLEF 2009 interpreted the task differently, which produced unwanted differences among the participants. Clearly, this issue has to be considered for future editions, and an intermediate solution could be that, for some topics, more than one EAT, previously agreed, could be accepted.

7 Overview of participation

Although we had almost 30 registrations of interest for GikiCLEF, in the end only 8 participants were able to submit. For the record, they are displayed in Table 3 by registration order.

Table 3: Participants in GikiCLEF 2009

Name	Institution	System name
Ray Larson	University of California, Berkeley	cheshire
Sven Hartrumpf &	FernUniversität in Hagen &	GIRSA-WP
& Johannes Leveling	& Dublin City University	
Iustin Dornescu	University of Wolverhampton	EQUAL
TALP Research Center	Universitat Politécnica de Catalunya	GikiTALP
Gosse Bouma & Sergio Duarte	Information Science, University of Groningen	JoostER
Nuno Cardoso et al.	GREASE/XLDB, Univ. Lisbon	GreP
Adrian Iftene	Faculty of Computer Science	UAICGIKI09
Richard Flemmings et al.	Birkbeck College (UK) & UFRGS (Brazil)	bbk-ufrgs

We received 17 runs, and their results are presented in Figure 8.

---

[6]This raises the problem, not yet satisfactorily solved, that culturally-laden questions are not exactly parallel, and that therefore the set of (multilingual) answers ultimately depends on the language the question was asked.

System/Run	#answers	#Corrects	#Unjustified	Precision	Score
EQUAL 1	813	385	105	0.4736	181.9329
GREASE/XLDB 1	1161	332	56	0.2860	96.0070
Cheshire 1	564	214	14	0.3794	80.9247
GIRSA-WP 1	38	31	0	0.8158	24.7583
GIRSA-WP 3	985	142	10	0.1442	23.3919
GIRSA-WP 2	994	107	7	0.1076	14.5190
JoostER 1	638	36	14	0.564	2.4053
GikiTALP 3	356	26	0	0.0730	1.9018
GikiTALP 2	295	20	0	0.0678	1.3559
GikiTALP 1	526	18	2	0.0342	0.6964
bbk_ufrgs 1	726	8	0	0.0110	0.0882
UAICGIKI09 2	6420	8	0	0.0012	0.0156
bbk_ufrgs 2	734	3	0	0.0041	0.0123
UAICGIKI09 1	1133	2	0	0.0018	0.0062
JoostER 2	272	0	3	0.0000	0.0000
bbk_ufrgs 3	686	0	1	0.0000	0.0000
UAICGIKI09 3	4910	0	0	0.0000	0.0000

Figure 8: Final scores in GikiCLEF 2009

Figure 9 presents the participation detailed for each language. The last row indicates how many participants per language, and the last column the number of languages tried in that run. Eight runs opted for all (10) languages, four tried solely 2 languages, and five one only.

While this seems a modest amount of work, in fact it produced a sizeable amount of material to deal with, as Table 4 shows.

Table 4: Numbers on the assessment process.

No. of answers received	21,251
No. of different answers	18,152
No. of different answers with complex justification	215
No. of different manually assessed answers	6,974
No. of manual assessments	10,332
No. of automatically assessed answers as incorrect	10,588
No. of automatically assessed answers as correct	283
No. of answers resulting in conflicts	383
No. of correct and justified answers	1,327
No. of correct but not justified answers	1,415

The reason why there were considerably more manual assessments than manually assessed answers is due to the important fact that 2,131 answers had more than one assessor (often two, but they may have been assigned up to four different ones), to test the soundness and coherence of the assessment process. Note, anyway, that this does not include repeated assessments by the same assessor, nor assessments done by the organizers during conflict resolution, so that in practice the work involved was substantial, even with 29 assessors.

Turning now to the comparative weight and/or performance of the different languages involved, at face value, all languages participated in the answer gathering.

Figure 10 provides an overview of the total number of answers per language, while Figure 11 shows the distribution of only the correct answers.

The number of answers provided per language, as well as the amount of the correct ones, seems to demonstrate that the GikiCLEF systems could be used with the same level of success in all GikiCLEF

System / run	BG	NL	EN	DE	IT	NO	NN	PT	RO	ES	Total	Languages
EQUAL 1	9.757	18.980	34.500	25.357	17.391	17.254	9.308	15.515	14.500	16.695	181.932	10
GREASE/XLDB 1	6.722	8.258	13.657	12.007	8.533	11.560	9.557	7.877	6.720	11.115	96.007	10
Cheshire	1.091	9.132	22.561	9.000	11.200	7.043	3.368	4.891	7.714	4.923	80.924	10
GIRSA_WP 1	1.333	2.25	1.800	1.125	2.250	3.000	2.000	3.000	3.000	3.000	24.758	10
GIRSA_WP 3	3.030	1.798	1.390	3.661	1.988	2.526	3.064	2.250	1.684	2.000	23.391	10
GIRSA_WP 2	2.065	1.299	0.496	1.540	1.429	1.723	1.841	1.350	1.029	1.306	14.519	10
JoostER 1		0.964	1.441								2.405	2
GikiTALP 3			1.635							0.267	1.901	2
GikiTALP 2			1.356								1.356	1
GikiTALP 1			0.668							0.028	0.696	2
bbk_ufrgs 1								0.088			0.088	1
UAICGIKI09 2	0.000	0.002	0.002	0.002	0.002	0.002	0.000	0.002	0.000	0.006	0.015	10
bbk_ufrgs 2								0.012			0.012	1
UAICGIKI09 1									0.000	0.006	0.062	2
JoostER 2										0.000	0.000	1
bbk_ufrgs 3								0.000			0.000	1
UAICGIKI09 3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	10
Total Participations	8	9	12	8	8	8	8	11	9	12		

Figure 9: Results in GikiCLEF 2009

Figure 10: Answers per language returned by the set of all systems

Figure 11: Correct answers per language returned by the set of all systems

languages. In Figure 12 also the precision per language is shown.

However, it does not say anything about whether there were languages which were necessary to check in order to have a (crosslingually) justified (and therefore correct answer). For this we tried to see if some languages had a large amount of correct answers due to other languages, that is, we wanted to check language dependence or interdependence. A language should be more authoritative the more answers it provided without requiring proof in other languages.

Figure 13 presents those numbers, as well as contrasting the number of correct answers per language, with the pre-assessed (correct) ones.

Interestingly, and contrary to our expectations concerning English, that figure again does not allow one to infer that English has more information or more detailed justifications in pages written in that language. This must be an artifact of our topic choice, which was on purpose geared toward languages different from English. Still, and even with our initial guidelines, many of the topics chosen were more international than really national (even if they did relate to specific individuals of a non-English-speaking nationality), and therefore one would expect that they would have equally developed pages in English as well.

A more thorough investigation of the different GikiCLEF topics regarding language spread should thus take place, such as the one done by [4], who claim that, of the 50 topics of this year's GikiCLEF only 25 had a (justified) answer in the Portuguese Wikipedia, vs. 47 in English. If this is true, systems that processed only the Portuguese Wikipedia and then tried to follow links into the other languages would be in definite disadvantage compared to others that did the opposite, even for answering in Portuguese.

Also, it remains to be investigated which topics might be popular (or even asked at all) regarding different language populations. Of course our organizers' sample was not representative, and, in addition,
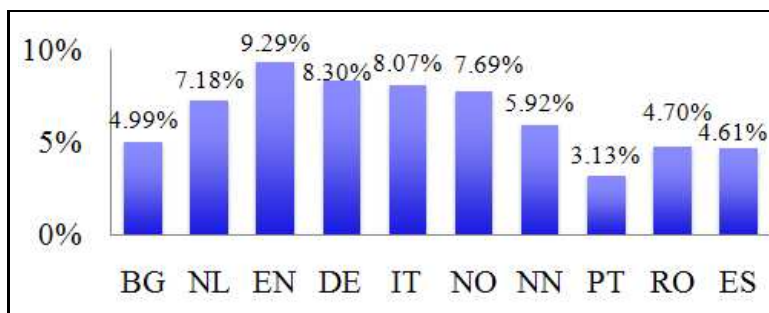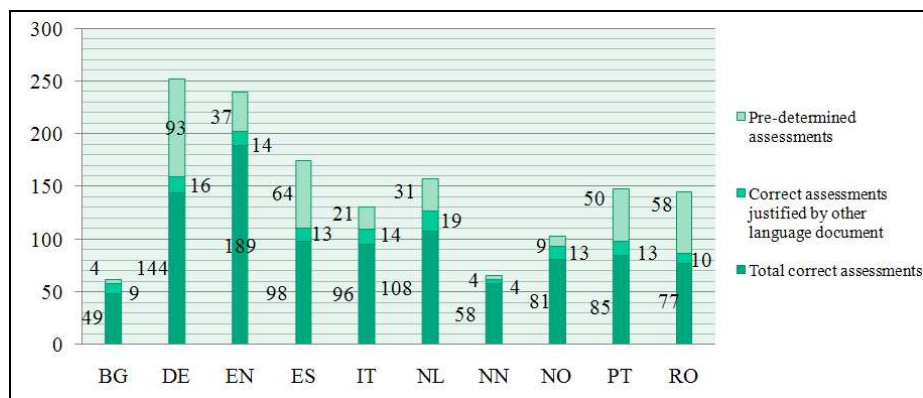
Figure 12: Language precision

Figure 13: Language authority: Languages where answers were found and also justifications

and due to the random choice, some topic owners (proposers) received more topics than others. In fact, a cursory examination of the final topics shows that language or culture distribution was quite skewed, with a predominance of Romanian and German topics, on the one hand, and a scarcity of Portuguese and Norwegian ones, on the other.

If we look at the topics per language, then the relative importance of English finally emerges: for the vast majority of topics the language with higher number of correct hits is English. Table 5 shows a selected sample of the topics per language. Most of the remaining ones did feature English as the decisive winner (the full table is available from the GikiCLEF site). One other thing that remains to be done is an

Table 5: Correct answers per language

Topic	BG	NL	EN	DE	IT	NO	NN	PT	RO	ES
GC-2009-07	8	24	23	16	15	15	12	16	16	16
GC-2009-09	3	8	7	8	7	7	8	7	7	7
GC-2009-19	0	2	17	5	3	1	0	6	0	7
GC-2009-27	1	0	1	0	0	0	0	0	0	0
GC-2009-34	15	20	21	22	20	20	20	27	5	20
GC-2009-48	1	0	0	1	0	0	0	0	0	0
GC-2009-50	0	0	4	10	0	0	0	0	0	0

investigation of how really different the several answers are, that is, are the answers relative to the same individuals or places, or rather different?

So, while GikiCLEF was able to demonstrate that there are systems that can answer (although still with poor performance) questions in these nine languages, the real utility for each language of processing also the other nine collections has not yet been established.

We present, for comparison, at the end of the paper, a description of the different monolingual Giki-

CLEFs, by presenting precision per topic for some of the languages. We would like to emphasize, however, that these views are somehow artificial on several counts: not only they represent the joint performance of the several participants, but for some languages, such as for example Italian or Norwegian, their collections were not processed in any way, that is; the hits came from processing e.g. the German, the Dutch and the English collections... So these could be called, in GikiCLEF parlance, "parasitic languages", and in fact it is interesting to note that they do attain better precision than other languages whose collections were processed, such as Spanish and Portuguese.

8 Investigating the difficulty of GikiCLEF 2009

As a general opinion it is fair to say that GikiCLEF was universally considered too difficult or ambitious, which resulted in that several prospective participants gave up and not even sent in results.

Many people strove hard to just be able to process the huge collections and minimally parse the topic renderings, and did not even consider cultural differences and/or crosslinguality. Our impression is that most participants did the bulk of processing in one main language, and then used naive and straightforward procedures to get answers in other languages. So, neither crosslinguality (differences in conveying related information) or multilinguality (the fact that different Wikipedias might produce different results) were really investigated by the first GikiCLEF participants.

Figure 14: Number of (total and correct) answers per topic

If we make a more detailed inspection of the topics and the systems' behaviors, we can identify the easiest and most difficult topics, through the display, in Figures 14 and 15, of the number of answers and the conjoined precision (taking all participants together) attained.

Figure 15: Precision per topic

Another feature we were expecting people to make use of was the justification field, which could in a way display the reasoning done. However, very few participants (only two) used a justification field, and apparently it was not very successful either, see Figure 16. In fact, the proportion of justified answers was only considered correct ca. 50% of the times. But we believe that if further justifications had been given by the participants their score would increase.
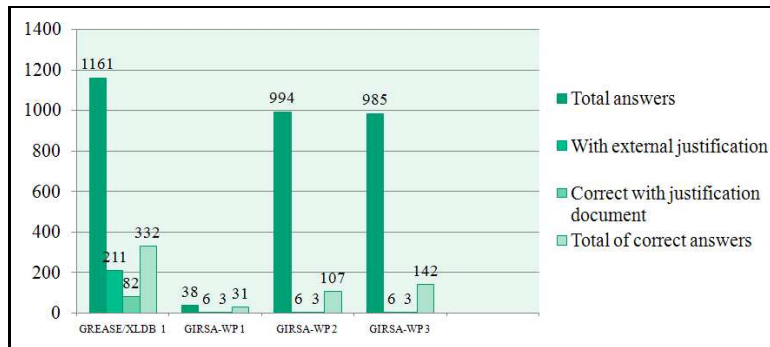

Figure 16: Extra justification by GikiCLEF participants

Several other things did not work out as expected, and in particular we believe now that some quizz-like or relatively strained topics ended up in the final topic list, while topic managers in general shied away from the formidable task to convey things peculiar to their own cultures to a set of foreigners, and decided for simpler topics to begin with.

Finally, we would have liked to see more practically oriented systems with a special purpose and an obvious practical utility to try their hand at GikiCLEF. Apparently most if not all participants were simply considering GikiCLEF too hard and had no independent system of their own to try out there. Again, this may prove the complexity of the task, or the fact that the audience was not appropriate. We hope that training with GikiCLEF materials, all of them made available on due course, may in any case help future systems to perform difficult tasks with semi-structured multilingual material, which we still believe is something required out there.

All resources compiled under GikiCLEF, as well as collections and Web access to SIGA, can be reached from http://www.linguateca.pt/GikiCLEF.

Acknowledgements

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007. Proceedings*, pages 722–735. Springer, 2008.

[2] Karen Bennett. Galileo's revenge: Ways of construing knowledge and translation strategies in the era of globalization. *Social Semiotics 17*, 17:171–193, June 2007 2007.

[3] Melissa Bowerman. The origins of children's spatial semantic categories: cognitive versus linguistic determinants. In J. Gumperz and S. C. Levinson, editors, *Rethinking linguistic relativity*, pages 145–76. Cambridge University Press, Cambridge, 1996.

[4] Nuno Cardoso, David Batista, Francisco Lopez-Pellicer, and Mário J. Silva. Where in the Wikipedia is that answer? the XLDB at the GikiCLEF 2009 task. In Francesca Borri, Alessandro Nardi, and Carol Peters, editors, *Cross Language Evaluation Forum CLEF 2009 Workshop*, 30 September - 2 October 2009. This volume.

[5] Silviu Cucerzan. Large Scale Named Entity Disambiguation Based on Wikipedia Data. In *The EMNLP-CoNLL Joint Conference*, pages 708–16, 2007.

[6] Marieke De Mooij. *Consumer Behaviour and Culture: Consequences for Global Marketing and Advertising*. London: Sage, 2004.

[7] Evgeniy Gabrilovich and Shaul Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence IJCAI 2007*, pages 1606–1611, 2007.

[8] Peter Gerrand. Estimating Linguistic Diversity on the Internet: A Taxonomy to Avoid Pitfalls and Paradoxes. *Journal of Computer-Mediated Communication*, 12(4), 2007.

[9] Frederic Gey, Ray Larson, Mark Sanderson, Hideo Joho, and Paul Clough. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Acessing Multilingual information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Revised Selected papers*, volume 4022 of *LNCS*, pages 908–919. Springer, Berlin, 2006.

[10] Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bishoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Di Nunzio, and Nicola Ferro. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Revised selected papers*, volume 4730 of *LNCS*, pages 852–876. Springer, Berlin, 2007.

[11] Valentin Jijkoun and Maarten de Rijke. WiQA: Evaluating Multi-lingual Focused Access to Wikipedia. In *The First International Workshop on Evaluating Information Access (EVIA)*, Tokyo, Japan, May 15 2007.

[12] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, Chicago and London, 1980.

[13] Thomas Mandl, Paula Carvalho, Fredric Gey, Ray Larson, Diana Santos, and Christa Womser-Hacker. GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Viviane Petras, editors, *Evaluating Systems for Multilingual*

and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Springer, 2009.

[14] Patricia O'Neill-Brown. The need for culturally adaptive agent behavior. In *Computational Models for Mixed Initiative Interaction: Papers from the 1997 AAAI Spring Symposium*, pages 117–21. AAAI Press, 1997.

[15] Philip Resnik and Noah A. Smith. The Web as a parallel corpus. *Computational Linguistics*, 29:349–380, September 2003.

[16] Diana Santos. *Translation-based corpus studies: Contrasting English and Port uguese tense and aspect systems.* Rodopi, 2004.

[17] Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Viviane Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, pages 894–905. Springer, 2009.

[18] D. I. Slobin. From "thought and language" to "thinking for speaking". In J. Gumperz and S. C. Levinson, editors, *Rethinking linguistic relativity*, pages 70–96. Cambridge University Press, Cambridge, 1996.

[19] Leonard Talmy. Lexicalization patterns: semantic structure in lexical forms. In Timothy Shopen, editor, *Language Typology and Semantic Description: Grammatical Categories and the Lexicon*, volume 3, pages 57–149, Cambridge, 1985. Cambridge University Press.

[20] Tony Veale. Enriched Lexical Ontologies: Adding new knowledge and new scope to old linguistic resources. In *(ESSLLI 2007)*, 2007.

[21] Fei Wu and Daniel S. Weld. Autonomously Semantifying Wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM on Conference on Information and Knowledge Management*, pages 41–50, 2007.
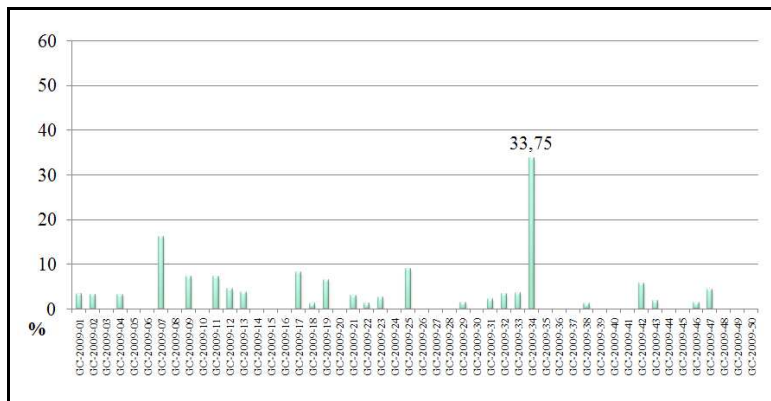
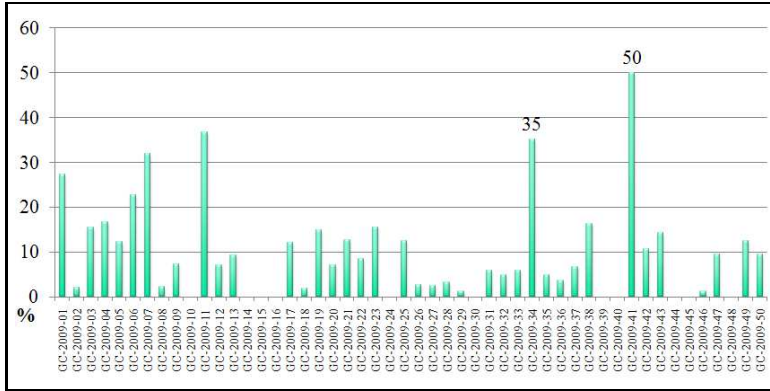Figure 17: Portuguese GikiCLEF: precision per topic

Figure 18: English GikiCLEF: precision per topic
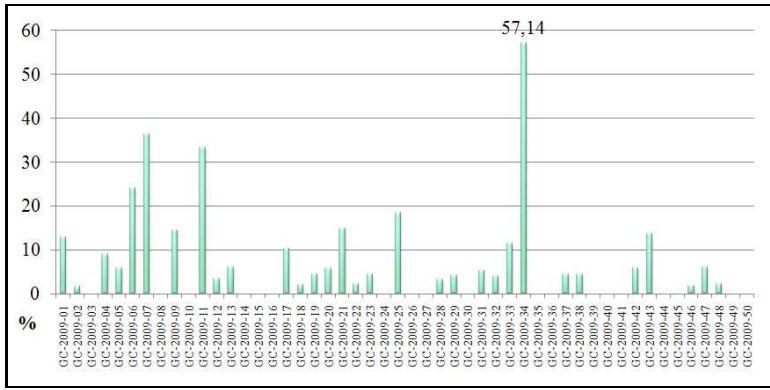

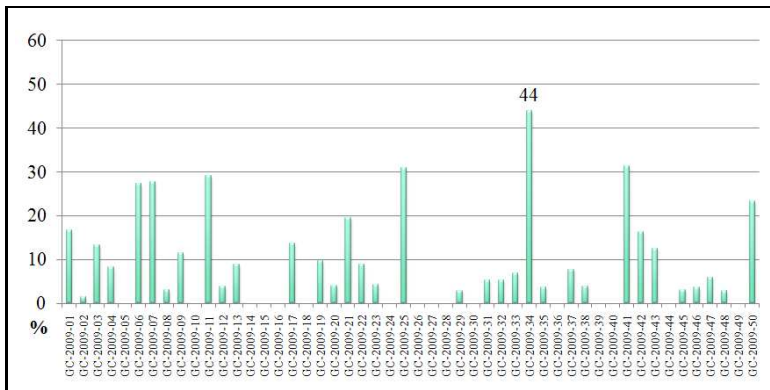Figure 19: Dutch GikiCLEF: precision per topic

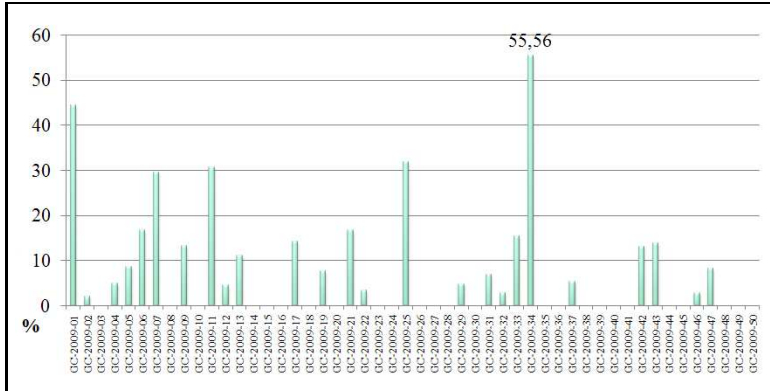Figure 20: German GikiCLEF: precision per topic

Figure 21: Italian GikiCLEF: precision per topic
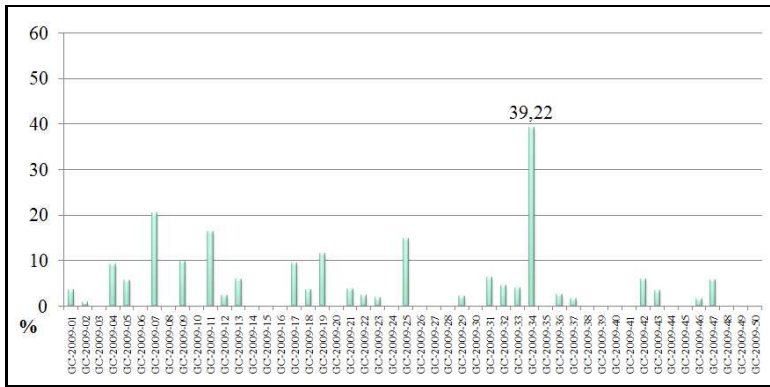

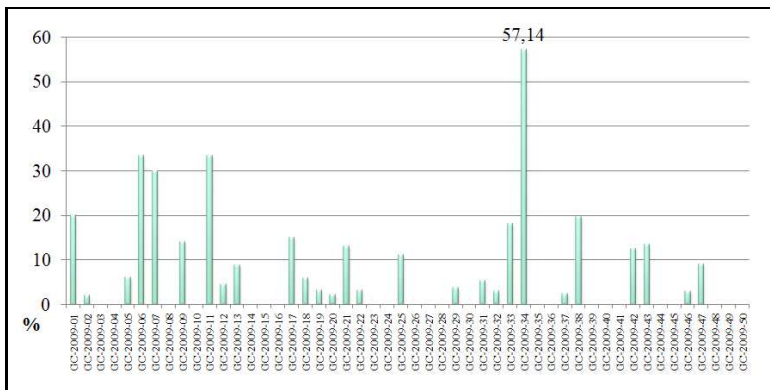Figure 22: Spanish GikiCLEF: precision per topic

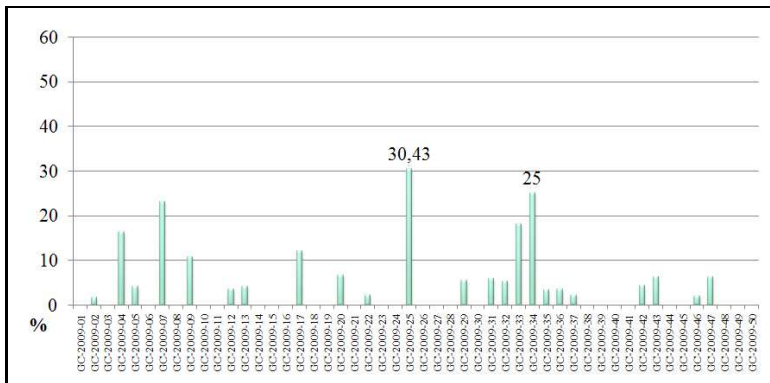Figure 23: Norwegian GikiCLEF: precision per topic
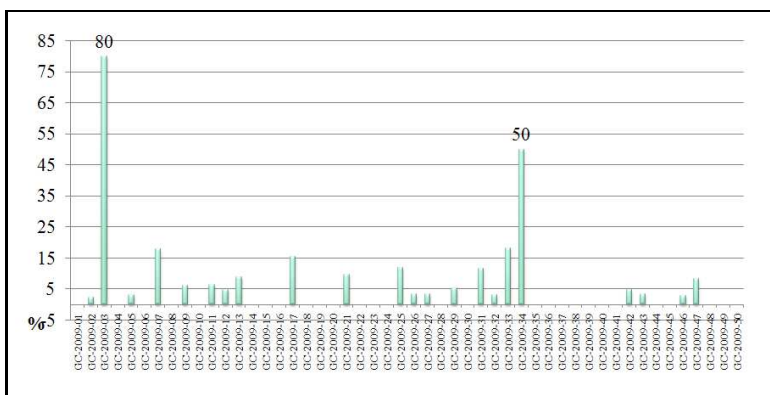
Figure 24: Romanian GikiCLEF: precision per topic

Figure 25: Bulgarian GikiCLEF: precision per topic