


## How geographic was GikiCLEF? A GIR-critical review

Diana Santos, Linguateca, SINTEF ICT  
Nuno Cardoso, Univ. of Lisbon, Faculty of Sciences, LaSIGE  
Luís Miguel Cabral, Linguateca, SINTEF ICT

POS CONHECIMENTO  
União Europeia FEDER  
UMIC  
FCCN

SINTEF Information and Communication Technologies



## GikiCLEF in a nutshell (1)

- Asking open list questions to Wikipedia
- A difficult task for people and for computers
- A realistic situation where crosslingual and multilingual skills may be of real interest


<http://www.linguateca.pt/GikiCLEF>

- A merger of QA and GIR, a follow-up of the GikiP pilot
- 10 Wikipedia collections, 50 culturally motivated topics
- 8 participants in 2009

SINTEF Information and Communication Technologies

## Examples of GikiCLEF-like systems in use

- In the middle of a football match on TV in a foreign station




- What south American countries have yellow in their football team?
  - Brazil
  - Colombia
  - Ecuador

SINTEF Information and Communication Technologies

## Examples of GikiCLEF-like systems in use

- Where was this American museum which had a Picasso painting on which we saw that program last Winter?




- Which American museums have Picasso works?
  - Museum of Modern Art (MOMA)
  - Museum of Fine Arts (Boston)
  - Solomon R. Guggenheim Museum
  - Metropolitan Museum of Art
  - National Gallery of Art
  - Art Institute of Chicago
  - Denver Art Museum
  - (...)

SINTEF Information and Communication Technologies

## Examples of GikiCLEF-like systems in use

- Which Romanian poets published volumes with ballads before 1931? ... which may have influenced Mircea Eliade? Preliminary research for a MA Thesis in Romanian literature...



Dimitrie Bolintineanu Vasile Alecsandri George Coșbuc Elisabeta de Neuwied

SINTEF Information and Communication Technologies

## The task as seen by a participant system

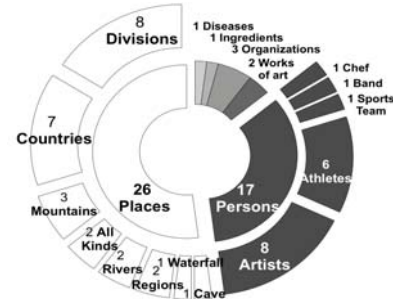
- In order to provide a correct answer, a system had to produce **justification** in at least one of the languages returned
- Even if correct, an answer would not be rewarded in GikiCLEF if its correctness were not possible to assess by a human reader
- Justification might be in the page itself, or in a set of other pages provided together
  - Example question: Name places where Goethe fell in love.
  - One correct answer is Leipzig. But in the article about Leipzig this is not stated ©
  - Where to find the justification? In "Johann Wolfgang von Goethe" page we find the following excerpt: "In Leipzig, Goethe fell in love with Käthchen Schönkopf and wrote cheerful verses about her in the Rococo genre."
  - Correct and justified GikiCLEF answer: **Leipzig**, Johann Wolfgang von Goethe

SINTEF Information and Communication Technologies

## Results from the topic group work (1)

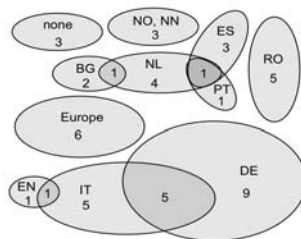
- 14 people (Corina Forascu, Pamela Forner, Danilo Giampiccolo, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Erik Tjong Kim Sang, Diana Santos, Julia Schulz, Yvonne Skalban, Alvaro Rodrigo Yuste + Paula Carvalho, Christian-Emil Ore) gathered 70 (> 50) “culturally-laden” topics
  - Some of them **cross-cultural**, in fact (like the influence of Italy in Hemingway or the spread of Picasso’s influence in the American continent)
  - Most around themes such as thinkers/writers/famous people, and of course places

## Expected answer type distribution of the 50 topics

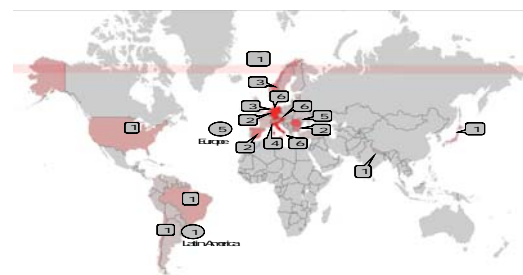


## Language/culture bias of topics

- Although we had eight non-English languages and one at least of each in the topic group, the culture bias was not uniformly distributed



## Geographical distribution of topics



## Results from the topic group work (2)

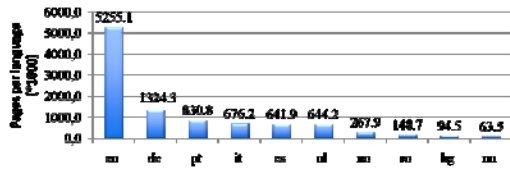
- How much insight did the topics gathered provide in terms of displaying truly geographical cross-cultural problems?
  - Alps are **subdivided differently** in different countries/languages
  - Flemish towns** turn out to be a not clear concept internationally
  - Not all language Wikipedias have the category **ski resort**
- Physical geography vs. human and political geography
  - Different relationship with culture and language
  - Temporal restrictions only used in connection with human geography
  - Nationality is a difficult issue
- Only two cases with unspecified spatial granularity (places where...)
- We allowed **visual** clues to decide (snow, colours and map left/right)

## Overview of topics (3)

- How was the universe of the collections?
- How were the topics related to hits (and justifications) in the collections?
- Was there a correlation between language/collection and topic bias?
- A lot of little-developed and inaccurate Wikipedia pages were found a posteriori, raising doubts on the initial assumption that searching Wikipedia crosslingually is a good idea

## Size of collections (pages \* 1k)

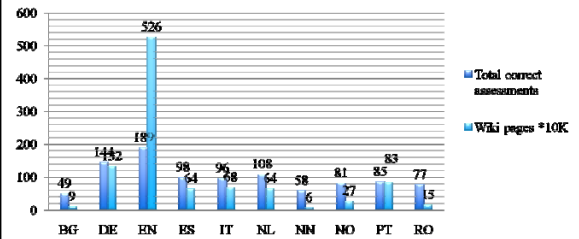
- Topic choice was **not** correlated with collection size



Collections dated June 2008, except for English SQL dump, from May and July 2008

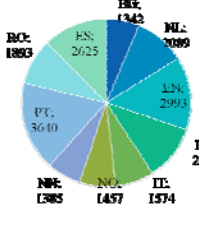
## Number of authoritative (correct and justified) answers per collection size

- Little correlation with collection size

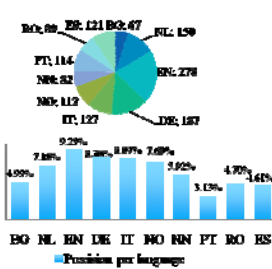


## Overview of results per language

### Total answers per language

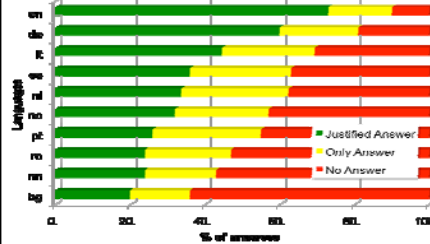


### Correct answers per language



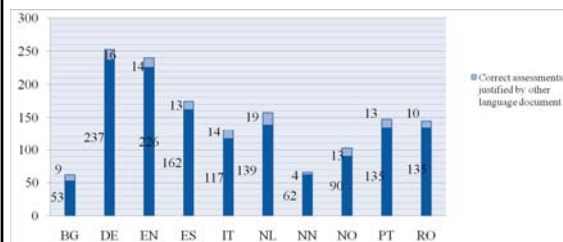
## Topic coverage in the pool, per language

- Still, some languages had a definite advantage over others: one could get 70% of the results in English alone against 20% only in Bulgarian



## Authoritative languages: where is the justification

- For each language, was the justification found in the corresponding Wikipedia?



## Inconsistencies and problems

- The language of the Vatican
  - Italian and/or Latin?

Official languages	Italian and Latin <sup>118</sup>
Andsprache	Italienisch (de facto) + Latin
Lingua official	latim
	italiana (de facto)
Offisielle språk	Latin, italiensk språk
	mest bruk
Официален език	italijanski <sup>119</sup>
Offisjele landsspråk	Latin
Limbă oficială	italiană
Lingua official	latino <sup>120</sup>

- Height of a Norwegian waterfall
  - EN: Rjukanfossen is a waterfall of 104 meter...
  - NO: Fossen har en total fallhøyde på 238 meter og høyeste lodrette fall er 104 meter
- Who died of tuberculosis?
  - Charlotte Brontë did not
- If a party changes name... was it founded when?

## Thanks to the GikiCLEF assessors

- Thanks to the **30** assessors: Sören Auer, Anabela Barreiro, Gosse Bouma, Luís Miguel Cabral, Nuno Cardoso, Leda Casanova, Luís Costa, Justin Dornescu, Ana Engh, Corina Forascu, Pamela Forner, Fredric Gey, Danilo Giampiccolo, Sven Hartrumpf, Katrin Lamm, Ray Larson, Laska Laskova, Johannes Leveling, Thomas Mandl, Cristina Mota, Constantin Orasan, Petya Osenova, Anselmo Peñas, Erik Tjong Kim Sang, Diana Santos, Julia Schulz, Yvonne Skalban, Rosário Silva, Kiril Simov, Alvaro Rodrigo Yuste

## Thanks to the GikiCLEF participants

- Richard Flemmings *et al.* (BBK\_UFRGS)
- Gosse Bouma & Sérgio Duarte (JoostER)
- Nuno Cardoso *et al.* (GREASE/XLDB)
- Justin Dornescu (EQUAL)
- Ray R. Larson (CHESHIRE)
- Sven Hartrumpf & Johannes Leveling (GIRSA-WP)
- Daniel Ferrés & Horácio Rodríguez (GIKITALP)
- Adrian Ifteene *et al.* (UAICGIK109)

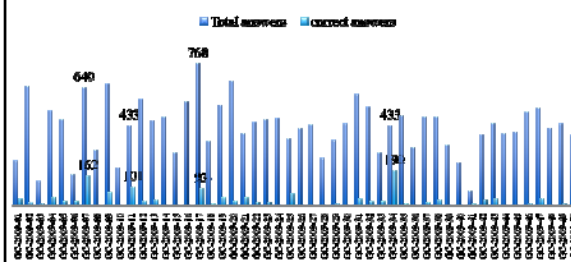
**17 runs submitted**

## Public resources

The GIRA package at <http://www.linguateca.pt/GikiCLEF/>

- Topic lists per language, in XML and text format
- Topic descriptions (in English, together with assessment decisions)
- Lists of correct answers
  - Correct and justified in some language: GikiCLEF\_answers\_correct\_justified.txt
  - Correct (because we know): GikiCLEF\_answers\_correct.txt
- Results (global, per language) and general statistics
- SIGA system, open source
- GikiCLEF collections

## Number of answers per topic



## Final discussion: Was GikiCLEF worthwhile?

- Recurrent comment: **too difficult!**
- Often not clear which (Wikipedia) pages were sought
  - Flag pages or country pages?
  - Team pages or country pages?
- Most (organizers) were probably not expecting so much work

↑ It is possible to create a system capable of answering many languages

↓ It pays to process English: The amount provided by each language, even in culturally-aware topics, is negligible

↓ The quality of other languages' Wikipedias is in strong need of improvement

## Concluding remarks

- We are offering GikiCLEF results to the GIR community
- We know that considerable empirical work remains to be done
  - Pool improvement
  - Study of the changes to Wikipedia after 2 years as far as the GikiCLEF topics are concerned
  - Generalization of the kinds of topics and problems tackled
- Measuring the "geographicity" (the degree or connection it has to geographical concepts and reasoning) of an evaluation contest, or of a collection, is not trivial, but more should be done in GIR
  - There are a lot of obvious statistics that can be gathered, from density of place-denoting NEs to number of related geographical queries for elements of a collection, etc