Providing Multilingual Access to FLICKR for Arabic Users

Paul Clough[1], Azzah Al-Maskari[1], and Kareem Darwish[2]

[1] Sheffield University, Sheffield, UK
[2] IBM, Cairo, Egypt
p.d.clough@sheffield.ac.uk

Abstract. In this paper we describe our submission for iCLEF2006: an interface that allows users to search FLICKR in Arabic for images with captions in a range of languages. We report and discuss the results gained from a user experiment in accordance with directives given by iCLEF, including an analysis of the success of search tasks. To enable the searching of multilingual image annotations we use English as an interlingua. An Arabic-English dictionary is used for initial query translation, and then Babelfish is used to translate between English and French, German, Italian, Dutch and Spanish. Users are able to modify the English version of the query if they have the necessary language skills to do so. We have chosen to experiment with Arabic retrieval from FLICKR due to the growing numbers of online Middle Eastern users, the limited numbers of interactive Arabic user studies for cross-language IR to date, and the availability of resources to undertake a user study.

1 Introduction

FLICKR[1] is a large-scale, web-based image database based on a large social network of online users. The application is used to manage and share personal (and increasingly more commercial) photographs. Currently FLICKR contains over five million accessible images, which are freely available via the web and are updated daily by a large number of users. The photos have multilingual annotations generated by authors using freely-chosen keywords (known as a folksonomy). Similar systems are also emerging for collections of personal videos[2] (e.g. youtube.com and CastPost).

Despite the popularity of FLICKR on a global basis, to-date there has been little empirical investigation regarding multilingual access to FLICKR. A common remark of Cross Language Information Retrieval (CLIR) is why would users want to retrieve documents that they (presumably) cannot read. Of course, in the case of image retrieval the motivation for CLIR is much stronger, because for many search tasks users are able to judge the relevance of images without the

[1] http://flickr.com
[2] http://www.techcrunch.com/2005/11/06/the-flickrs-of-video/

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 205–216, 2007.

need of additional text, thereby eliminating the need for translation of search results. This linguistic-neutrality of images makes text-based image retrieval an ideal application for CLIR.

For our submission to iCLEF 2006, we wanted to experiment with providing an interface which would enable users to query FLICKR in Arabic. We selected Arabic because of the availability of resources to us locally, the limited number of interactive Arabic user studies so far in CLIR, the growing number of online Middle Eastern users and the limited availability of online material in Arabic. According to Reuters and ABC Science Online[3] there are currently only 100 million Arabic web pages, constituting 0.2% of the total pages on the web. There is no doubt that a system designed for Arabic users that is able to search English documents and other languages would open new possibilities both in terms of the quantity of accessible topics and the quality of the items retrieved.

The remainder of the paper describes the system developed for Arabic users, the experiments, results and conclusions. The three main aims of our work were: (1) to analyse the tasks offered by iCLEF, (2) to analyse our initial interface: query translation and use of English as an interlingua, and (3) to observe the searching behaviour of users.

2 The System: FLICKRArabic

We developed an Ajax-based online application called FLICKRArabic (shown in Fig. 1) to provide query translation to the FLICKR API[4]. The system centres on translating user's queries from Arabic into English (the interlingua) and then consequently translating into French, Spanish, German, Italian or Dutch. This is necessary because many translation resources provide only Arabic to English translation. The English translation is shown to users who can modify the query if they have sufficient language skills. In the case of polysemous Arabic query words, translations for all available senses are displayed to the user and the user is able to ignore or select the correct translations. Translation between Arabic-English is performed using a bilingual dictionary (described in Section 2.1). Translation between English and other languages is performed by using a custom-built wrapper for Babelfish[5], which is an online Machine Translation (MT) system. This means that beyond English, users have little control over translation. The design and implementation is described further in Section 2.2.

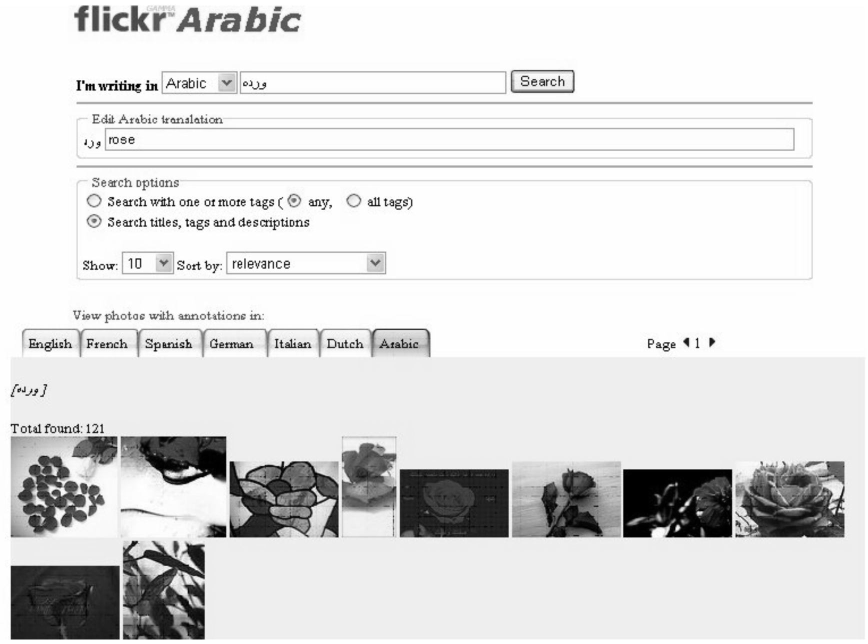
2.1 Language Resources

Two translation resources have been used to create the application. The first is an Arabic-English bilingual dictionary, the second is the online MT tool Babelfish. To create the bilingual dictionary, two bilingual term lists were constructed

[3] http://www.abc.net.au/news/newsitems/200604/s1624108.htm

[4] http://www.flickr.com/services/api/

[5] http://babelfish.altavista.com



I'm writing in Arabic ورده

Edit Arabic translation

ورد rose

Search options

○ Search with one or more tags (● any, ○ all tags)

● Search titles, tags and descriptions

Show: 10 Sort by: relevance

View photos with annotations in:

[ورده]

Total found: 121

Fig. 1. FLICKRArabic interface

using two Web-based MT systems, namely Tarjim[6] and Al-Misbar[7]. In each case, a set of unique isolated English words found in a 200 MB collection of Los Angeles Times news stories was submitted for translation from English into Arabic [4]. Each system returned at most one translation for each submitted word. In total, the combined bilingual term lists contained 225,057 unique entries. In preprocessing the Arabic text, all diacritics and kashidas (character elongations) were removed, the letters *ya* and *alef maqsoura* were normalised to *ya* and all the variants of *alef* and *hamza*, namely *alef*, *alef hamza*, *alef maad*, *hamza*, *waw hamza*, and *ya hamza*, normalised to *alef*, and lastly all words were stemmed using Al-stem[8].

2.2 Interface Design and Functionality

Design of the system has emerged from an interactive evaluation design process. A user-centered approach was implemented, where five Arabic potential users were involved during the design and implementation phases of the system (following the advice of [7]). During the pilot session, users were observed and questioned about their cross-language actions (e.g. editing the translation of the

[6] `http://tarjim.ajeeb.com`, Sakhr Technologies, Cairo, Egypt.

[7] `http://www.almisbar.com`, ATA Software Technology Limited, North Brentford Middlesex, UK.

[8] `http://www.glue.umd.edu/~kareem/research/`

Arabic query and flipping through the results of other languages). Previous research which has considered user interaction in the formulation of multilingual queries includes the Keizai system [9], ARCTOS [10], MULINEX [2], WTB [8], MIRACLE [5], EUROVISION [3] and CLARITY [11]. The basic functionality of this system is as follows:

- Users can search FLICKR using initial English or Arabic queries
- If searching in English, the system calls the FLICKR API and displays results
- If searching in Arabic:
  - The query is first converted from UTF8 to CP1256 (using *iconv -f utf8 -t cp1256*) and stemmed (using the *stem_cp1256.pl* Perl program)
  - The query is then translated into English using the Arabic-English dictionary
  - The system returns all dictionary matches (senses and synonyms) for each translated query
  - Users can modify the English translation by deleting unwanted terms or adding their own. Previous research has shown it useful to enable this for most bilingual users
  - The English query is then used for searching FLICKR using the API[9]
  - The user can additionally view photos annotated in Arabic only (for comparison with other languages)
- Results for each language are displayed in separate tabs with the total number of images found displayed (when each language selected)
- The user can view photos with annotations in any one of the five languages: French, Spanish, German, Italian and Dutch[10]
- Users can select the following search options:
  - Display 10, 20, 50, 100 or 200 images per page
  - Sort images by a relevance score or interestingness
  - Search all annotation text (titles, tags and descriptions) or tags only (all query terms or any)

Results from an English search were displayed to users first (left-hand tab), because during initial tests most users were able to make use of images with English annotations (ordering of languages was therefore arbitrary). Results for Arabic were also provided to enable users to compare results from searching FLICKR using purely Arabic.

3 The Experiment

To obtain feedback on the implemented system, we recruited 11 native Arabic speakers to carry out the tasks specified by iCLEF [6]. The subjects were undergraduate and postgraduate students with a good command of English [11]. The mean age of the 11 users was 28 years old, and 85% stated they typically

[9] Only photos posted before 1/6/2006 were returned.

[10] If the query is not translated, it remains in English and results for this are returned to the user.

[11] We are currently planning experiments with monoglot users: those who can only make use of Arabic.

searched the Web using English. They also had the following characteristics: 82% used the Internet several times a week, all had a great deal of experience with point-and-click interfaces, 46% searched for images very often and 46% of those people often found what they were searching for.

Subjects were asked to perform 3 tasks: (1) a classical ad-hoc task: “find as many European parliament buildings as possible, pictures from the assembly hall as well as from the outside” (parliament); (2) a creative instance-finding task: “find five pictures to illustrate the text - the story of saffron - with the goal being to find five distinct instances of information described in the narrative: saffron, flower, saffron thread, picking the thread/flower, powder, dishes with saffron (saffron); and (3) a visually orientated or known-item task: given a picture, find the name of the beach on which the crab is resting (crab). More details of the tasks can be found in [6].

In these experiments, users first completed a preliminary questionnaire, then spent 20 minutes on each task. Tasks were assigned randomly to users to reduce the effects of task bias on the results (e.g. user 1 performed task 2, 3, then 1, user 2 performed task 3, 1, then 2). We also asked subjects to perform a final search where they were able to search for images on any topic. Finally, we asked users to complete a questionnaire to establish their overall satisfaction with and impressions of the system. During the experiment, we recorded some attributes of the task such as time taken and queries input, as well as taking notes of the user’s searching behaviour during each task.

4 Results and Observations

4.1 User Effectiveness

We first discuss how well users were able to perform the tasks[12]. Almost all users (10 out of 11) were able to perform task 3 (crab) successfully[13]. Table 1 shows the results for task 1 (parliament). For this task, from the total number of images found, we deemed which ones are *correct* and of these we counted the number of *unique* European parliament buildings. To compute recall we divided unique by correct; for precision we computed correct divided by total. We also divided the total number of pictures found between those of the inside of the building versus the outside. Across all users we obtained a recall of 0.69 and precision of 0.84. This varied between users with some scoring higher recall (e.g. user 3) and others achieving higher precision (e.g. users 7 and 9). Of the images found, almost twice as many were of the outside of buildings.

Table 2 shows the results for task 2 (saffron). In this task users were asked to find 5 images to illustrate each part (or instance) to the story of saffron. Users were given one point for retrieving each instance (*counted*) and no credit

[12] The system effectiveness and the correlation between user and system effectiveness is explored further in [1].

[13] Despite users not being familiar with German, they were able to recognise the name of the beach.

Table 1. Search results for task 1 (parliament)

User	total found	correct	unique	recall	precision	inside	outside
1*	–	–	–	–	–	–	–
2	20	11	3	0.27	0.55	9	11
3	11	8	7	0.88	0.73	4	7
4	20	17	11	0.65	0.73	4	13
5	13	12	8	0.67	0.92	4	9
6	12	10	9	0.90	0.83	3	9
7	12	12	8	0.67	1.00	5	9
8	8	7	6	0.86	0.88	5	3
9	12	12	9	0.75	1.00	0	12
10	13	10	7	0.70	0.77	4	9
11	10	9	7	0.78	0.90	3	7
Total	131	108	75	0.69	0.84	41	80

*The system did not function correctly for this user during this task

for repeated instances (i.e. no additional credit for selecting two images of the same aspect of the story). Results show the number of images found for each aspect/instance and the precision (counted divided by total). Overall precision was 0.70 for this task and again, precision varied between users as some were good at instance-finding (e.g. users 3 and 9) and others were less successful (e.g. users 5 and 11).

Table 2. Search results for task 2 (saffron)

User	flower	thread	food	powder	picking	total	counted	precision
1	3	2	–	–	–	5	2	0.40
2	1	1	2	–	1	5	4	0.80
3	1	1	1	1	–	4	4	1.00
4	1	2	2		–	5	3	0.60
5	3	–	–	–	–	3	1	0.33
6	2	1	1	–	1	5	4	0.80
7	1	1	1	1	1	5	5	1.00
8	2	1	1	–	1	4	5	0.80
9	1	1	1	1	1	5	5	1.00
10	–	4	1	–	–	5	2	0.40
11	2	1	2	–	–	5	3	0.60
Total	18	17	11	2	4			0.70

Table 3 shows the number of users who judged results for each language and task as highly relevant, partially relevant or not relevant. It would appear that users found relevant images with annotations in most of the languages, except Arabic. Users commented that they were disappointed with the Arabic results. This would suggest that multilingual access to FLICKR could improve retrieval. For task 2 (saffron), most users found relevant images in Italian which is likely due to the narrative mentioning Italy.

Table 4 shows the user's satisfaction with the accuracy and coverage of the search results. Overall it appears that users were very satisfied with the accuracy of search results from FLICKRArabic for tasks 1 and 3, but less satisfied with the accuracy of results for task 2. Similarly, users appear in general satisfied with the coverage of results (again less so for task 2). We asked users whether

Table 3. User relevance per language

	Task 1 (parliament)			Task 2 (saffron)			Task 3 (crab)		
Language	Highly	Partially	Not	Highly	Partially	Not	Highly	Partially	Not
English	9	2	0	7	4	0	3	5	3
French	4	5	2	7	4	0	2	7	2
Spanish	5	4	2	7	4	0	4	6	1
German	4	5	2	8	3	0	2	6	3
Italian	5	4	2	6	5	0	3	3	5
Dutch	2	5	4	3	5	3	3	3	5
Arabic	0	0	11	0	0	11	5	3	3

they felt accuracy or coverage was more important for the tasks 1 and 2. In task 1 (parliament), 7 users favoured accuracy and 4 preferred coverage; in task 2 (saffron) 8 users favoured accuracy and 3 preferred coverage. Overall, it would appear that users would prefer more accurate results, which likely reflects the precision-orientated nature of the tasks.

Table 4. User's satisfaction with accuracy and coverage

	Accuracy			Coverage		
Task	Highly	Partially	Not	Highly	Partially	Not
1 (parliament)	10	1	0	8	2	1
2 (saffron)	5	6	0	5	6	0
3 (crab)	10	1	0	9	2	0

We asked users about the usefulness of the results for each task and overall 70% of users were very satisfied with the results (30% partially satisfied). Table 5 indicates how user rate the importance of factors in helping them to determine the usefulness of the images for each search task. Users rated these as very important (v. imp.), important (imp.) and not important (unimp). For task 1, users found textual information very important in addition to the image itself. We expected this as users need to check the annotations to determine whether a parliament building is European or not. In task 2, users found the image and caption to be the most important for determining relevance. Users were able to identify possible pictures of saffron, but needed the captions to confirm their decision. In task 3, as expected, users found the visual content of the photos most useful. This reflects the fact that this task is more visual in nature. Users also found foreground and background text useful to determine the beach where the crab was placed.

4.2 Users' Search Behaviour

The following observations regarding search behaviour were observed during the experiment: users typically viewed initial results in English before trying other languages. This is because they were able to read annotations in English. Two main strategies for searching prevailed: some users input fewer queries and looked through many pages of results; others input many queries and if no relevant found

Table 5. User-rated usefulness of image attributes

	Task 1 (parliament)			Task 2 (saffron)			Task 3 (crab)		
	V. imp.	Imp.	Unimp.	V. imp.	Imp.	Unimp.	V. imp.	Imp.	Unimp.
Image only	8	3	0	8	2	1	11	0	0
Image and caption	10	0	1	8	3	0	5	6	0
Comments	7	3	1	5	5	1	1	6	4
Foreground details	7	1	3	4	5	2	6	3	2
Background details	1	3	7	1	3	7	6	4	1
Previous knowledge	2	2	7	5	4	2	2	1	8

in the first page of results, they reformulated the query. For the search results, some users would systematically look through results for each language from left to right; others would start with the languages which returned the least number of results (testing each language first). Most users selected 100 images at a time to view in the search results suggesting they are able (and willing) to view a large number of thumbnails.

4.3 Users' Comments on the Tasks

To determine the success of each task, we gathered users' comments on different aspects of the tasks as shown in Tables 6 and 7. Overall (from Table 6) it would appear that tasks 1 and 3 were the clearest, with task 1 being the easiest, task 3 being the most familiar, and tasks 2 and 3 being the most interesting to users. Interestingly, the majority of users did not find any of the tasks relevant to them. This is primarily because these topics were not designed specifically for Arabic users who are likely to search for different topics than Europeans and the tasks themselves were not entirely realistic (e.g. searching for a crab on a beach to find a specific location). It was also interesting to find that users were reluctant to search for crab, because the equivalent Arabic word has another sense, namely cancer. Also, the query for saffron was alien to most of the male searchers who did not cook and therefore were unsure what saffron was or looked like.

Table 6. User's assessment of the search tasks (1)

	Task 1 (parliament)			Task 2 (saffron)			Task 3 (crab)		
	Highly	Partially	Not	Highly	Partially	Not	Highly	Partially	Not
Clear	9	1	1	4	4	3	11	0	0
Easy	8	3	0	7	4	0	6	4	1
Familiar	7	3	1	3	8	0	9	2	2
Interesting	5	3	2	6	5	0	7	3	1
Relevant	2	4	5	1	4	6	1	4	6

In Table 7 that compares between tasks, users found task 1 to be the most interesting and easiest, tasks 1 and 3 to be the most enjoyable and task 1 to be the most realistic out of the three tasks. Users commented that task 3 was very unrealistic and did not represent the type of search task they would perform. For iCLEF organisers, this might indicate that concentrating on adhoc search is more likely to represent users' tasks and is less likely to be artificial.

Table 7. User's assessment of the search tasks (2)

	Interesting			Easiest			Enjoyable			Realistic	
Task	Most	Somewhat	Least	Most	Somewhat	Least	Most	Somewhat	Least	Yes	No
Task 1	6	1	4	7	1	3	4	5	2	6	5
Task 2	3	6	2	1	7	3	2	6	3	3	8
Task 3	2	4	5	3	4	4	4	3	4	1	10

4.4 Free-Search Task

If users could search for their own topics, what would they search for? We asked users to submit their own queries to determine the types of topics that would be representative or interesting to this user group. Table 8 shows the queries submitted by each user, the query language used, and the language of the annotations viewed in the results. As expected, most users typed queries more related to their culture (e.g. names of places such as Oman and Leptis and objects such as a mosque) and interests (e.g. welding). Most users searched using Arabic but found results in other languages helpful or useful. For queries containing out-of-vocabulary terms (e.g. Leptis), users had sufficient language skills to search in English.

Table 8. User's queries for free-search task

User	Query	Language of query	Languages of viewed results
1	Damascus	Arabic	Arabic, German, then English
2	Leptis (a city in Libya)	Arabic (not in dict), then English	Arabic, then English
3	Shef Uni, Mecca, Jeddah	Arabic	English, Arabic, then right-left
4	Tower, Skyscraper, Suspension bridge	Arabic	Arabic, English, then left-right
5	Damascus, Mecca	Arabic	Arabic
6	Orientalism	Arabic (not in dict), then English	English
7	Welding, Pyramid	English	English
8	Mosque	Arabic	English, German, Italian, Dutch, Arabic
9	Castles in Oman, andalus, Hamraa palace	Arabic	Arabic, then English (few results)
10	Cats, Muscat	Arabic and English	English, Arabic, French; English, French, left-right
11	Manchester, England, Libya	English	English

4.5 Overall User Comments

Table 9 indicates overall users' comments about the system we implemented. This represents users' satisfaction as recorded by the categories: very satisfied, partially satisfied and not satisfied. User's were very positive about the system and definitely found the provision of multilingual access to be useful to them (the ability to view pictures with annotations in various languages). Most users (9 out of 11) were very satisfied with the multilingual search results (compared against using FLICKR as is). However, the majority of these users were not happy with the query translation (7 partially satisfied and 1 not satisfied). Indeed we found that because this user group had good English language skills, the translation from Arabic to English was actually an unnecessary step and most preferred to formulate and modify queries in English (10 out of 11 users were willing to modify the English version of the query, and all users would enter the English version of a query term if not in the dictionary and add synonyms).

Table 9. User's overall satisfaction rating of the system

	Very	Partially	Not
Overall success	7	4	0
Multilingual usefulness	11	0	0
Multilingual satisfaction	9	2	0
Use system again	8	0	3
Recommend system to friend	11	0	0
Easiness of use	11	0	0
Quality of translation	3	7	1
Willingness to modify query	10	0	1
Entry of synonyms	11	0	0

Most users (8 out of 11) would use the system again and all users would recommend the system to a friend or colleague. All users were very satisfied with the ease with which the system could be used. Many users said they would have viewed most non-English annotations if translations had been provided in English (or Arabic). This suggests that some form of document translation could improve a users' search experience.

5 Discussion and Conclusions

The goal of our work was to build and test a simple Arabic query interface to FLICKR enabling users to view images with annotations in a range of languages. To enable Arabic translation into multiple languages, we first translated into English (interlingua) using a bilingual Arabic-English dictionary. From initial user testing, we decided to show users the English translations and allow them to edit as desired. The English version of the query was then translated into other languages as users requested to view results in those languages using the Babelfish MT system.

Overall with the group of users recruited for this experiment, we found that providing Arabic as an initial query language was unnecessary and caused more frustration than usefulness due to poor translation or out of dictionary words. Users were much happier submitting and reformulating queries in English (particularly for the tasks set which were orientated to Europeans as opposed to Arabs). Some users expressed the need for the ability to search in Arabic in some cases (e.g. when they are unable to formulate a query in English), but this was not the case for most of these tasks. However, users did comment that being able to start the search in Arabic to obtain some terms in English was a useful way to begin their search. Some users also suggested that being able to combine Arabic and English queries would be useful.

Compared to the current FLICKR system, it would seem that being able to submit an English query and translate it into multiple languages is considered as very beneficial to end users. It was particularly apparent with some queries where search results are very much language-dependent and different (e.g. searching for *car* typically produces British-built cars for English and *voiture* produces French-built cars). It would, therefore, seem more important to focus on this part of the system than initial query translation. Users were generally complementary of our system and they were able to carry out the search tasks set with reasonable success: overall precision of 0.84 for task 1, precision of 0.70 for task 2 and 10 out of 11 users completed task 3 successfully. Further work is planned in the following areas:

- Running the experiment again with users who are less proficient in English and would be less likely to reformulate English versions of queries.
- Improving query translation by increasing the size of the dictionary, handling the translation of phrases, and enabling the user to correct erroneous dictionary entries.
- Presenting results to users with different ranking and clustering strategies to reduce the number of images users view to find relevant images.
- Performing simultaneous searches in differnet languages to provide users with a summary of the number of results in each language. This might help users who typically view languages which exhibit the fewest number of results first.
- Translating image annotations in FLICKR results.

Acknowledgments

This work has been partially supported by the MultiMatch project (FP6-2005-IST-5-033104) under the EU Sixth Framework.

References

1. Al-Maskari, A., Clough, P., Sanderson, M.: User's effectiveness and satisfaction for image retrieval. In: Workshop Information Retrieval, University of Hildesheim, Germany (October 9–11, 2006)
2. Capstick, J., Diagne, A., Erbach, G., Uszkoreit, H., Cagno, F., Gadaleta, G., Hernandez, J., Korte, R., Leisenberg, A., Leisenberg, M., Christ, O.: Mulinex: Multilingual web search and navigation (1998)
3. Clough, P., Sanderson, M.: User experiments with the eurovision cross-language image retrieval system. 57(5), 679–708 (2006)
4. Darwish, K., Oard, D.W.: Clir experiments at maryland for trec 2002: Evidence combination for arabic-english retrieval (2002)
5. Dorr, B.J., He, D., Luo, J., Oard, D.W., Schwartz, R., Wang, J., Zajic, D.: iclef 2003 at maryland: Translation selection and document selection. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 435–449. Springer, Heidelberg (2004)

6. Gonzalo, J., Karlgren, J., Clough, P.: iclef 2006 overview: Searching the flickr www photo-sharing repository. In: Proceedings of Cross-Language Evaluation Forum (CLEF2006) Workshop, Alicante, Spain (2006)
7. Hackos, J.T., Redish, J.C.: User and task analysis for interface design. John Wiley & Sons, Inc., New York (1998)
8. Penas, A., Verdejo, F., Gonzalo, J.: Website term browser: Overcoming language barriers in text retrieval. Journal of Intelligent and Fuzzy Systems 12(3–4), 221–233 (2002)
9. Ogden, W., Cowie, J.: Keizai: An interactive cross-language text retrieval system(1999)
10. Ogden, W., Davis, M.W.: Improving cross-language text retrieval with human interactions. In: Proceedings of the Hawaii International Conference on System Science (HICSS-33), vol. 3 (2000)
11. Petrelli, D., Levin, S., Beaulieu, M., Sanderson, M.: Which user interaction for cross-language information retrieval? design issues and reflections. Journal of the American Society for Information Science and Technology (JASIST) 57(5), 709–722 (2006)