Concept Hierarchy across Languages in Text-Based Image Retrieval: A User Evaluation

Daniela Petrelli **Paul Clough**

Department of Information Studies, University of Sheffield,
Regent Court, 211 Portobello Street,
S1 4DP, Sheffield, UK
{d.petrelli, p.d.clough}@sheffield.ac.uk

Abstract. The University of Sheffield participated in Interactive ImageCLEF 2005 with a comparative user evaluation of two interfaces: one displaying search results as a list, the other organizing retrieved images into a hierarchy of concepts displayed on the interface as an interactive menu. Data was analysed with respect to effectiveness (number of images retrieved), efficiency (time needed) and user satisfaction (opinions from questionnaires). Effectiveness and efficiency were calculated at both 5 minutes (CLEF condition) and at final time. The list was marginally more effective than the menu at 5 minutes (no statistical significance) but the two were equal at final time showing the menu needs more time to be effectively used. The list was more efficient at both 5 minutes and final time, although the difference was not statistically significant. Users preferred the menu (75% vs. 25% for the list) indicating it to be an interesting and engaging feature. An inspection of the logs showed that 11% of effective terms (i.e. no stop-words, single terms) were not translated and that another 5% were ill translations. Some of those terms were used by all participants and were fundamental for some of the tasks. Non translated and ill translated terms negatively affected the search, hierarchy generation and, results display. More work has to be carried out to test the system under different setting, e.g. using a dictionary instead of MT that appears to be ineffective in translating users' queries that rarely are grammatically correct. The evaluation also indicated directions for a new interface design that allows the user to check query translation (in both input and output) and that incorporates visual content image retrieval to improve result organization.

Track: Interactive Image CLEF

Categories: H.1.2 [**User/Machine systems**]: Human factors. H.3.3 [**Information Search and Retrieval**]: Clustering. H.5.2 [**User Interfaces**]: Evaluation/methodology, Interaction styles.

Keywords: Italian, English, user evaluation, concept hierarchy, text-based image retrieval.

1 Introduction

Providing an intuitive summary of the search results is considered a benefit for users of IR systems. Different types of result summaries have been proposed in the past (see (Hearst, 1999) for a survey) and a variety of clustering techniques have been developed to group documents into topically-coherent sets. This is expected to help users in browsing through the search results, obtain an overview of the main topics/themes and help focus their inspection (i.e. limit exploration to only those clusters likely to contain relevant documents).

Organizing a set of documents automatically based upon a set of categories (or concepts) derived from the documents themselves is an obviously appealing goal for IR systems: it requires little or no manual intervention (e.g. deciding on thematic categories) and, like unsupervised classification, depends on natural divisions in the data rather than pre-assigned categories (i.e. requires no training data). Concept hierarchy generation (Sanderson and Croft 1999) is one such method: it automatically associates terms extracted from a document set and organizes them into a hierarchy, each term representing a group of documents.

This technique has been successfully employed to help users search and browse textual documents (Joho et al. 2004) and as a way of organizing retrieved images (Clough et al. 2005). For this year's Interactive Image CLEF at Sheffield University we decided to explore the use of concept hierarchies across languages and in the context of image retrieval. The language pair used was Italian as source and English as destination.

2 The Experimental Setting

Our participation in the Interactive Image CLEF track focused on evaluating the idea of concept hierarchy across languages: the retrieved images were organised into a hierarchical menu based on concepts automatically extracted from the image metadata and translated from English into Italian. This interaction mode was compared to a simple listing of the result, used as baseline. In both cases, the system used a version of CiQuest (Joho et al. 2004) developed initially for investigating interactive query expansion with a standard textual document collection (TREC) and modified to satisfy cross-language needs. The query entered by the user (in Italian) was sent to AltaVista for automatic translation into English via BabelFish. The returned translated query was then used to search the image collection provided by Image CLEF: historic photographs from St. Andrews University Library. The standard version of the Okapi search engine was used to perform retrieval: a probabilistic retrieval model based on the BM25 weighting function (Robertson et al. 1995).

In the *list interface*, the baseline, results were displayed as a ranked list (Figure 1). Images were ordered by the BM25 score between terms in the captions and query. The entire CiQuest interface (including results) was then translated into Italian using the Web page translation service offered by AltaVista, and finally displayed to the user. Clicking on the image title showed a larger version of the image with caption.

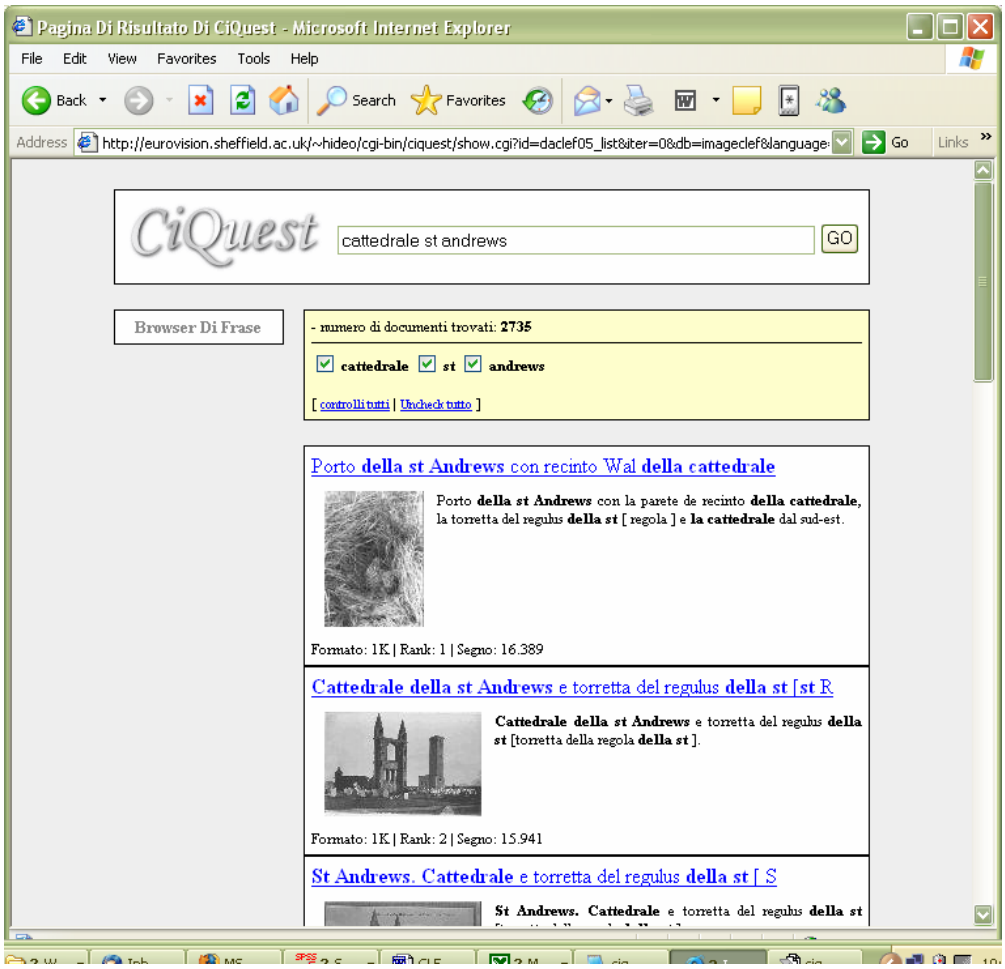

Fig. 1 The basic cross-language image retrieval system (list interface) in Italian.

Limitations in the interaction were due to both the use of the CiQuest interface and AltaVista's translation service: only the first 200 images could be seen whichever the number of retrieved images; and the query translation was not displayed so the user was not aware of which terms had been translated and which were not, or which translation was actually used to search. This lack of feedback and control contradicts our previous findings (Petrelli et al. 2005), however it was considered more important to run the evaluation even under limited conditions as this could help in better understanding CLIR in the context of interactive image retrieval and provide useful directions for future research, as it was for our participation in CLEF 2001 (Petrelli et al. 2004).

Figure 2 shows the *menu interface* tested in the comparative user evaluation. On the left of the result list, a DHTML menu that represented the concept hierarchy is displayed. It is dynamically generated from the captions of the set of retrieved images. Following Sanderson and Croft (1999), words and noun phrases (called concepts) are extracted from the captions and organized into a hierarchy of terms. The selection of concepts is based upon term co-occurrence (the same term occurring in multiple captions), term frequency, and statistical relation. The hierarchy is then used to generate the menu: each term is displayed together with an image randomly selected from the set associated with the term. By clicking on the image/term the user selects the group of images that corresponds to it (the group size is the number in parenthesis). Groups are not mutually exclusive and the same image may appear in more than one group depending on its caption. As the menu is generated by the captions of retrieved images, it is only loosely related to the entered query, e.g. the terms in the hierarchy may not be those issued by the user. Figure 2 shows how the result of a query (the same used in Figure 1) is displayed. Here the user has selected the menu item “timpano orientale” (Italian translation for “oriental gable”) and the 12 images in the group are displayed in the results list.

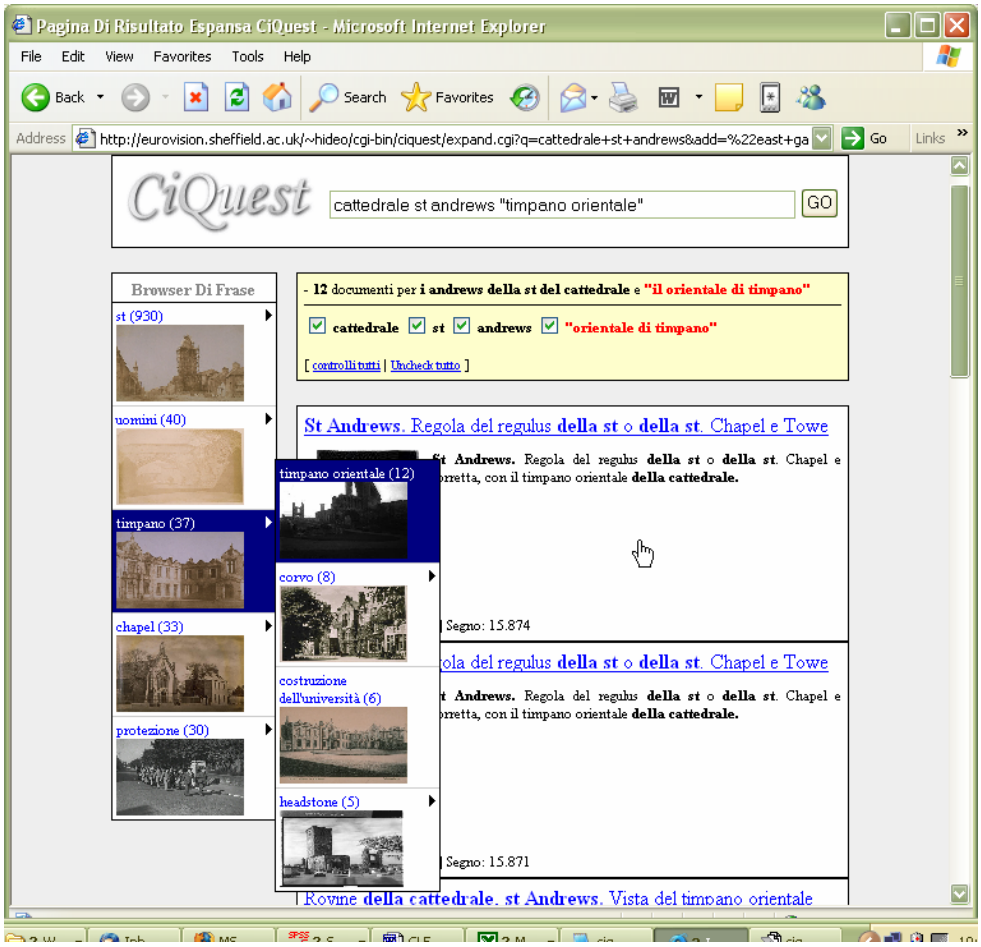

Fig. 2 The system with the cross-language concept hierarchy (menu interface).

3 The User Evaluation

Following Image CLEF directives, a within-subject experimental design was adopted, i.e. each participant tests both interfaces. A Latin-Square was used to fully counterbalance systems and tasks assuring data collected was unbiased. A precision task was set for Interactive Image CLEF 2005: users were required to retrieve a given image. A total of 16 images were used in the experiment, 8 with each interface; 2 more were used in training sections, 1 with each interface.

An initial briefing explained the experiment to participants and the basic mechanisms of CLIR. An online questionnaire to collect user profiles and searching attitudes was filled in. A training session with each system

followed prior to starting the actual tasks. Participants were presented with the image to search and were required to state their familiarity with it[1]. They were also required to type 3 queries they could use for searching the image: the idea was to compare those hypothetical queries against those actually used during the following task in order to see if browsing the hierarchy had an impact on query reformulation.

Participants performed the tasks individually and were observed by an experimenter. They were asked to type queries in Italian to retrieve images with English captions; results were back-translated into Italian before being displayed. Participants' activities were recorded by logging system events and videoing user's actions for further behavioural analysis. Users were given 5 minutes to find the given image (although the system did not interrupt searching after that time and more time was granted if participants wished). This fact was taken into consideration when data was analyzed.

Questionnaires to collect participant's opinions were filled in after each session (questionnaires were a variation of Chin et al.'s (1998) proposal for system usability testing). They encompassed questions on interface layout and cross-language functionality; further space was left for personal comments. An additional comparison questionnaire was completed at the end: participants were required so state which system they preferred and how different they felt the two systems to be. They were also asked to state what they liked and disliked about each system. Before leaving participants were invited to express any other opinion or comment on their search experience. The whole evaluation lasted 3 hours at most.

4 Data Analysis

Participants were 8 Italian native speakers, 5 male and 3 female, recruited through a Sheffield University mailing list for volunteers. Participants were all students or researchers at the University of Sheffield and bilingual (although their level of English language knowledge and UK culture awareness varied[2]). The profile questionnaires showed only 1 participant was less than 25 years old and had a BSc (studying for an MSc); the others were between 26 and 44 and had an MSc (working on their PhD) or a PhD (working as research associates).

All were computer literate and searched the Web daily; search through the library or commercial software was less popular with 50% using both rarely and 17% never using commercial search engines. Only 33% had received searching formal training (as part of university courses) but all felt confident in retrieving the information they needed.

All participants stated they were aware of what machine translation is, though no test was done to check their real understanding. Also, all stated they had previously used image search on the Web.

4.1 Quantitative Analysis

As described in section 3, participants were allowed more time to search than the 5 minutes required by Interactive Image CLEF; performance data was therefore analysed at 5 minutes as well as at final time.

All the 3 usability measures: effectiveness, efficiency and user satisfaction (van Welie et al. 1999, Frokjaer et al. 2000), were used to analyse the results. *Effectiveness* is set to the number of images retrieved; *efficiency* is then the average time required to find them; and *user satisfaction* is the participants' opinion as from the comparison questionnaire.

The global effectiveness was surprisingly identical with 64% of images found with both interfaces. If only those images found in 5 minutes are considered (CLEF condition), then the list is more effective with 53% of images found against 47% of the menu. This difference shows that the menu needs more time to be effectively used. The menu success rate includes also cases when the image was found because it was at the top of the list and no interaction with the menu was required to find it: in 53% of the cases the image was found in the results list, in 31% of the cases the image was displayed in the menu and found whilst browsing it, and in 16% of the cases it was found via selection (i.e. the participant has clicked on a sub-menu and the image was found there). By observing interaction it was found that participants were particularly pleased when the image was displayed in the menu indicating that the menu has a value as visual summary as well as content summary.

The list interface proved to be the most efficient in both cases. When measured at 5 minutes the list had a performance of 113 sec. on average to find the relevant image, whilst the menu scored 139 sec. Final measures show 170 sec. (min 10 sec.; max 643 sec.; median 123 sec.) for the list and 221 sec. (min 22 sec.; max 617 sec.;

[1] The goal was to record how confident users felt in retrieving the given image, though it was discovered during the evaluation that the intent was not clear and participants had interpreted the question as if they had previously seen the image.

[2] A wider variation was registered with respect to culture awareness depending on the time spent in the UK.

median 188) for the menu interface. A Mann-Whitney U Test[3] was conducted to compare the performance time for list and menu; it showed there is no statistically significant difference in both conditions (Z=-1.47, p=0.14 at 5 minutes; Z=-1.75, p=0.08 at final time). In the menu condition, time seems spent exploring the cluster as the number of queries issued is lower (256 queries) than within the list condition (282 queries). However the difference is not statistically significant (Mann-Whitney U Test Z=-0.472, p=0.64). How much the user browses the results impacts on the total interaction time and can be measured only by inspecting the recorded video as a comprehensive measure must include both clicking-on and browsing-through; this is set for future analysis.

Although effectiveness and efficiency of the two conditions were similar, user satisfaction was clearly in favour of the menu: 75% of participants favouring it compared to the list. The menu was also considered easier to use (75% vs. 25%), while the list was stated as easier to learn (75% vs. 25%). The two systems were seen as just slightly different by 87% of participants (13% completely different) but no one rated them as equal showing that the menu is perceived as an important feature. The two participants who favoured the list said: "I found the labels of the images confusing, to the point that I would not know which one to follow" and "sometimes the menu drove me on the wrong path and sidetracked my thought". Among those favouring the menu, the compact format and the (perceived) fester interaction were commented on as important. However, a few participants complained about unclear labels and unrepresentative images for the clusters.

Two questionnaires collected opinions on specific features of each interface. Images were interesting for 86% of participants (14% neutral) while opinions differ with respect to the captions: captions were considered useful by 43% but not useful by 43% (14% neutral); same numbers for the quality of the translation considered good enough by 43% but not good by 43% (14% neutral). However 72% agreed that captions were useful to see details in the images (14% disagree, 14% neutral) then partially contradicting the previous numbers.

The menu was considered easy to get to grips with (87% agree, 13% neutral) and navigate (72% agree, 38% disagree). The majority (62%) considered useful both text and images, while 25% favoured images and 13% text. Images were considered appropriate (87% agree, 13% neutral); labels were useful to explore the result (75% agree, 25% disagree) and suggested new search terms (63% agree; 12% neutral; 25% disagree). The opinions on the organization of labels (i.e. the concept hierarchy) is less positive with only 37% thinking they are organized in an intuitive progression (38% neutral, 25% disagree) and 37% considering them to be in a manageable number (37% neutral, 26% too many). This could be due to the hierarchy construction method, but also to the poor translation of the terms in the menu, a fact we discovered when inspecting the system behaviour, as discussed below.

4.2 Qualitative

The interactions recorded in the logs were analysed looking for interesting phenomena in user input, in query and menu translation.

Confirming previous studies (Petrelli et al. 2005), in 11 queries (2% of the total) participants inputted English words in the attempt to overcome (real or perceived) system limitations. Examples include "bagpipes" and "lighthouse" entered after the system failed in translating the Italian words, but also "cottage" or "clubhouse" entered straightforward by participants. In another 3% (15 cases) queries contained proper names (e.g. Plymouth, Robert Burns, Wallace) or nouns (i.e. "ballgown", "temple", "golfers") picked up from the displayed results. In a further 1% of queries the picked up terms were ill translations, for example "randello" shown as translation of golf club (the correct Italian term is "mazza"). Very often the terms picked up by participants in the result display and used in follow-up queries failed in being correctly translated; this was always the case for ill translations, but it happened also when the English-Italian translation was not too bad, for example the English "bridge" was translated as "ponticello" –little bridge- that failed when back-translated into English in a query. This is just one example of dictionary asymmetry that negatively affected the search.

All of the above behaviours show how users are flexible and able to adapt their interaction to the system capabilities in order (sometimes just in the hope) to improve its performance.

The translated queries were compared with the original ones to see how effective the machine translation step had been. Of the total 892 valid different keyterms[4], 84% were correctly translated, but for 100 terms (11%) translations failed, a further 46 terms (5%) were ill translations. Ill translations were due to selecting the wrong

[3] This test was preferred to the more commonly used t-test because the time distribution was not normal.

[4] This number is the sum of all the valid unique terms used by each participant in each task; stop words and repetitions of the same term by the same user in a task have not been counted.

sense for multiple senses words or preferring verbs over nouns for the same spelling, but some ill translations were quite inexplicable and bizarre; a summary is given in Table 1. The result was at times unexpected and confusing as, for example, “signora vestito bianco” (lady white dress) retrieved very many portraits of man as the translated query was “mrs. dressed white man”. This behaviour puzzled the participants who could not understand why portraits of men were retrieved as the query translation was not displayed.

Query	Meaning(s)	MT translation	Reason (supposed)
bianco	white [adjective]	white man	very colloquial sense
signora	madam, lady, ms., mrs., woman	mrs.	multiple senses, mrs. is used generally in formal written communications
vestito	dress, dressed	dressed	multiple senses, noun an verb (past tens)
reale [famiglia reale]	real, royal [royal family]	real family	multiple senses
lanterna	lantern	spider	
prato	lawn	Prato	Prato is an Italian city
riva	seaside, (river) bank	river	
sala	hall, sitting room	it knows it	inexplicable
cappelli	hats	nails head	inexplicable
coppia	couple	brace	
primo piano	foreground	Association of Bologna	inexplicable (Bologna is an Italian city)
bianco e nero	white and black	R-bianco.e.nero	inexplicable
ingresso	entry, entrance	income	
macchine	machines, cars	it blots some	inexplicable

Table 1. Summary of ill translated terms; all were check in Altavista while the correct translation is from an online Italian dictionary (Garzanti Linguistica www.garzantilinguistica.it).

Some ill translated or not translated terms were quite frequent in the corpus. Tables 2 and 3 summarise those used by more than one participant, in brackets the topic for which they were issued. A few terms were used by all or almost all users and correspond to important image features; failing to translate those properly surely impacted on the retrieval performance and forced the users generating new terms. Future analysis will compare issued queries (and non/ill translated terms in particular) with the captions to measure the impact of this misbehaviour.

Term	Ill translation	Correct translation	Users who used it (number of tasks)
macchine	it bolts some	cars	1 (2), 3 (2)
faro	beacon	lighthouse	1 (6), 2 (6), 3 (6), 4 (6), 5 (6), 6 (6), 7 (6), 8 (6)
bianco	white man	white	1 (6, 12), 2 (12), 3 (12, 14), 4 (12), 5 (12), 6 (12), 7 (12), 8 (12)
riva	river	shore/bank	2 (7), 3 (1)
letti	read	beds	3 (9), 4 (9), 8 (9)
reale/reali	real	royal	1 (10), 2 (10), 5 (10), 6 (10), 7 (10), 8 (10)
ingresso	income	entrance (hall)	3 (11), 5 (3), 8 (3, 11)
signora	Mrs.	madam, lady, woman	2 (12), 3 (3)
bianca/bianche	white woman	white	1 (12), 2 (6), 3 (6)
vestito	dressed	dress	5 (12), 6 (12)
coppia	brace	couple	1 (14), 3 (10)

Table 2. Summary of the ill translations by users and tasks; only those terms used by more than one person have been listed.

Term	Translation	Users who used it (tasks list)
celtica	Celtic	1 (1), 2 (1), 4 (1), 5(1), 7 (1)
citta'	city, town	1 (2, 15), 3 (2)
tempio	temple	1 (5), 2 (5), 3 (5), 4 (5), 5 (5), 6 (5), 7 (5), 8 (5)
vagone/vagoni		1 (7), 2 (7), 3 (7), 4 (7), 5 (7), 6 (7)
carrozza	coach	1 (10), 2 (10), 3 (10), 4 (10), 5 (7, 10), 6 (10), 7 (7, 10), 8 (10)
vetrata/vetrate		1 (11), 2 (12), 6 (11)
lampadario		1 (11), 3 (11), 6 (11), 7 (11), 8 (11)
candelabro		2 (11), 6 (11)
gotico/gotica	Gothic	2 (11), 3 (11), 7 (11)
ritratto	portrait	1 (12, 14, 16), 3 (16), 4 (12, 16), 5 (14), 7 (14), 8 (12, 14)
cornamuse	bagpipes	1 (13), 2 (13), 3 (13), 5 (13), 6(13), 7 (13)
tamburi	drums	1 (13), 2 (13), 3 (13), 4 (13), 6 (13)
lungomare	seashore	4 (15), 5 (15)

Table 3. Summary of the missed translations by users and tasks; only those terms used by more than one person have been listed.

Ill translations negatively affected the menu generation as well as sometimes labels did not make sense and were discarded by participants even though the wanted image was in there. For example, images of children walking on the seashore were grouped in a set labelled as "remare di bambini" literally "rowing of children" while the original text was "children paddling". In this case there was also a conflict between the label and the image as no boat could be seen in the image that justifies the rowing. Further research should be done to determine if and how much the use of a dictionary would improve the translation for searching as well as the hierarchy generation.

Observation of interactions showed the effect of translation asymmetry. Some terms were correctly translated from English to Italian, e.g. "portrait" into "ritratto", but the translation failed when the Italian term was used in the query as it was not in the BabelFish dictionary. This negatively affected the interaction as often specific terms (e.g. "croce Celtica" for "Celtic cross") were picked up by participants and used in follow up queries to focus the search but failed in improving the result as the term was not translated.

Useful comments were collected outside the formal questionnaires. The need to better control the search mechanism by forcing the use of all the terms simultaneously was a shared need. A few participants commented that the menu did not reflect their query and the relation was not straightforward. The two comments must be considered as a pair: forcing an AND retrieval is likely to impact on generation of the hierarchy and consequently on the displayed menu.

Comments on the images collected in a set were interesting: participants expected to see similar images but because the retrieval was text based this was not the case. A further step of visual content clustering would likely satisfy this need, but different interface design could be explored. Indeed a preliminary analysis of interaction behaviours shows two different attitudes in browsing through the menu: some participants used a horizontal approach and looked at all the children before moving to the next one, other proceeded vertically comparing siblings terms and selecting the one to explore next. Both behaviours may result in ignoring part of the menu that might include the wanted image. More effective alternative layouts to represent the result summary many be explored in future research.

5 Discussion

The menu feature did not prove to be more effective or efficient than a simple list display; however it was central in user satisfaction. It seems to engage users more than the list but requires more interaction time. However this additional time is not perceived by users as a waste of time; tediously scrolling the list is.

Some work is needed to make the menu more robust and effective. For example multiple senses need to be displayed or at least notified to avoid the discard of good sets because of ill translations. Users also expect to see their query reflected in the menu: some form of query-biased hierarchy should be explored. Moreover users

positively commented on the menu as a summary and this feature should be exploited. An improvement would be to dismiss the random selection of the image representing a sub-set in favour of a visual-content analysis to cluster images inside each group; this could generate a range of prototypical images that better represent the content. The menu would then become a summary of textual and visual content. In addition different layout of the summary could be explored, for example as a table instead of a menu.

As a more general result, the evaluation showed once more that good and consistent (bi-directional) translations are fundamental for CLIR. Furthermore allowing the user to check (and change) the translation is mandatory as users are flexible and able to adapt their interaction to get around system limitations.

It also became apparent during the experiment that the hierarchical summary of results is perhaps better suited to exploring a document collection or set of results rather than a high precision task as prescribed by Image CLEF. Further study is planned to determine the effectiveness of image organisation using concept hierarchies for other types of search task.

6 Conclusions and Future Work

The evaluation presented is the first step in an investigation of the use of concept hierarchies to cluster results in cross-language image retrieval. Results were encouraging, particularly as the users were very positive in their comments about the hierarchical menu.

Further analysis on the collected data is planned. This includes comparing queries across participants for the same query to see how similar/different terms are used, and which was most effective with respect to the image captions. Queries will be also compared to see if list of key terms were used more than fully structured phrases. This is expected to help in identifying which tools could better support searching in the context of images where the text is short and lacks redundancy.

Further work must be carried out to determine the impact of mis-translation. The collected corpus of queries will be used in investigating different translation mechanisms (e.g. dictionary) and search algorithm (e.g. AND vs. BM25). The length and complexity of queries issued by the user will be part of the investigation.

A new interface will be designed to give the user more control over the query translation and images clustered by visual content as well as by text will be tested.

Acknowledgements

The authors would like to thank Hideo Joho and Mark Sanderson for their work on concept hierarchies that allowed us to expand across languages. We are grateful to the 8 participants for spending their time with us and patiently dealing with bizarre translations and puzzling system's behaviours.

References

1. Chin, J. P., Diehl, V. A., Norman, K. L. (1998) Development of an instrument measuring user satisfaction of the human-computer interface. CHI '98. ACM Press, 213-218.
2. Clough, P., Joho, H., Sanderson, M. (2005) Automatically Organising Images using Concept Hierarchies. Workshop on Multimedia Information Retrieval, held in conjunction with 28th annual ACM SIGIR conference.
3. Frokjaer, E., Hertzum, M., Hornbaek, K., Measuring Usability: Are Effectiveness, Efficiency, and User Satisfaction Really Correlated? CHI 2000, 345-352
4. Joho, H., Sanderson, M., and Beaulieu, M. (2004) "A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool". In: McDonald, S. & Tait, J. (eds), *Advances in Information Retrieval, 26th European Conference on Information Retrieval*, 42-56.
5. Petrelli, D., Beaulieu, M., Sanderson, M, Demetriou, G., Herring, P. (2004) Observing Users Designing Clarity: A Case Study on the User-Centered Design of a Cross-Language Information Retrieval System. JASIST Journal of the American Society for Information Science and Technology, 55 (10), 923-934.
6. Petrelli, D., Levin, S., Beaulieu, M., Sanderson, M. (2005) Which User Interaction for Cross-Language IR? Design Issues and Reflections. JASIST special issue on Multilingual Information Access, in press.
7. Robertson, S.E., Walker, S., Beaulieu, M.M., Gatford, M. & Payne, A. (1995). "Okapi at TREC-4". In: Harman, D.K. (ed.), *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, Gaithersburg, MD. pp. 73-97
8. Sanderson, M. and Croft, B. (1999) "Deriving concept hierarchies from text" In: *Proceedings of the 22nd ACM Conference of the Special Interest Group in Information Retrieval*, pp. 206-213

9. Van Welie, M., van der Veer, G. C., Eliens, A. (1999) Breaking down Usability. Proc. INTERACT99, 613-620.