International Journal of Advanced Computer Science, Vol. , No. , Pp.

Building a Multilingual Research Platform through Usability Testing

Jiangping Chen & Min Namgoong

Manuscript
Received:

Revised:

Accepted:

Published:

Keywords
Multilingual Web System, Usability testing, Multilingual users, User behavior

***Abstract*— This paper describes usability testing performed on a Web-based multilingual platform for human evaluation of machine translation (HeMT). We developed a workflow and a usability testing model in order to evaluate HeMT's functionalities, especially its interaction with users. Usability measures including easy to learn, effective to use, efficient to use, and subjectively pleasing were applied to understand the user-friendliness of HeMT. Usability tests were conducted in laboratory settings with international students. We demonstrate that multiple-session usability tests are necessary and effective for successful implementation of a research platform like HeMT. Our experience suggests that usability testing should be well planned and flexibly conducted to achieve its purposes.**

1. Introduction

Usability testing has become an indispensable process of building a successful information system, whether it is a sophisticated organizational Website or a small research prototype system. Developers depend on usability testing to ensure a system can be used easily.

We conducted a two-year international collaborative project, the Metadata Records Translation (MRT) Project (Http://txcdk-IIA.unt.edu/MRT/), to evaluate the extent to which current machine translation (MT) technologies generate adequate translations for metadata records and to identify the most effective metadata records translation strategies for digital collections [1]. The MRT Project extracted 2,010 English metadata records from two digital collections: the UNT Catalog (http://iii.library.unt.edu/), and the Portal to Texas History (http://texashistory.unt.edu/). Each metadata record included up to six elements: title, creator, abstract, subject, publisher, and coverage. These extracted records were then translated by 3 Internet MT systems into Chinese and Spanish. To evaluate the quality of these translations, we built a multilingual research platform HeMT (Human evaluation of Machine Translation) so that users who understand Chinese and Spanish could participate in activities such as generating reference translations for English metadata records and evaluating MT results of these records. HeMT (http://txcdk-IIA.unt.edu/HeMT/) is a database-driven Web system that contains six major functions in three languages. Its major users are evaluators who are native speakers of Chinese or Spanish.

This paper describes the usability testing we conducted in order to make HeMT more user-friendly, especially for the evaluators. In the following sections we discuss related literature that guided our tests, the functionalities of HeMT, the workflow and evaluation model for the usability testing, the actual testing sessions and observations, actions taken after each session, and the analysis of the post-test survey. Finally, we review the whole experience and summarize our understanding of usability testing on multilingual systems.

2. Related Literature

A user interface refers to the part of a computer system that allows human users to interact with the computer, or an information channel that conveys information between users and computers [2]. User interfaces are important because "To most users, the interface is the system" ([3], p. 104). Five criteria or dimensions, including easy to learn, efficient to use, easy to remember, low error rate, and user satisfaction have been generally accepted to assess usability of the user interface of a computer system [4][5]. These criteria have been considered the foundation for Web usability, which can be simply defined as the degree to which users can use the functionality and the content of a Website to accomplish their tasks.

Zhang and Chen [6] suggested that Web usability should consider three interrelated components of a Website: Website utility usability (how well users can use the functionality the Website provides), website content usability (how well users can use the information provided by the website to accomplish a task), and browser's utility usability (how well users can use the functionality the browser provides). The usability of website content and browser's utility were emphasized in Web environments.

HeMT contains more than 20 pages that interact with its users. Some of the pages are informational while others are task-based. Zhang and Chen's Web usability model fits well for a system like HeMT.

Jeng [7] proposed an evaluation model for assessing the usability of academic digital libraries. The model contains four dimensions: Effectiveness, efficiency, satisfaction, and learnability. She developed measures and

Jiangping Chen, University of North Texas, Jiangping.Chen@unt.edu, Min Namgoong, University of North Texas, MinNamgoong@my.unt.edu.

instruments and tested the model through evaluating two Websites. The correlations among the four dimensions were also discussed. We adapted Jeng's four dimensions when developing HeMT's evaluation model.

Koohang [8] evaluated users' current views about applied e-learning usability and users' perceived importance of e-learning usability design features with consideration to the variable of experience – namely, users' prior experience with the Internet and amount of time users spent weekly on the e-learning courseware. He discussed the importance of experience in the e-learning instructional design process regarding usability. Some of Koohang's user experience variables were applied to HeMT's pre-test questionnaire.

Bilal and Bachir [9] examined Arabic-speaking children's interaction with the International Children's Digital Library (ICDL) and provided recommendations for assessing the cross-cultural usability of the ICDL and suggestions for system design improvements. Assessment of the ICDL to Arabic-speaking children as a culturally diverse group was grounded in "representation" and "meaning" rather than in internationalization and localization. The utility of the ICDL navigation controls was judged based on the extent to which it supported children's navigation. Most of the ICDL representations and their meanings were found to be highly appropriate for older children but inappropriate for younger ones. Sindhuja and Dastidar [10] investigated the satisfaction level of users from an interface usability perspective. Researchers found that information content, format, consistency, and ease of navigation were significant factors in the satisfaction level of the users. As a multilingual system, HeMT should be tested by native-speaking Chinese and Spanish participants. The Chinese and Spanish information content should be tested so that cross-cultural usability is possible.

Becker and Yannotta [11] presented a model for creating a new academic library Website by pairing usability testing and the design process. They conducted four rounds of usability testing using talk-aloud methods. They found that testing throughout the design process is an effective way to build a Website that reflects user's needs and accommodates ever-changing digital environments. Similarly, four rounds of testing throughout the design process were conducted for HeMT. Using the above literature, we developed a testing workflow, an evaluation model, and pre-test and post-test questionnaires for HeMT usability testing.

3. Human Evaluation of Machine Translation (HeMT)

HeMT is a multilingual Web system developed to carry out human evaluation of machine translation. It was one of the deliverables of the MRT project. HeMT was developed with open-source tools including PHP and MySQL. HeMT is expected to be released and used by future projects. Open-source programming languages are preferred for the convenience of libraries, museums, and researchers interested in using HeMT for their own purposes.

A. HeMT System Functionality

Our goal for HeMT was to design a system so that different types of research participants could seamlessly work together to perform high quality manual translation and evaluation of MT. Fig. 1 presents HeMT's functions, connections among different functions, input/output, and its users.

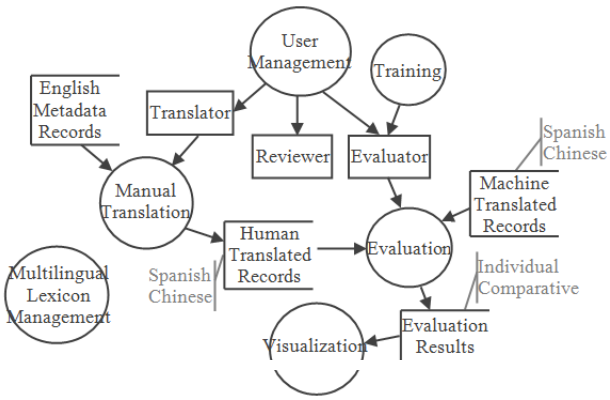


Fig. 1 HeMT functionality

HeMT consists of six functional modules, shown as circles in Fig. 1. Each module interacts with the users in multiple languages. The six modules include: (1) User Management—this module allows users to register, login, and revise their profiles. It also requests the reviewers to approve or deny new users; (2) Multilingual Lexicon Management—this module allows reviewers to generate multilingual terms that are used by HeMT webpages; (3) Manual Translation—this module generates reference translations for evaluation. It also allows reviewers to approve, edit, or deny a translation; (4) User Training—this module explains the background of the project, registration, evaluation types, measures, and methods of dealing with specific situations; (5) Evaluation—this module allows evaluators to perform two types of tasks: Individual evaluation based on traditional measures such as accuracy and fluency, and comparative evaluation of MT results generated by 3 MT systems; and (6) Result Visualization—this module presents evaluation results visually, which allows the research team to monitor the evaluation process as well as present the project to the public after the evaluation is completed.

In summary, HeMT accepts English metadata records and presents them to translators for translating them manually into Simplified Chinese and Spanish. It also accepts the translation results of these records from three MT systems. The manual translations and MT results are then presented to evaluators for assessing MT quality.

B. HeMT Users

HeMT was used by three major types of users: (1) Translators—translators conduct manual translation of selected metadata records. Each record needed to be translated into Simplified Chinese and Spanish. The manual translation served as reference translation for evaluating MT

results; (2) Evaluators—evaluators were recruited from China, Mexico, and United States with the assistance of the partners of the MRT project. Evaluators were mainly college students and librarians. They were not required to understand English, but were to be native speakers of Chinese or Spanish; (3) Reviewers—reviewers were consultants or members of the research team who monitored the manual translation process and reviewed/edited manual translation results.

C. HeMT Website

The HeMT Website provides the following interactions to its users: (1) allow users to register and login; (2) present metadata records for human translation; (3) train evaluators with examples; (4) present machine translation and corresponding manual translations to evaluators; and (5) display statistics of evaluations. Among them, (1), (2), (3) and (4) should be presented to the users in one of the three languages: English, Spanish, or Simplified Chinese. Fig. 2 is the site map of HeMT.

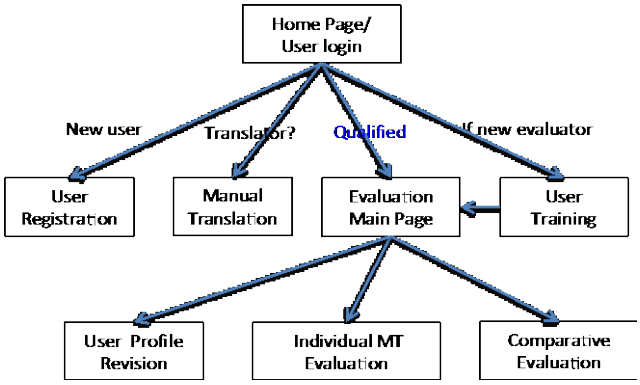


Fig. 2 Site map of HeMT

Fig. 3 is a screen shot of the English homepage of HeMT. It provides links to the Chinese and Spanish homepages.

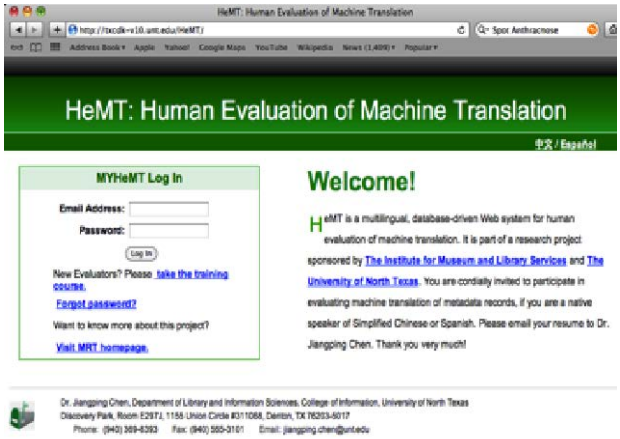


Fig. 3 HeMT English homepage

4. Usability Testing Methodology

The workflow we used for the tests is illustrated in Fig. 4. The usability testing involved seven steps, three of which were repeated with new or prior participants. These steps are explained in detail below.

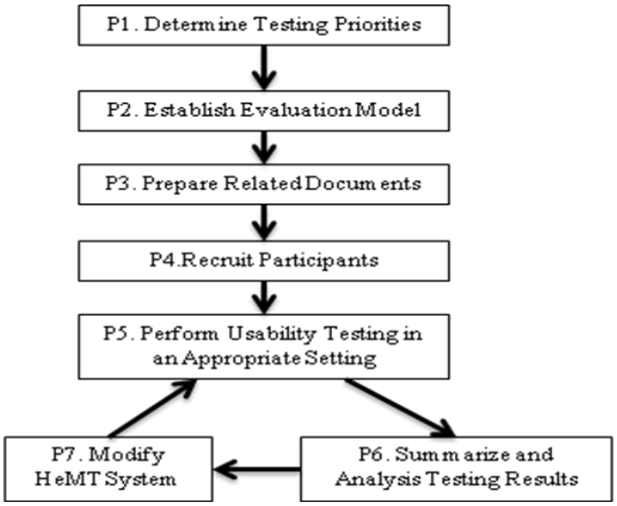


Fig. 4 HeMT usability testing workflow

A. Determining Testing Priorities

The process began with establishing a clear understanding of HeMT – the system to be tested, and the purposes of testing. As discussed in Section 3, HeMT was designed to interact with three types of international and multilingual users. The main purposes of the testing were to: (1) detect errors or problems in evaluation of modules such as User Management including Registration, User Training, and Evaluation; (2) understand the ease of use of the above related modules, including their content, function, and visual appearance. The focus of the testing would be on whether the participants could successfully and easily perform the two evaluation tasks.

B. HeMT Evaluation Model

Our study followed an evaluation model integrating Nielson [4], Zhang and Chen [6], and Jeng [7] by considering the unique characteristics and functions of HeMT and its users. Table 1 presents the model we used to evaluate HeMT as a research platform. Two check marks indicate a higher weight assigned to the variable.

TABLE 1
HEMT EVALUATION MODEL

	HeMT Utility	HeMT Content
Easy to Learn/Read	✓	✓
Effective to use	✓✓	✓✓
Efficient to use	✓✓	✓✓
Subjectively pleasing	✓	✓

HeMT utility usability evaluates how well users can use the functionalities that HeMT provides through its 6 modules. For the usability testing described in this paper,

we were particularly evaluating the following functionalities:

- Training lesson including the quiz after the lesson. As a research platform, HeMT would be used for evaluators to carry out evaluation of machine translation tasks. These tasks would be quite difficult in the beginning for the intended users. HeMT's training lesson has Chinese, English, and Spanish versions. It was intended to be used by international evaluators. Fig. 5 is the Chinese version of the training lesson homepage. The training module teaches evaluators the procedures and evaluation measures for the evaluation of MT.

Fig. 5 Chinese homepage of HeMT training lesson

- Registration function provided by the User Management module. Once the evaluators passed the training lesson –they took a quiz and had to score 70% or above to continue to the registration page (Http://txcdk-IIA.unt.edu/HeMT/UserManagement/Registration.php). The registration page also contains pre-test questions which collected participants' information for the MRT project as well as the usability testing.

- The evaluation module. Evaluation was the major function of HeMT. It is hence the focus of this usability testing. This module enabled evaluators to perform two sub tasks: Individual evaluation and comparative evaluation. Individual evaluation compared MT results of each metadata record with two reference translations using two measures: Adequacy and Fluency; while comparative evaluation compared the performance of three MT systems (Systems 1, 2 and 3) to determine which systems provided the best and worst translations. Both evaluation tasks allowed evaluators to provide comments on the evaluation process and translation quality.

HeMT content usability concerns how well users can use the information provided at HeMT Website to carry out their tasks.

HeMT utility usability and content usability can be measured in four dimensions;

- Easy to Learn/Read—how fast the new users can learn the functionalities or how easily the users can understand the information presented by HeMT;
- Effective to use—can the users perform tasks, such as registration or individual evaluation, with minimal incorrect actions;
- Efficient to use—how quickly the users perform the tasks or understand the information content; and
- Subjectively pleasing—whether the task is interesting or the site is visually pleasant to the users.

The above four dimensions were weighted differently in our model. In Table 1, "Effective to use" and "Efficient to use" are marked with "✓✓", which indicates these two criteria carried more weight than the other two.

C. *Data Collection and Analysis Methodology*

We decided to apply an iterative, formal-setting strategy for the usability testing. As Krug [12] suggested, usability testing should be performed iteratively (p. 135) so that detected problems can be fixed before the next test. Specifically, we performed 4 rounds of usability testing. In the test session, participants were taken to a computer. They were then asked to upload the HeMT homepage to a computer browser, follow links on the page to take the training lesson, register in HeMT, and perform evaluation on the designated computer. The whole process was observed and recorded. Additionally, participants took a post-test survey to report their experience. In the first three sessions, the authors briefly interviewed the participants to understand their experience. The four dimensions in the HeMT evaluation model were measured in terms of the duration of time the participants spent on each task, the questions they asked during the session, the mistakes they made, and their responses to the post-test questions.

After each test session, the research team and the observers met together to discuss and analyze the information collected during the sessions. Observation notes and videos recorded through the sessions were reviewed and discussed. The meetings also determined problems that could be fixed immediately and problems that needed further testing.

We then revised HeMT, fixing the problems that could be fixed immediately, and prepared for the next round of testing until we were satisfied with the usability of HeMT.

D. *Pre-test and Post-test Instruments*

We developed pre-test and post-test instruments based on our testing purposes, evaluation model, and usability testing literature. The pre-test questionnaire contained 21 questions asking participants' demographic information, their language skills, and computer literacy skills. Appendix A lists these questions. They constitute the main questions at the HeMT registration page, which all HeMT users answered.

The post-test questionnaire contains 18 multiple-choice questions and 5 open-ended questions. It was developed specifically for the usability testing based on the evaluation

criteria. Appendix B presents the questions and their options where applicable. We expected the participants to complete the post-test questionnaire in less than 15 minutes.

The instruments were approved by the Institutional Review Board (IRB) of our university prior to the usability testing.

E. Recruiting Participants and Prepare for Testing

Announcements were sent to the mailing lists of the UNT Linguistics Department, the UNT Libraries, and the Department of Library and Information Sciences to recruit participants. The announcements were also emailed to UNT international students. We recruited twelve Chinese-speaking students and four Spanish-speaking students as participants. Four participants were male and twelve participants were female (see Table 2 for the profile of the participants).

TABLE 2
DEMOGRAPHIC DATA

	Number of Students
Native language	
Chinese	12
Spanish	4
Second Language: English	16
Gender	
Male	4
Female	12
Education Level	
Undergraduate student	5
Post-graduate student	9
Non-student	2
Total	**16**

Prior to the testing, the research team secured one computer classroom with 24 computers in addition to our research laboratory, which has 3 computers for use. All the computers in the two rooms were examined to make sure they were functional and could display and receive input in both Simplified Chinese and Spanish. Usability testing materials including: Usability testing procedures, list of items for testing, usability testing protocol, sign-up sheet, payment sheet, and usability testing observation notes were created to guide participants and researchers in the test procedures.

F. Usability Test Sessions

Four rounds of usability testing were conducted on November 4, 10, 17, and December 1, 2011. Table 3 lists the numbers of participants and observers present in each session. The first three sessions were comprised of one or two participants. These small sessions enabled the research team to gain experiences on usability tests and closely observe the participants. The fourth session had the largest number participants and was treated as the final exercise before the release of HeMT for formal evaluation.

TABLE 3
TESTING SESSIONS

Testing Session	**Date (in 2011)**	**Number of Participants**	**Number of Observers**
1	Nov. 4	2	3
2	Nov. 10	2	3
3 (2 sub-sessions)	Nov. 17	1 2	4 4
4	Dec. 1	9	4

During each testing session, we assigned observers to one or more participants in addition to using a video camera to record the session. The observers were members of the research team and HeMT developers. We added additional helpers to assist the participants in sessions 3 and 4. For example, the project's Spanish consultant attended session 3 and assisted participants in Spanish.

Each testing session took about 80 minutes. Researchers who also served as observers arrived at the test location 30 minutes earlier to prepare for the session. The researchers checked all the documents to be used, (such as the informed consent form, sign-up sheet, and payment sheet), and the materials for recording and observation (such as a video camera, a tripod, pens, notepads, papers, whiteboard markers, and highlighters). They also set up the computer(s) for the participants.

When a participant arrived, he or she was given a procedure sheet, which had been emailed to him or her two days before the testing. The procedure sheet is shown in Table 4.

TABLE 4
PROCEDURES FOR PARTICIPANTS IN A TEST SEESION

Step	**Tasks/Activities**
1	Sign in.
2	Read and sign the informed consent form.
3	Take the training lesson and the quiz which tests the participants' understanding of the training lesson. This step takes about 22 minutes on average.
4	Register in HeMT and take the pre-test.
5	Wait for approval of the registration. HeMT would send a message to its administrator who would then check the registration information and approve/disapprove the participant as a valid evaluator.
6	Login HeMT and start to evaluate machine translation of metadata records. The participants were given about 40 minutes to perform the evaluation.
7	Take the online post-test survey.
8	Receive and sign payment sheet. Each participant was awarded $20 for his or her time and effort.
9	Receive a business card for future contact.

The researcher explained the above procedure to the participants and instructed them at each step. Participants were encouraged to ask any questions during the session. After Step 2, an observer took notes regarding the duration the participant took for each step (steps 3, 5, 7, and 8), any

questions he or she asked, and other observations. At the end of the session, each participant was paid $20 in cash for his/her participation.

5. Results and Analysis

In this section, we report the four rounds of usability testing including a summary of the observations and actions taken after each round. Also, we present a participants profile and summarize the results of the post-test survey.

A. The Four Rounds of Usability Testing

The first test session was conducted on November 4, 2011 with two native-speaking Chinese, female participants. The participants completed the training lesson in two minutes and ten minutes respectively. But they came back to the lesson several times when taking the quiz. While performing Step 6 (as indicated in Table 3), both participants' HeMT sites crashed once. One participant evaluated 4 records in 46 minutes, and the other evaluated 9 records in 53 minutes. While there was no time limit for each step, we stopped the evaluation process (Step 6) so the participants could complete the post-task survey. There was an issue with the post-task survey's numbering. The test demonstrated that HeMT could function in general as a platform for MT evaluation in spite of a couple of issues. We then fixed the bugs in the system, changed the wording that caused confusion, and corrected the post-test survey's numbering problem.

The second usability test session was conducted on November 10, 2011, from 1:00 p.m. to 3:00 p.m. in our computer laboratory. One male and one female Chinese participant attended the session. Different from the first test, we reminded the participants to study the training lesson carefully, and we asked participants to evaluate records as soon as possible after becoming familiar with the evaluation process for the first record. We allowed the participants only 30 minutes for actual MT evaluation to see the number of records they could evaluate. During the session, one of the participants failed to register at first because he did not click the "Yes" button after reading the informed consent form. Both participants sometimes skipped to the next step before they had completed all the required items in the registration and the evaluation sub tasks. Based on our observations, brief conversations with the participants after the test, and the responses of the post-test survey, we believed the usability of HeMT had been improved since the first test. The corrections we made after the first session were effective. The Chinese version of HeMT was working quite effectively. For the overlooking of the "Yes" button, we later added a reminder on the registration page.

The third usability test focused on the Spanish version of HeMT. It was conducted with two sub sessions on November 10, 2011. Three Spanish participants were recruited for the test. The first sub session was held from 11:00 a.m. to 1:00 p.m. with one male student. He found several problems with the Spanish quiz questions. He tried to register without finishing the training lesson. His post-task survey results were not stored in the system (due to a bug in the survey program), so we printed a hard-copy of the survey for the participant to complete. The second sub session was held from1:00 p.m. to 3:00 p.m. with one male and one female participant. One participant made a similar mistake—tried to register without finishing the training lesson. Both participants asked questions when taking the quiz. The testing showed that the HeMT Spanish version had issues such as inconsistent use of words or phrases, unclear sentences, and culturally incorrect expressions. We then significantly revised the Spanish version.

From the three sessions, we observed that the individual evaluation page caused confusion from time to time. We had to explain the items on that page to the participants before they could correctly perform the evaluation, even though the training lesson had provided instructions. We decided to observe further before we made any changes to the evaluation pages.

The fourth usability test was conducted in a larger computer classroom. One Spanish student and eight Chinese students were recruited for the test. Four researchers observed the session. Two participants had difficulty with the login due to not realizing their usernames and passwords were case sensitive. The Spanish participant found an inconsistency in one quiz question. Several participants got confused after completing each task and asked what the next step was. Even with the various questions, the participants on average evaluated more records than those in the previous tests. The participants seemed less pressured than those at the previous session where each participant was observed by one or two observers. This might explain why they evaluated more metadata records on average in 30 minutes than the participants of the previous sessions.

B. The Profile of the Participants

The pre-test questions (See Appendix A) implemented at the registration page were related to the roles of the participants, their language skills, and their computer literacy competencies. The 16 participants are bilingual and English is their second language. They considered themselves as having excellent native language skills on listening, speaking, reading, and writing. Most of the Chinese-speaking participants were undergraduate students while all 4 Spanish-speaking participants were graduate students or post-graduates. All the participants were familiar with using a computer, the Internet, and the Web. These backgrounds match the evaluators we recruited later for the formal evaluation. Table 5 presents the average scores of the participants' language skills and computer literacy competencies.

TABLE 5
PARTICIPANTS' SELF-PERCEIVED LANGUAGE SKILLS AND COMPUTER LITERACY

Proficiency in Native Language		**Average Score**
Listening	(1, 2, 3, 4, 5)	5.00
Speaking	(1, 2, 3, 4, 5)	4.93
Reading	(1, 2, 3, 4, 5)	4.93
Writing	(1, 2, 3, 4, 5)	4.87

Proficiency in Second Language	
Listening (1, 2, 3, 4, 5)	3.87
Speaking (1, 2, 3, 4, 5)	3.53
Reading (1, 2, 3, 4, 5)	4.07
Writing (1, 2, 3, 4, 5)	3.67
Computer Literacy: I feel confident in…	
working on a personal computer (microcomputer) (1,2,3,4,5)	4.53
learning to use a variety of programs (software) (1, 2, 3, 4, 5)	4.20
browsing/surfing the World Wide Web (WWW) (1, 2, 3, 4, 5)	4.73
finding information on the World Wide Web (WWW) (1, 2, 3, 4, 5)	4.67
taking an online survey or fill an HTML form (1, 2, 3, 4, 5)	4.33
joining a social media such as Facebook or LinkedIn (1, 2, 3, 4, 5)	4.20

C. Post-testing Survey Results

Only fifteen participants completed the post-test survey. Table 6 summarizes the average scores of the answers to the first 16 questions.

TABLE 6
HEMT USABILITY: POST-TEST SURVEY RESULTS

Question ID	**Question**	**Average Score**
1	This system is simple to use, even when I was using it for the first time.	4.27
2	I was easy to learn how to use this system.	4.13
3	I felt comfortable using this system.	4.47
4	I was able to complete the evaluation tasks effectively using this system.	4.27
5	I was able to complete the evaluation tasks quickly using this system.	3.60
6	This system displayed error messages that include clear instruction on that to do next.	3.60
7	It was easy to recover from error.	3.73
8	Training lesson provided with this system was clearly understandable.	4.47
9	Training lesson was helpful in completing the evaluation tasks.	4.40
10	The organization of information on the system screens was clear.	4.40
11	It was easy to find the information I needed.	4.13
12	The design of interface is visually attractive.	3.53
13	The text on the system screen was displayed in a way that was easy to read.	4.33
14	This system has all the functions and capabilities I expected it to have.	4.00
15	I liked using the interface of this system.	4.07
16	Overall, I am satisfied with this system.	4.27

Table 6 shows that most questions received responses that had a mean score above 4, but mean scores for questions 5, 6, 7 and 12 were lower than 4. These lower scores indicated that more efforts were needed to reduce the difficulty of the evaluation task and to add more messages to help evaluators to handle mistakes.

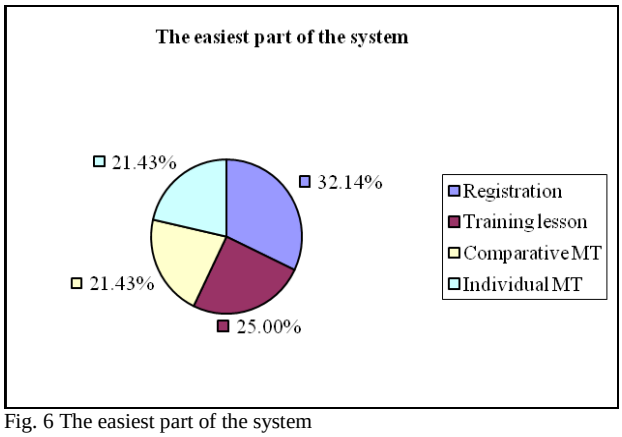


Fig. 6 The easiest part of the system

For questions 17 and 18, participants perceived that registration was the easiest part of HeMT (Fig. 6) and comparative MT evaluation was the most challenging part of HeMT (Fig. 7). These results explain the low mean score for Question 5 in Table 6, which suggested more effort was needed to reduce the difficulty of evaluation tasks. We therefore made significant changes to the individual and comparative evaluation pages.

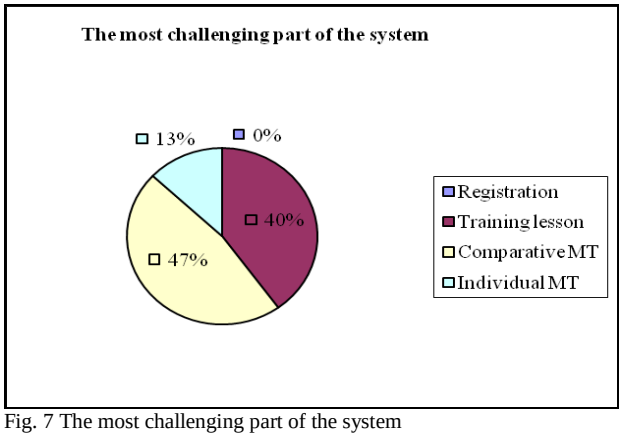


Fig. 7 The most challenging part of the system

In the responses to Questions 19 to 23, participants suggested that HeMT provide hints and tips during the

evaluation process to help them make judgments. Concerning the interface design, participants said that a different font and more colorful background would be more attractive. In addition, participants said they had to save the information inserted in the registration page after submission so they would not have to retype the information when they wanted to go back and edit the information.

6. Discussion

It was a challenging task to build HeMT as an integrated multilingual platform interacting with international users for the MRT Project in 8 months. The four rounds of lab-setting usability testing found various problems and issues. In each round new issues or problems popped up and were then managed. Table 7 summarizes the major problems/issues found during the testing and the respective revisions we made to HeMT. The iterative process as presented in Fig. 1 improved the usability of HeMT. Shortly after the usability testing, HeMT was first released for evaluating Chinese MT. Its Spanish version took much longer to be modified before it was released to Spanish evaluators.

TABLE 7
MAJOR PROBLEMS FOUND AT USABILITY AND ACTIONS TAKEN

Major Issues Found at Usability Testing	**Actions Taken**
Broken links	Reviewed all links and fixed broken ones
Fewer error handling messages	Added more messages to hand various situations, revised unclear error messages
Registration page: top information was ignored	Added a reminder message and increased font size
Registration page: left out questions	No actions taken – system limitation
Training Lesson: Inconsistent expressions in training lesson	Reviewed the whole lesson in three languages and fixed identified issues
Ambiguous quiz questions and options	Modified quiz questions and options
Missing connections between different functions	Added statements after each step to remind the evaluators of the next step
Individual MT evaluation: confusing presentation of information	Reviewed and revised information content on the pages
Comparative MT evaluation: Difficult evaluation cases	Added more evaluation tips in the training lesson
Need faster error handling or help	Revised HeMT to let the researcher know an evaluator's comments immediately

Given limited time and funding, our testing focused on improving HeMT's functionalities – providing correct information to its users, collecting users' interactions including evaluation scores accurately, and storing the information in the database. The evaluation model reflected the order of priorities for the testing. It kept us focused on the most important usability dimensions and measures.

For a research platform like HeMT, usability testing is an integral component of system design. We let HeMT developers observe and help participants during the testing sessions. We found this improved their understanding of HeMT from its users' perspectives. Usability testing for a system like HeMT does not require many participants. Using only a few participants is even better for observation and data analysis. The first three testing sessions had 1-3 participants with 2-4 observers. This situation might have put pressure on the participants, which was a disadvantage, but this enabled us to quickly understand usability problems and develop solutions. The recorded videos were only occasionally used in after-session discussions as we took notes on many details of the testing sessions.

Just like other Websites or information systems, usability testing should be a continuous process for a research platform as long as the system is active. After the first round of MT evaluation, we made further changes to the system based on the comments and feedback from the evaluators. For example, we added a new option for comparative evaluation to handle the situation where two MT systems provided identical translations. HeMT was under constant review and minor revision during the two rounds of MT evaluation.

This study has its limitations. We were unable to recruit more Spanish participants. The study indicated that the Spanish version of HeMT was not as user friendly as the Chinese version. We later hired one of the Spanish participants to help revise the Spanish version. Also, during the formal MT evaluation after the usability testing, the research team worked closely with Spanish evaluators to assist them in performing their tasks. As a result, two rounds of MT evaluation were completed successfully by December 2012.

The Internet browsers used in the usability testing were Internet Explorer (IE) version 8, IE version 9, or Mozilla Firefox 5. We did not test HeMT on other browsers such as Safari or Google Chrome during the testing. These browsers were only tested later by MRT team members. Luckily, we hadn't received complaints from evaluators regarding use of different browsers.

Nielsen [13] suggested usability tests for sites developed for international users should be conducted in the countries where users were expected to use the sites. We could not afford to carry out usability testing in China or Mexico even though HeMT is expected to be used by evaluators in these two countries. However, some of the participants were international students from the two countries.

7. Conclusion

Web usability testing has been considered very

important for developing effective Internet application systems. Procedures and models for conducting usability testing need to be designed based on targeted user groups by following general usability principles. To develop a user-friendly multilingual research platform, we established a workflow and a straightforward evaluation model to guide our usability testing based on characteristics of the system to be tested and related literature. Four rounds of usability testing were conducted and greatly helped in designing HeMT into a multilingual research platform for metadata records MT evaluation.

We agree with Becker and Yannotta [11] that usability testing throughout system design was an effective way to build a successful system. Our usability testing approach as described in this paper was appropriate and effective. It helped us to build HeMT into a successful platform within a limited time frame.

Acknowledgment

The constructing of HeMT including its usability testing was supported by U.S. Institute of Museum and Library Services (http://www.imls.gov/) under the National Leadership Grant LG-06-10-0162-10. The authors thank our other research team members Miriam Rodriguer, Olajumoke Azogu and Wenqian Zhao for their hard work on assisting the development of the pre-test and post-test questionnaires, serving as observer of some of the testing sessions, analyzing testing results, and revising HeMT. We thank Ryan Knudson for editing this manuscript.

References

[1] J. Chen, O. Azogu, & W. Zhao, "An integrated participatory platform for human evaluation of machine translation," (2012) *In JCDL'12 Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 421-422.

[2] Thimbleby, H., User Interface Design – ACM Press Frontier Series. New York: ACM Press, 1990.

[3] H. R. Hartson, "Human-computer interaction: interdisciplinary roots and trends," (1998) *The Journal of System and Software*, vol. 43, no. 2, pp.103-118.

[4] Nielsen, J., Usability Engineering, First Edition. Boston, MA: Academic Press, 1993.

[5] Shneiderman, B., Design the user interface: Strategies for Effective Human-Computer Interaction. Third Edition. Addison Wesley Inc., 1998.

[6] P. Zhang & J. Chen, "What is web usability anyway? A conceptual study on usability in the Web environment", (May, 1999) Presented at Mid-year conference of ASIS.

[7] J. Jeng, "Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability," (2005) *Libri: International Journal of Libraries and Information Services*, vol. 55, no. 2-3, pp. 96-121.

[8] A. Koohang, "Expanding the concept of usability," (2004) *Informing Science Journal*, vol. 7, pp. 129-141.

[9] D. Bilal & I. Bachir, "Children's interaction with international and multilingual digital libraries. I. Understanding interface system design representations," (2007) *Information Processing & Management*, vol.43, no. 1, pp.47-64.

[10] P P. N. Sindhuja & S. G. Dastidar, "Impact of the factors influencing website usability on user satisfaction," (2009) *The IUP Journal of Management Research*, vol.8, no. 12, pp. 54-66.

[11] D. A. Becker & L. Yannotta, "Modeling a Library Website Redesign Process: Developing a User-Centered Website Through Usability Testing," (2013) *Information Technology and Libraries*, vol. 32, no. 1, pp. 6 – 22.

[12] Krug, K., Don't Make Me Think! A Common Sense Approach to Web Usability, Second Edition. Berkeley, CA: New Riders, 2006.

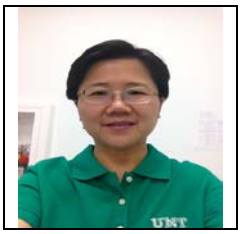
[13] J. Nielsen, "International web usability," (1996) (http://www.nngroup.com/articles/international-web-usability/)

APPENDIX A
PRE-TESTING QUESTIONNAIRE

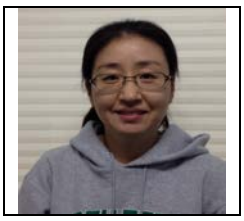
QID	Question
1	First Name:
2	Last Name:
3	Expected Role in this Project: Evaluator, Reviewer, or Translator
4	Education: (Highest degree obtained)
5	Native Language: Chinese, English, or Spanish
Proficiency in Native Language: (Choose one from the 5 scales: 1 Not at all – 5 Extremely Well)	
6	Listening (1, 2, 3, 4, 5)
7	Speaking (1, 2, 3, 4, 5)
8	Reading (1, 2, 3, 4, 5)
9	Writing (1, 2, 3, 4, 5)
10	Second Language: Chinese, English, or Spanish
Proficiency in Second Language: (Choose one from the 5 scales: 1 Not at all – 5 Extremely Well)	
11	Listening (1, 2, 3, 4, 5)
12	Speaking (1, 2, 3, 4, 5)
13	Reading (1, 2, 3, 4, 5)
14	Writing (1, 2, 3, 4, 5)
15	You are currently a (an):
Computer Literacy: I feel confident in (Choose one from 1 never – 5 always): (1, 2, 3, 4, 5)	
16	working on a personal computer (microcomputer) (1, 2, 3, 4, 5)
17	learning to use a variety of programs (software) (1, 2, 3, 4, 5)
18	browsing/surfing the World Wide Web (WWW) (1, 2, 3, 4, 5)
19	finding information on the World Wide Web (WWW) (1, 2, 3, 4, 5)
20	taking an online survey or fill an HTML form (1, 2, 3, 4, 5)
21	Joining a social media such as Facebook or LinkedIn (1, 2, 3, 4, 5)

APPENDIX B
POST-TESTING QUESTIONNAIRE

QID	Question
Choose from the 5 scales (1 strongly disagree to 5 strongly agree) for Questions 1 – 16	
1	This system is simple to use, even when I was using it for the first time.
2	It was easy to learn how to use this system.
3	I felt comfortable using this system.
4	I was able to complete the evaluation tasks effectively using this system.
5	I was able to complete the evaluation tasks quickly using this system.
6	This system displayed error messages that include clear instruction on what to do next.
7	It was easy to recover from error.
8	Training lesson provided with this system was clearly understandable.
9	Training lesson was helpful in completing the evaluation tasks.
10	The organization of information on the system screens was clear.
11	It was easy to find the information I needed.
12	The design of interface is visually attractive.
13	The text on the system screen was displayed in a way that was easy to read.
14	This system has all the functions and capabilities I expected it to have.
15	I liked using the interface of this system.
16	Overall, I am satisfied with this system.
17	What are the easiest part(s) of this system? A. Registration B. Training Lesson C. Individual MT System Evaluation D. Comparative MT System Evaluation E. Other, please specify
18	What are the most challenging part(s) of this system? A. Registration B. Training Lesson C. Individual MT System Evaluation D. Comparative MT System Evaluation E. Other, please specify
19	What are your suggestions on improving training lesson?
20	What are your suggestions on improving the evaluation process?
21	What are your suggestions on improving the interface?
22	What features that you would like to see on this system to complete evaluation effectively?
23	Do you have any other comments about the system?



Author I Jiangping Chen is an Associate Professor at the Department of Library and Information Sciences in the College of Information, University of North Texas. She conducts research in multilingual information access for digital libraries. She can be reached via email Jiangping.chen@unt.edu.



Author II Min Namgoong is a Ph.D. candidate in the interdisciplinary Ph.D. program of the Department of Library and Information Sciences in the College of Information, University of North Texas.