A Successful Field Test of a Mobile and Multilingual Information Service System COMPASS2008

Hans Uszkoreit[1], Feiyu Xu[1], Weiquan Liu[3], Jörg Steffen[1], Ilhan Aslan[2], Jin Liu[4], Christel Müller[4], Bernhard Holtkamp[5], Manfred Wojciechowski[5]

[1+2] DFKI GmbH, [1]LT Lab/[2]IUI Lab, Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
{ilhan.aslan, feiyu, steffen, uszkoreit}@dfki.de

[3] Capinfo Co. Ltd., Yu Min Dong Lu, Xi Cheng District
Beijing, 100029 P. R. China
liuweiquan@capinfo.com.cn

[4]T-Systems, Advanced Voice Solutions, Goslarer Ufer 35,
10589 Berlin, Germany
{Christel.Mueller, Jin.Liu}@t-systems.com

[5]Fraunhofer-Institut für Software- und Systemtechnik ISST
Emil-Figge-Straße 91, 44227 Dortmund, Germany
{holtkamp,wojcie}@do.isst.fhg.de

Abstract. We will describe a field test and its evaluation of a truly novel type of mobile computer programs that will assist foreign tourists in their communication with Chinese people. The software is an electronic phrase book and a translation aid but at the same time a powerful multilingual information system connected to numerous services via the Internet. It effectively helps visitors to navigate through the streets, temples and shopping centres of the Beijing megalopolis. It was developed in the German-Chinese project COMPASS 2008, a research action within the Digital Olympics framework. The subjects of the field test were fifteen tourists from seven countries. The test concentrated on usability and acceptance. The applied methodology adapts recognized standards and widely accepted best practice to the specific application type.

Keywords: field test, multilingual, multimodal, mobile applications, context-sensitive information service, usability and acceptability evaluation

1 Introduction

The main goals of COMPASS2008 are to help foreigners to overcome language barriers in Beijing during the Olympic Games event and to assist all visitors in finding information anywhere and anytime they need it. COMPASS2008 works on a large-scale and open-service platform supporting multimodal and multilingual information access, available both for mobile devices and PC. The supported languages of the prototype are Chinese, English and German. But the design of the system

accommodates the rapid addition of further languages. The mobile, multimodal and multilingual information service system framework worked out within COMPASS2008 includes an open service platform for user and service profile registration, translation services, multilingual service and information access, and context-sensitive service access.

In July 2006 the system was successfully tested by tourists from several countries in a field test in Beijing under realistic conditions. Equipped with the new software on a palmtop computer the test subjects faced typical challenges for a nonnative visitor. After the test they all praised the system. Without speaking a word of Chinese themselves, they could give directions to cab drivers, search for sights, order food and ask for directions. Since the system "knows" the location of the user through GPS satellite connection, it can adapt the advice to the place. Even in emergency situations, when the lack of language skills can really be critical, the electronic assistant could offer effective help. The field test consists of a list of tasks associated with the special contexts, e.g., restaurant and shopping. A systematic evaluation and assessment of the usability and acceptability of the system functionalities have been worked out taking the filled questionnaires after each task and the whole test as input material. The overall evaluation result is very satisfying and promising.

In section 2, we will give an overview of the field test organization. Section 3 describes the relevant technologies developed in the pilot system. Section 4 defines our evaluation approach and shows the evaluation results. In section 5 we close off with a short conclusion.

2 Overview of the Field Test Organization

During the field test, we aim to test the applicability of the following concepts and technologies developed within the COMPASS2008 project: multilingual information access, crosslingual communication help, multimodal interaction, personalization, semantic service description and organization, and location and situation awareness. We decided that the field test should take place in a natural environment, namely, a district in Beijing. We choose the area around the headquarters of our Chinese partner, since there we had access to the direct technical support and on the other hand examples for all task-relevant location types can be found in this area. Three days are planned for three user groups: one user group for one day. The user groups are native English speakers without Chinese knowledge, non-native English speakers without Chinese knowledge and Native Chinese users. They are mainly foreign students and young tourists, the most potential future user group of the mobile services. Each group has five members. Each test user is accompanied by a test instructor who ensures the proper test procedure. Each test run takes around three to four hours. The subjects have to fulfil a list of tasks that refer to functionalities of the COMPASS2008 services. Each task is associated with one location: a cafe, a shopping mall, a point of interest or a restaurant. The transfer from one location to another is via taxi, where the subjects have to use the taxi communication help. The tasks included but were not restricted to: planning daily activities, communicating with local people (e.g. taxi drivers, restaurant staff), finding location based information (e.g. restaurants in the area), getting information on points of interest,

bargaining for good price and searching information on sport sites, museums and exhibitions. An example of a task is shown in the following table:

Table 1. An Example of a Task Specification

Task	Subtasks	Services	Situtaion
Dining	a. order table and order dish	dining dialog smart dining	Eating
	b. express wishes and payment	dining dialog	Eating

Table 1 shows the final step of our field test: dinning in a restaurant. The subtasks are 1) ordering table and dish 2) expressing wishes and payment. The services needed for accomplishing the task are dining dialog and smart dining service. The dining dialog helps users to communicate with the waiters and the smarting dining service provides menu list and other additional information about a specific restaurant: opening time, map, contact information and ingredient information of a dish. The dinning dialog and smart dining information service are integrated with each other. Therefore, the subjects can test several services within one test session. The access channels for the mobile devices rely mainly on GPRS and WLAN, when available. A short tutorial is planned to introduce the subjects the device usage. Two different questionnaires are defined: post task and post test. After every task, subjects are asked to provide comments on possible problems or suggestions of system improvements. In addition the subjects have to fill out post task questionnaires and post test questionnaires. We also have questionnaires for local people with whom the subjects interacted to complete their tasks. The evaluation data in form of video material, paper based questionnaires and log files are collected.

3 Relevant Technologies

The scientific and technical innovation of the COMPASS2008 system is to combine translation technology, multilingual information access technology, multimodal technology, personalization, situation-aware services and an open semantic structure oriented architecture, so that users of COMPASS2008 are supplied anytime and anywhere with the best information in their own language. In the following, we will explain the relevant technologies applied in the pilot system:

3.1 Crosslingual Communication Help

The special contribution of COMPASS2008 for the multilingual information systems is an innovative combination of several methods ranging from scientifically less sophisticated but carefully designed techniques for phrase-book based translation to most sophisticated state-of-the-art methods for machine translation ([6], [7], [8]). A novel crosslingual dialog has been developed to allow controlled communication between two persons who do not share a common language. The three main approaches employed in COMPASS2008 for crossing the language boundaries are multilingual generation, phrasebook translation and free machine translation. For each

approach, we exploit existing resources whenever possible. Among such resources are monolingual and bilingual dictionaries, terminologies, tourist phrasebooks, and existing MT systems.

Multilingual Template-based Generation
In COMPASS2008, the input for the template-based generation [2] can be delivered by various components: location-based search based on GPS information, food search, restaurant search or user input. For example, a user query such as "Take me to a Chinese restaurant close to here" will trigger the positioning service and location search for restaurants within the smallest radius. Then a record will be filled with Chinese restaurant names and their addresses. The multilingual template generation will translate it into Chinese. The corresponding person, in this case, most probably, the taxi driver will obtain the request, which is spoken in Chinese by the automatic speech generation.

Phrase-book Translation
Whenever utterances in a relevant limited domain need to be translated for communication partners who do not master the source language, we try to use a phrasebook-like specialized translation scheme. In some cases, the user can select the source phrase through a menu, in other cases an input utterance is mapped into the semantically equivalent or closest source phrase. The phrase-book approach has the advantage that the translations are correct. It works because in most situations the choice of possible input utterances is limited by the context. In order to exploit this dependence, we have developed dialogue models for the relevant situation types.

Free Text Translation
Free translation by a rule-based state-of-the-art machine translation system is applied when users want to translate texts at their own risk using the so-called *transearly* service. We have assessed a few translation systems based on availability and on a simple evaluation procedure. Since there do not exist many systems with Chinese as a source or target language, the number of considered systems was small for translations into or out of Chinese. The best system for translations involving Chinese[1] turned out to be a commercial system from Beijing, provided by our Chinese project partners.

COMPASS Translation Center
The COMPASS Translation Center provides a list of translation services that are relevant for the stay of foreigners. It combines the free textual translation mentioned and application-specific translation services:

- **transearly**, online translation service
- **smart dining**, assistance with restaurant and menu selection and ordering
- **taxi talk**, assistance in dialogues with taxi drivers
- **resc you**, emergeny assistance
- **shopping assistant**, assistance in dialogues with sale persons

[1] we have applied the BLEU score evaluation method and subjective confirmation of the BLEU results.

Smart dining, *taxi talk* and *shopping assistant* combine multilingual generation with phrasebook translation. *Resc you* combines phrasebook translation with a free general text translation for selected situations. *Transearly* combines phrasebook translation with free text translation.

3.2 Multimodal User Interface

COMPASS2008 is supposed to be used in everyday life for different situations and contexts. Therefore, it seems to be difficult to provide a static interaction modality that is always suitable [1]. For example the usage of speech input within a crowded stadium seems to be difficult, because of the background noise. The usage of a stylus to tap or write on the handheld device will only be possible if the user has both hands free. The interaction modalities that are supported by the COMPASS2008 client are based on the results of a user study describe in [6]. This study investigated preferences of users, which had no or little experience with handheld PCs, considering multimodal interaction with a mobile and multimodal interactable shopping assistant in a public environment. The result of the study was that subjects preferred (in addition interaction with only one modality) to interact with speech in combination with gesture. In COMPASS2008 we focus on fast and intuitive interaction that we provide by focusing on gesture (e.g. pointing) input and optional speech input and output. Also, we provide interfaces that are one-handed operable and developed in cooperation with interaction designers. Because COMPASS2008 is designed to support natural communication (e.g. communication between two people who do not speak a mutual language), we emphasize multilingual speech interaction. In addition to the textual presentation of information we allow users to receive auditory information.

3.3 Semantic-oriented Service Platform

The information logistical service platform (ILSP), as a backend part of the COMPASS2008 system, provides basic functionalities needed for a situation aware selection and provision of information and services [4]. This includes demand-specific evaluation and provision of services and information regarding their relevance for the users. It also includes evaluation and execution of strategies regarding delivery of such services and information. The key technologies used and combined by the platform contain the following:

- **Personalization:** basis for personalized information and service supply is an extensible user profile, comprising personal data, preferences and interests.
- **Context / Situation awareness:** Context information includes any relevant information about the user's state and his environment and can be used to retrieve the information needed at the location or in the situation the user currently is in. Location and time are typical context information but not the only ones.
- **Open distributed service infrastructure:** a semantic service registry enables the usage and selection of already existing information and services (e.g., Web services), which can be included dynamically at run-time into the platform.

- **Information-logistical evaluation**: All described key technologies are combined by application-specific evaluation knowledge. The evaluation component of the platform provides knowledge for the appropriate selection of services and information, the definition of appropriate presentation and interaction forms, and for realization of a suitable delivery strategy.

ILSP is implemented as a DOTNET-based multilingual open service platform. Openness here refers to the fact that ILSP aims at managing an arbitrary number of services from multiple service providers. The service registry also maintains service semantic descriptions, based on extended OWL-S [5] to support a semantic matching of user demand and available services [3], i.e the COMPASS2008 services and third party services for e.g. transport information (e.g. flight schedules, metro map) or sightseeing information on points-of-interest. User registration and user profile management is based on LDAP. For the field test some 15 users are registered. ILSP also supports services "customizing", i.e. transfer of individual user demand information with a service request to enable personalized, situation aware content provision. Figure 1 shows the logical system architecture.

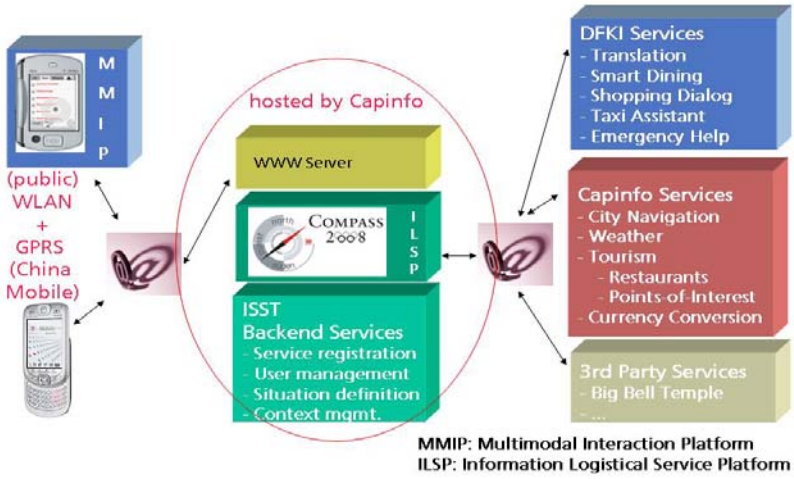

Figure 1. COMPASS20008 pilot system architecture

3.4 Context- and Situation-Awareness

The ILSP context subsystem uses abstraction and aggregation mechanisms for all context dimensions to obtain a set of instances of higher-level situation concepts. Such situation detection is triggered when context sensors report context changes. The result is an abstract context, where each context value is replaced by a set of higher-level (i.e. more abstract) logical values. This abstract context, together with user profile information and possibly other sources serves as an implicitly constructed situation request profile.

The resulting situation request profile is then semantically matched with all the situation profiles known to the system, leading to a (possibly empty) set of situations

fulfilling the request profile. The situations are then used for constructing a service request profile that provides for a semantic matching against the profiles of all known services found in the semantic service registry [10, 11]. Matching evaluates different types of semantic relationships for all profile properties, like subclass, instance and containment relationships. For the fieldtest situation awareness means that a subject, visiting the Big Bell temple, gets offered sightseeing information. When walking to a nearby restaurant, COMPASS2008 recognizes a situation change from "Sightseeing" to "In a restaurant". The offered service portfolio automatically adjusts to the newly recognized demand, i.e. the service offer includes now the "smart dining" service, for instance.

4 Evaluation

The objectives of the field test evaluation are to test the usability and acceptability of the COMPASS2008 services and the overall system. Main Criteria for usability and acceptability are 1) Usefulness of single services and their combination/bundling 2) Effectiveness of service usage: 3) Efficiency of usage 4) User satisfaction and prospective acceptability**.**

We have looked at several different test methods. The *"Classical" Laboratory User Test* selects a small number of users who perform representative tasks with the new application. The Usability problems are detected via observation by usability experts. Therefore, they obtain mostly qualitative data. Post-test interviews will take place supported by questionnaires with open or (standardized) closed-ended questions on user experience. The second method is *Field Test* that takes more than ten users to test the application in a real environment. The usability problems are collected by subjective reports of test users (e.g. online questionnaires). Logging data can be used to assess usage duration and quality. The *Acceptability Test* can be run in real and laboratory environments. Acceptability is measured by test users' self reports (questionnaires, interviews). In our field trial test, we combine the "field test" and the "acceptability test" methods mentioned above. Two standard evaluation instruments have been used as references for the development of our own evaluation methods: 1) SUMI[2] (Software Usability Measurement Inventory) 2) ISO NORM 9241/10[3]. SUMI emphasizes the five dimensions of subjectively experienced usability (Efficiency, Affect, Helpfulness, Control, Learnability), while ISO NORM 9241/10 checks compliance to ISO Norm 9241/10 (ergonomic requirements for screen work places) and tests the following additional features: suitability, self-descriptiveness, controllability, conformity with user expectations, error tolerance, suitability for individualization and learn-ability. For COMPASS2008, we use the standardized sources to formulate questionnaires for usability test. The questionnaires for the acceptability test are developed based on the individual tasks and scenarios. In the following, we give a sample item of our questionnaire:

[2] www.sumi.org or http://sumi.ucc.ie/
[3] www.sozialnetz.de/ca/pq/mdl/

	Dimension	Sample Item	Scale
Post Task	Task completion / goal attainment	Could you complete your task with the service?	Y/N
	Time to task (subjective)	It took me to much time to complete the task	5-step Likert Scale (Agreement)
	Adequacy of perceived effort (five point scale rating)	There have been times in using this service when I have felt quite tense.	5-Step Likert Scale (Agreement)
	Short description of occurred Usability problems (open ended question)	Did there occur any problems while using the service? If yes, please give a short description of the problem	Y/N, Open
Post Test	Functionality / Effectiveness	The service provides all functions I need	5-Step Likert Scale (Agreement)
Usability	Comfort/ Efficiency	There are too many steps required to get something to work.	5-Step Likert Scale (Agreement)
	Control	I feel in command of this service when I am using it.	5-Step Likert Scale (Agreement)
	Orientation	I sometimes don't know what to do next with this service	5-Step Likert Scale (Agreement)
	Reliability	The service has done something unexpected at some time.	5-Step Likert Scale (Agreement)
	Aesthetics	The service has a very attractive presentation.	5-Step Likert Scale (Agreement)
Acceptability	Satisfaction and attitude towards service	Working with this service is satisfying.	5-Step Likert Scale (Agreement)
	Willingness to use the service	I would like to use this service every day.	5-Step Likert Scale (Agreement)
	Reasons/Conditions for usage	What has to be changed with the service to make you using it?	Open

The evaluation results can summarized as follows:

- Task completion is 100% for all tasks. With most tasks, 10 to 20% the users reported little usability problems.
- On the average of 60% subjects did not feel that it takes too much time to complete it. The percentage of subjects, who complain about a long task duration is quite high (40%) in some tasks.
- Most subjects (73%) felt well orientated while solving the different tasks.
- The main factors for the acceptability of the COMPASS2008 services are rated positive by 62%. Still there is potential for optimization. In particular, the controllability and completeness of the service functionality should be improved in order to increase user satisfaction.
- All COMPASS2008 services are judged to be useful or rather useful by more than 80 Percent of the subjects.
- The percentage of users willing to use COMPASS2008 is a above 50-80 % for almost all services, except “Currency Converter”
- There are significant differences between the test groups. Native Chinese and native English tend to rate the services as more useful and show a greater willingness to use COMPASS2008 services than non native English.

Regarding the quality of the situation awareness feature we obtain a recall result of 95,8% (correct situation identified, only relevant services offered). That looks quite satisfying for the first deployment in the field. An analysis of the reason for wrong situation identifications turns out that deviation of the GPS signal leads to mismatches of a user’s position and the geographical locations of points-of-interest. The impact is illustrated by the following figure. The horizontal axis in the figure shows the different situation types. The green bar on the right shows the time intervals for the situations of user 1. The colored dots indicate the detected situations by COMPASS2008. The colored squares indicate the valid real situations.

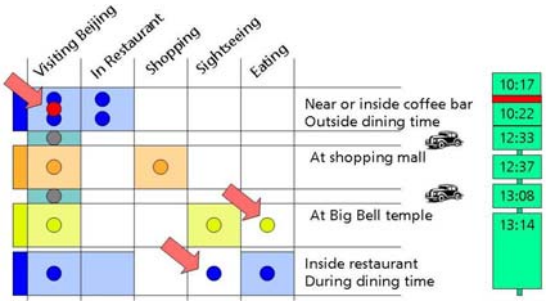

Figure 2. COMPASS2008 situation recognition (user 1)

In the first situation user 1 is in a small coffee shop. The COMPASS2008 system recognizes the matching of the user position with the geographic area of the coffee shop and assigns the corresponding situation (here: "in a restaurant"), including the provision of corresponding services (blue dots). For about 2 minutes during his staying there the GPS signal moves out of the geographic rectangle that is assigned to the coffee shop. Hence, COMPASS2008 detects a situation change from "in a restaurant" to "visiting Beijing" (red dot) although the user does not move. Two minutes later the GPS signal moves back into the geographic area of the coffee shop. COMPASS2008 detects the situation change again and assigns the "in restaurant" situation again (blue dots). Thus, the instability of the GPS signal causes situation changes and not the mobility of the user.

When the COMPASS2008 project is introduced to the field test subjects, one issue is to provide them with information about the system's feature of situation aware service provision. To get user feedback on this feature we include a corresponding question in the post test questionnaire. The evaluation of the questionnaires shows the following results: the test users confirm that situation related provision of services eases the use of COMPASS2008 services (67%) and they sometimes would prefer provision of services more adjusted to personal needs (87 %). In summary, situation-aware service provision is considered useful by most test users, most people even desire to extend the feature.

5 Conclusion

The evaluation of the field test results showed that the applied test methodology was well suited for gathering the desired insights on usability and acceptance of the application. The collected data turned out to be highly informative. They are now serving as a guide for the further perfection of the system.

Such informative and conclusive data could not have been obtained by tests in a lab or at simulated environments. Two important observations are: (i) The subjects were not only able but also quite happy to fill out the detailed questionnaires after each task. Without such immediate recording of the user experience, the results would not have been as trustworthy and pointed. (ii) Without the field test under realistic conditions, we could not have obtained the large number of valuable

reactions from the Chinese communication partners of our subjects such as waiters, cab drivers and sales persons. If we had to do the test again, we would seriously consider to also ask these co-users for their written judgements.

Acknowledgments. The research reported here has been conducted in the Sino-German cooperation project COMPASS2008 (www.compass2008.org), funded by the German Ministry for Education and Research grant no. 01IMD02A, 01IMD02B and 01IMD02C, and by the Chinese Ministry of Science and Technology. Special thanks to German Telekom for their co-funding too.

References

1. Aslan, I., Xu, F, Uszkoreit, H., Krüger, A., Steffen, J.: COMPASS2008: Multimodal, Multilingual and Crosslingual Interaction for Mobile Tourist Guide Applications. In: INTETAIN 2005, Italy, 2005.
2. Busemann, S. , Horacek, H.: A Flexible Shallow Approach to Text Generation. In: Eduard Hovy: 9th INLG '98, August 5-7 pp. 238-247, Niagara-on-the-Lake, Ontario, Canada, 1998
3. Holtkamp, Gartmann, Weissenberg, Han, Gang: Situation-Aware Service Selection from Large Offers, Proc. e2004 Conference, Vienna, Oct. 2004
4. Meissen, U., et al.: Context- and Situation-Awareness in Information Logistics, Proc. of EDBT Workshop on Pervasive Information Management, Springer, 2004.
5. OWL Service Coalition, OWL-S: Semantic Markup for Web Services, http://www.daml.org/services/owl-s/1.0/owl-s.pdf 2004 (accessed Oct. 30, 2006).
6. Uszkoreit, H., Xu, F., Steffen, J., Aslan, I.: The pragmatic combination of different cross-lingual resources for multilingual information services. In Proc. of LREC 2006, Genova, Italy , 2006.
7. Uszkoreit, H, Xu, F., Aslan, I., Steffen, J.: COMPASS2008: An Intelligent Multilingual and Multimodal Mobile Information Service System for Beijing Olympic Games. In Proceedings of KI2006 Demo Collection, Germany, 2006.
8. Wahlster, W. (ed.): Verbmobil: Foundations of Speech-to-Speech Translation. Springer-Verlag. Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo, 2000.
9. Wasinger, R., Krüger, A., Jacobs, O.: Integrating intra and extra gestures into a mobile and multimodal shopping assistant. In Pervasive, pages 297–314, 2005.
10. Weißenberg, N., Voisard, A., Gartmann, R.: Using Ontologies in Personalized Mobile Applications, in Proc. 12th ACM Intl. Symposium on Advances in Geographic Information Systems, D. Pfoser, & I. Cruz (Eds.), ACM Press, New York, pp 2-11, 2004.
11. Weissenberg, N., Voisard, A., Gartmann, R.: An Ontology-based Approach to Personalized Situation-aware Mobile Service Supply, GeoInformatica, An International Journal on Advances of Computer Science for Geographic Information Systems, Springer US, Volume 10, Number 1, March 2006.