Supporting Global Design Through Data Mining and Localization

Barbara Rita Barricelli
Department of Computer Science and Communication
Università degli Studi di Milano
Milan, Italy
barricelli@dico.unimi.it

Malte Ressin
Centre for Internationalization and Usability
University of West London
London, UK
malte.ressin@uwl.ac.uk

Abstract **Localization can be an important work step for software development with a considerable impact not only on success, but also on cost and quality. To facilitate localization, a number of tools exist. In particular in the area of computer-assisted translation, data analysis is used to aid in the work of the translator. In this paper, we propose to apply data mining to assist in the localization of software elements beyond text translation, such as colors, symbols and images. In particular, we propose to apply data mining to make the most of available resources on the internet and treat them as distributed databases.**

Keywords-global design; localization; data mining; translation; internationalization; globalization; culture

I. Introduction

The World Wide Web plays the fundamental role of medium for international communication, participation, and transaction. The characteristics of the Web, its tools and technologies support and stimulate the evolution of methods and techniques for interface design for multi-cultured environments [1]. However, in order to guarantee an international usability, the software applications have to be designed in a culture-oriented way [2][3].

To design and develop global products means in fact to a) extend it to different international contexts, b) to make it able to handle various languages and conventions, c) to localize it according to specific cultures, and d) to translate it in the proper languages.

Up to now many efforts have been made in the field of machine-based translation, and data mining techniques are widely used to this end. However, other aspects related to cultures that are not related to the languages are not yet taken into account.

The contribution of this paper is twofold. First, we highlight the challenges that emerge from the current asset of Web, its global access and the spread of its technologies. Second, we propose a data mining application for the localization of software applications which is able to make the most of available resources on the internet and treat them as distributed databases.

This paper is organized as follows. Section II illustrates the theoretical and practical backgrounds in the field of global design and localization. Section III presents a review of the current tools of Computer-Assisted Translation. In section IV, the challenges that arise from global software design are highlighted and a proposal about the implementation of data mining tools is presented. Section V closes the paper proposing open questions to be addressed in future development of this research.

II. Global Design and Localization

In the global software design literature, many definitions of culture have been given. Yeo [5] defined it as “behavior typical of a group or class (of people)”; Bødker and Pedersen [6] defined culture as “a system of meaning that underlies routine and behavior in everyday working life”; while for Borgman [7], culture includes “race and ethnicity as well as other variables and is manifested in customary behaviors, assumptions and values, patterns of thinking and communicative style”.

The literature about culture and its meaning shows that for a long time, this topic has been discussed in the field of computer science. However, the outcomes of these discussions are still limited and have not been applied concretely and completely to software design and development. Our effort in this research area is to study, propose and develop methods and techniques able to respond to the challenges that emerge from the global design context in semi-automatic ways. We consider it in fact necessary to involve experts in localization to validate the results of automatic tools, because the human knowledge and expertise has to be reinforced and not replaced. Our work stems from the study of the literature about culture, cultural dimensions and localization methodologies and best practices that already exist and that we describe in what follows.

A. Cultural Dimensions

In literature, various cultural models have been proposed and each of them is described by a set of cultural dimensions [8][9][10][11][12]. The most adopted and discussed classification of cultural dimensions is the one proposed by Hofstede [12], by which he recognized five main dimensions:

- Small vs. large power distance: measures the extent to which the less powerful members of organizations and institutions accept and expect that power is distributed unequally.
- Individualism vs. collectivism: measures the degree to which members of organization and institutions are integrated into groups.
- Masculinity vs. femininity: measures the distribution of roles between the genders.

ISBN: 978-1-61208-162-5

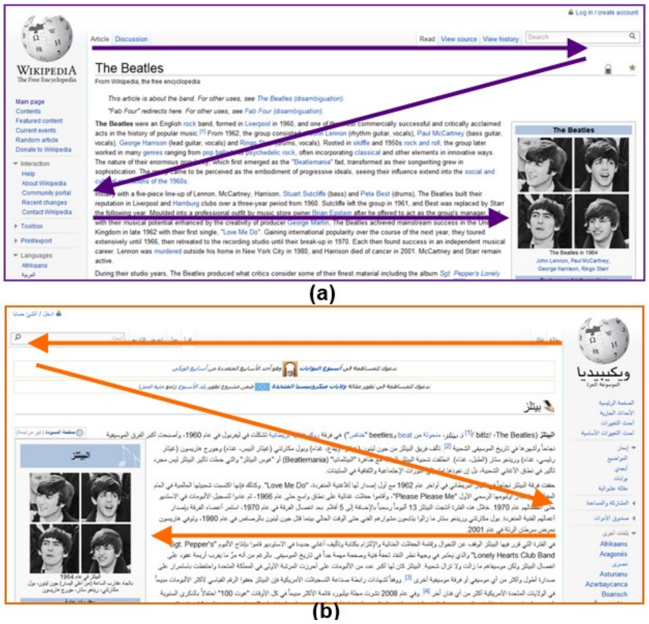
- Weak vs. strong uncertainty avoidance: measures to what extent a culture prepares its members to feel either uncomfortable or comfortable in unstructured situations. Uncertainty-avoiding cultures try to minimize the possibility of unknown and surprising situations by strict laws and rules, safety and security measures, and on the philosophical and religious level by a belief in absolute Truth. Uncertainty-accepting cultures are more tolerant of opinions different from what they are used to; they try to have as few rules as possible, and on the philosophical and religious level they are relativist and allow many currents to flow side by side.
- Long vs. short term orientation: measures to what extent a culture respects values associated with long term orientation or short term orientation. Long term orientation values are thrift and perseverance, while values associated with short term orientation are respect for tradition, fulfilling social obligations.

Several studies have shown how these five cultural dimensions, which classify a person's cultural background into certain scores, relate to certain aspects of a user interface [13][14][15]. One of the most interesting categorizations is the one given by Yeo [5] that categorizes the factors needed to be addressed in global design processes into covert and overt. Overt factors are tangible, straight forward and publicly observable elements. Some examples are date, calendars, time, address formats, character sets, punctuation, and currency. Covert factors are those elements that are intangible and culture-dependent. Colors, sounds, metaphors are examples of covert factors.

B. *GILT Methodological Model and Best Practices*

The design and the development process of global products passes through the performance of four distinct activities [16]: globalization, internationalization, localization, and translation. These four activities constitute the so-called GILT methodological model. Internationalization is an activity that is performed independently by localization and translation, because it affects the structure of the product under design and development and not its content. Translation is included in localization because it represents just one of the actions required to localize a product. An example of internationalized and localized Website is Wikipedia, as shown in Figure 1.

The localization activity is detailed in [17], by separating it into two distinct components, content and package. Content is defined as the linguistic structures, while package is the set of all the non-textual elements and the media through which the content is distributed. After Esselink [4], globalization "addresses the business issues associated with taking a product global. In the globalization of high-tech products this involves integrating localization throughout a company, after proper internationalization and product design, as well as marketing, sales, and support in the world market". Globalizing means therefore the extension of a product to different international context, with the aim of making it usable by the different potential users.



The Beatles

From Wikipedia, the free encyclopedia

This article is about the band. For other uses, see The Beatles (disambiguation).

"Fab Four" redirects here. For other uses, see Fab Four (disambiguation).

The Beatles were an English rock band, formed in Liverpool in 1960, and one of the most commercially successful and critically acclaimed acts in the history of popular music.[1] From 1962, the group consisted of John Lennon (rhythm guitar, vocals), Paul McCartney (bass guitar, vocals), George Harrison (lead guitar, vocals) and Ringo Starr (drums, vocals). Rooted in skiffle and 1950s rock and roll, the group later worked in many genres ranging from pop ballads to psychedelic rock, often incorporating classical and other elements in innovative ways. The nature of their enormous popularity, which first emerged as the "Beatlemania" fad, transformed as their songwriting grew in sophistication. The group came to be perceived as the embodiment of progressive ideals, seeing their influence extend into the social and cultural revolutions of the 1960s.

Initially with a five-piece line-up of Lennon, McCartney, Harrison, Stuart Sutcliffe (bass) and Pete Best (drums), The Beatles built their reputation in Liverpool and Hamburg clubs over a three-year period from 1960. Sutcliffe left the group in 1961, and Best was replaced by Starr the following year. Moulded into a professional outfit by music store owner Brian Epstein after he offered to act as the group's manager, and with their musical potential enhanced by the creativity of producer George Martin, The Beatles achieved mainstream success in the United Kingdom in late 1962 with their first single, "Love Me Do". Gaining international popularity over the course of the next year, they toured extensively until 1966, then retreated to the recording studio until their break-up in 1970. Each then found success in an independent musical career. Lennon was murdered outside his home in New York City in 1980, and Harrison died of cancer in 2001. McCartney and Starr remain active.

During their studio years, The Beatles produced what critics consider some of their finest material including the album Sgt. Pepper's Lonely

Figure 1. Two localizations of Wikipedia: (a) English and (b) Arabic. The organization of the page follows the writing direction of the language (from left to right for English and from right to left for Arabic).

Internationalization is "the process of generalizing a product so that it can handle multiple languages and cultural conventions without the need for re-design. Internationalization takes place at the level of program design and document development". Localization "involves taking a product and making it linguistically and culturally appropriate to the target locale (country/region and language) where it will be used and sold", while translation is "only one of the activities in localization; in addition to translation, a localization project includes many other tasks such as project management, software engineering, testing, and desktop publishing". While translation is aimed at maintaining the meaning of original information by exposing them in different languages, localization transforms the information in equivalent ones but adapted to a different culture.

As suggested by [18], in order to develop software suitable for the global market, a two-step process is needed: internationalization of the software first and its localization next. In [19], a cross-cultural checklist that should be considered by interface designers is given. The authors consider several key factors:

- Text: a simple translation is not enough, many aspects should be taken into account (e.g. jargon, character sets, numbers, date, time formats).
- Images: images represent the visual language of a culture and therefore not only image recognition but also image acceptability problems should be considered.
- Symbols: as for images, also symbols have to be acceptable for the target culture.
- Colors: as pointed out in [20][21][22] interpretation of colors varies in the various cultures. Colors play a

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-162-5

fundamental role in interface design because they convey information, and therefore they need to be chosen very carefully.

- Flow: the writing system of a language and therefore its reading/writing direction affects the way in which the information is recognized by users on a screen. Hence, the logical flow of what is represented on an interface should follow the proper directions.
- Functionality: sometimes functionalities implemented in software application are not accepted in some cultures because they do not respect the cultural conventions that the user needs.

III. Current Software Tools

A number of software tools are available to assist translators in their work. These tools are exclusively aimed at text translation. Their use is commonly called Computer-Assisted Translation (CAT), the software suites incorporating them are usually referred to as Translators' Workbenches. Most of these software tools are centered on extended data collection and database searches and fall into the following two rough categories:

- Translation Memories (TMs), also called repetition manager, store previously translated content and keep it accessible for future use by the translator.
- Machine Translation (MT) employs computer algorithms to derive translations.

A. *Translation Memory*

Nowadays, the use of TMs is widespread throughout commercial translation [23][24]. A number of commercial as well as free translator workbenches are available [25]. Often, the TMs in these come with tools which automatically scan the text of source documents and provide close matches found in its database. Provision of such suggestions can save the translator precious time not having to look up previous translations, while at the same time increasing translation consistency. Accordingly, Schäler [26] has found that usage of such tools provides the following benefits:

- Speed up the translation process.
- Improve translation quality.
- Reduce translation cost.

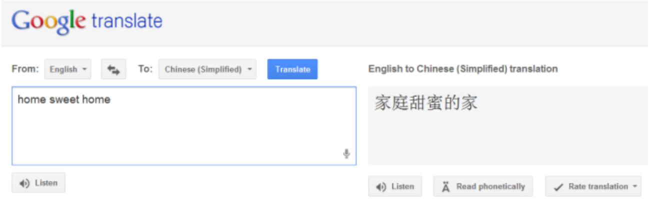
However, Ottman [27] argues that the use of TMs introduces an additional source of translation errors through incorrect context. As a consequence, she argues that additional quality assurance might be necessary.

B. *Machine Translation*

Broadly speaking, there are two approaches to MT. Rule-Based Machine Translation (RBMT) aims to translate text through the use of dictionary and grammar encoded as a program. Statistical Machine Translation (SMT) compares the source text with existing bilingual text to derive a translation of the source text. Combinations of both approaches exist.

For the context of this article, only SMT is of interest. This approach relies on analysis of text which is already available in different languages, similar to the Rosetta stone. The results of this analysis are then used to translate new source texts.

The potential use of MT as tool for translation has been understood early on [28][29]. However, pure machine translation still requires extensive human review due to prevalent quality issues [30]. Elsen [31] states that MT has the potential to increase translation speed while at the same time reducing translation cost. Although he asserts that MT should be particularly applicable to short and simple text, a requirement satisfied by typical user interface text, he concludes that placeholders would pose additional difficulty. An example of MT tool is given in Figure 2.



From: English To: Chinese (Simplified)

English to Chinese (Simplified) translation

home sweet home

家庭甜蜜的家

Figure 2. Google translate, one of the most used MT tools available online.

IV. Challenges in Data Mining and Localization

While the tools mentioned in the previous chapter can be employed for software localization, they will be of assistance only for text translation. The remaining visual elements described in chapter II are not covered by these tools.

From our experience in localization and internationalization and from the critical analysis of the literature, the need of semi-automatic tools to support the software developers in choosing the right visual elements is strongly emerging.

However, we consider the localization process as an activity that should be managed by an interdisciplinary team: developers are in fact not able to deal with cultural issues alone and they need to work together with translators, ethnographers, and other experts. This is the reason why we are not promoting the development of automatic systems but of semi-automatic systems that should be designed to help the team and not to substitute the human tasks. The experience and background of all the stakeholders in the localization team should be exploited and applied to the cases at hand. In the category of visual elements, we consider all the elements that could be part of an interface and are not textual, e.g., colors, pictures, symbols.

Ryan, Anastasiou and Cleary (2009) suggested use of a Localization Knowledge Repository (LKR) to facilitate localization of such elements.

The goal of this paper is to propose the adoption of data mining techniques to address the challenges that emerge from the state of the art in localization and internationalization. Data mining could be in fact the good means by which to derive from existing applications the rules to be used for the design and development of new ones.

ISBN: 978-1-61208-162-5

We identified three main challenges that are related to the choice of colors, symbols, and images. A data mining tool could be used to make comparisons between the various localized instances of internationalized application in order to derive some cultural rules that regard the use of colors and the appropriate symbols and images to be used. For instance, by comparing two instances of the same Website, one localized for a culture and the other localized for a different one, we could observe the different choices made in order to deliver the same meanings but using different visual elements. Clearly, such kind of analysis could be significant only if applied to a large number of software applications. Another help could be given by the analysis of the meta-tags included in the code of the websites which could add some information about context and content of the application.

For our proposed application, we suggest the use knowledge discovery and known techniques of software mining on the code level. Specifically, software is mined for patterns at the user interface and data level, possibly also at the business code and statement level. Treatment of individual elements would differ on element complexity. For example, color can be normalized into RGB codes. More complex elements such as symbols and images would require additional processing, for example via image recognition. Similarly, mining internationalized software would differ marginally from mining non-internationalized software insofar as different releases or (language) versions might have to be processed in case of the latter. However, both could feed into the resulting data sets.

This software mining will enable the creation of association rules between software properties or elements marked through markups, meta-tags, and localization implementations. Additionally, it is conceivable to use surrounding code as marker. After normalizing those association rules, they can be applied to new, yet unlocalized applications.

V. CONCLUSIONS AND OPEN QUESTIONS

In this paper, we discussed software localization and its facilitation by software tools. We suggested leveraging data analysis to provide culture-conforming colors, symbols, images etc. already during the design phase.

The approach we suggested is obviously heavily influenced by statistical machine translation. As such, it shares its downsides, i.e. the requirement of identical documents for different languages or cultures. As we elaborated above, in order to be useful, these documents need to provide sufficient context information, for example via the use of meta-tags. Such explicit information can only be avoided if there is sufficient implicit data to ensure the appropriateness of an element's or item's translation in the given context.

We firmly believe that the potential of data analysis methods for localization has not yet been fully realized, and that there are a plethora of opportunities to further simplify localization beyond the realm of text translation.

ACKNOWLEDGMENT

The work of Barbara Rita Barricelli was supported by the Initial Training Network "Marie Curie Actions", funded by the FP 7-People Programme with reference PITN-GA-2008-215446 entitled "DESIRE: Creative Design for Innovation in Science and Technology".

REFERENCES

[1] W. Barber and A. N. Badre, "Culturability: The Merging of Culture and Usability," Proc. 4th Conference on Human Factors and the Web, June 1998.

[2] K. Reinecke and A. Bernstein, "Predicting user interface preferences of culturally ambiguous users," Proc. of 26th Conference on Human Factors in Computing Systems (CHI'08), ACM Press, Apr. 2008, pp. 3261-3266.

[3] E. M. del Galdo and J. Nielsen, International Users Interfaces. New York, NY: John Wiley & Sons, 1996.

[4] B. Esselink, B. A Practical Guide to Localization. Amsterdam, NL: John Benjamins Publishing Company, 2000.

[5] A. Yeo, "Cultural user interfaces: a silver lining in cultural diversity," SIGCHI Bull., vol. 28, no. 3, 1996, pp. 4-7.

[6] K. Bødker and J. Pedersen, "Workplace Cultures: Looking at Artifacts, Symbols and Practices," in Design at Work: Cooperative Design of Computer Systems, J. Greenbaum and M. Kyng, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991, pp. 121-136.

[7] C. L. Borgman, Cultural diversity in interface design. SIGCHI Bull., vol. 24, no. 4, 1992, p. 31.

[8] N. L. Hoft, "Developing a cultural model," in International Users interfaces, E. M. del Galdo and J. Nielsen, Eds., New York, NY: Wiley & Sons, 1996, pp. 41-73.

[9] E. Hall, The silent language. New York, NY: Doubleday, 1959.

[10] F. Trompenaars, Riding the waves of culture. London, UK: Nicholas Brealey publishing, 1993.

[11] D. Victor, International business communications. New York, NY: Harper Collins, 1992.

[12] G. Hofstede, Cultures and organisations: software of the mind. New York, NY: McGraw Hill, 1991.

[13] C. Dormann and C. Chisalita, "Cultural Values in Web Site Design," in Proc. 11th European Conference on Cognitive Ergonomics (ECCE11), Sep. 2002, pp. 8-11.

[14] A. Marcus, "Cultural Dimensions and Global Web Design: What? So What? Now What?," in Proc. 7th Conference on Human Factors and the Web, June 2001, pp. 1-15.

[15] A. Smith and Y. Chang, Y. "Quantifying Hofstede and Developing Cultural Fingerprints for Website Acceptability," in Proc. 5th International Workshop on Internationalisation of Products and Systems (IWIPS 2003), P&SI, July 2003, pp. 89-102.

[16] P. Cadieux and B. Esselink, "GILT: Globalization, Internationalization, Localization, Translation," LISA Globalization Insider, vol. 1, no. 5, 2002.

[17] M. O'Hagan and D. Ashworth, Translation-mediated communication in a digital world Facing the challenges of globalization and localization. Clevedon, UK: Multilingual Matters LTD, 2002

[18] T. Madell, C. Parson, and J. Abegg, Developing and localizing international software. Upper Saddle River, NJ: Prentice-Hall, Inc., 1994.

[19] P. Russo and S. Boor, "How fluent is your interface?: designing for international users," in Proc. Conference on Human Factors in Computing Systems (INTERCHI '93), IOS Press, Apr. 1993, pp. 342-347.

[20] L. G. Thorell and W. J. Smith, Using Computer Color Effectively: An Illustrated Reference. Englewood Cliffs, NJ: Prentice Hall, 1990.

[21] K. Garland, "The use of short term feedback in the preparation of technical and instructional illustration," in Proc. Conference on Research in Illustration, 1982.

[22] A. J. Courtney, "Chinese Population Stereotypes: Color Association," Human Factors, vol. 28, no. l, 1986, pp. 97-99.

[23] K.-H. Freigang and U. Reinke, "Translation-Memory-Systeme in der Softwarelokalisierung [Translation-Memory-Systems in software localization]," in Einführung in die Softwarelokalisierung [Introduction to Software Localization], D. Reineke and K.-D. Schmitz, Eds. Tübingen, Germany: Narr, 2005, pp. 55-71.

[24] E. Yuste, "Corporate Language Resources in Multilingual Content Creation, Maintenance and Leverage," in Proc. 2nd International Workshop on Language Resources for Translation Work Research and Training, Aug. 2004, pp. 9-15.

[25] S. Falcone, "Translation Aid Software - Four Translation Memory Programs Reviewed," Translation Journal, vol. 1, no. 2, 1998.

[26] R. Schäler, R. "A Practical Evaluation of an Integrated Translation Tool during a Large Scale Localisation Project," in Proc. 4th Conference on Applied Natural Language Processing (ANLC'94), 1994, pp. 192-193.

[27] A. Ottmann, "Lokalisierung von Softwareoberflächen [Localization of Software User Interfaces]," in Einführung in die Softwarelokalisierung [Introduction to Software Localization], D. Reineke and K.-D. Schmitz, Eds. Tübingen, Germany: Narr, 2005, pp. 101-115.

[28] C. Brace, "Trados: Ten Years On," Language Industry Monitor, July-August 1994.

[29] R. W. Collins, "Software Localization: Issues and Methods," in Proc. 9th European Conference on Information Systems, June 2001, pp. 36-44.

[30] J. Yao, M. Zhou, T. Zhao, H. Yu, and S. Li, "An Automatic Evaluation Method for Localization Oriented Lexicalised EBMT System," in Proc. 19th Conference on Computational Linguistics (COLING 2002), Aug. 2002.

[31] H. Elsen, "Maschinelle Übersetzung in der Softwarelokalisierung. [Machine Translation in Software Localization]," in Einführung in die Softwarelokalisierung [Introduction to Software Localization], D. Reineke and K.-D. Schmitz, Eds. Tübingen, Germany: Narr, 2005, pp. 89-99.