S. Boll, S. Maaß & R. Malaka (Hrsg.): Workshopband Mensch & Computer 2013
München: Oldenbourg Verlag, 2013, S. 463–471

A Corpus for Intercultural Comparison of Web Sites

Thomas Mandl, Christa Womser-Hacker

Information Science and Natural Language Processing, University of Hildesheim

{mandl, womser}@uni-hildesheim.de

Abstract

The scientific comparison of intercultural differences of Web sites lacks a standardized collection for which results can be reproduced. Results are often anecdotal and based on few items which change very dynamically. We developed a corpus of Web sites from different cultures for future research which is stored and which can be accessed through an interface developed for that purpose. We show how the preconditions for corpora as they are well known in linguistics and retrieval are adopted for international Web design research. The collection policy is described. The challenges for Web crawling and storing pages are discussed.

1 Introduction

Scientific research requires that experiments are repeatable and that the results can be reproduced. The research on Web design and especially on the comparison of internationally different Web design solutions are typically limited in their object of study to a small or very small set of Web sites. The selection of these sites remains often unclear or is directed by economic interest to a few companies. Results are sometimes restricted to anecdotal evidence.

As a consequence, the validity of these studies for the Web as a whole is completely unclear. We have no guarantee that some sites represent the Web well. Especially, quantitative analysis suffers from a lack of rigor. Because most people select their own set of sites for their study, comparisons between studies are hard. In addition, the results cannot be reproduced by others due to the dynamics of the Web. The content and the design of Web sites change quickly and the same experiment might lead to other results some months later.
These challenges need to be overcome in order to reach the scientific requirements for the repeatability of experiments and in order to allow the comparison between the outcomes of research. It would be necessary to create a Web corpus which stores pages and allows intercultural comparisons. Such a corpus would need to integrate Web sites from several countries and domains and would need to be stored at several times.

Corpus research and corpus creation are not yet part of Human-Computer Interaction and intercultural research on Web systems. However, they are accepted scientific areas within Computational Linguistics (Meyer 2002) and Information Retrieval (Voorhees & Harman 2005) and much can be learned from these fields. Also, the National Libraries of several countries store the Web as part of the cultural heritage. The desire to archive Web pages for different reasons (Hockx 2011) has led to much development of technologies like tools for crawling pages. We describe how our corpus was designed, how existing technology was used and how components were developed.

2 Intercultural Differences of Web Sites

Information systems should to be adapted to the culture of the user. This adaptation process is referred to as *localization*. It needs to consider issues like formats, reading direction, colors, icons or symbols (Esselink 1998, George et al. 2010). In addition, localization needs to be aware of the deeper layers of culture (Choong et al. 2005). Understanding a particular culture and the resulting needs in relation to the design of information systems require an understanding of culture itself and the factors that contribute to its existence. There are many definitions of culture (Kroeber & Kluckhohn, 1952). The influential Dutch anthropologist Hofstede defined culture as learned patterns of ”thinking, feeling, and potential acting” that form the mental program or the ”software of the mind” (Hofstede 1997) of an individual. This particular ”software” affects our way of thinking and acting in the world. National or social cultures show how people interact with each other.

Cultures are often classified in accordance with their relative positions on a number of polar scales which cultural anthropology commonly calls cultural dimensions. The position of a culture on those scales is determined by the dominant value orientations. Hofstede (1997) defined four dimensions of culture:

6. **Power distance** measures the extent to which subordinates (employees, students) respond to power and authority (managers, teachers) and how they expect and accept unequal power distribution.
7. **Individualism vs. Collectivism**: these value orientations refer to the ties among individuals in a society.
8. **Uncertainty avoidance** describes the extent to which individuals feel threatened by uncertain or unknown situations.
9. **Masculinity vs. Femininity**: these two extreme values of this dimension focus on the differences between the social roles attributed to genders.

Other researchers discussed further dimensions concerning low and high context communication, proxemics (attitude toward space) and chronemics (perception of time) (Lustig & Koester 2003). Cultural dimensions like the ones defined by Hofstede have often been used in research on cultural differences due to the appeal of their quantitative approach. Nevertheless, these dimensions have drawn criticism within intercultural research.

An early study of Barber & Badre (1998) collected typical cultural markers in an inductive approach. The approach of Marcus & Gould (2000) started with knowledge on cultural dimensions in general and intended to locate effects within web sites. Marcus & Gould (2000) presented examples for differences for all cultural dimensions which are convincing. However, their findings are based on a small and pre-selected set of web sites. Their studies like many others illustrate the need for a reliable corpus for this kind of research.

Cultural markers within Web sites were also procured by Sun (2001). His study which included interviews about certain homepages showed that the presence of cultural markers increased the aesthetic satisfaction with a web site.

The methodology for intercultural research is especially challenging. What is measured in a human-computer interaction experiment in an intercultural setting? Can good vs. bad design be determined or can usability or typical design for one culture or another be identified? Empirically convincing studies are difficult to set up from a methodological point of view. In common quantitative human-computer interaction studies, two versions of a user interface are presented to two user groups who are selected from the same culture and who are believed to be homogeneous. For comparative studies in international web design analysis, the user groups are different and their reaction to the system is under investigation. However, it is difficult to leave the system constant. The system cannot be presented to two groups of users from different countries without modification. The system needs to be translated and culturally adapted. For example, the investigated task may be embedded completely differently in the two cultures. Typical user groups like university students may have quite different features like social group in different societies (Schmitz et al. 2008). Hence, the system often needs to be changed significantly in order to be adequate for a real-life experiment which makes comparability difficult (Evers 2002).

Cultural differences have also been investigated in different domains like touristic websites or e-learning systems (Kamentz & Womser-Hacker 2002). Therefore, a corpus should allow studying domains in comparison among cultures.

However, it is not clear how cultural dimensions may contribute to research on intercultural web design (Mushtaha & Troyer 2009). Some authors noted that the assumptions made on the basis of cultural dimensions may be misleading because they have not been developed for Web design. Most important, findings from most studies could not be repeated because the sites had changed or even disappeared from the Web.

3 Corpus Design

A corpus for intercultural Web site design needs to be planned with typical requirements for a corpus. Despite and moreover due to the dynamic nature of the Web corpora are useful. The corpus should be sufficiently large in order to be somewhat representative for the Web. Because Web design greatly differs between domains (consider computer games and banking sites), the selection should be controlled to encompass sites from several domains in several

countries. The selection policy and issues on the site storage and domain assignment are presented further down in this section. That way, even for current Web sites the corpus allows the researcher quick access to Web sites of different categories even if one cannot understand the language of the Web site.

The pages should be downloaded and stored in their original form so they can be displayed as they looked like. This is an issue for Web archives in general and there are still some problems. In addition, meta data needs to be stored. The corpus needs to record the domain of the page, the language, the number of links it is away from the homepage and the date of the download as well as if it is connected to other pages in the corpus. The crawling is further described in the following section.

Another requirement for a corpus is access for the users. Due to the unclear copyright situation, the access to the whole collection can only be granted on our venue which is not an optimal solution. Access to the list of sites is provided openly. The interface needs to allow an overview of the crawled pages, the display of the individual pages and navigation between the connected pages. Our corpus interface also allows primary visual analysis but access for automatic tools is important for quantitative analysis. It will be presented in section 5.

Corpus creation and crawling require a selection policy in order to control the content of the corpus. The assignment of sites to a domain is not a trivial task and needs to be carried out by humans. We decided that native speakers of a language should control this assignment. They use tools like the Open Directory Project (dmoz.org) which provides a hierarchical classification of Web sites to topics. As a consequence, the corpus is still limited to 9 languages and needs extension in the future.

Domains	Sports, News, Universities, Tourism, Restaurants/Food
Languages	Chinese, English, Spanish, Russian, French, Portuguese, German, Bulgarian, Czech

Table 1: List of Domains and Languages in the Corpus

For each pair of language and culture, 100 sites were added to the corpus except for the East European languages Bulgarian and Czech for which only 50 sites for each domain were selected. In cases when the Open Directory did not contain a sufficient number of sites, other sources like search engines were also used. Also sites personally known were added.

Sites that were obviously not modified during the last six months were excluded in order to have more recent sites in the collection. Blogs were also excluded because they form a specific kind of site. Another reason for exclusion was the size of the site. If a site did not encompass three levels of navigation and seemed to be small, it was not included. Sometimes, even three levels might not be sufficient for an analysis of, for example, differences of information architecture. However, there has to be a trade-off between size and potential of the corpus.

Some food sites contained information about alcohol and were protected in order to restrain minors from them. These sites were omitted as well. Sites were deleted from the collection if malware (trojan, virus) was detected on them. Due to the individual checking of each site, the process of the seed list creation was very time consuming. The manually created seed lists for the selected domains were the starting points for the crawler.

The corpus is created from a list of URLs (sites) which can be considered as homepages. The corpus need to be assembled as a list of pages which can be found under the address starting with the homepages but considering also other pages on the site. Crawling tools allow to parameterize the number of links that are followed when downloading pages from a site. Currently, our corpus considers sub pages that are a maximum of two internal links away from the homepage or starting page. For hierarchically constructed sites, this corresponds to the hierarchical level of the page.

The pages should be archived in a way which allows the reconstruction of their original appearance. In order to assure that, images (jpg, gif) and style sheets (CSS) were included. Advertisement was not excluded and also archived because it could be of interest in intercultural analysis. Large binary data types like pdf were not included in order to keep the size of the corpus as small as possible. Often zip and pdf files are to be printed or installed and do not contribute much to the web design.

4 Crawling the Sites and Database Backend

Several crawling tools like Jobo and Crawler4J were evaluated. The final decision was made for HTTrack which was designed to copy web sites to the file system in order to allow offline browsing. It is open source, includes a graphical user interface and enables access directly from JAVA programs. HTTrack allows the interruption and continuation of crawling processes which was a very important function. Filters for file formats as necessary for our project are supported as well as the number of links to follow. HTTrack also comes with an indexer to enable text search on the offline data. During crawling the system follows the Netiquette rules and Robots exclusion Standard. As such, it does not access parts of the site that are excluded in the robots.txt file.

After some testing, the crawling process could be started on a SuSe Linux Server and all pages fulfilling the conditions were saved in a MySQL Database. Because of the large volume of data, the crawling lasted several weeks and some unforeseen problems had to be solved. Complete archival of 50 test pages took typically 36 hours.

A client-server-architecture was established where crawler and database were connected via a JAVA backend. The user interacts via the interface with the corpus which was realized as a browser based application.

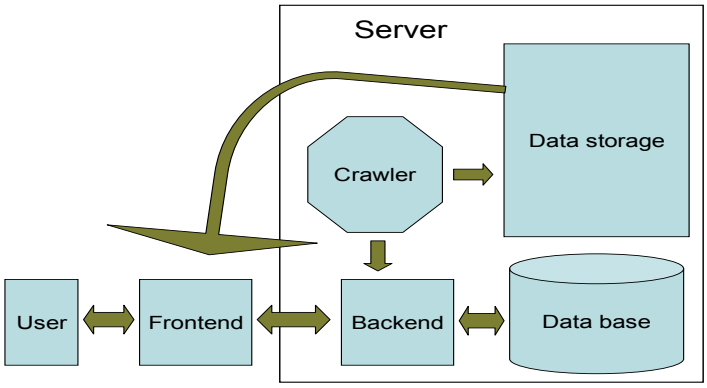

Figure 1: Technological components of the corpus (Bertram et al. 2012: 44)

The model of the database is shown in fig. 2.

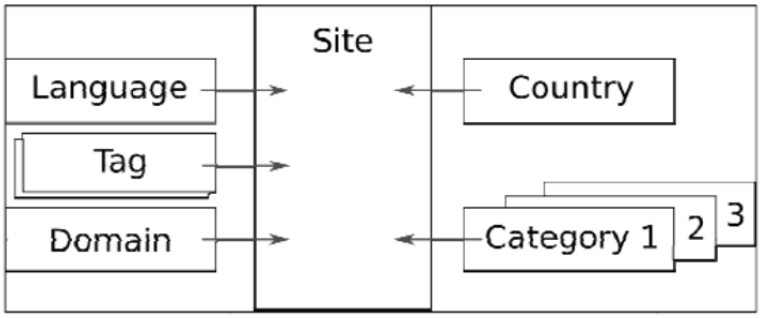

Figure 2: Database Structure (Bertram et al. 2012: 52)

For every Website saved by the crawler some metadata was added: time stamp, title, original URL, language code (ISO 639-2/B), country code (ISO-3166-1), main category and subcategories (e.g. restaurants & food → drink → non-alcoholic), domain. Free tags assigned by the seed list creator are also allowed.

5 Frontend and Tools for Accessing the Corpus

For gaining insight into the information needs of potential users of the corpus, interviews were conducted. Participants mentioned mainly three tasks for working with the corpus:

- Simple selection of web pages on the basis of various conditions, e.g. URL, language, country, category, subcategory etc.
- Comparing pages by putting them in parallel on the screen

- Statistics based on HTML codes to be presented together with the pages, e.g. number of words, images and internal and external links, size of images, colors, fonts, menu items, dates etc.

Based on these needs, a prototype of a user interface was developed providing access to the corpus. It is shown in figure 3.

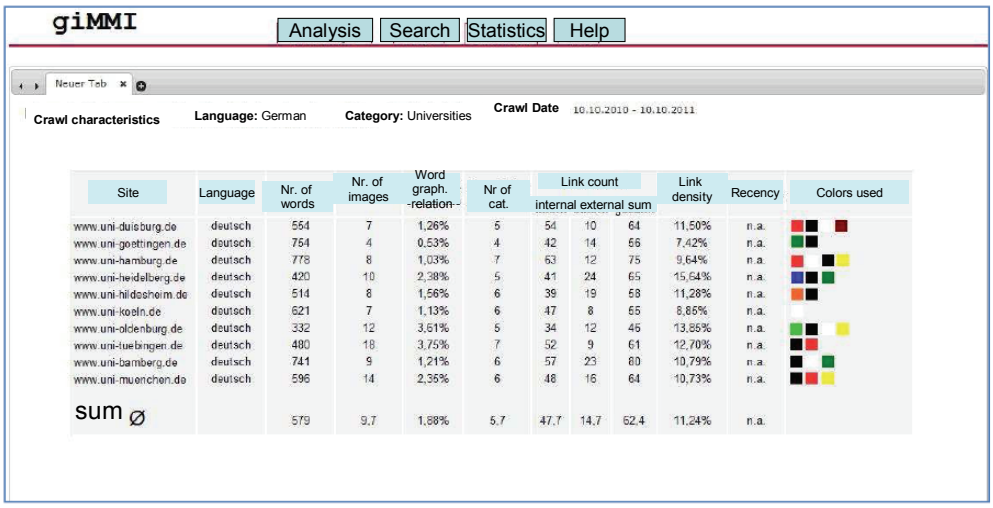

Figure. 3: Prototype of the user interface (Bertram et al. 2012, 56)

The tool developed intends to give some basic information without enforcing a certain type of analysis. As research on cultural differences of Web interfaces shows, very many aspects can be analyzed both intellectually and automatically. Our interface is primarily a tool to explore the corpus and develop hypotheses. A statistical overview provides some basic information about the pages, e.g. number of words, number of images, relation between images and words, number of internal and external links, link density and main colors used on the site. To be able to compare two or more Web pages, the Tabbed Browsing Concept was applied. One active Web page is shown whereas the other ones can be reached very quickly by clicking on the tabs. Since the selection of Web pages can lead to big amount of results, a simple dropdown menu with auto suggest search as known from Google was integrated.

We are aware that it is not possible to analyze high level features of cultural dimensions automatically. Nevertheless, the user should be supported to fulfill this task. Therefore, an evaluation scheme was developed which can be used as a basis during the analytical work. The items are related to the cultural dimensions of Geert Hofstede and other cultural models proposed by several scientists. It is possible to generate an online questionnaire which can support the analysis. Another important result of the project is a collection of usability guidelines in different countries which can be added to the corpus toolbox. It could be shown that the official portal provided by the US government www.usability.gov has a lot of influence. On the other hand other countries like the UK, Germany or Russia have been developing

proper guidelines with respect to their specific characteristics. In a first pretest of the corpus, these guidelines were applied and a small set of websites was evaluated.

6 Outlook

The corpus and the interface developed will allow detailed analysis of Web designs under different perspectives and with repeatable results. We expect that scientists and students will be able to carry out comparisons. Future work will be needed to extend the corpus by including other countries and domains. In addition, we intend to store the corpus at different points in time and to develop a strategy on how the corpus can be increased. These extensions will also require additional functions in the user interface. The tools needs to integrate a text based search engine and a user administration. Furthermore, a user centered evaluation is necessary to test and improve the interface.

7 References

Barber, W. & Badre A. (1998). *Culturability: The Merging of Culture and Usability.* In Proceedings of the 4th Conference on Human Factors and the Web.

Bertram, J., Block, M., Deiloff, K., Fischer, J., Gätzke, N., Heimsoth, M., Jatho, E., Kastner, S. Koniger, V., Lahousse, S., Maleshkova, K., Petersen, T., Rasche, C., Scharnhop, C. (2012). *Master-Projektseminar Internationale Mensch-Maschine-Interaktion.* Project Report, University of Hildesheim.

Choong, Y-Y., Plocher, T.A., Rau, P-L.P. (2005). *Cross-cultural Web Design.* In Procter, Robert (Ed.). Handbook of Web Design. Lawrence Erlbaum Associates.

Esselink, B. (1998). *A practical guide to software localization.* Amsterdam, Philadelphia: John Benjamins Pub.Co.

Hockx-Yu, H. (2011). *The Past Issue of the Web.* In Proc. ACM WebSci'11. Koblenz, pp. 1–8

Evers, V., (2002). Cross-Cultural Applicability of User Evaluation Methods: A Case Study amongst Japanese, North-American, English and Dutch Users. In *Proceedings ACM CHI Conf*. pp. 740-741.

Funke, N.; Hong, H.; Kiefer, V.; Klobassa, K.; Suckow, K. ; Zidek, M. (2010). *Korpusentwicklung für interkulturelle Informationssysteme.* Project Report, University of Hildesheim.

George, R.; Nesbitt, K.; Gillard, P.; Donovan, M. (2010). *Identifying cultural design requirements for an Australian indigenous website*. In Proc Eleventh Australasian Conf. on User Interface, Jan., Brisbane, Australia.

Hall, E. T. (1976). *Beyond Culture.* New York: Doubleday.

Hofstede, G. (1997). *Culture and Organizations: Software of the Mind.* London: McGraw-Hill.

Kamentz, E., Womser-Hacker, C. (2002). *Cross-Cultural Differences in Academic Styles and Learning Behavior in the Context of the Design of Adaptive Educational Hypermedia.* Proc 6th World Multi-conference on Systemics, Cybernetics and Informatics (SCI 2002) Orlando. pp. 402-407.

Kroeber, A. & Kluckhohn, C. (1952). *Culture: a critical review of concepts and definitions.* New York: Random House

Lustig, M. W., Koester, J. (2003). *Intercultural competence: interpersonal communication across cultures.* Allyn & Bacon.

Marcus, A., Gould, E. (2000). *Cultural Dimensions and Global Web User-Interface Design: What? So What? Now What?* Proc 6th Conference on Human Factors and the Web in Austin, Texas, 19 June.

Meyer, C. F. (2002). *English Corpus Linguistics.* Cambridge University.

Mushtaha, A.; Troyer, O. D. (2009). *Cross-Culture and Website Design: Cultural Movements and Settled Cultural Variables.* In Internationalization, Design and Global Development [LNCS 5623] Springer. pp. 69-78.

Romberg, M.; Röse, K.; Zühlke, D. (1999). *Global Demands of non-european Markets for the Design of User-Interfaces.* MMI-Interaktiv Nr. 1, March.

Sun, H. (2001). *Building a culturally-competent corporate web site.* In SIGDOC '01, Proc 19th annual intl. conf. on Computer documentation (pp. 95-102). New York, NY, USA: ACM.

Schmitz, A. K.; Mandl, T.; Womser-Hacker, C. (2008). *Cultural Differences between Taiwanese and German Web Users: Challenges for Intercultural User Testing.* In Proc 10th Intl. Conf. on Enterprise Information Systems (ICEIS) 12 - 16, June Barcelona. pp. 62-69.

Trompenaars, F., Hampden-Turner, C. (1997). *Riding the Waves of Culture: Understanding Cultural Diversity in Business.* London: Nicholas Brealey.

Voorhees, E. M.; Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press.

Contact Information

Thomas Mandl, Christa Womser-Hacker

Institute for Information Science and Natural Language Processing (IWIST)

Universität Hildesheim
Marienburger Platz 22
31141 Hildesheim
Germany

http://www.uni-hildesheim.de/iwist

mandl@uni-hildesheim.de, womser@uni-hildesheim.de,