Localisation Focus

THE INTERNATIONAL JOURNAL OF LOCALISATION

ISSN 1649-2358

Centre for Next Generation Localisation

science foundation ireland

fondúireacht eolaíochta éireann

The peer-reviewed and indexed localisation journal

VOL. 6 Issue 1

EDITORIAL BOARD

PUBLISHER INFORMATION

Editor: Reinhard Schäler, *Director*, Localisation Research Centre, University of Limerick, Limerick, Ireland
Production Editor: Karl Kelly, *Manager* Localisation Research Centre, University of Limerick, Limerick, Ireland
Published by: Localisation Research Centre, CSIS Department, University of Limerick, Limerick, Ireland

AIMS AND SCOPE

Localisation Focus – The International Journal of Localisation provides a forum for localisation professionals and researchers to discuss and present their localisation-related work, covering all aspects of this multi-disciplinary field, including software engineering, tools and technology development, cultural aspects, translation studies, project management, workflow and process automation, education and training, and details of new developments in the localisation industry. Proposed contributions are peer-reviewed thereby ensuring a high standard of published material. Localisation Focus is distributed worldwide to libraries and localisation professionals, including engineers, managers, trainers, linguists, researchers and students. Indexed on a number of databases, this journal affords contributors increased recognition for their work. Localisation-related papers, articles, reviews, perspectives, insights and correspondence are all welcome.

To access previous issues online go to http://www.localisation.ie/resources/locfocus/pdf.htm and click on the issue you wish to download. Use the following logon details - username: locfocsub and password: V610808

Members of **The Institute of Localisation Professionals (TILP)** receive Localisation Focus – The International Journal of Localisation as part of their membership benefits. Membership applications can be filed electronically from ***www.tilponline.org*** Change of address details should be sent to LRC@ul.ie

Subscription: To subscribe to Localisation Focus - The International Journal of Localisation visit www.localisationshop.com (subscriptions tab). For more information visit www.localisation.ie/lf

Localisation Focus – The International Journal of Localisation (ISSN 1649-2358) is published and distributed annually and has been published since 1996 by the Localisation Research Centre, University of Limerick, Limerick, Ireland. Articles are peer reviewed and indexed by major scientific research services.

FROM THE EDITOR

The Editors and the Editorial Board of Localisation Focus - The International Journal of Localisation are proud to present to you our journal in its new format, reflecting our aim to provide you with solid, well-researched, and detailed information on advances in localisation research.

In this issue, researchers in academia and industry report on their findings in web localisation, efficiency improvements in multilingual localisation testing, linguistic technologies for native South African languages, games localisation and approaches to the cultural adaptation of digital content. The variety of these themes and topics is an excellent reflection of the multi-disciplinary nature of localisation research.

Miguel Jiménez of Rutgers University reports on his research in the area of web genres. Jiménez believes that the industry has not yet fully researched which characteristics, conventions or language have been developed and established in each target locale. Having investigated localisation, texts and digital genres mainly relating to corporate websites, he concludes that while websites are supposed to look like those produced in the receiving locale, there is a need for the localisation industry and Translation Studies to produce a clear comparative model of the specific textual, terminological, discourse or structural aspects of the major target locales.

Arthur, Hannon and Ward of Brandt report on the results of their experiments aimed at improving the simultaneous release of multilingual versions of software ("simship") by performing a comparative analysis of two automated software testing tools, WinRunner and ShadowTM. Having examined some of the advantages and disadvantages of the different modes of testing software applications, the authors conclude that in most cases a mix of manual and automated testing will yield the best results. The authors see the main advantages of the approach used by ShadowTM in the separation of testing expertise from specialist product knowledge and hardware setup, i.e. testers without specialist product knowledge or the linguistic background associated with a specific target locale can still perform adequate testing of localised products.

Bosch, Jones, Pretorius and Anderson of the University of South Africa discuss the development of computational morphological analysers for six South African Bantu languages which, due to their rich agglutinating morphological structures, pose particular challenges. In their contribution, the authors focus on the lessons learnt during the development of a number of prototypes of morphological analysers and the development of standardised XML machine-readable lexicons for South African Bantu languages.

Miguel Bernal of Roehampton University in his contribution, What's in a 'Game'?, highlights the need for research and specialisation in translation studies in the area of multimedia interactive entertainment software product localisation. His research shows that in this highly specialised area translators are required to deal with many different textual types and to have access to a wide variety of technical and analytical skills.

Schäler introduces the notion of reverse localisation as a challenge to the widely adapted cultural adaptation approach in mainstream localisation where one criterion to determine whether a localisation project has been successful is that the origin of the localised digital content can no longer be identified: users in Italy believe that the content is of Italian origin, users in Denmark believe it is of Danish origin and so on. The author provides examples that demonstrate how digital content can be more attractive and successful across a variety of locales, if it highlights (rather than hides) cultural differences and recommends a new approach to cultural adaptation of digital content.

Localisation Focus has been published since 1995. Due to financial constraints, the first issues were produced by the LRC researchers themselves. It soon became one of the most widely read and distributed industry newsletters in printed form. The LRC then started to develop a dedicated journal section, allowing for longer and more detailed research papers to be published. Today, you are holding in your hands the first version of Localisation Focus - The International Journal of Localisation that is exclusively dedicated to peer-reviewed articles that will be indexed by some of the world's most prestigious scientific indexes to ensure wide visibility and recognition of the work published in our journal.

Reinhard Schäler

Web Genres in Localisation: a Spanish Corpus Study

Jiménez, Miguel A.
Rutgers University,
USA
miguelji@rci.rutgers.edu

Abstract

Web site localisation, a process that was developed adapting procedures that were already established for software localisation, has grown exponentially during recent years. According to the localisation industry the goal of this process is to produce websites that are received as if "it was originally developed in the target country". Nevertheless, the industry has not yet fully researched which characteristics, conventions or language have been developed and established in each locale. Corporate websites were selected for this study since they are the most conventionalised web genre according to digital genre research, and therefore could show some aspects that have been distinctively conventionalised in the various locales.

Keywords: *localisation of websites, genre, hypertext structure, web site comparative studies.*

I. INTRODUCTION

DURING the last two decades there has been an exponential growth in the field of localisation. This new discipline has opened a new area for translation research (Folaron 2006, pp.195-222), (Pym 2003). Parallel to the expansion of the localisation market, the divide between the localisation industry and Translation Studies has to some extent widened, since the industry established its own business models and processes largely without reliance on knowledge of conventional translation (Quah 2006), (O'Hagan and Ashworth 2003, p.130). Nevertheless, several scholars have helped bridge this gap lately (Dunne 2006), (Bouffard and Caignon 2004, pp.806-23), (Reinke 2005), (Quiron 2003, pp.546-58), (Pym 2003b). This article is part of a wider study on web localisation, one of several localisation processes[1] that have evolved during the last two decades, and its main goal is to study the impact of the localisation process in the final product, the localised text. The theoretical model combines basic Translation Studies concepts such as text, genre and corpora analysis with established localisation terms such as locale in compiling corpora. This article establishes a comparative base for the contrastive study of web textuality anchored on genre theory, and this is applied to an extensive monolocale corpus of Spanish websites.

II. LOCALISATION, TEXTS AND DIGITAL GENRES

A. Localisation

The translation process during web localisation is immersed in a global development cycle that is usually well defined (LISA 2004, p.15), and varies according to several factors, such as the level of localisation. The localisation process is complex and multidisciplinary, and in the midst of technologies in constant flux, it is difficult to establish exactly how to define localisation (Folaron 2006, pp.195-222). A quick review of the literature in the field shows that its technical aspect has been regarded by the industry as the clear divide between "traditional translation" and localisation (Pym 2003a). From a translation perspective, some scholars have indicated that this technical aspect represents just one of the current and future components of the profession (Quirion 2003, pp.546-58). A review of the diverse proposed definitions of localisation can shed some light into the different aspects of this process[2], even when most of them have their origin in the industry and not a translation perspective. The most prevalent aspects in these definitions are: the existence of both a cultural and a linguistic adaptation (Pym 2003a), (LISA 2007), (Esselink, 2000), (O'Hagan and Ashworth

Manuscript was submitted on April 4th, 2007. Miguel A. Jiménez is the coordinator of the MA in Spanish Translating and Interpreting at Rutgers University, The State University of New Jersey. He recently completed his PhD in Translating and Interpreting with a doctoral dissertation on web localization. He also taught Localization at Wake Forest University, USA. The author can be reached at Department of Spanish and Portuguese, Rutgers University. 105 George st., New Brunswick, NJ, 08901. (email: miguelji@rci.rutgers.edu)

1 Other areas include game localisation , small device localisation, software localisation etc. (LISA 2007)

2003), (Quirion 2003, pp.546-58), the existence of a "product" or "digital content" that needs to be localised (LISA 2007), (Dunne 2006), (Yunker 2003), (Depalma 2003, pp.69-77), (Esselink, 2000), the existence of a receiving "locale"[3] (Dunne 2006), (Pym 2003a), (LISA 2007), (Esselink, 2000), (O'Hagan and Ashworth 2003), (Quirion 2003, pp.546-58), and often, the use of the term "translation"[4] is avoided, even when localisation historically evolved from translation. In fact, localisation was only developed once practitioners and localisation industry leaders recognized the need for further technical adaptations in the process. Nowadays, this global process can be referred to as GILT (Globalisation, Internationalisation, Localisation and Translation)[5].

Any research into the web localisation process needs to take into account some factors such as the presence of other interdependent processes, i.e. Globalisation[6] and Internationalisation, and therefore needs to be contextualised in reference to them (Dunne 2006). Globalisation is generally regarded as the processes that enable companies to conduct business globally, and focuses mainly on management issues (LiSA 2004). Internationalisation takes place at the stage of product development and document design with the main objective of making sure that a product can handle multiple languages and cultural conventions without the need to perform important technical changes (LISA 2007). The most interesting aspect of these two interconnected processes is that their absence or presence can clearly influence the translation process itself, since a product that has not been properly internationalised will present additional challenges to translators/localisers. An example would be the lack of context in the segments provided to a translator, or a program that cannot handle dates in different formats.

This particular aspect, the presence or absence of a clearly defined GILT process, can be directly linked to the concept of localisation level, since it is the commission or skopos (Reiss and Vermeer 1984), or the importance of economic and social considerations in the localisation field, that will determine the level of adaptations that will be commissioned. Several classifications of the levels of localisation have been proposed, both for software (Brooks 2000, pp.42-59) and web localisation (Singh and Pereira 2005), (Yunker 2003). From the point of view of the translation process, the levels that these classifications propose can be divided between those that deal only with the translation or the front-end[7], or the localisation of both the front-end and the back-end (Yunker 2003), including changes or adaptations in the actual programming behind it. In the context of the GILT cycle, this is equal to the presence or absence of a quality internationalisation stage, and in second place, to the level of content adaptation that might be required for a website to be received as if it have been originally developed in the target language or "with the look and feel of locally made products" (LISA 2007, p.5)

From the point of view of the translation process itself, the main question to discern is what aspects and practices affect and change the actual translation process that takes place during localisation. Does a badly internationalised software product or website affect the translation process itself? This aspect is of great importance since most localisation providers usually send up to 80% of their translation work to independent freelance translators (LISA 2007), and many scholars have indicated that the lack of a clear context in string localisation is one of the greatest challenges translators might encounter. This has lead to an increased importance of the editing stage, or the linguistic QA in localisation, that is not as important and time consuming as in other translation processes.

B. Digital Genres in Translation Studies

The specific objective of this paper focuses on the text or genre structure of web sites. In Translation Studies, it is accepted that text structure is culturally bound (Neubert and Shreve 1992). In the case of websites, text structure changes normally require reengineering at the internationalisation stage, and therefore, the presence or absence of an effective internationalisation stage might result in the production of websites that might not comply with the norms and conventions in a specific locale. As an example, a translator that would be translating into Spanish the British online instructions for any electronic device would need to delete the specific page or paragraph that deals with adapting the device plug to British specifications (Pérez 2001). In web localisation, the translator is somewhat more limited in

2 Nineteen different proposed definitions of localisation were found in the research stage for this paper. 3 A locale is defined in terms of a language, geographical area and encoding (Esselink, 2000). 4 "Translation" can be found in Bert Esselink (Esselink, 2000) definition of localisation. 5 Keiran Dunne (Dunne 2006) indicates that it should be more precise to reverse the acronym, GILT to "TLIG" to reflect the historical evolution of the industry and its sequential development. 6 LISA (LiSA 2004):"….making all the necessary technical, financial, managerial, personnel, marketing, and other enterprise decisions necessary to facilitate localisation." 7 The front-end of a webpage or software program is what the user actually sees; the back-end is the programming behind it, such as the source code for a webpage.

introducing structural changes in the localised version of a website since it would likely require technical adaptations, such as removing or adding an item to a navigation menu and the corresponding web pages. Our previous research showed clear structural differences in a comparable corpus of original Spanish web sites and localised ones, such as the almost inexistence of "terms and conditions" pages in original Spanish websites (Jiménez 2005). In order to account for these text structure changes, the concept of genre was introduced in web localisation since research has shown that genres might show interlinguistic and intercultural differences at the structure, intratextual, communicative, and sociocultural levels that need to be accounted for (Trosborg 1997, pp.3-23).

The concept of genre has usually been studied in conjunction with text typology and-or register (Trosborg 1997, pp.3-23). It was introduced in Translation Studies from the fields of literary studies and English for Specific Purposes (Swales 1990), (Bhatia 1993). Lately, it has been the center of a great amount of research, with the appearance of research groups exclusively dedicated to its introduction in Translation Studies, such as the GENTT[8] group in the University Jaume I in Spain (Izquierdo and Nebot 2003, pp.83-97). Genres represent communicative acts that express themselves through conventionalised forms of texts, therefore increasing the communicative efficiency in a recurring particular social occasion. Hatim and Mason (Jiménez 2006) defined genre as:

> *conventionalised forms of texts which reflect the functions and goals involved in particular social occasions as well as the purposes of the participants in them.*

This definition combines formal aspects, such as prototypical structure, social and cultural aspects, since genres are determined by a specific culture and social occasion, and cognitive ones, since it represents the purposes and expectations of both the sender and the receiver of the text. In the specific case of web sites, the notion of genre implies that receivers or users interact with websites with a generic mental model of how they are supposed to work and look, accumulated through the prior visit to thousands of other websites over the years (Nielsen and Tahir 2002). Additionally, this generic mental model is usually culturally-bound and determined to some extent by each specific locale or culture. This is one of the reasons why this concept was introduced in Translation Studies; different languages and cultures can potentially show different prototypical structures and different textual and linguistic conventions (Pérez 2001). As mentioned earlier, the localisation industry has as a goal to produce localised versions of its products that are received as if they had been originally developed in the target culture (LISA 2007). Nevertheless, it has not fully researched which structural, textual, and linguistic conventions have been established in each locale to which the localised versions are supposed to comply to. As an example, in our previous web study the corpus of localised web pages showed that the frequency of terms such as "política de privacidad" [privacy policy] was four times higher than in the corpus of corporate web pages originally produced in Spanish (Jimenez, 2005). The explanation was traced back to the fact that this conventionalised term in English corresponds to term behind a communicative block that has not been fully conventionalised in the target culture. The results of this different degree of conventionalisation was also present in the term variability in the localised corpus: "política de privacidad", "cláusula de privacidad" or "declaración de privacidad", "política de confidencialidad". The lack of a highly conventionalised block in a given genre could lead to the translator creating more diverse and creative translations. This could lead to greater linguistic variability both between localised websites and original websites or localised websites in different countries (Bouffard and Caignon 2004, pp.806-23).

Several models of genre characterisation have been developed in Translation Studies. They usually take into consideration several aspects, such as conventions, textual functions, the communicative situation, the social and cultural context and intratextual elements (Pérez 2001). In the specific case of web genres or cybergenres[9], the functionality needs to be added to the characterisation model for these genres, since it has been found to be the main force behind genre evolution and development on the World Wide Web (Shepherd and Watters 1998, pp.97-109), (Crowston and Williams 1999). The functionality was the main aspect that would separate "general" translations from localisation in the earlier stages of the localisation industry (Uren et al. 1993). Genres are in constant evolution (Miller 1984, pp.151-67),

8 GENTT.- Géneros Textuales para la Traducción, [Textual Genres for Translation]. www.gent.uji.es. 9 The genres that developed in the new medium, Internet, have been called "cybergenres", "digital genres" or web genres. Nevertheless, the Internet and the WWW are different communicative situations since the WWW is only one of the many communicative situations that can be studied in Translation Studies, such as chat interpreting, teletranslation etc. (O'Hagan and Ashworth 2003).

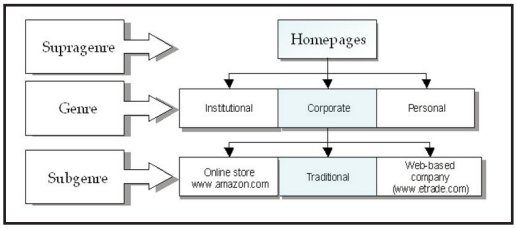
and the evolving functionalities of the new medium has been the reason behind both the appearance of a number of new genres, as well as the adaptation of pre-existing genres to the web, such as printed vs. online newspapers.

The cybergenre model proposed by Shepherd and Watters (Shepherd and Watters 1998, pp.97-109) presents an evolving genre characterisation that starts with genres that already existed in paper and were made available online without any adaptations (extant genres), to genres that appeared in the new medium and are totally independent of those in any other medium (novel-spontaneous genres), such as corporate homepages or blogs. These new novel genres not only show functional aspects that separate them from genres in other mediums, but also different conventions, structure, and a new textual model that defines the kind of language or intratextual elements in them (Crystal 2001). On the other side, extant genres that are simply made available online do not represent special instances of web texts since they conform to the characteristics and language found in other mediums, such as a copy of a contract on a website, and therefore do not need to be studied independently.

Due to the nature of evolving online genres, the current paper focuses on the first cybergenre to establish itself as such, the corporate homepage (Askehave and Nielsen 2004, pp.120-41). This cybergenre also shows a more conventionalised form and language than some others, such as portals or institutional websites (Kennedy and Shepherd 2005). One of our basic hypotheses is that corporate homepages have been established as a genre for many years and might therefore show more variation between different cultures in their textual conventions and prototypical structures. The identification of the prototypical structure in a given genre in Translation Studies has been used to compare it with the structure in different languages (Pérez 2001), or in our case, different locales.

The prototypical lineal structure in specific printed genres[10] has evolved to a mutlilinear structure in hypertexts (Janoshka 2003), and therefore the concept of text or document structure in genre theory had to be adapted to hypertexts. Askehave and Nielsen (Askehave and Nielsen 2004, pp.120-41) have indicated that the different links embedded in homepages represent the different prototypical blocks, stages, moves or sections[11] that make up the document structure. Consequently, the methodology to establish the prototypical structure of the corporate homepage genre was based on identifying all links in all homepages and assigning them to a specific block or section. Each of these blocks and sections might have a specific function that complements the overall genre, and they were established according to the characterisation factors of genres indicated above: conventional aspects, textual function, elements of the communicative situation and functionality.

Two other concepts are used in genre theory in order to further limit and characterise the genre object of study, "supra-genre" and "sub genre". Supragenres engulf a group of genres that share some common characteristics but that do not belong to a specific genre. In our case, the homepages would be a "supragenre" that would include corporate, institutional and personal homepages (Kennedy and Shepherd 2005). Sub genres can be found in a specific genre whenever the topic or the function might slightly change (Biber 2004, p.170). For example, corporate homepages could be subdivided between those that are mainly directed towards on-line sales, www.amazon.com, traditional ones represent the additional communicative platform for bricks and mortar companies, such as www.microsoft.com, or those that solely exist on-



line, such as www.e-trade.com

FIG.1. HOMEPAGE GENRES

The specific genre that was compiled in our comparable corpus was therefore traditional corporate websites that have a bricks and mortar presence and do not exclusively specialise on selling products online.

C. Texts in Localisation

The notion of text has been central in Translation Studies since the first theories and paradigms were developed. Some of them place a special emphasis on

10 With the exception of the so-called "printed hypertexts", such as dictionaries or encyclopedias.
11 All these terms have been used in genre theory to indicate the different "parts" that a text can possibly show and compare its order or existence with the same document in another language.

texts as the foundation of translation theory such as Albert Neubert, or even as the minimum translation unit (Neubert and Shreve 1992). Scholars have stressed the importance of establishing what constitutes a text, since the "original text" has to somehow be represented in a "target text". Questions about coherence and cohesion, intertextuality, situationality (Neubert and Shreve 1992), or the function or functions of a given text are of special interest and have been a recurring topic in Translation Studies (Nord 1997, pp.43-66), (Nord 1996). However, in localisation it has not been clearly defined what "makes" a text and how to define its boundaries. The localisation industry has produced most of the research in this area, therefore placing special emphasis on the technical aspects, and the linguistic concept of text is not usually found in these studies. Instead of the concept of "text", we find "material" (Esselink, 2000), "linguistic part" (LiSA 2004), "content"[12] (Dunne 2006), (Folaron 2006, pp.195-222); (Pastor 2005, pp.187-252), (Depalma 2003, pp.69-77), (O'Hagan and Ashworth 2003), or "information elements" (Lockwood 2000, pp.187-252). The reasons why the notion of "text" has been put aside in localisation research may be due to the lack of clear limits in these types of texts, its multiple authoring (Pym 2003a), the intensive reuse of translation memory that breaks up and stores previous texts or the lack of a clear hypertextual model.

In the case of web content, the "texts" that translators work with are usually hypertexts, and these have evolved from the original "web pages" (Landow 1992) to hyperlinked "web sites" that represent a new textual model on the Internet, an evolving medium. Web pages have become content and storage units (Nielsen 2000), and they are immersed in a specific website that contextualises them and provides the necessary cohesion and coherence[13] to function as such. As an example, a single bogus page imitating an E-bay site would not be considered as a valid and credible text by the receiver since it lacks the necessary coherence and cohesion with a complete real website. This leads us to consider entire websites as the basic textual unit, including all typographic, tables, graphics, videos or multimedia presentations that it might include. Our proposed definition of text in localisation would be "any textual unit that is developed or presented to the receiver as such". Anthony Pym proposed a very similar definition in his study about localisation (Pym 2003a, p.17), and it is rooted in the importance of textual distribution as well as resistance to it: *"a text is quite simply whatever unit is distributed as a unit"*.

From the different hypertext classifications that have been developed so far, Angelika Storrer (Storrer 2002, pp.157-68) presents a classification of hypertexts that in our opinion is essential in order to establish what constitutes a text in the new medium and what kinds of hypertexts represent new textual forms. In first place, Storrer introduces E-texts, those texts with a sequential structure that are usually copies of documents written for another medium, such as thesis, research articles or a newspaper articles. Hypertexts are electronically published texts with a non-linear structure, a recognizable textual function and thematic consistency, they are also open since authors can update them and add more nodes[14]. These hypertexts are interconnected through hyperlinks, and users can access them through activating links on each page or through deep-linking (Nielsen 2000, p.179) or in other words, accessing the hypertext through any of it pages and not necessarily through the start page. Corporate homepages represent a clear example of this new textual model since they are limited, they represent a unit of production and distribution and most hyperlinks are usually internal, that is to say, directed only toward pages inside the same hypertext (Janoshka 2003, p.179). Finally, the hyperweb interconnects all E-texts and hypertexts through hyperlinks; the author mentions that to some extent the WWW as a whole is an interconnected hyperweb. As an example, any Google search could demonstrate to which extent millions of hypertexts or websites are interconnected in this global network.

The objects of the study, corporate web pages, are therefore hypertexts associated with one company that can be accessed through one single domain, such as www.telefonica.com. The proposed model establishes hypertexts (Storrer 2002, pp.157-68) as a new textual model that represents the unit of development, distribution (Pym 2003a), and therefore translation, even when the translation process might be not be carried out by a single localiser in the light of the new GMS or Global Management Systems (LISA 2006).

12 "Content" has been defined as "any digitalised information - that is, text, document, image, video, structured record, script, application code, or metadata - used to convey meaning or exchange value in business interactions or transactions (Depalma 2003, pp.69-77). In our opinion, as in technical documentation, videos, images, visual presentation, typography etc. is part of the global texts, in our case, the global website, and consequently the introduction "content" in order to account for specific textual parts is not needed. 13 Coherence in hypertext research indicates that coherence is its single most important textual aspect.
14 In Hypertext theory web pages can be considered nodes, lexia or hyperdocuments.

III. METHODOLOGY

Corpus linguistics was introduced in Translation Studies to study both the product and the process of translation itself (Baker 1995, pp.223-43), (Kenny 2001), (Laviosa 2002). Additionally, introducing corpus linguistics in localisation is a new development in this area (Shreve 2006, pp.309-331), (Jiménez 2006), (Jiménez 2005), mostly as an answer to the constant reuse of previously translated material and terminology, the golden rule of the localisation industry (Schäler 2002). This reuse of translation memories could lead to limiting the resources available to translators during problem solving tasks in the translation process (Shreve 2006, pp.309-331). Using carefully designed corpora for specific purposes could increase the number of available resources for translators (Shreve 2006, pp.309-331), and could also show the degree of conventionalisation of different terms and textual structures in any specific translation, such as the localisation of corporate homepages.

In order to base this study in solid theoretical principles, it was necessary to establish which specific genre and text represent the object of this empirical investigation. The previous review of genre and text in localisation were developed as an answer to compiling a corpus that would include complete texts (Kenny 2001) that correspond only to one delimited digital genre. For these purposes, a representative monolingual or monolocale corpus of the population object of study was designed and collected. Theoretical considerations in corpus design are of utmost importance (Biber 1998), (Kenny 2001), (Zanettin 2000, pp.105-118), and to date, it is common in web textuality and genre research to collect web corpora without sound theoretical bases, and therefore not following clear principles such as representativity, standarisation and text or genre types.

In our case, the population that needed to be represented is the corporate websites of a specific locale, es-Es, Spanish-Spain. The concept of locale, as opposed to language, was chosen in order to limit and exclude any dialectal or cultural variation in the present study, and also to bridge the gap between translation and the localisation industry. Our corpus is therefore a monolocale, es-ES, corpus that was compiled synchronically in one day, May 6th 2006.

The Google directory **World>Español>Regional>Europa>España>Economia_y_Negocios** was used, and the first corporate website that was originally produced in the chosen locale was selected from each subdirectory. The Google directory was used since it is the most comprehensible directory of Spanish corporate homepages on the WWW and it has been used previously by scholars such as Biber (Biber 2004). One of the most important aspects in website selection was to check that the website was originally produced in Spain and was not the localisation of another website. Due to the different models of website localisation, such as centralised or decentralised (Yunker 2003), (O'Hagan and Ashworth 2003, p.74), some clear guidelines where developed in order to establish the origin of each website. For this purpose the country of origin of the company or the language in which the comments or the variables are written in the source code of the web page were used.

IV. RESULTS

The final compiled corpus includes 172 original Spanish corporate websites from all possible economic areas. The basic characterisation of this text shows that it comprises an average of 161.93 pages per website and 205.9 words in the text body of each page. Nevertheless, the total amount of translatable words per page is 356.50, including all different textual elements in each web page, such as texts included in <meta>, <alt>, <OnMouse>, <input> and <select>, Html tags or text included in Scripts. The

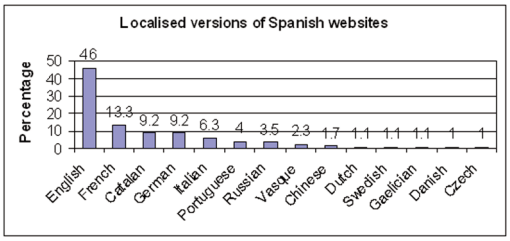
Global corpus statistics: 172 corporate Spanish (es-ES) websites		
Type	**Total**	**Average**
Web pages	27,852	161.93 Pages/site
Words in main body text	5,737,289	205.99 Words/page
Total words	9,929,302	356.50 Words/page
Words in <Meta>	934,361	33.54 Words
Words in <alt>	317,605	11.40 Words
Words in <input> and <select>	1,359,017	48.79 Words
Words in Scripts	242,720	9.71 Words/page
Words in <OnMouse>	1,338,310	48.05 Words/page
Number of links	936,975	33.64 Links/page
Number of Images	852,938	30.62 Images/page

statistics are presented in the following table:

FIG.2. WEB CORPUS STATISTICS WITH WEBBUDGET

Any localised version derived from the original Spanish websites was also included in the corpus, partly since due to a clear localisation structure model: localised versions can be included in the same directory structure of the original website, as a different directory under the same domain, or in a different domain. It was therefore impossible to separate these

localised versions from the original ones for a synchronic compilation in a single day. The influence of English as the de-facto international business language around the world, as well as the lingua franca on the Internet (LISA 2007), was confirmed in the analysis of the present corpus, with 46% of Spanish websites offering an English version of the website, followed by 13.3% offering a French localised version, German and Catalan both with 9.2%, Italian



with 6% and Portuguese with 4.2%.

Fig. 3. Localisation of Spanish websites into other languages

Since this study was conceived as the initial stage of a more in depth research in textual variation between original and localised texts, the main corpus analysis at this stage was centered on obtaining the prototypical superstructure of this genre in a specific locale. Each genre is formed by a series of textual segments, the communicative blocks, and in written genres they are usually organised hierarchically into a linear structure. These communicative blocks are typical of each genre (Swales 1990), and in web hypertexts can be identified with different links that produce a mutlilinear superstructure (Askehave and Nielsen 2004, pp.120-41). These textual blocks make up the different parts of a global text, the hypertext, and each of them conveys a specific function in the global multifunctional text (Swales 1990). As an example, the block "The company" in a corporate website expresses a expositive function since it describes its history, organisation and experience, and at the same time is expresses a exhortative secondary function since it needs to establish a trust and confidence relationship with the potential customer that might lead to a transaction. The concept of genre was introduced in order to observe text structure differences between cultures and locales. In our previous study of navigation menu terminology (Jiménez 2005), the term "Privacy Policy" or "Política de Privacidad" appeared in 42.4% of American websites localised into Spanish, while in our current research on Spanish websites, only 13.6% showed this term. Spanish corporate websites usually include a "legal" block under the term "avisos legales", and this block usually includes any privacy legal provisions. This aspect is indicative of cultural differences in the conventionalisation of genre structure, an important aspect to take into consideration if we need to produce localised websites that are received as if produced in the target locale (LISA 2007).

At the same time, each block might be divided into communicative sections, and they also represent a specific function inside each communicative block (Pérez 2001). The section "location" inside the block "The company" shows the user where the premises of the company are or how to physically get to it through maps or directions.

Once the analysis was performed, the prototypical superstructure of the web genre shows eight possible communicative blocks: Start pages, 100%, contact pages 86.4%, company information pages 54.65%, products and services pages, 54.05% and 44.76% respectively, news pages 54.05%, legal content pages 45.34%, specific user areas 22.09%, and interactivity pages, those pages based on the interaction between the website and the user, such as search pages, registration pages or faqs.

The following prototypical superstructure of Spanish corporate websites includes the total of communicative blocks and its sections. It needs to be mentioned that the percentages are based on the appearance of a block or section as an independent web page in the global website. For example, independent contact pages are present in 86.4% of all websites. Nevertheless, this block is considered "compulsory" in this genre and the remaining 13.6% of websites would show contact information on its start page. These percentages are therefore just indicative of the prototipicallity of different blocks and sections on this genre in a specific locale, and this could be used as a comparative base with other locales (Pérez 2001). The following table shows the level of prototipicallity of each block and section in the Spanish corporate homepage genre:

Identified communicative blocks and sections

1.	**Start page [Página de inicio]**	**100%**
2.	**Contact [Contacto]**	**86.04%**
	2.1.Contact forms [formularios de contacto]	31.42%
3.	**Company information (La empresa)**	**75 %**
	3.1. Location [Localización]	28.48%
	3.2.Company Experience [Experiencia de la empresa]	17.44%
	3.3.Mission [Misión]	10.46
	3.4.Quality [Sistema de calidad]	9.30%
	3.5.History [Historia]	8.13%

	3.6.Premises-Offices [Instalaciones]	7.55%
	3.7.Logistics [Logística]	6.49%
	3.8.Projects [Proyectos]	4.65 %
	3.9.Image Galeries [Galerías de Imágenes]	4.65%
	3.10.Research [I+D]	4.06%
	3.11.Divisions [Divisiones]	1.16%
	3.12.Exports[Exportación]	0.58%
4.	**News- Current events [Noticias]**	**54.06%**
5.	**Product- Services [Productos -Servicios-Soluciones]**	
	5.1.Products [Productos]	53.48%
	5.2.Services [Servicios]	44.76%
	5.3.Offers-Promotions [Ofertas-promociones]	9.88%
	5.4.Technical info [Infor. técnica]	5.81%
6.	**Legal information [Legal]**	**45.34%**
	6.1.Legal notes [Aviso-Nota legal]	27.90%
	6.2.PrivacyPolicy[Declaración de privacidad]	13.37%
	6.3.Terms and Cond. [Condiciones generales]	4.65%
7.	**Client areas [Zonas de Clientes]**	**22.09%**
	7.1.Jobs [Trabajo]	19.76%
	7.2.Advice [Consejos]	9.30%
	7.3.Education [Formación]	5.81%
	7.4.Prices [Tarifas]	5.81%
	7.5.Orders [Pedidos]	5.23%
	7.6.Publications [Publicaciones]	4.65%
	7.7.Professionals [Profesionales]	2.90%
	7.8.Budgets [Presupuestos]	2.90%
	7.9.Investors [Inversores]	2.90%
	7.10.Franchises [Franquicia]	2.32%
	7.11.Presents [Regalos]	0.58%
	7.12.Financing [Financiación]	0.58%
8.	**Website Interactivity [Interactividad con sitio web]**	
	8.1.Search [Buscar]	14.53%
	8.2.Questions or FAQs [Preguntas o ayuda]	2.20%
	8.3.Links [Enlaces]	11.62%
	8.4.Registration [Regístrese]	9.30%
	8.5.Glossary [Glosario]	0.74%

FIG. 4. LEVEL OF PROTOTIPICALLITY OF THE DIFFERENT BLOCKS AND SECTION IN THE TRADITIONAL CORPORATE WEBSITE

This structure shows the level of prototipicallity of the basic blocks, as well as the degree of conventionalisation of the possible sections that could be included. Our hypothesis is that to some extent the degree of conventionalisation of different elements in different languages and cultures will vary (Singh and Pereira 2005). The sections found in the communicative block "Company" are indicative of the degree of conventionalisation of different information sections this block might include. In order of importance; location, experience, mission, quality and history are the basic sections included in this block. This model of hypertext description based on genre is established as a comparative base for textual variation between different locales or between "original" website and "localised" ones. Corpus linguistics research in Translation Studies has shown that translated texts are less creative (Kenny 2001), that translators use fewer words in their work (Laviosa 2002) or that translations are usually longer than the original texts. The next step in this project will be to compare this structure and the terminology used for each block and section with websites originally developed in different locales, mainly English and French. In the above mentioned Spanish language localisation analysis they showed as the most important locales in Spain.

Additionally, and due to some basic characteristics of hypertext structures, this block and section structure could be used in order to construct representative subcorpora, such a "contact" subcorpora or "Legal notes" subcorpora for further research into the conventionalised structure of each block, its terminology or phraseology. Subcorpora could also be compiled for the "invisible" recurring textual elements that need to be localised, such as <Meta>, <Select>, <OnMouse>, <Input> or <Select>, that are usually repeated throughout the website.

A. "Contact us": a compulsory block in the corporate homepage genre.

The Internet opened a new platform for information distribution that lead to a new model of communication, the so-called Interactive Mass Communication Model (Janoshka 2003). In this new model the flow of information does not only flow from the sender to the receiver; the communicative acts can be established between the medium and the sender or the receiver through the interaction with the website, such as filling a form or receiving a "wrong password" message, and between the receiver and the sender, through interactive forms, help chats forums etc. The Internet has multiplied and sped this interaction (Crystal 2001), and therefore receivers expect to contact the sender with the immediacy that the new medium allows. In this respect, the evolution of this digital genre has meant that 31.42% of Spanish websites include a contact form, a percentage that will probably increase in this evolving genre.

The main function of this block is expositive[16], since users will get to this block with the clear intention of obtaining contact information. At the same time it

16 The contextual focus or functions presented by Hatim and Mason (Jiménez 2006) are used in this article.

had an exhortative secondary function since it needs to foster possible interaction with the sender. In the case of forms, its primary function is exhortative since it encourages the user to "act", whether by contacting the company or filling the form.

The use of corpora as an aid for terminology extraction has been the object of several studies (Faber et al. 2005, pp.167-197). In these studies the organisation of knowledge structures or ontologies establishes a base for terminology extraction in a specific domain. In our study, among other possible uses, this prototypical structure constitutes the base for terminology extraction. In the case of a corporate English website, Nielsen and Tahir (Nielsen and Tahir 2002) indicate that 89.9% of North American websites use "Contact us" in their web pages, a very high degree of conventionalisation. In the case of Spanish websites, "Contacto", with 49% of use, appears to be the most conventionalised terms, followed by "contactar". The possible terms found in this block in the corpus are:

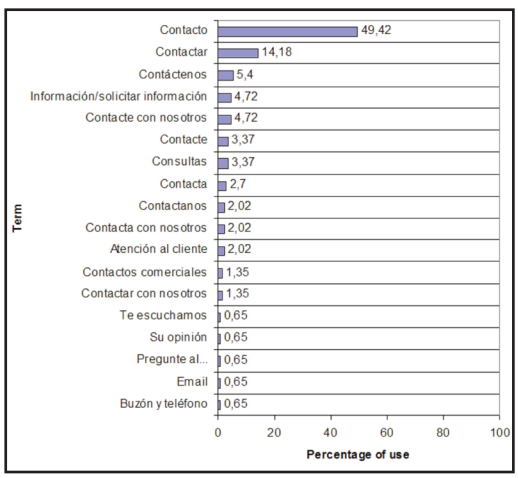


FIG. 5. PERCENTAGE OF TERM USE IN THE "CONTACT" BLOCK IN SPANISH

In our previous study, localised pages from English into Spanish showed a clear preference for "Contáctenos", which shows the influence of the original term, Contact us, followed by several other options such as "contacte con nosotros", "contacto", "contacta" (Jiménez 2005). This shows a clear influence of the original English text in the Spanish localised texts, and therefore implies the need for comparative studies in this area. The table shows the level of conventionalisation of different terms used in Spanish websites to indicate a link to the "contact" block in this genre. All of them are valid since for any convention to exist, there has to be a possible alternative (Lewis 1969), otherwise, a convention could not exist. This genre model can show the most conventionalised options for any recurring term in corporate websites. This type of statistical analysis of terminology extracted from a clearly defined corpus could help translators by providing additional options that go beyond the constraints of TM or terminology bases (Shreve 2006, pp.309-331). It could also help justify translators' decisions beyond the single option of pre-translated segments or terminology banks.

There are other important aspects that could be extracted from this table, such as the digital tenor or the level or formality used to address the receiver. This is of great importance in languages with formal and informal grammatical markers. Language in the Internet has been usually described as showing a lower level of formality than that used in other mediums (Crystal 2001). Some authors have called this tendency, "conceptual orality" (Janoshka 2003), written texts that resemble oral ones. Most languages deal differently with formal/informal markers in websites: when localising a website into Spanish the use of "tú" or "usted" is an important decision to take since that marker is inexistent in English websites. In this case, the infinitive form of the verb is the most used, 42%, followed by "usted", 40% and "tú", 18%. These results are comparable to the analysis of another section that uses mostly verbal forms, "Register", but surprisingly, one section "Jobs", showed that "tú" was the most used form, with 70% of use against 20% of use of "usted". This finding may point to different levels of formality inside a specific genre, since in this case the section "jobs" needs to break the power relation between the company and customer: the receiver in this case could possibly be part of the company in the future, and therefore the website adresses the receiver as "tú". This also points to the validity of our analysis since this prototypical genre structure could be of great use for translators and localisers.

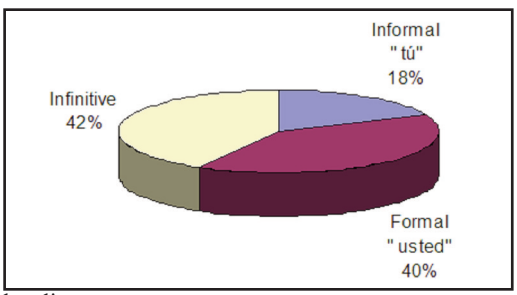


FIG.6. LEVEL OF FORMALITY IN BLOCK "CONTACT"

IN SPANISH.

V. CONCLUSION

Localised websites are supposed to look like those produced in the receiving locale (LISA 2007), but to date the localisation industry and Translation Studies have not produced a clear comparative model to study what specific textual, terminological, discourse or structural aspects have been established in each major locale. Some efforts are currently under way in Canada, especially in the Quebec region (Bouffard and Caignon 2004, pp.806-23), (McDonough 2006, pp.7-14), centered mainly in the cultural aspect of localisation. Our study focuses in the textual aspects of web localisation and applies one of the main principles of Translation Studies: translated texts can show language and textual structures that are different from texts originally produced in the target language. In this context, we have applied genre theory in order to develop a comparative base that can be used to clearly compare the structure and language of websites. Furthermore, this base can be used in order to isolate blocks of texts, such as "legal notes" in websites, and compare them to the same block in a specific genre. For example, the "privacy" terms of a corporate website might be totally different to those of an institutional website, and these differences need to be taken into account. Furthermore, digital genres show intercultural or interlocale differences in the macro and micro textual levels, and a genre-based model is ideal for these comparative studies.

Cybergenre studies normally include each of the genre blocks described as a separate genre, such as FAQ pages or Flash presentations (Paolillo et al. 2007). Nevertheless, from a translation standpoint, FAQs or privacy term pages in different genres will include different textual conventions, and more importantly, these blocks might not be conventionalised to the same extent in the receiving locale. The translator therefore will not have a clear established textual reference to produce a translation that reads as if it has been originally produced in the target locale.

The next stage in this study, which is already under way, is centered in compiling a parallel corpus of localised North American corporate websites in order to study the differences between localised vs. original produced websites in Spanish. This study can show the extent to which localised websites have departed from the established conventions in the Spanish-Spain locale. Language variation between Spanish locales is also an area of special interest in this field since American companies such as American Express are trying to find an "International" Spanish for their websites, an objective that might be elusive in the ever changing world of Internet texts.

ACKNOWLEDGMENT

I would like to thank Dr. Maribel Tercedor Sánchez from the School of Tranlating and Interpreting, University of Granada, Spain, for her committed direction of my doctoral dissertation. Her encouragment, support and advice has been a vital contribution to the development of my research.

REFERENCES

I. Askehave, I. and A. E. Nielsen, "Digital genres: a challenge to traditional genre theory", Information Technology and People, vol. 18 (2), 2005, 120-141.

M. Baker, "Corpora in Translation Studies: An Overview and some Suggestions for Future research". Target, vol. 7 (2), 1995, 223-243.

V. K. Bhatia, Analysing genre. Language use in professional settings. London: Longman, 1993.

D. Biber, 2004. "Towards a typology of web registers: A multi-dimensional analysis". Paper presented at Corpus Linguistics: Perspectives for the Future, October, 2005, University of Heidelberg, Available at http://jan.ucc.nau.edu/~biber/Web%20text%20types.ppt).

D. Biber, Variations across speech and writing. Cambridge: Cambridge University Press, 1988.

P. Bouffard, P. and P. Caignon, "Localisation et variation linguistique. Vers une geolinguistique de l´espace virtuel francophone", Meta, vol. 51 (4), dec. 2006, 806-823.

D. Brooks, "What Price Globalisation? Managing Costs at Microsoft", Translating into Success. Cutting-edge strategies for going multilingual in a global age, in R. C. Sprung, Ed., Amsterdam-Philadelphia: John Benjamings, 2000, 42-59.

Crowston, K. and M. Williams. "The effect of Linking on Genres on Web Documents". Actas del la XXXIII Annual Hawaii International Conference onSystem Sciences, Kilea, Hawaii. Los Alamitos, CA: IEEE-Computer Society, January, 1999.

D. Crystal, Language and the Internet. Cambridge: Cambridge University Press, 2001.

D. DePalma, "Rage against the content management machine." In Proccedeings of the LRC 2003: The 8th Annual Localisation Conference and Industry Showcase, Localisation Research Centre: Limerick, Ireland, 2003, 69-77.

K. Dunne, Perspectives on Localisation, Amsterdam-Philadelphia, John Benjamins. 2006.

B. Esselink. A Practical Guide to Localisation. Amsterdam - Philadelphia: John Benjamins, 2000.

P. Faber, C. I. López and M. Tercedor, "Utilización de técnicas de corpus en la representación del conocimiento médico". Terminology, vol. 7 (2), 2005, 167-197.

D. Folaron, "A discipline coming of age in the digital age", in Perspectives on localisation, K. Dunne, Ed., Amsterdam-Philadelphia: John Benjamins, 2006, 195-222.

S. Gamero Pérez. La traducción de textos técnicos, Barcelona: Ariel, 2001.

I. García Izquierdo and E. Monzó Nebot, "Una enciclopedia para traductores. Los géneros de especialidad como herramienta privilegiada del traductor profesional". In R. Muñoz Martín (ed.). Actas del I Congreso Internacional de la Asociación Ibérica de Estudios de Traducción, Granada, Asociación Ibérica de Estudios de Traducción e Interpretación, 2003, 83-97.

A. Janoschka, Web Advertising. Amsterdam-Philadelphia: John Benjamins,

2003.

M. A. Jiménez, "La localizacion de hipertextos: el género y la tipología textual en los sitios web corporativos". Pre-doctoral dissertation [Trabajo de Investigación Tutelada]. Department of Translating and Interpreting, Unviersity of Granada, Spain, 2006.

M. A. Jiménez, "Las peculiaridades textuales de las páginas web localizadas al español". In Proceedings of the 46th Anual Conference of the American Translator Association, Seattle, EEUU, 2005, ed. Marian Greenfield, 2005, 275-286.

A. Kennedy and M. Shepherd, "Automatic Identification of Home Pages on the Web", in Proceedings 38th Hawaii International Conference on System Sciences. Los Alamitos, CA: IEEE Press, 2005.

D. Kenny, Lexis and Creativity in Translation. A corpus-based study. Manchester: St. Jerome, 2001.

G. Landow, Hypertext: The convergence of contemporary Critical Theory and Technology. Baltimore: The John Hopkins University Press, 1992.

S. Laviosa, Corpus-based Translation Studies. Amsterdam: Rodopi, 2002.

LISA, Localisation Industry Primer, 3rd edition.. Lommel, A., ed., Geneva, the Localisation Industry Standards Association (LISA), 2007.

LISA. Localiztion Industry Primer, 2nd Edition. A. Lommel, A. ed., Geneva: The Localisation Industry Standards Association (LISA). 2004.

LISA, LISA Best Practice Guide: Managing Global Content. Globan Content Management and Global Translation Management Systems, 2nd edition. A. Toon, A. Draheim, A. Lommel y P. Cadieux, Eds., 2006.

K. D. Lewis, Convention. A Philosophical Study. Cambridge, MA: Harvard University Press, 1969.

R. Lockwood, "Machine Translation and Controlled Authoring at Carterpillar", Translating into Success. Cutting-edge strategies for going multilingual in a global age, in R. C. Sprung, Ed. Amsterdam-Philadelphia: John Benjamings, 2000, 187-202.

M. Mata Pastor, M., "Localización y traducción de contenido web". In Reineke, D. (ed), Traducción y Localización. La Palmas de Gran Canaria: Anroart Ediciones, 2005, 187-252.

J. McDonough, "Beavers, Maple Leaves and Maple Trees. A study of National symbols on Localised and Domestic Websites". Localisation Focus, vol. 5, (3), 2006, 7-14.

C. R. Miller, "Genre as Social Action". Quarterly Journal of Speech. vol. 70, 1984, 151-67.

J. Nielsen, Designing Web Usability: the practice of simplicity. Indianapolis: News Riders, 2000.

J. M. Nielsen and M. Tahir. Homepage usability: 50 Websited deconstructed. Indianapolis: News Riders, 2002.
A. Neubert and M. Shreve. Translation as Text. Kent, Ohio: Kent State University Press, 1992.

C. Nord, "A functional typology of translations". In A. Trosborg Ed., Text Typology and Translation. Amsterdam- Philadelphia: John Benjamins, 1997, 43-66.

C. Nord, Translating as a Purposeful Activity. Functionalist Approaches Explained. Manchester: St. Jerome. 1996.

M. O´Hagan and D. Ashworth, Translation-Mediated Communication in a digital World: facing the challenges of Globalisation and Localisation. Clevendon, England: Multilingual Matters, 2003.

J. C. Paolillo, J. Warren and B. Kunz, "Social network and genre emergence in amateur flash multimedia", in Proceedings 40th Hawaii International Conference on System Sciences. Los Alamitos, CA: IEEE Press, 2007.

A. Pym, The Moving Text. Amsterdam-Philadelphia: John Benjamins, 2003.

A. Pym, "What localisation models can learn from Translation Theory". The LISA Newsletter. Globalisation Insider, vol. 12, (2/4), 2003.

C. K Quah, Translation and Technology. Hampshire, Inglaterra: Palgrave Mcmillan, 2006.

M. Quirion, "La formation en localisation à l'université : pour quoi faire?". Meta, vol. 48 (4), 2003, 546-558.

D. Reinke, Traducción y Localización. La Palmas de Gran Canaria: Anroart Ediciones, 2005.

K. Reiss and J. Vermeer. Grundlegung einer Allgemeinen Translationshtheorie. Tubinga: Niemeyer, 1984.

R. Schäler, R., "The Irish Model in Localisation". Conference Presentation at LISA Forum Cairo 2005: Perspectives from the Middle East and Africa. [Online], 2005, Available: http://www.lisa.org/utils/getfile.html?id=61136686
R. Schäler, "The Cultural Dimension in Software Localisation". Localisation Focus, vol. 1 (2), 2002.

Shepherd, M. y C. Watters, 1998. "The evolution of cybergenres". In Sprague R. (ed) Proceddings of the XXXI Hawaii International Conference on System Sciences. Los Alamitos, CA: IEEE-Computer Society, 97-109.

G. M. Shreve, "Corpus Enhancement and localisation". In Dunne, K. (ed.), Perspectives on Localisation. Amsterdam-Philadelphia, John Benjamins, 2006, 309-331.

N. Singh and A. Pereira. The culturally customized Web site: customizing web sites for the global marketplace. Oxford: Elsevier, 2005.

A. Storrer, "Coherence in text and Hypertext". Document Design, vol. 3 (2), 2002, 157-168.

J. M. Swales, Genre Analysis. English in Academic and Research Settings. Cambridge: Cambridge University Press, 1990.
M. Tercedor Sanchez, "Aspectos Culturales en la localización de productos multimedia". Quaderns. Revista de Traducció, vol. 12, 2005, 51-160.

A. Trosborg, "Text Typology: Register, Genre and Text Type". In A. Trosborg, ed., Text Typology and Translation. Amsterdam-Philadelphia: John Benjamins, 1997, 3-23.
E. Uren, R. Howard and T. Perinotti, Software Internationalisation and Localisation: An Introduction. New York: Van Nostrand-Reinhold, 1993.

J. Yunker, Beyond Borders: Web Globalisation Strategies. Indianápolis, Indiana: New Riders, 2003.

F. Zanettin, "Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis", in Maeve Olohan (ed.) Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects. Manchester: St. Jerome, 2000, 105-118.

SimShip software testing using Shadow™

K Arthur, D Hannan, M Ward
Brandt
6 Faughart Terrace,
St. Mary's Road,
Dundalk, Co. Louth.
karthur@brandttechnologies.com, mward@brandttechnologies.com, dhannan@brandttechnologies.com
www.brandttechnologies.com

Abstract

We believe that our approach to automated software testing is novel. We can test several language instances of a product simultaneously, either through direct engineer interaction or by a record/playback script. The Shadow™ application can manage a situation where the user interface of the product under test is slightly different in layout, either due to localisation of the different versions, or due to the original language version running on different platforms. In our pilot studies, we examine the effect of separating out the functions of a test engineer into a product specialist and QA specialist. Our testing methodology outputs a set of screenshots of the products under test in each language. The screenshots can be used by a translator for linguistic/consistency QA or in product documentation. We performed a comparative analysis of the automation tool Winrunner with the Shadow™ testing process.

Keywords: *Automated testing, Quality Assurance testing, QA., Localisation, Localisation Testing*

1. Introduction

The purpose of this paper is to describe a new automated testing tool called "Shadow™" developed by Brandt and to illustrate how it works with the aid of case studies. Shadow™ is a software application that allows the user to control one or more operating systems (PCs, VMWare instances) simultaneously. The idea of one computer controlling another is not new and there are many products available for this purpose such as VNC (RealVNC Ltd. 2002). However, the idea of one computer controlling many computers simultaneously appears to be novel.

SimShip is the process of shipping the localised product to customers at the same time as the original language product, or within a couple of weeks. This can result in increased market share (Common Sense Advisory 2004) and possibly increased revenues. However, SimShip can also result in unrecoverable costs due to factors such as those described by Langewis (2003): localised software not taking off in foreign markets, localised software being substandard, or localised software having been created inefficiently.

In this article we propose to examine software testing, explain what Shadow™ is and how it is relevant to testing and, finally, discuss some case studies. In the first section we will introduce Shadow™ and describe in general terms what the product does. We will give an overview of the Shadow™ functionality and describe how it might be used to test original products and localised products.

2. Software testing and software quality

In this section we will examine software testing and software quality in general terms. Towards the end of this section we will introduce Shadow™.

2.1 What is quality?

In this section we are interested in the definitions of quality as applied to software development. The Crosby (1980) definition of quality defines quality as "conformance to requirements". An alternative definition states that quality is "fitness for use" (Juran 1951). The value of these context free definitions is that they can be used in domains other than software (Mass 2004). For our purposes, software quality will be defined as software that conforms to customer driven requirements and design specifications. Categories under which quality can be assessed include (King 1996): functionality, reliability, usability, efficiency, maintainability, and portability.

Software testing should be viewed as a scientific process, wherein an application is placed under known conditions of setup, hardware and configuration, where it accepts some known input and where it should result in an expected outcome. We expect that the quality of the test process and effort will have an impact on the quality of the software. Unexpected performance, errors or bugs are introduced into software in a large variety of ways during the development and localisation process. Software "testing involves operation of a system or application under controlled conditions and evaluating the results. The controlled conditions should include both normal and abnormal conditions. Testing should intentionally attempt to make things go wrong..." (Hower 1996). Bugs are typically of the following types:

- Logical errors - where the software runs, but produces unexpected results.
- Crash bugs - where the software fails in a catastrophic manner.
- Failed functionality - where the application fails to meet its specifications.
- Layout issues - widgets/text not displayed correctly.
- Linguistic issues - spelling, grammar.
- Localisation issues - date/time format, number format.

2.2 Software testing

Software testing is performed to find defects, to ensure that the code matches the specification and to estimate the reliability of the code. The output of testing is a defect list and the effectiveness of the testing process can be measured through the number of defects (Voas 2004). As the number of defects is reduced, there is an enhanced confidence in the quality of the product. We believe testing should be viewed as an integral part of the software development process, rather than an activity performed after the core development. Internationalisation testing targets unexpected performance in a software product when used with different character encoding schemes; it also ensures that a product is localisable. Internationalisation testing should be performed early in a product's lifecycle to identify and remedy any issues.

Testing adds to the cost of production. When balanced against the idea that the most expensive defect to fix is the one found by the customer, it is highly desirable to capture defects as early as possible in the software development lifecycle. In 2002, a US government agency suggested that software bugs cost $59.5 billion annually (Tassey 2002).

There are two approaches to software testing, namely manual testing and automated testing. Manual testing uses human engineers to aid in this process. Automated testing requires that the test script is coded, and then executed using some application. Successful software development and localisation should use both manual and automated testing methodologies. The balance between each can be decided using budgetary and schedule constraints. Shadow™ can be used to complement manual testing or it can be used as a test automation solution.

2.3 What is Shadow™?

Shadow™ is a software-testing tool for performing automated and manual tests on original or localised software products. Shadow™ allows the user to record and play back scripts that run and control multiple machines at the same time. It also allows the user to directly interact with multiple machines simultaneously, making the manual test effort more efficient. Shadow™ allows the user to simultaneously test localised software applications running:

- Different language operating systems, or
- Original language products running in different configurations.

It is in this context that Shadow ™ is ideal for use in a project where the localised and original language products must ship together, that is "SimShip".

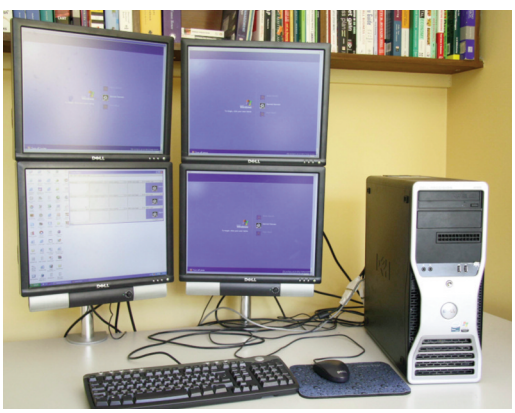


FIGURE 1: SHADOW™ SETUP

Our development philosophy has been that we avoid operating system dependencies, where possible. We want to avoid getting information about the application under test that is not available on all operating systems in the same manner. In fact, Shadow™ con-

centrates on getting information about the application under test that is available to the human user.

Shadow™ uses the keyboard, mouse and timing information from the user, and employs intelligent technology to identify the location of the interactions in the screen. During the replay of scripts Shadow™ finds the appropriate location in the screens under test. This allows for widgets to have changed position and size between tests. We are currently adding character recognition to the Shadow™ suite of tools. This has particular relevance to localisation, where testing is not just a matter of finding the appropriate text, but where a mapping exists from one language to another. Testing of the localised products can take place at the same time as original language product testing. The key to SimShip is having the localisation and development cycles in parallel, now with Shadow™ the test cycles of original and localised products can also take place in parallel.

The specific problems we address with Shadow™ are:

- Making automated software testing easier to use with less programming.
- Separating the roles of Test Engineer into the complementary roles of "Product Specialist" and "Quality Assurance Specialist". See the glossary of terms for proposed specifications of the two roles.
- Making software testing more like the actions of a human user.
- Making automated testing easier - reducing the barrier to automation.
    - o Reduction of the cumbersome nature of script creation (Dustin 1999).
    - o Reduction of the training time necessary for the tool to be useful.
- Increasing compatibility with 3rd party application components (widgets), see (Dustin 1999).
- Accelerating the manual testing process through Shadow™'s unique user interface.
- Recording screenshot data by default. In other automated testing applications this has to be implemented programmatically.
- Reducing script maintenance. This is accomplished by examining the developer's release documentation to identify those user interface areas that have changed in the new build. We can then compare screenshots from the newly run tests with previously recorded screenshots in those areas that have been changed.

3. Shadow™

Automated testing tools currently come in three general categories with different operating modes, namely (Skrivanek 2005):

- Simple capture - record and playback.
- Object Oriented Automation - API calls "under the hood".
- Image-based component discovery.

Simple capture utilities: Using a "simple" capture utility an engineer can create a test script containing an exact sequence of keystrokes, mouse movements and/or mouse commands. However, with the slightest change to the GUI of the software under test the recorded script becomes invalid requiring the engineer to possibly re-record the script.

Object Oriented Automation: Makes API calls to the operating system to identify information about the control being interacted with. Once the automation tool has the handle for the object, it can then manipulate the control. It is the reaction of the code to a function that makes an API call that is being tested, not the actual user interface. This is an important distinction.

Image-based component discovery: In this mode of automated testing, images are taken of regions around the mouse at the time of some mouse interaction. This image is then stored. At a later time the image is used to identify where the mouse action should be taken either during playback of a script or where the action should be taken on some other system. This is the method used by Brandt in the Shadow™ application. The characteristics of image based component discovery are similar to those of Object Oriented automation, without the need for any of the "under the hood" information.

3.1 Shadow™ setup

Shadow™ is a piece of software that can control several machines simultaneously. It can do so in several different ways, such as:

- Shadow™ can make a group of machines perform exactly the same actions at the same time.
- Shadow™ can make a group of machines perform "nearly" the same action at the same time.
- Shadow™ can record and playback scripts ("Mimic" and "Exact Match" modes).
    - o "Mimic" mode is where Shadow ™runs as a "simple capture utility".
    - o "Exact match" mode is where Shadow runs

as an "image based component discovery" application.

4. Case studies

In this section we will look at case studies involving the use of Shadow™. We will examine three case studies where Shadow™ was used, and the purpose for which it was used.

Client	Profile
A	Multinational software publisher
B	Supplier of technical authoring, documentation and localisation services
C	Brandt

Table 1: List of case studies

4.1 Client A

Client A produces enterprise resource planning (ERP) software for managing and analysing corporate spend. Their clients include "Fortune 100" multinationals. Brandt provides translation and engineering services to this client.

4.1.1 Task specification

Client A requires that the translated and localised user interface of its products undergo linguistic testing and functional testing. For this test case, the client used a combination of methods to perform the linguistic and functional testing, that is using Shadow™ and WinRunner. The engineer performed the following tasks:

- Wrote test scripts.
- Updated test scripts.
- Set up the hardware and software.
- Executed the test script on the machines, using both Shadow™ and WinRunner.
- LQA performed by linguists using the screenshots.
- Localisation functional QA using Shadow™ and WinRunner.

4.1.2 Shadow™ usage

In this section we examine how the output of Shadow™ is used in linguistic testing and how Shadow™ itself is used in functional testing.

Linguistic testing

Software is translated in a tool that does not show the translator the context of the strings. No matter how familiar the translator is with the product, and how much experience they have with previous versions, there is no substitute for having the translator seeing the translated strings appearing in a running build in their proper context. The translator can then make the appropriate changes, if necessary, to the software strings. There are at least two methodologies that can be used in this situation such as having the translator go through the running build with a test script to bring up every screen, or providing the translator with the screens in the form of screenshots (or MHT files in this case study). We use the second method in this case study. Some advantages of giving the translator only screens to review include that the translators do not have to have spend time negotiating their way through the product to find the areas in which an update occurs and there are cost savings in setup time. This is especially useful for small and frequent updates to a product.

Functional testing

The engineer used Shadow™ with three machines for functional testing, each machine running a different language operating system. The output of this was a list of bugs documented by the engineer. In this process the engineer manually went through the test script on one machine, with the others automatically following in Shadow™, noting bugs as they progressed. As with the linguistic testing, the process was then repeated with a different language set.

4.1.3 Results

Table 2 and Table 3 below show the time spent using both Shadow™ and WinRunner to perform the same tasks for a QA cycle. The engineer has detailed the time taken for the average amount of days spent on each task. The tables show the complete list of tasks.

The engineer notes that WinRunner can fail if the Internet connection is slow and fails to bring up a widget in a "reasonable" time. A key element of Shadow™ functionality is "Wait for feature". This means that Shadow™ waits a configurable amount of time for a feature to appear on the screen before proceeding or declaring a failure.

4.1.4 Conclusions

From the tables above we can see that Shadow™ and WinRunner take approximately the same time to setup and run a test cycle in which there are a small number of screenshots required. Where a larger number of screenshots are required, the time taken to run Shadow™ is less than the time taken to write the code and execute it in WinRunner. The engineer notes in his report that WinRunner requires the build of software under test to be specially prepared in

40 screenshots	*Shadow™*	*WinRunner*	*Total*	
Task	*Days*	*Days*	*Days*	*Comment*
Write LQA script			3 - 4	Tool independent
Update LQA script			1 - 2	Tool independent
Write TSL script		1 - 2		WinRunner only
Execution using tool for screenshots	2	1		Both Shadow™ and WinRunner
LQA by translators			1 - 2	Tool independent
Functional QA			1 - 2	Tool independent
Total days	8 - 12	8 - 13		

Table 2: Results of using Shadow™ Vs WinRunner for 40 screenshots

400 screenshots	*Shadow™*	*WinRunner*	*Total*	
Task	*Days*	*Days*	*Days*	*Comment*
Write LQA script			20 - 25	
Update LQA script			10 - 15	
Write TSL script		25 - 30		WinRunner only
Execution using tool for screenshots	16	8		
LQA by translators			5 - 6	
Functional QA			9 - 10	
Total days	60 - 72	77 - 94		

Table 3: Results of using Shadow™ Vs WinRunner for 400 screenshots

order to function with it, whereas Shadow™ does not. Each resource on the page has to have an AWL name (AWL is a client proprietary web language). The AWL name identifies the widget without reference to the text on it, so that the same button in French and German will have the same AWL name, but different text. This is an "under the hood" requirement of WinRunner that Shadow™ does not have. In summary, Shadow™ was used as a QA tool in this study. Shadow™ was more efficient than WinRunner on the QA of a larger product. Shadow™ did not require special preparation of the product build before its use and as a result it could be used "out of the box".

4.2 Client B

Client B is a provider of documentation, consultancy and recruitment solutions.

4.2.1 Task specification

Client B wanted Brandt to prepare a document containing screenshots of their software that could be sent to translators for review. There were 203 screenshots to be taken of the software opened in a software engineering tool. This tool is a visual localisation environment allowing engineers to view the localisable resources of software. The final deliverable to Client B was one PDF per language, each containing the 203 screenshots of the English and localised software side by side.

4.2.2 Shadow™ usage

Shadow™ was installed on 5 VMWare machines each running Windows XP Professional. The appropriate tools were also installed. Each localised software file was opened in the visual editor on one virtual machine, and then the English, giving a total of 5 connected clients to Shadow™. Usually, Shadow™ takes a screenshot of the whole screen, or if configured to do so, just the window in focus. In this project, the window in focus was the visual editor application and the required dialog was part of a screenshot that would need to be cropped. The Shadow™ code was modified, using an extension of the existing "exact match" technology so that a screenshot could be adaptively cropped to leave only a feature of interest. This project is unique among the test cases, as it showed how the Brandt software development team was integral to completing the project using Shadow™ by making appropriate modifications to the application.

4.2.3 Results
The engineer took about 1.5 hours in total to setup and screen shoot 4 languages. The integration of the screenshots into the final document MS Word document and production of the PDF are tasks independent of Shadow™ and would be the same length irrespective of how the screenshots were taken.

4.2.4 Conclusions
Shadow™ was used as a QA tool for this project, where the output was a set of screenshots for linguistic QA performed by the translation team. This project involved use of a software application that requires significant processing resources of the host operating system. We found that the process using Shadow™ was faster than the manual process, directly as a result of the ability to perform tasks in parallel.

4.3 Brandt
Brandt uses Shadow™ for the purposes of testing and as an automation tool to perform tasks that need to be repeated frequently. Brandt has found that there are tasks in the production of multimedia tours in Adobe Captivate® that are repetitious and prone to human error, such as audio integration, text integration, and font assignment. We have written short scripts, called macros, which can be activated using different keystrokes that perform these tasks.

4.3.1 Task specification
Adobe Captivate® is a tool for rapid authoring of multimedia tutorials. These tutorials contain screenshots, text, animations and audio that all have to be localised. The localised elements have to be integrated into each slide in the tour. Tours can vary in size from 30 slides to in excess of 100 slides, and are localised in a number of European and Asian languages. Brandt has run this project several times over the past eighteen months and improved the process to make it more efficient, with the result that the output is more consistent and of a higher quality.

Audio integration
This involves importing a single WAV file per slide. The WAV file name is numbered in sequence, for example "0001.wav". To perform this task manually, the engineer runs the risk of importing the wrong file into a slide. For an engineer who is not familiar with the language they will not be able to test that the audio is appropriate to the slide. This has to be repeated up to 100 times without error, for up to eight languages. That is a total of 800 cut-and-pastes.

Text Integration
There is localised text on every slide in the tour. On most slides, the text is unique. The engineer has a MS Word document with the text ordered sequentially for every slide. Once again, this is a straightforward process, but it is prone to error. This has to be repeated up to 100 times without error, for up to eight languages. That is a total of 800 cut-and-pastes. One possible error is that the text sequence goes out of line with the slide sequence. Once this happens, the process will have to be repeated from the place where the error occurred.

Font assignment
The localised text appearing on every slide has to be a certain font for European languages, the SimSun font for Simplified Chinese and MS Mincho for Japanese. Each slide has to have the font individually set. The engineer must go to every slide, bring up the text properties and set the correct font. Again, this has to be repeated for each slide and then for each language.

4.3.2 Shadow™ usage
Shadow™ is used in the following way. Three virtual operating systems are set up in VMWare, in our case MS Windows XP Professional. Shadow™ is set up to view the 3 operating systems, generally referred to as (virtual) machines. Each machine runs Adobe Captivate®, MS Word and possibly MS Notepad. Each virtual machine is set up with the correct script (French, German or Japanese as required). The macro is set to run the correct number of times, usually equal to the number of slides, and it is then started. Each run of Shadow™ can process one tour in three languages in parallel. While this is running, the engineer can usually go and perform another task.

4.3.3 Results

Shadow™ was vital for this project, as some of these tasks are repetitive and subject to human error. Table 4 shows the times taken for Shadow™ to run the individual tasks versus the manual time for the same task. Note that Shadow™ can run at least three tours in parallel so there is a further efficiency present.

Task	Automation per 30 slide tour - minutes	Manual time per 30 slide tour - minutes
Audio integration	10	25
Text integration	15	25
Font assignment	10	20

TABLE 4: RESULTS FOR SHADOW™ AUTOMATION VS MANUAL IMPLEMENTATION

Whereas one Shadow™ run will perform audio integration (or another task) for three tours in parallel, one engineer performs the manual equivalent process in a serial fashion.

4.3.4 Conclusions

Shadow™ was used as an automation tool for this project and it was essential to the effectiveness of the engineering team. For some tasks, the time it takes to perform the task manually is not much different from the time the automation takes. In this case the automation is more efficient when it is possible to perform the tasks in parallel. It might not be worth investing time in automating some once off tasks. For this project, Shadow™ was invaluable, because of the number of tasks that are repeated.

5 Conclusions

In this paper we have examined some of the advantages and disadvantages of the different modes of testing software applications. We have come to the conclusion that a mix of manual and automated testing is essential to the success of a project. One type of testing will not replace the other. The ratio of the mix between automated and manual testing can be dictated by the nature of the project, the budget and the schedule. Shadow™ can help make automated testing more efficient by separating the QA from specialist product knowledge and hardware setup. As part of the localisation process, Shadow™ can be used to take screenshots of the running software that can be given to linguists for review. Shadow™ can also be used by the engineer, with specialist product knowledge, to walk through the different language versions of a product at the same time.

References

RealVNC Ltd. (2002) RealVNC remote control software [online], available at: http://www.realvnc.com/ [accessed 28 August 2007].

Common Sense Advisory (2004) Common Sense Advisory explains how companies will become "world enterprises" [online], available at: http://commonsenseadvisory.com/news/pr_view.php?pre_id=8 [accessed 05 September 2007].

Langewis, C. (2003) Localization and ROI: Increasing Value by Eliminating Pink Ink [online], available at: http://www.ableinnovations.com/pdf/pinkink-I.pdf [accessed 05 September 2007].

King, M. (1996) The ISO 9126 Standard [online], available at: http://www.issco.unige.ch/ewg95/node1.html [accessed 30 August 2007].

Hower, R. (1996) Software QA and Testing Resource Center [online], available at: Http://www.softwareqatest.com/qatfaq1.html#FAQ1_1 [accessed 28 August 2007].

Voas, J. (2004) A Few Assertions on Information Hiding [online], available at: http://www.cigital.com/papers/download/qualitytime1.pdf [accessed 28 August 2007].

Juran, J. (1951) "Quality control handbook", New York: McGraw Hill.

Maas, K.F. (2004) Introduction to Quality [online], available at: http://www.kfmaas.de/qintroed.html [accessed 28 August 2007].

Dustin, E. 1999, Lessons in Test Automation [online], available at: http://www.stickyminds.com/sitewide.asp?ObjectId=1802&ObjectType=ART&Function=edetail [accessed 06 September 2007].

Skrivanek, J. (2005) Testing GUI Applications [online], available at: http://wiki.java.net/bin/view/Javapedia/TestingGUIApplications [accessed 06 September 2007].

Computational Morphological Analysers and Machine-Readable Lexicons for South African Bantu Languages

Sonja Bosch, Jackie Jones, Laurette Pretorius, Winston Anderson
University of South Africa
PO Box 392, UNISA, 0003, South Africa
boschse@unisa.ac.za, jackiej@stthomas.co.za, pretol@unisa.ac.za, winston.anderson@btgroup.co.za

Abstract

In this paper the development of computational morphological analysers for six South African Bantu languages is discussed. Due to the rich agglutinating morphological structures of these languages, the morphological processing poses particular challenges. These challenges are of an orthographical, a morphological as well as of a lexical nature. The current status of the project is reported on, firstly in terms of the development of prototypes of morphological analysers for the various languages, and secondly in terms of the development of standardised XML machine-readable lexicons for the South African Bantu languages, based on an appropriate general data model.

1. Introduction

It is well known that localisation initiatives are supported by language translation, and in particular machine assisted and machine translation. However, there is much more to machine translation than meets the eye, especially in highly agglutinating languages such as the Bantu languages. The importance of morphological analysis is recognised as a basic enabling application for further kinds of natural language processing (NLP), including part-of-speech tagging, parsing, semantic analysis and information retrieval, and also for high-level applications such as spelling checking, lexicography, language teaching, text-to-speech systems, question answering and last but not least machine translation.

In order to be of practical use, such analysis needs to be automated and be based on underlying machine-readable lexicons that conform to common lexical specifications and de facto international standards to ensure their compatibility at international and multilingual level. The morphological analyser is regarded as the first in a series of text processing components.

Human language technologies (HLT) and NLP enable the electronic handling of both spoken and written language, and are aimed (amongst other things) at improving communication between humans and machines, as well as communication among humans. This is especially important in a country such as South Africa with its eleven official languages, which need to be developed technologically so that automated services can be rendered to citizens in their language of choice.

The development of computational morphological analysers for South African Bantu languages is linked to a project funded by the National Research Foundation in South Africa. The main research question in the project concerns the development of finite-state morphological analysers for five Bantu languages, namely Zulu, Xhosa Swati and Ndebele (belonging to the Nguni group of languages), and Northern Sotho and Tswana (belonging to the Sotho group of languages).

2. Challenges posed by Morphological Analysis of Bantu Languages

Automated morphological analysers exist for many European languages, but the development of morphological analysers has only been reported for a few Bantu languages, such as Swahili (Hurskainen 1992) and a few others in southern Africa (for example, Bosch & Pretorius 2003). Due to the rich agglutinating structures of these languages, the morphological processing poses particular challenges. These challenges are of an orthographical, a morphological as well as of a lexical nature.

In the case of the Nguni languages, a conjunctive system of writing is adhered to with a one-to-one correlation between orthographic words and linguistic words. For example, the Zulu orthographic word siyakuthanda (si-ya-ku-thand-a) 'we like it' is also a

[1] An earlier version of this paper was published in the Proceedings on the Workshop on Networking the Development of Language Resources for African Languages, 5th International Conference on Language Resources and Evaluation, 22 May 2006, Genoa, Italy, pp 38-43.

linguistic word. The Sotho languages on the other hand, are disjunctively written, and the above mentioned single Zulu orthographic word is written as four orthographic words in Northern Sotho, namely re a go rata (re a go rat-a) 'we like it'. These four orthographic entities constitute one linguistic word. It should be noted that, in contrast, the English orthographic words 'we like it' are three independent words that each have their own meaning and can stand alone.

The **orthographical challenge**, which lies in the writing conventions of the Bantu languages, may according to Hurskainen and Halme (2001, p.399), be ascribed to the fact that disjunctive writing systems "require a special treatment, before they can be analysed successfully". Pre-processing of the text in order to identify linguistic words, before morphological analysis takes place, is one of the options of addressing this challenge.

The morphological challenges in computational morphological analysis are twofold and comprise the modelling of two general linguistic components, namely morphotactics (word formation rules) as well as morphophonological alternations:

- The **morphotactics component** includes word formation rules, which determine the construction of words or word forms from an inventory of morphemes. This inventory of morphemes consists of word roots and affixes. Morphemes that make up words cannot combine at random, but are restricted to certain combinations and orders. A morphological analyser is required to recognise valid combinations of morphemes of the language in question.

- The morphophonological alternations component deals with the morphophonological changes between lexical and surface levels. A morphological analyser should identify the correct form of each morpheme since one and the same morpheme may feature in different ways depending on the environment in which it occurs.

The main **lexical challenge** in the building of morphological analysers for the Bantu languages is the fact that machine-readable lexicons, which are fundamental resources, are not readily available in any form. Although online dictionaries for Bantu languages are reported on by de Schryver (2003), such dictionaries available for Zulu and Xhosa for instance, contain a maximum of 2000 to 3000 lemmas and do not include explicit linguistic information, which is essential for a word root dictionary of the analyser. In the case of Northern Sotho, a bilingual online dictionary SeDiPro 1.0 (de Schryver 2003, p.10) containing over 20,000 entries is available with linguistic information. However, such online dictionaries are only accessible for look-up of individual words or word stems, and are not accessible as a whole.

3. Meeting the Challenges

3.1 Orthographical Challenges

The Sotho languages pose a pre-processing challenge in that the disjunctive orthographical tradition isolates as separate "words" what are essentially affix morphemes of a lexical unit. Thus in order to correctly analyse multi word lexical units morphologically without causing excessive ambiguity, a multi word tokeniser is required. For Northern Sotho, this was addressed by first constructing regular expressions to deal with all verb constructions and they were then extended to address all predicate constructions. These cater for the most complex multi word tokens in the Northern Sotho language.

The grammars historically cover the verbs and copulatives reasonably adequately but other research theses and more modern study grammars (for example, Louwrens 1989) had to be consulted to get consolidated views of these rules. None of the sources adequately covered what Ziervogel and Mokgokong (1985) term "deficient verbs", but in other texts are referred to as auxiliary verbs. A new linguistic research project is now under way to examine these in more detail.

There are various computational alternatives to producing a tokeniser (Hurskainen & Halme 2001). The Northern Sotho morphological analyser team chose the approach of using finite state software to construct the tokeniser (Beesley 2004). The tokeniser for Northern Sotho now adequately deals with all predicate clauses (verbs, auxiliary verbs and copulatives). For more information see Anderson and Kotze (2006).

3.2 Morphological Challenges

Since human language technology is a novel field of research in South Africa, especially in the field of Bantu languages, a team approach was decided on for the morphological analysis project. Each language team consists of a computer scientist and one or two linguists.

Morphological analysis in this project is based on a finite-state computational approach, using the natural language independent Xerox Finite-State Tools (Beesley & Karttunen 2003). This integrated set of tools is used to model and implement the complexities of word-formation rules as well as morphophonological alternations by means of finite-state networks. The latter are subsequently combined algorithmically into larger networks that perform morphological analysis.

The Xerox tools provide a declarative programming language, lexc (Lexicon Compiler) for specifying the required natural language lexicon and for modelling the morphotactic structure of the words in the language concerned.

Alternation rules are subsequently needed to map the abstract lexical strings into properly spelled surface strings, as they occur in the natural language. The alternation rules are formulated as regular expressions, and are then compiled into a finite-state network by means of the Xerox tool xfst.

In practical terms this means that all morphemes in the natural language need to be arranged in a cascade of LEXICONs (in a lexc description), while each entry in a LEXICON consists of morphological information and either a continuation class (the name of the next LEXICON in the cascade) or the end symbol #, which indicates the end of a valid morpheme sequence, as shown below in the example of a lexc description:

```
...
LEXICON NounPrefixes
...
i[NPrePre7]si[BPre7]:^I^SI
NStem;
i[NprePre8]zi[Bpre8]:^I^ZI
NStem;
...
LEXICON  NStem
...
gubhu                              NClass7-
8;
...
LEXICON NClass7-8
@U.CL.7-8@                          NomSuf;
...
LEXICON NomSuf
ana[DimSuf]:ana
#;
...
```

The **lexc** source file is then compiled into a finite-state network. This network recognises morphotactically well-formed, but still abstract morphophonemic or lexical strings such as

```
i[NPrePre7]si[BPre7]gubhu[NRoot]ana[Dim].
```

Alternation rules are subsequently needed to map these abstract lexical strings into properly spelled surface strings, as they occur in the natural language. The alternation rules are formulated as regular expressions, and are then compiled into a finite-state network by means of the Xerox tool **xfst.**

The orthographic changes that manifest between the lexical and surface words when morphemes are combined to form new words or word forms are described as illustrated in the following example:

b h [o|u] -> j || _ a n a

This alternation rule models the change of a bilabial sound -bh- appearing in the final syllable of a noun stem such as -gubhu to a palatal sound -j- when the diminutive suffix -ana is added to the noun stem.

The final step in the development of the morphological analyser is the combination of the **lexc** and **xfst** finite-state networks by means of composition (cf. Beesley & Karttunen 2003) into a single network, a so-called lexical transducer. This transducer constitutes the morphological analyser and represents all the morphological information about the language being analysed.

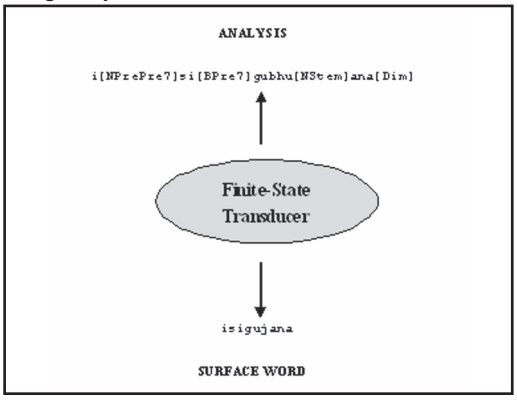


FIGURE 1 GIVES A SCHEMATIC REPRESENTATION OF THE APPLICATION OF A MORPHOLOGICAL ANALYSER.

In Figure 1 the morphological analyser maps the Zulu morphemes i-, -si-, gubhu and -ana to isigujana 'little calabash'. In other words, if the surface word isigujana constitutes the input string to the finite-state transducer, the output string is the morphological analysis which consists of the following morphemes in combination with their morphological feature tags: i[NPrePre7]si[BPre7]gubhu[NStem]ana[Dim].

The arrow in Figure 1 indicates the bidirectionality of the transducer and shows that analysis takes place in the upward direction while generation takes place in the downward direction. For more details regarding the Zulu morphological analyser prototype (ZulMorph) see Bosch and Pretorius (2003) as well as Pretorius and Bosch (2003a) and (2003b).

3.3 Lexical Challenges

In addressing this problem of unavailability particularly for Zulu and Xhosa, a lemma list in electronic format was extracted from a Zulu paper dictionary (Doke & Vilakazi, 1964). For Xhosa however, the resources available in terms of lemmas were even more limited. This therefore demanded a time consuming exercise of extraction of lemmas from existing Xhosa paper dictionaries. Lemmas were retyped from a number of dictionaries and the scanning and proof reading of these resources increased and contributed to the development of the lemma lists substantially. These various sources yielded data in largely varying formats and forms containing many inaccuracies and errors. The non-existence, but urgent need for lemma lists for Xhosa also created the opportunity for researchers to devise a practical compilation procedure in accordance with appropriate standards in order to ensure reusability. The procedure for producing a large and reliable collection of Xhosa nouns and verb stems from this data consisted of a semi automated data validation phase and, in the case of nouns, an automated generation phase. Data inconsistencies were identified by means of Perl style pattern recognition, then scrutinised and corrected by the linguists in the team in the data validation phase. The validated data formed the input to the automated generation phase. Nouns were generated in two formats. The first of these was for human readability and the second was in an XML document. The second of these is particularly important in the creation of reusable lexical resources for future applications. The only available Swati paper dictionary is being scanned and proofread in stages and then included into an electronic lemma list. Similarly for Tswana a paper dictionary has been scanned and is in the process of being proofread also to be developed into a lemma list for use in morphological analysis.

Regarding word lists for Northern Sotho, the major dictionaries were examined. The largest, the Comprehensive Northern Sotho dictionary (Ziervogel & Mokgokong 1985) includes support for the extra vowels (beyond the five standard vowels) marked with a circumflex in Northern Sotho, as well as support for the letter š and its capitalised form. Furthermore, the comprehensive dictionary includes tone markings on each main entry. In order to obtain an accurately scanned word list these characters needed to be recognised by optical character recognition (OCR) software. No standard Northern Sotho OCR packages are available, so standard language settings were used. The scanning errors are consistent with the incorrectly scanned characters. Therefore, Perl scripts were developed to automatically correct the incorrectly optically recognised text. A further process of human editing is now underway to confirm all corrections. Subsequent to the dictionary scan, many other works have been scanned to add to the Northern Sotho test corpora. Eastern European language recognisers, such as Czechoslovakian, have proved most effective in recognising characters due to their adequate handling of š and its capitalised form (cf. ABBYY FineReader 2007).

In terms of the lexical challenges our ultimate aim is to develop from these above-mentioned word lists and paper dictionaries machine-readable lexicons according to a standardised data model in XML format that would be applicable to all the languages under investigation.

4. Current Status of the Project

The current status of the project is reported on, firstly in terms of the development of prototypes of morphological analysers for the various languages, and secondly in terms of the development of machine-readable lexicons for the South African Bantu languages, based on the above-mentioned proposed data model.

4.1 Analyser Prototypes for the various Languages

The Zulu analyser prototype (ZulMorph) at present covers most of the morphotactics and morphophonological alternations required for the automated analysis/generation of nouns of all classes, the positive and negative forms of verbs (including object concords, tense morphemes, aspectual prefixes and verbal extensions), pronouns, the demonstrative and copulative demonstrative, underived adverbs, relatives and adjectives, possessives, conjunctions and ideophones. Word categories that still need to be completed are compound tenses of the verb, and derived adverbs. Preliminary testing of the current prototype was done on a test corpus consisting of 30,000 types. The application of the morphological analyser to the test corpus results in the recognition of approximately 77% of the types in the corpus. This result may be ascribed partially to morphological constructions that have not been dealt with completely, but mainly to

roots that do not yet occur in the root lexicon. By design the morphological analyser includes LEXICONs (lists) of noun stems, verb roots, relative stems etc. and therefore it only analyses words based on roots or stems that feature in this so-called underlying lexicon.

Since individual words in wordlists are analysed in the morphological analysis component, the ambiguity rate is high, and each word is assigned all its possible readings or analyses. For an application such as a second generation spelling checker that has some form of automatic morphological analysis implemented as a part of the spelling checker (without grammar checking), such ambiguity poses no problems. The reason is that the emphasis is on lexical recall or the recognition of correctly spelled words by the spelling checker irrespective of their context (cf. Bosch & Eiselen 2005).

In the case of running text being analysed, the challenge in the subsequent phase of the project is the elimination of ambiguity or contextually inappropriate readings such as the following:

```
bakhe   ba[PossConc2]khe[PronStem]
bakhe   ba[PossConc14]khe[PronStem]
bakhe   ba[SC2]akh[VRoot]e[VerbTermPerf]
bakhe   bu[SC14]akh[VRoot]e[VerbTermPerf]
```

These analyses of bakhe illustrate ambiguity not only regarding class concord information, i.e. classes 2 and 14, but also stem and root information. The first two examples are analysed as possessive pronouns while the latter two are analysed as verbs.

The research aims for the other Nguni languages in the project, i.e. Xhosa, Swati and Ndebele closely follow those for Zulu, since all these languages follow a conjunctive writing system. This enables the fast-tracking of the development of the morphological analysers for the Nguni languages by adapting the Zulu continuation classes and rules. Implementation and testing of the model in terms of the Xerox finite state tools are already in progress.

Regarding Northern Sotho a framework is under development for the nominal and verbal structures of Northern Sotho, with special emphasis on establishing the order of verbal extensions, reduplication patterns in nouns and verbs, as well as formalising rules for the derivation of morphological processes that involve the phonological process palatalisation in the formation of passives and diminutives. Implementation and testing of the Northern Sotho prototype (NsoMorph) is based on a limited, though representative lexicon, while cleaning up a scanned version of a Northern Sotho dictionary is in progress. The first prototype of a morphological analyser for Tswana (TsnMorph) is being developed with nouns being treated first, while other word categories are added systematically.

Progress with the development of analyser prototypes for the various Bantu languages in the project has been reported in a number of publications (cf. University of South Africa 2007).

4.2 Development of Machine-Readable Lexicons

By definition the analyser can only recognise and analyse words of which the roots/stems have been explicitly included in its embedded lexicon. Ideally, a comprehensive machine-readable lexicon in the form of an XML document should be available for each language as a basic resource from which word roots/stems may be obtained.

As stated previously, electronic lemma lists for Xhosa and Zulu have been developed albeit on a small scale. To date lemma lists for these two languages, extracted from paper dictionaries, contain a total of over approximately 28,000 entries each.

In order to eventually arrive at an XML lexicon structure, the underlying standardised data model needs to be formulated and verified first. This is the subject of recent work in this regard (Bosch et al. 2006) where a data model towards a standardised machine-readable lexicon for all languages in the project is developed and formulated as an XML DTD. This model aims to ensure maximum inclusiveness of all linguistic information and to provide flexibility and handle the various representations applicable to Bantu languages in particular. It is therefore applicable to diverse uses of electronic lexicons ranging from research in numerous areas resulting in publication. Included in this data model are particular requirements for complete and appropriate representation of linguistic information as identified in the study of available paper dictionaries. As starting point the extent to which the Bell and Bird (2000) data model may be applied to and modified for the above-mentioned languages was investigated. It was found that changes to this data model were necessary to make provision for the specific requirements of lexical entries in the relevant languages.

Our model differs in various ways from the Bell and

Bird model. The latter model was originally designed for descriptive purposes while our model is primarily for computational use, where the emphasis is on marking up lexicon and linguistic information for logical structure in order to provide essential information for the computational language processing task concisely, precisely and unambiguously. Examples are the representation of class information, singular and plural, locative formation (derivation) in the case of nouns, and verbal extensions (derivations) in the case of verbs. Further examples are the identification of specific socio-linguistic features in Xhosa and Zulu such as isiHlonipho sabafazi (married women's language of respect) and Xhosa isiKhwetha (male initiates' language) both features of which would also necessitate explicit representation in the lexicon. Another area where the Bell and Bird model seemed inadequate to accommodate the South African Bantu languages was the exclusion of the appropriate nesting of derived forms so prevalent within these languages. Other aspects of interest include our stem entry approach, the reflexive form of the verb, and the desirability or not of recursion in machine-readable lexicons.

The proposed data model seems to provide flexibility and handles the various representations applicable to Bantu languages in particular and is therefore applicable to diverse uses of machine-readable lexicons. Our hope is that it will contribute to the further discussion and development of a common scheme for storing lexical data not only for the South African Bantu languages, but for the Bantu language family as a whole.

5. Conclusion and Future Work

Morphological analysis is generally recognised as a technology that enables the development of more advanced tools and practical applications in various areas of natural language processing, such as part-of-speech tagging, syntactic parsing, text-to-speech systems, information extraction, and machine translation. Research in this project concerning the development of computational morphological analysers for South African Bantu languages has confirmed the importance of comprehensive machine-readable lexicons as fundamental resource of the morphological analysers. The current project in computational morphological analysis includes research into the development of automated morphological analysers for Zulu (ZulMorph), Xhosa (XhoMorph), Swati (SswMorph), Ndebele (NblMorph), Northern Sotho (NsoMorph) and Tswana (TsnMorph), using finite-state methods in computational morphology. It is envisaged that the project will eventually cover all South African Bantu languages.

Further aims of the project are:

- Wider distribution of the intermediate versions of the electronic lexicon for constructive feedback from the broader community of lexicographers and other users/speakers of the relevant languages.
- Investigation into a disambiguation component, the task of which is to eliminate contextually inappropriate readings.
- Research into lexicon design and development in order to contribute to the international definition of standards as envisaged by the International Standards Organisation ISO/TC37/SC4, whose goal it is to develop a platform for the design and implementation of linguistic resource formats and processes in order to facilitate the exchange of information between language processing modules (Romary and Ide 2002).
- Research into place names as occurring in the various languages of the project, for inclusion in the relevant machine-readable XML lexicons.

6. Acknowledgements

This material is based upon work supported by the National Research Foundation under grant number 2053403. Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Research Foundation.

7. References

ABBYY FineReader. (2007). [O]. Available at: http://www.abbyy.com/finereader_ocr/ [Accessed on 31 May 2007].

Anderson, W.N. & Kotze, P.M. (2006). Finite State tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho. In Proceedings of the 5th International Language Resources and Evaluation Conference, Genoa, Italy.

Beesley, K.R. & Karttunen, L. (2003). Finite-state morphology. Stanford, CA: CSLI Publications.

Beesley, K.R. (2004). Tokenizing Transducers. Xerox Research Centre. Europe. Unpublished course notes presented in Pretoria, South Africa, September 2004.

Bell, J. & Bird, S. (2000). A Preliminary Study of the Structure of Lexicon Entries. [O] Available at: http://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html [Accessed on 19 September 2005].

Bosch, Sonja E & Roald Eiselen. (2005). The effectiveness of mor-

phological rules for an isiZulu spelling checker. South African Journal of African Languages 25(1) pp. 25-36.

Bosch S.E. & Pretorius, L. (2003). Building a computational morphological analyser/generator for Zulu using the Xerox finite-state tools. In Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the Association for Computational Linguistics, April 13-14 2003, Budapest, Hungary. ACL. pp. 27-34.

Bosch, S.E. & Pretorius, L. (2004). Software tools for morphological tagging of Zulu corpora and lexicon development. In Proceedings of the 4th International Language Resources and Evaluation Conference, Lisbon: ARTIPOL , vi, pp. 1251-1254.

Bosch. S.E., Pretorius, L. & Jones, J. (2006). Towards machine-readable lexicons for South African Bantu languages. In Proceedings of the 5th International Language Resources and Evaluation Conference, Genoa, Italy.

De Schryver, G-M. (2003). Online Dictionaries on the Internet: An Overview for the African languages. Lexikos, 13, pp. 1-20.

Doke, C.M. & Vilakazi, B. (1964). Zulu-English Dictionary. Johannesburg: Witwatersrand University Press.

Hurskainen, A. (1992). A two-level formalism for the analysis of Bantu morphology: an application to Swahili. Nordic Journal of African Studies, 1(1), pp. 87-122.

Hurskainen, A. & Halme, R. (2001). Mapping between Disjoining and Conjoining Writing Systems in Bantu Languages: Implementation on Kwanyama. Nordic Journal of African Studies, 10(3), pp. 399-414.

Louwrens, L.J. (1989). Northern Sotho. Study guide for Grammar. University of South Africa: Pretoria.

Pretorius, L. & Bosch, S. (2002). Finite-State Computational Morphology - Treatment of the Zulu Noun. South African Computer Journal, 28, pp. 30-38.

Pretorius, L. & Bosch, S. (2003a). Finite-State Computational Morphology: An Analyzer Prototype for Zulu. Machine Translation, 18, pp. 195-216.

Pretorius, L. & Bosch, S. (2003b). Towards technologically enabling the indigenous languages of South Africa: the central role of computational morphology. Interactions of the Association for Computing Machinery, 10(2), pp.56-63.

Romary, L. & Ide, N. (2002). Standards for Language Resources. In Proceedings of the 3th International Language Resources and Evaluation Conference, 1, pp. 59-65.

University of South Africa. Department of African Languages. (2007). [O]. Available at: http://www.unisa.ac.za/africanlanguages [Accessed on 29 May 2007].

Ziervogel, D & Mokgokong, P.C. (1985). Comprehensive Northern Sotho dictionary. Second corrected edition J.L. van Schaik: Pretoria.

What's in a 'Game'?

Miguel Bernal Merino
Roehampton University
London, UK
m.bernal@roehampton.ac.uk

Abstract

This article highlights the growing use of video games in modern society and the level of penetration in today's entertainment habits. The demand for entertainment software has prompted game publishers to translate more of their products into more languages. However, the nature of multimedia interactive entertainment software products seems to require a particular kind of translation. The development of new professional practice calls for new research within translation studies and a new area of specialisation. The present article explains the many different textual types that translators might find when working for the multimedia interactive entertainment software industry, and how different video games may require a variety of skills from translators, such as being a proficient TMT user, having good research skills, and being inventive.

Keywords: *video game, localisation, translation, game localisation, localization, game localization, entertainment software, multimedia interactive software*

"What's in a game? That which we call 'original'
by any other language would play as sweet"
Shakespeare, had he been a video game translator

1. Introduction:

Entertainment software products have become such a worldwide phenomenon that many public organisms, companies, and artists are starting to explore the market to see how they can utilise the interest generated by these products, most commonly known as video games. An even higher level of interest is shown by the entertainment industry, which is looking at expanding their successful franchises into the interactive software sector. Needless to say, game developers and publishers are blooming almost everywhere.

Most games are normally developed in either English or Japanese, but they are often translated into other languages, the main ones being French, Italian, German, and Spanish (abbreviated as FIGS). The translation process has to deal with the same problems that most other areas of translation do, but video games combine a variety of characteristics that make their translation rather unique and, at times, troublesome. Unfortunately, the linguistic and cultural aspect of this customisation does not seem to be a relevant issue for the industry, most companies outsource this part of the process and they very rarely mention localisation in their international conferences. However, I would like to highlight its importance for the game industry, as well as for gamers, and translation studies, because I think that everybody would benefit from further research into this complex process. The following article will try to explain, in a simple manner, the main characteristics of the translation of video games.

2. The spread of video gaming.

It is difficult to know why, but video games have spread rapidly and become one of the leisure activities of choice of children and adults around the world. The interactive factor probably plays an important part in it since it empowers its audience, who become players, agents in control of the creation. We could argue endlessly about the bad influence of violent games and the potential of multimedia interactive software for education, but there is no doubt that is this format is becoming very popular, and it is here to stay. Evidence of this is, for example, the dozens of game conferences taking place yearly around the world, and that in October 2006 BAFTA (British Academy of Film and Television Arts) opened a dedicated section to video games with eighteen categories.

Video games combine characteristics of other arts, and disciplines like film studies, literature, and com-

puter science in one audiovisual interactive product. Perhaps video games are, in a way, the epitome of 21st century pop culture entertainment where people can find the theme they like, adjust it to their skills, save their progress, and achieve their goal at their own pace. We have seen interactivity before, where the actions of the receiver influence his/her particular experience of the product, for example in game books, like the series started by Ian Livingston and Steve Jackson, in interactive theatre like 'La fura dels bous', and in TV programs like 'Big Brother'. Entertainment industries have been working together for many years, but it is only now that we can start to appreciate the success of their joint ventures. For people outside of the game sector, it is still difficult to realise the size of the far-reaching game market. The following examples will help readers comprehend the extent of video game penetration in today's world:

On the cinematic side:
- Many cinema blockbusters nowadays have a video game adaptation of the film, for example: The Godfather, The Incredibles, Torrente.
- Likewise, there are games that are transformed into movies, such as Tomb Raider, Silent Hill, or Dead Or Alive.

On the Children and Young adults' Literature side:
- Popular books are turned into video games such as The Lord of the Rings, The Chronicles of Narnia, or Harry Potter.
- Popular comic books are also made into video games such as Superman, Astérix, X-Men.

On the television side:
- Sportspeople give their names to a game franchise, for example: Tiger Woods, Colin McRae
- Singers contribute with their music and media persona to the success of some video games, for example, 50 Cent: Bulletproof, Kiss: Psycho Circus, Britney's Dance Beat.
- There are TV series and game shows that inspired video games, such as: The Simpsons, Buffy the Vampire Slayer, and Pop Idol.

On the more serious side:
- The American Ministry of Defence is using a game to attract people into the armed forces called America's Army.
- Some companies and public organisms use simulators (life-like video games) in the first stages of their staff training, programs like Flight Simulator, Rail Simulator, Ship Simulator.
- There are some government initiatives to help fight crime using video games, like CrimeStoppers' Gameover4knives.
- The BBC and some other channels are using video game-type applications to educate youngsters, for example: BBC Bitesize.

In the planning stages of a new project, game designers have a more or less clear idea of the kind of game they want to make, and so they create a video game with a target audience in mind, normally that of their country, which is the one they know best. Worlds, storylines, characters, and features are generated to make the perfect game. Online games have evolved into not only a very social gaming experience, but also into a surprisingly lucrative business, as Castronova (2005) explains in his Synthetic Worlds: The business and Culture of Online Games. However, due to the high cost of the development, and the nature of today's global market, it is almost mandatory to release the game in as many territories as possible. This calls for a full internationalisation of the product, i.e., the game has to be able to incorporate in its design all the changes importing countries might require. This customisation (normally referred to as 'localisation') will secure ROI (return on investment) and further profits in the highly competitive interactive entertainment software market. It comprehends, among other things, the adjustment to different hardware requirements (PAL, NTSC), multi-network configuration (for international online playability), legal framework, cultures and languages.

However while the spread of video games over international markets is self-evident, as well as being a multimillion pound market, the translation of multimedia interactive entertainment software has not been implemented into translation studies. I write about these issues in my forthcoming article "Training translators for the video game industry" (Bernal-Merino 2007)

3. The translation of video games in context

As the international demand for video games rises, successful titles depend on their adaptations for various cultures in a slightly different way to that which other audiovisual creations have, up until now, needed. These products tap into a very emotional activity within society: "play". It is "play" that first bonds us to our own culture and history, to what we see as normal, fun, appropriate, or funny. Video games, unlike any other entertainment product, aim at motivating

and challenging players at their own level and pace. They do this by various means, for example, a customisable avatar, an adjustable difficulty level, and relative freedom of movement and interaction within the virtual world. The country and language of destination may also affect the game itself (Bernal 2006), especially when dealing with violence, historical events, bad language, or sex, since different cultures are more sensitive than others to these matters. But there is also what Sutton-Smith (1997:99) calls 'counterludic identity', which says that sometimes the country importing the game refuses to play them the way the exporting ones do, putting more emphasis on their own way of playing. As a result, the same game released simultaneously in the US, France, Germany, China, and Japan, might highlight different features to adjust to fans' expectations, as well as the cultural and legal framework.

When games are more story orientated rather than action-driven, making them culturally acceptable for different locales can be challenging because of the premises the designers and the story are taking for granted. Asian gamers seem to prefer more child-like characters, while western countries might emphasize adult features, think of the difference between Zelda (Nintendo 1986-2006) and Lara Croft (Core 1996-2006). An example of the changes that are likely to happen during localisation is Fatal Frame (Tecmo 2001). In the original Japanese version the female protagonist, Miku, was 17 years old, in the American and European version she was 19, had western features, and was not wearing the original Japanese school uniform. A similar thing happens with depiction of blood or historical events, everything has to be readjusted to fit the country's tolerance and taste so as not to hurt sensibilities. This is probably one of the reasons why so many games take place in imaginary worlds. This customisation effort will draw on the knowledge of geopolitical strategists, like Tom Edwards from Englobe. He explained during the 2006 Game Developers Conference in California the importance of being culturally aware when internationalising games with a very clear presentation called "Fun vs. Offensive: Balancing the 'Cultural Edge' of Content for Global Games". Both developers and publishers want to please their clients. Gamers are not particularly interested on where the game comes from, or who created it any more than someone buying a new car or DVD player. A product for mass consumption only keeps the branding features of the trade mark, all the other characteristics are subject to customisation, due to the need to appeal to the local market, therefore the translation will be in some cases an actual recreation, or a 'transcreation' (Mangiron & O'Hagan 2006), where translators will be expected to produce a text with the right 'feel' for the receiving market. It is important for translators to be aware of the logic behind this. Video games are a software product, so they will have manuals and instructions, as well as menus and help files. This will call for technical translation. On the other hand, we will also find texts full of narration and dialogue with a more inventive quality closer to literature but, unlike literature, translators are allowed to treat equivalence in a more flexible sense, always highlighting fun and playability.

Translators have for centuries worked from written texts in order to produce other written texts. These texts were firstly, and primarily, aimed at other scholars and the privileged that could actually read, normally religious leaders and thinkers. Popular books, from the religious to the scientific or the literary ones, were translated and distributed in other nations. These first translators started writing about the language transfer process and communication between cultures centuries ago, as we can see in anthologies and historical studies on translation (see Vega 1994, and Deslile & Woosworth 1995). Some of their findings are still applicable nowadays.

With the advent and popularisation of new modes of entertainment and communication, (like cinema, radio and TV), translators have had to adapt and learn new techniques, since the product to be translated required a slightly different approach. Good old practice was revised and adjusted to the demands of modern products, for example: space and time constraints in subtitling (Díaz-Cintas 2003), lip-syncing for dubbing (Agost & Chaume 2001), or the translation of seemingly 'impossible' audiovisual puns (Bernal-Merino 2002), to name but a few. These changes did not transform the art of translation, but they did add to its complexity and the debate in Translation Studies, since the translation of a text that is part of a multi-channel (image, sound, and text) product has added difficulties due to its audiovisual nature.

A similar development has taken place with the spread of multimedia interactive computer technology and the popularisation of video games. Translators working for the entertainment software industry have to deal with the same complexities belonging to the written and audiovisual medium (to different degrees) and adapt to the specific needs of an interactive digital product. As shown by Mangiron and O'Hagan (2006), the translation of video games nor-

mally allows for and encourages a more inventive approach than other translation areas to enhance players' immersion. This may seem to clash with the well established translation principle of 'equivalence', based on the long-established practice of translating canonical (and non-canonical) texts, but I think it may highlight levels in the degree of equivalence, as well as the need to reconsider the boundaries of its applicability (Bernal-Merino 2002). Video games are developed by a team of creative people, there is no single author, and they do not necessarily broadcast their nationality, in fact, the game will morph into whatever form publishers consider appropriate for the receiving culture to guarantee its appeal and a high level of market penetration.

Video games have a variety of texts, such as manuals, dubbing scripts, and subtitles that need translating, but they also have other type of texts in a format only common to utility software, like a word processor application, or an Internet browser. All these programs have one thing in common: information and commands are readily available at the click of a button. It is what we call 'interactivity'. Interactivity allows readers to navigate the text in a different manner, sometimes non-linear (when going through the menus), and at other times choosing the progression of our story the way we consider appropriate (when playing a role playing game). Aarseth explores in his book Cybertext (1997) the aesthetics and textual dynamics of what he calls "ergodic" literature, where "nontrivial effort is required to allow the reader to traverse the text." (Aarseth 1997:1). Another publication worth reading is First Person. New Media as Story, Performance, and Game edited by Wardrip-Fruin and Harrigan. This book is organised as a series of discussions on a variety of topics from "cyberdrama" to "ludology" among game creators and theorists. The conversational structure (partly done through a web site created in conjunction with Electronic Book Review www.electronicbookreview.com) inspired contributors to revise, update and expand their arguments as they prepared them for the book. The fact is that we do not seem to have a theoretical framework where we can locate textual products that have a variety of interactive features, some of which link paragraphs that read as literature.

The interactive element of computer programs, has serious consequences for translators because it means that access to texts and information is random, i.e., each user will activate a particular message or command at a different point, or not at all. An arbitrary sequence of events does not allow for linear texts and contextual information, therefore, translators lose one of the most important sources needed in the decision making process. Context is still available, but it has to be understood in a much wider sense. In these cases, translators have to rely on existing manuals or the actual technical team that created the software. Esselink (2000) is probably one of the best reference for the localisation of utility software and web pages.

The problems of interactivity for translators come to the fore in video games, which, unlike other types of interactive software, are meant to narrate stories; players' adventures in a virtual world. These stories are told in more or less literary terms, through both narration and dialogue. Texts will be stored in different parts of the game code, from where the program will retrieve them and present them seamlessly to the player, whatever the order of his/her actions. Due to the nature of interactivity, scripts are written in a non-linear manner. Translators, therefore, have to work with words and sentences with very little or no context (unless they have access to and can read game code). The combination of all these factors highlights the translation of video games as a new specialisation that deserves further and deeper research within translation studies.

4. Translatable assets generated by the game industry

Nowadays, more games have a "localisation friendly" development process, although, unfortunately, it is still an afterthought in many cases (Chandler 2005b). The most commonly used file types are .txt, .rtf, .doc and .xls, which are compatible with most systems. More and more the translation industry is turning to CAT (Computer Aided Translation) Tools due to the fact that they can increase productivity and consistency if used correctly. These type of programs are a must in the localisation industry, and translators are expected to make the most of them if they are going to meet the tight deadlines typical of this fast-paced industry. The only downside of using these programmes is that the most popular ones tend to be quite expensive and freelancers will need to weigh carefully the pros and cons of acquiring them. Nevertheless, there are shareware and freeware programs that can also help. Quah (2006) provides a great insight into how new technologies and initiatives may benefit language and translation professionals.

Unfortunately, the software localisation industry has

not been able to create a GUI (General User Interface) localisation tool, such as the ones used in the translation of utility software, for translators to use with video games. These programs (for example Alchemy Catalyst and Passolo) allow users to work directly but safely with the game code, generating a visual representation of the final product, which means that translators can see exactly what the end result will look like and adjust the text or the interface to suit the space available and general look. The LRC (Localisation Research Centre) and LISA (The Localization Industry Standards Association) have ample information on these programmes.

When a game has followed a localisation-friendly development, translators will be able to work from a less cryptic file, if still non-linear. The localisation engineer will provide translators with a 'localisation kit' of the game that would often include: the instructions for the project, the strings that need translating (normally Word and Excel files), a glossary of known (or previously used) terminology, and a 'translation memory' (TM) file. If it there is no previous TM file, translators will be expected to create one for future reference. Translatable strings will normally be organised in tables with independent columns and rows for each piece of information. Part of this data will be for the programming team (coders, audio and video engineers, etc.) and part of it will be for language professionals. The spreadsheet format helps to organise data in an easy-to-find way, which is ideal for a non-linear, multi-threaded storyline, but it also means that translators might get very little context, if any at all.

People unrelated to the entertainment software industry might find the quantity and quality of translatable assets it produces quite surprising. Whether in combination with other entertainment sectors or not, most video games will require translation for thousands, or even hundreds of thousands of words from the beginning of the project until the end. The workload will increase depending on the number of languages aimed for, as well as the number of platforms being developed for since they all have different specifications. Current platforms include PC, PS2, PSP, Xbox, Xbox 360, GC, GBA, Nintendo DS, mobile phone, PS3 and Nintendo Wii. With 11 possible platforms the linguistic and cultural adaptation of a game is a huge technical undertaking that adds to the complexities of hardware and software adaptation, and translators will need to become acquainted with all these different brand glossaries and hardware specifications.

Linguistic assets will be utilised in a variety of ways at different times throughout the creation, development and launch of the game, and they will be found in different formats, mainly:

- The game itself, which has a variety of texts in multiple formats, from the packaging and manual, to the installer programs and readme files, UI (user interface), as well as audio and video files
- The official web site of the game, which will normally use HTML or Java Script. Many websites use content management programs, which can be a very effective tool for regular updates.
- Promotional articles and merchandising in general.
- Game patches. They are downloadable programs that enhance game functionality.
- Game updates. Periodical downloadable augmentation of game features, storylines, and characters.

Within these products, there are different textual types, each of which has its own characteristics and purpose. Because we are dealing with a multimedia product, the challenges the translators are going to face are also multimedia. Within the same project we will have to deal with a wide variety of issues such as reproducing the oral quality of dialogues in writing, lip-synching for dubbing, space and time constraints for subtitling, number of characters for captions and UI, etc. The following table is an attempt to detail the textual types in video games:

TRANSLATABLE TEXTS	FORM	DESCRIPTION
1. Manual	Written	Normally includes legal, technical, literary, didactic, and corporative texts.
2. Packaging	Written	Mixes a promotional text with a literary one.
3. "Read me" files	Written	Technical text.
4. Official Web Site	Written	Mixes a promotional text with a literary one, but it will also have technical information like minimum requirements, etc.
5. Dialogues for dubbing	Spoken	Oral text where registers, accents, and idiosyncrasies have to be conveyed into another languages.
6. Dialogues for voice-over	Spoken	The narrator's voice where an oral text does not need lip-syncing but it needs to be cued with the visuals.
7. Atmospheric utterances	Spoken	Many games will include people talking or reacting to the player's actions. No synchronisation required normally, but we have to maintain the orality.
8. Dialogues for subtitling	Written	Oral text in written form. Not all languages allow for the same licences when writing. We also have time and space constrains.
9. UI (User Interface)	Written	Space is at a premium and redesigning is rarely an option.
10. Online help	Written	Briefness and clarity are needed.
11. Graphic art with words	Graphic	Use for titles and in-game signs and ads.

Due to this variety of textual types and file formats, translators who specialise in video games are advised to be computer and console literate, as well as to be able to switch from one type of text to another. Freelancers and translators in general, are used to dealing with these kinds of issues, but there is another level of difficulty. Despite the fact that entertainment software products can generate more than 50% from their translated versions, many video games do not have a 'localisation-friendly' development (Chandler 2005a:115). In these cases, the translation process will actually be more challenging, time-consuming, and expensive. Translators might have to deal with 'hard-coded strings', that is, actual game code where translatable text will normally be signalled by characters such as '$', '[ ]', '{ }', '%', '< >'. The system was reasonable enough at the beginning because games tended to have very few translatable sentences (think of Pac-Man, Space Invaders, or Pong), but it made the process very arduous, and introduced code and text mistakes (generally called 'bugs' in the industry). However, most of today's games have several hundreds (if not thousands) of lines of text, with RPGs (Role Playing Games) being the most text intensive, and hard-coded text for translation would interfere and delay the whole process.

I will now try to explain in more detail the difficulties translators working for the entertainment software industry have to deal with, since it is highly advisable for them to not only know about game mechanics, but also to understand the logic behind the game code in order to be able to fulfil their role correctly. There are mainly four issues to take into account when translating video games: the addressee, text fragmentation, translating tables, and translating variables.

4.1. The addressee:

One of the important characteristics of the translation of video games is, as mentioned earlier, the variety in textual types. Games address the player at three different levels, and the textual type changes accordingly. These levels are:

- As a client of the developing, and the publishing companies. This texts will normally be produced by marketing departments, and they will be used in the packaging, promotional articles and websites, advertisements and commercials. These texts are used to pitch the product and attract customers.
- As a legal owner and user. We will see this type of content in the installer program, the EULA (End User License Agreement), the 'Readme' file, and the manual. The textual type here pivots between legal, technical, and pedagogical.
- As the protagonist of the story. This is the most important level, the one that ultimately convinces

buyers, and immerses them into a virtual world where they can venture in and live out extraordinary stories.

4.2. Text fragmentation

Language professionals that specialise in audiovisual translation often have to work from only a written script, or with a bad copy of the actual programme and no script, or even with only a portion of the needed information. This is far from ideal but not uncommon. As we all know, lack of context (and co-text) affects the act of communication (Cutting 2002) and, consequently, translation, because isolated linguistic items tend to polysemy, i.e., they have various possible meanings. Fragmentation, however, is one of the features of texts within the interactive entertainment software industry. This is not to say that there is no story, or that games have a random and chaotic sequence of events, but that the story is also dependant on the individual performance of the player, and the underlying structure is provided by the game code which makes interactivity possible, and not the storyline. In video games things happen as and when players trigger them through their actions, there is no unique and compulsory sequence of events. In fact, this interactivity is part of their appeal, the relative freedom to resolve situations in the way and at the pace the player chooses to. This feature of entertainment software products has a great influence in its design and, more importantly for translators, in the way scripts are written (Chandler 2006) and prepared for them to work on.

Everything in a video game has to be programmed through the game code, which is an artificial language that is used to give instructions to the computer. Programming languages have been optimised to produce the best result with a minimum of commands, memory use, and variables. Readers can see an example of source code if the go to their Internet browser and click on 'View' and then 'Source'. Here is an example from the beginning of the home page of L4G (groups.msn.com/L4G):

```
<HTML>
<HEAD>
<TITLE>L4G</TITLE>
<LINK REL = "stylesheet" TYPE = "text/css"
HREF =
"http://sc.groups.msn.com/themes/R9c/pby/MS
Nframing.css"><LINK REL = "stylesheet"
TYPE = "text/css" HREF =
"http://sc.groups.msn.com/themes/R9c/pby/them
e.css">
<META NAME="TITLE" CONTENT="L4G">
<META NAME="DESCRIPTION" CON-
TENT="LANGUAGES FOR GAMES
LINGUISTS FOR GAMES
LOVE FOR GAMES
LOCALISATION FOR GAMES">
<META NAME="KEYWORDS" CON-
TENT="msnlang1, msncommengb, Language,
translation, localisation, localization, game,
videogame, video-games, linguist, linguists,
games, videogames">
```

This is the type of text that programmers have to generate for us to see the user-friendly version. It is not completely impossible to understand, but it is certainly very far from a novel, a manual, or a screenplay. Game source code is even more cryptic to the untrained eye. This is the reason why programmers (or localisation engineers in some cases), have to extract all the linguistic assets of the game and present them to translators in a format that is useful for all parties in the team, mainly the localisation, the programming, and the QA (quality Assurance) departments. The preferred format is the table and the spreadsheet. Information is fragmented but easier to find. By allocating each piece of information a separate column, the team is able to work with a more understandable source, and programmers can then safely insert the pertinent strings back into the game code, avoiding the problems of having non-technical people editing the source code, which would produce bugs.

4.3. Translating tables

Tables will normally have a column for the 'resource file ID/name', one for the original string, and another one for the translation. Often they will also feature a column for comments where the localisation engineer can insert extra information to help voice actors, translators, etc. Tables are an ideal tool to organise data but not to tell stories. However, as we said earlier, stories in video games are non-linear because they depend on players' decisions. If the project has been planned carefully, and enough time has been allocated to the localisation of the game, spreadsheets might contain other columns such as: 'name of character', 'situation', 'location in the game', 'format', and 'sound effect'. These tables have the advantage of organising the multi-threaded possibilities that are available to players so that everybody involved in the project can find exactly what they need. Contextual information will only come from other boxes in the spreadsheet, which won't necessarily have any chronological relevance to each other. Translators accustomed to a

more traditional kind of work will most likely resent this lack of context.

4.4. Translating variables:

Most games allow players to choose their name, gender, nationality, etc. which means that translatable strings will need to incorporate 'variables' (similar to the ones used in mathematics or physics) for the game code to be able to take that data into account and present the right text correctly phrased.

Variables are used in many complex ways to enhance players' immersion by addressing them and their chosen profile directly. The most commonly needed one is the variable for the player's name. For example, the winning message after the completion of a part of the game normally say: " /n player1 /n wins ! ". The string between the '/n' characters is the variable, which the programme will substitute depending on players' choices. So, for instance, in my case, the phrase would read " Miguel wins! ". There is no set way to indicate variables. I chose ' /n ' but this will depend on the SDKs (Software Development Kits) used and the lead programmer of the project. Translators have to be aware of the strings that belong to the game code, and the strings that belong to the localisable assets, since mistaking them would probably make the game 'crash' or even block the computer. For the above example, the Spanish translation would probably say: " ¡ Ha ganado /n player1 /n ! [" /n player1 /n has won !"]. As it usually happens, the translation is longer than the original text which could pose a problem. We have also added the opening exclamation mark character (mandatory in Spanish), and changed the order and tense of the verb to make it sound more natural. Note that blank spaces are also meaningful characters to computers, so adding or deleting one by mistake will affect the functioning of the programme. This example shows how a rather simple variable, that works perfectly fine with English, might prove to be complicated due to syntax, usage, and orthography, apart from the obvious space constraints.

Many games use variables for nouns as well, which may vary (depending on the language) in gender and number, affecting their inflection and, therefore, the translation of the whole sentence. If the game code does not take into account the grammar of the languages covered by the project many mistakes will appear. Mistakes that may be attributed to the translation but that actually show a problem in the way the game code deals with the grammar of natural languages. For example, strategy games may allow the player to choose from different nations to conquer the world. When a nation attacks you the message normally says something like: " /o nameofnation /o is attacking you! ". Names of nations change widely from one language to another. They may carry an article or not, they can be singular or plural, and masculine or feminine, so not only do we have to be careful with the syntax of the sentence and the possible relocation of the variable, we also need to be aware of potential changes due to the morphology of individual linguistic items. This formula could generate a sentence like: "Rome is attacking you!", but, if we are not careful, it could also produce "The Barbarians is attacking you!". Whenever possible, programmers and designers opt for rephrasing the sentence, to avoid this grammatical issues. The above message, for example, could be rewritten as " You are being attacked by /o nameofnation /o ! ", so the formula would allow for both "the Barbarians" and "Rome". But this might not an option for the target language.

Other games use concatenated strings with variables for nouns and adjectives to give feedback to the player. Guitar Hero (Harmonix 2006) does this through newspaper headlines. So after each part of the game the player is presented with the cover of a newspaper saying something like: "Incredible performance from the Boyz at the Plaza!". The coded string would look like this: " <ADJ> <NOUN> from <BAND> at <VENUE> ! ". The game code will include a list of variables where each 'adj', 'noun', 'band', and 'venue' will be allocated a name and a number to account for quality of performance, synonyms of performance, names of bands, and names of venues respectively. So the above formula could also generate: "Poor show from the Boyz at the Plaza!" or "Unique concert from Claxon5 at the Coliseum!". This formula works relatively well for analytic languages, like Chinese or English (English being one of the most analytic languages of the indo-European group). However, it is prone to errors when dealing with synthetic languages (like most of the other Indo-European languages) due to the high degree of concordance between articles, demonstratives, nouns, adjectives, and verbs in a sentence. If the game code does not allow for this morphological and syntactical agreement, translators will have to limit their options to one gender and one number, which could produce a very unnatural discourse.

Game programmers don't necessarily know grammar or languages, other than C++, Javascript, Assembly, etc., that is the languages they use to programme the video game. They are, however, starting to become

more aware of the complexities related to working with natural languages and their inclusion in game code. Heimburg (2003), an engineer for Turbine Entertainment, wrote: "[…] people don't even notice when the grammar is good, but they certainly notice when the grammar is bad."

5. Localisation and the software entertainment industry

If the localisation of the game has not been planned for from the very beginning it will end up being more costly both in time and money. The statistics produced by the European organisation Elspa (Entertainment and Leisure Software Publishers Association www.elspa.com/assets/files/0/20060505174657708_319.pdf), and North American ESA (Entertainment Software Association www.theesa.com/facts/sales_genre_data.php) show that video games attract all kinds of players whether young or old, male or female. Twenty years ago games would only be released in English, because the only real market was the US. The common practice nowadays for important titles is the localisation into FIGS (French, Italian, German, and Spanish) catering for a large percentage of the world population with these five languages. The more countries join the computer and IT revolution, the bigger the demand for video games will be. It is clear then that there is a case for preparing most games for international markets, and publishers as well as developers, are trying to find the best way forward. Chandler (2005, p.18) offers valuable advice and a very detailed breakdown of the game localisation process. Here is a brief overview of the three main phases on the localisation of a video game:

- Planning: In the pre-production stages the developer and the publisher will need to determine the level of localisation of for each language. Chandler highlights questions to consider: support of Unicode, international formats for currency, dates and time, subtitles, scalable UI, proprietary tools needed, who will actually do and build the localisation.
- Producing: when will translators receive originals?, how will the assets be organised?, what extra documentation will be in the localisation kit?
- Concluding: who is responsible for giving the OK to the localisation process? Will demos, patches, and updates be localised?

The model in place is still based on the old practice of developing one game and then thinking about its localisation, which was appropriate at the beginning of the industry since the only market to serve was the national market, i.e., Americans produce games for the US and the Japanese did the same for Japan. In the 21st century, this model is not sustainable. There is no question about the benefits of localisation, and there is no doubt that demand for localisation will keep on growing dramatically as more countries join the technical revolution. It is a question of how to organise it better in order to speed up the process while maintaining quality. There are several things the industry could do improve the localisation process, and the same rule applies as in any other industry: standardisation, regularisation, and automation.

a. Separate and streamline the different parts of the localisation process.
b. Eliminate steps in the linguistic and cultural localisation process. Empower translators so they can make corrections directly into the game text.
c. Avoid the idea that casual cheap inexperienced labour is a necessary evil. It takes more time to check, correct, report, and explain than to translate from scratch for an experienced translator.
d. Game developer companies could either produce their own localisation software, or help localisation software developers create an application for game localisation by agreeing file formats for UI, in-game text, and subtitles.
e. Keep bug reports to minimum technical problems that should only be dealt with by coders.
f. Equally, console companies could supply their own hardware and software to produce localisations, or outsource but monitor the development of a localisation tool kit.
g. Game developers, as well as publishers and localisation vendors could build up their own term banks through translation memory tools.

6. Conclusion

A video game is a multi-textual interactive entertainment product for mass consumption with shared authorship that is customised to attract audiences in a variety of countries. New generation video games are becoming more immersive than ever, with powerful engines that can control lifelike graphics and physics, as well as high quality surround sound. However, no matter how realistic environments are if players cannot understand the language, or if the translation of the texts has not been treated with the same care as

the rest of the assets, the game experience will be affected in a negative way. Inappropriate translation can easily break the gaming illusion and bring players out of their immersive experience, just like any other bug in the game. Helen Trainor (2003, p.1), a senior manager at the Symbio Group, says: "At a most fundamental level, games tell stories. The localizer's challenge is to make these stories resonate for different cultures." It is also worth highlighting that video games, as a rule, are not meant to be felt as belonging to a particular country, where a nationalistic perspective prevails (JFK Reloaded, America's Army, and Grand Theft Auto being popular exceptions). On the contrary, most games develop fictional worlds, where our real ideological, political or religious world has little or no bearing at all, although they might be loosely based on real human cultures.

We may assume that game localisation is not much more complicated than any other software application, after all, it is just another software product but that assumption would be wrong. The process of localising games for multiple countries has a unique mixture of challenges that are worth studying in detail. It is difficult to realise the amount of linguistic assets that go into a game and how varied these texts are. Because we are dealing with a multimedia product boasting innovative technology, the challenges translators face are highly technical and multifaceted. Within the same project, linguistic localisers usually have to deal with a wide variety of issues that can range from, for example: graphical constraints in menus and popups, the degree of technicality, lip-sync and orality for dubbing, to space and time constrains for subtitling, to name but a few.

From the point of view of translation, this is the only product in which the linguistic transfer is part of the development process and can affect the final version of the game for a particular locale. This level of impact, together with the variety of file formats and textual types, as well as the constant use of variables in localisable strings, makes the translation of video games different from any other type of language transfer.

Bibliography:

Aarseth, E. (1997) Cybertext: Perspectives on Ergodic Literature, Maryland: Johns Hopkins University Press.

Agost, R. and Chaume, F. (2001) La traducción en los medios audiovisuals, Castellón de la Plana: Universitat Jaume I.

Bernal-Merino, M. (2007) 'Training translators for the video game industry', in Diaz Cintas, J. eds., The Didactics of Audiovisual Translation. Amsterdam/Philadelphia: John Benjamins, pp.87-106

Bernal-Merino, M. (2006) 'On the Translation of Video Games', The Journal of Specialised Translation [online], p.6, pp. 22-36., available: http://www.jostrans.org/issue06/art_bernal.php

Bernal-Merino, M. (2002) La traducción audiovisual, Alicante: Publicaciones Universidad de Alicante.

Castronova, E. (2005) Synthetic Worlds: The Business and Culture of Online Games, Chicago: University of Chicago Press.

Chandler, H. (2005a) The Game Localization Handbook, Massachusetts: Charles River Media.

Chandler, H. (2005b) 'Start Game: Game development and Localization', Enlaso webinar, 12 June, available: http://www.translate.com/technology/multilingual_standard/gameslocalization.html [accessed 20 June].

Chandler, R. (2006) 'Screen/play: Documenting Voice Assets', Gamasutra [online], June 2006, available: http://www.gamasutra.com/features/20060608/chandler_01.shtml [accessed 1 July 2006]

Cutting, J. (2002) Pragmatics and Discourse, London: Routledge.

Delisle, J. and Woodsworth, J. (1995) Translators through History, The Netherlands/Philadelphia: John Benjamins.

Díaz-Cintas, J. (2004) Teoría y práctica de la subtitulación, Barcelona: Ariel.

Edwards, T. (2006) 'Fun vs. Offensive: Balancing the 'Cultural Edge' of Content for Global Games'. Paper presented at the Game Developers Conference, San Francisco, 24 Mar 2006, Presentation available: http://www.englobe.com/englobe/docs/GDC2006_Edwards_Tom_CulturalEdge.zip, [accessed 14 Apr 2006]

Esselink, B. (2000) A Practical Guide to Localization, Amsterdam/Philadelphia: John Benjamins.

Heimburg, E. (2003) 'Localizing MMORPGS'. Gamasutra [online], Sept 2003, available: http://www.gamasutra.com/resource_guide/20030916/heimburg_pfv.htm [accessed 3 Oct 2005]

Mangiron, C. and O'Hagan, M. (2006) 'Game Localisation: unleashing imagination with 'restricted' translation'. The Journal of Specialised Translation, 6, pp. 10-21, available: http://www.jostrans.org/issue06/art_ohagan.php [accessed 3 Aug 2006]

Quah, C. K. (2006) Translation and Technology, Hampshire/New York: Palgrave Macmillan.

Sutton-Smith, B. (1997) The Ambiguity of Play. Cambridge/London: Harvard University Press.

Trainor, H. (2003)'Games localization: Production and Testing', Multilingual Computing & Technology, #57, vol.14, issue 5, pp. 17-19.

Vega, M. Á. (1994) Textos clásicos de la teoría de la traducción. Madrid: Cátedra.

Wardrip-Fruin, N. and Harrigan, P. (2006) First Person: New Media as Story, Performance and Game. Massachusetts: MIT press.

Reverse Localisation

Reinhard Schäler
Localisation Research Centre (LRC)
Department of Computer Science and Information Systems, University of Limerick, Ireland
www.localisation.ie
Reinhard.Schaler@ul.ie

Abstract

This paper revises the general perception that localisation is about linguistic and cultural adaptation of digital content to the requirements of foreign markets; that localisation is successful if the origin of the material can no longer be detected. We will show that in a more and more globalised society (not just economy) publishers, and especially publishers of advertisements, play with 'strangeness' and stereotypes. For example, there are advertisements running completely in French on Irish television and radio advertisements in English-speaking countries that are completely in German (or in English with heavy German accents). Rather than adapting to the culture of the target country, rather than avoiding differences, in these cases publishers highlight the differences, focus on 'strangeness', introduce (rather than avoid) accents, embrace cultural diversity rather than avoid it - and all that to increase sales. As a complimentary, pleasant and valuable by-product, the entertainment value for the consumer increases significantly.

Uneasiness

When the localisation industry emerged in the mid 1980s, localisation was technically more complex than it is today. Applications were not properly internationalised; content was not separated from functionality; a full recompilation of the application after translation was almost always the norm, making extensive and labour-intensive testing obligatory. At the same time, however, the question of how to culturally adapt a word processor, a spreadsheet or a similar office-type application - the most common applications to be localised then - was not even asked. The idea was to use localisation as a vehicle to increase return on investment (ROI) in the original application by opening up huge new markets (mainly in rich, developed western European countries) through a relatively cheap and low-tech 'adaptation' of the products, which would then make them accessible to non-English speaking consumers.

Twenty years later the landscape has changed considerably. Much of the localisation effort has been reduced to simple translation tasks thanks to the use of sophisticated tools that automate much of the engineering and testing effort. Mainstream localisation is now far less technical than it used to be. However, what is being localised has changed so much that the question of how localisation should be done has to be re-visited.

Web localisation has been and will for the foreseeable future remain the one area in localisation with the highest growth rates. Web localisation deals not just with simple user interfaces but with more general digital content. This digital content can include material on a wide variety of topics, including history, education, politics, culture, entertainment and gaming. While the technical problems of localising this content have been solved in principlem, if not strategically, the question of how to adapt this content culturally has not yet been answered.

Although localisers such as McKenna (2005) and Singh (2004) and Sheridan (2001) have started to discuss the issue of cultural adaptation at conferences and in relevant industry publications, the solutions they are recommending largely follow old principles: design for a global audience, i.e. internationalise your service or your product, keeping the required localisation effort to a minimum; when localising adapt your digital product or service to the expectations of your target audience, i.e. give the Germans a 'German' product, the French a 'French' product and the Italians an 'Italian' product. "A successfully localized service or product is one that seems to have been developed within the local culture" (Diller 2008).

The general idea is to hide the origin of the original content, strive for the global common cultural denominator and make everyone believe the digital product or service they are dealing with was developed in their own country (Schäler 2005a). To back

up this strategy, experts invariably cite the godfather of cultural difference in the workplace, Geert Hofstede (1977 and 2005), without questioning or critically appraising the findings of his research which has its origins in the 1960s.

This paper aims to highlight uneasiness with this approach and to open up a better informed discussion about cultural adaptation as part of the overall localisation effort.

I18N - L10N - G11N

It might be hard to believe but it is true: after twenty years of localisation there is still no consensus on what internationalisation (I18N), localisation (L10N) and globalisation (G11N) mean and how they relate to each other. Definitions given by industry associations, such as Gala (www.gala-global.org) and LISA (www.lisa.org), and companies, such as Microsoft (www.microsoft.com) and IBM (www.ibm.com) on their web sites - although they vary considerably in detail - refer to localisation generally as the 'linguistic and cultural adaptation of products to the requirements of foreign markets'. Most surprisingly, the fact that all internationalisation and localisation deals exclusively with digital material is mostly overlooked - or, maybe it is so obvious that it is not even worth mentioning? What makes this oversight so important is that its implications have never been explicitly discussed. The fact that some digital material, be it simple text, a graphic, audio or video, is not being adapted in a traditional medium such as paper or celluloid (what Negroponte described as the world of atoms) but in digital format has important implications for the tools and technology used, the process employed, and the knowledge required by the professionals involved. These implications are important, but can, unfortunately, not be considered in more detail in the context of this paper.

For our purposes, we will use the terms as follows:

Internationalisation is the process of designing (or modifying) digital content (in its widest sense) and services so as to isolate the linguistically and culturally dependent parts of an application and of developing a system that allows linguistic and cultural adaptation supporting users working in different languages and cultures.

Localisation is the linguistic and cultural adaptation of a digital product or service to the requirements of a foreign market and the management of multilinguality across the global, digital information flow.

Globalisation, in contrast, is a business strategy (not so much an activity) addressing the issues associated with taking a product to the global market; this includes world-wide marketing, sales and support.

The underlying rationale for localisation, the principal driver behind the localisation effort is the interest of the developers of the original product or service to increase their return on investment in that service or product. When software publishers were looking for markets in the mid-eighties where they could sell their products, they realised that there were several potential markets in Europe, ready to absorb their products, with all the right ingredients, i.e. a well-educated population with a sufficiently high income to buy their products. The only problem was that they did not speak English. That moment the localisation industry was born: its mission became to adapt software at a relatively low cost generating a relatively high revenue. While the subject localisers are dealing with has developed and expanded - localisers today deal not just with software but also with more general digital content - the underlying rationale behind this effort has remained the same.

There are a number of factors cited by localisation experts when asked what makes a successful localisation project. As in other industries, the successful balance between quality, cost and time required to complete a project are crucial. In relation to the acceptance by users, experts largely agree that localisation has been successful when the localised products and services have been linguistically and culturally adapted to the point that users do not realise that the product or service they are using was developed originally in a different country for a different target group.

Therefore, and here we extend the definition of the term, localisation is the linguistic and cultural adaptation with the aim to produce digital products and services for which the country of origin can no longer be traced. In other words, the measure of success is I believe it's mine, you believe it's yours - but, in its essence, it is all the same.

Cultural adaptation and localisation

Cultural adaptation in the context of localisation can only be understood on the background of its (short) history and rationale (Schäler 2005b). Localisation is a tool used by digital publishers to sell products and

services into markets where the original product 'as is' would not sell. The adaptation process aims to ease the use of products and services by removing linguistic and cultural barriers inherent in some digital products.

These barriers are present at a shallow level, which is now mostly understood, and at a deep level, which localisers still struggle with.

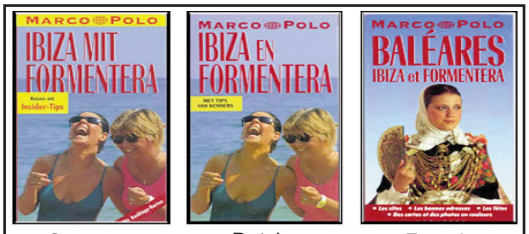
The shallow level includes the use of colours, symbols, sounds, and signals which have different meaning in a different cultural context. The deep level includes less evident but probably even more important aspects of the underlying value system, described, among others, by Geert Hofstede (1977) in his five categories of cultural differences which will be examined in more detail later in this paper.

One of the largest, oldest and most global organisations that has adapted the way it operates in different cultural spaces is the Catholic Church (Catechism of the Catholic Church 2008). While making clear "that diversity must not damage unity" (paragraph 1206), the Catechism states that "It is fitting that liturgical celebration tends to express itself in the culture of the people where the Church finds herself (…)" (paragraph 1207) and that "The diverse liturgical traditions or rites, legitimately recognized, manifest the catholicity of the Church, because they signify and communicate the same mystery of Christ." (paragraph 1208).

The Catholic Church has adapted not just its liturgical celebration to reflect changing beliefs and value systems in different cultural spaces, but also the images of the members of the holy family. For example, the image of the Virgin Mary looks distinctly local in Europe, Northern Africa/Middle East and South America.


Egyptian European Andean

Modern publishing houses have followed suit. The cover photograph of a recent guide book to the fun island of Ibiza, originally published in Germany, shows two young women in bathing suits on a beach having fun. This picture was kept for the Dutch version of the same guide book. When the French publishing house Hachette localised the guide book into French, they not only required the author to remove the unsuitable references in the guide book to the large gay community and the widespread use of drugs on the island, they also decided that the cover picture had to be changed. They believed that when French people go on holidays, they are looking for local costumes, folklore and traditions. Therefore, the two women on the beach had to make room for a middle-aged lady dressed in a traditional ibicenco dress. In essence, they adapted the guide book to match the expectations of their potential readers; the reality of island life was rather less important. (Communication to author 1992).



German Dutch French

Although the examples above are taken from the traditional world of paintings and printing presses, the same principles hold in the digital world. In fact, they are probably even more prevalent because changing or replacing images in the digital world is easy, in comparison with the world of traditional publishing.

In the following sections we will explain in more detail the difference between what we have defined as the shallow and the deep levels of cultural adaptation in localisation.

Shallow level

The shallow level of cultural adaptation has been of relevance in localisation since the introduction of the graphical user interface (GUI). It includes the use of:

- Colour
- Sensitive pictures and images
- Hand signals
- Symbols
- Sounds
- History
- Product names and acronyms

The following paragraphs provide some example for

each of these areas. (For a detailed analysis on international user interface design see Del Galdo and Nielsen (1996)).

Use of colour
For example, in many Western countries red is an alarming colour, white can indicate a pure or basic state, and black is sombre. This is different in Asian countries like China where red expresses joy, white indicates mourning and black is "the lucky colour". Green is associated with lush growth and ecology in Western countries, while it is the holy colour of the prophet in the Islamic world.

Sensitive pictures and images
For example, the national flag of a country is widely used to identify products aimed at specific markets and is, therefore, often printed on packaged software products. The Saudi Arabian flag contains holy symbols associated with the Koran, which Muslims are forbidden to destroy or dispose of.

Hand signals
Hand signals probably represent the most dangerous area of non-verbal communication. For example, a hand held up with the forefinger stretched out and the palm towards the viewer could be used to indicate "Danger!" or "Stop" in many countries - but in Greece it could cause serious offence. The "thumbs-up" sign, and "ok" sign (index finger and thumb forming a circle) used in many Western countries are regarded as sexual gestures in others.

Symbols
Icons related to system components (disk, printer, monitor etc.) or application-determined elements (drawing, writing, opening files etc.) usually do not cause problems. However, other symbols and icons that do not form part of the culture of the target country can cause serious problems for users in that country. For example, most users would not understand the use of the US-type post box with a flag to indicate that email has arrived; they would also probably not recognize the typical US yellow school bus as a symbol referring to education.

Sounds
Different cultures use sounds in different ways. For example, while a gong sound alerting a user that he made a mistake is perfectly acceptable in Western cultures, it should not be used in applications aimed at the Japanese market, where it would be seen as embarrassing for the user in front of colleagues.

History
Historical items frequently dealt with in multimedia encyclopaedias can be especially contentious. For example, which European was first to land on the American continent: was it St. Brendan, was it the Vikings, was it Columbus or was it a representative of the Mormons?

Product names and acronyms
Acronyms cannot be carried over into different languages and markets, even if they refer to international organisations. NATO is NATO in German, but it is OTAN in Spanish, for instance.

Deep level - entertainment, education, information, eContent
While at least some aspects of cultural adaptation at the shallow level are well understood, there are no strategies or guidelines helping localisers struggling with the deep level of cultural adaptation, probably best captured by Geert Hostede and his framework of cultural differences in the workplace.

Prof. Geert Hofstede conducted perhaps the most comprehensive study of how values in the workplace are influenced by culture. His study analyzed a large database of employee value scores collected by IBM between 1967 and 1973 covering more than 70 countries.

He first used the 40 largest countries only and afterwards extended the analysis to 50 countries and 3 regions. In the editions of his work since 2001, scores are listed for 74 countries and regions, partly based on replications and extensions of the IBM study on different international populations.

Subsequent studies validating the earlier results have included commercial airline pilots and students in 23 countries, civil service managers in 14 counties, 'upmarket' consumers in 15 countries and 'elites' in 19 countries.

From the initial results, and later additions, Hofstede developed a model that identifies four primary Dimensions to assist in differentiating cultures: Power Distance - PDI, Individualism - IDV, Masculinity - MAS, and Uncertainty Avoidance - UAI. He added a fifth Dimension after conducting an additional international study with a survey instrument developed with Chinese employees and managers. That Dimension, based on Confucian dynamism, is Long-Term Orientation - LTO and was

applied to 23 countries. These five Hofstede Dimensions can also be found to correlate with other country and cultural paradigms. (See http://www.geert-hofstede.com/ for more details. Singh (2004) is one example of how Hofstede's work has been used and referenced in localisation.)

Power Distance Index (PDI) focuses on the degree of equality, or inequality, between people in the country's society. A High Power Distance ranking indicates that inequalities of power and wealth are prevalent within the society. A Low Power Distance ranking indicates that the society de-emphasizes the differences between a citizen's power and wealth.

Individualism (IDV) focuses on the degree society reinforces individual or collective achievement and interpersonal relationships. A High Individualism ranking indicates that individuality and individual rights are paramount within that society. A Low Individualism ranking typifies societies of a more collectivist nature with close ties between individuals.

Masculinity (MAS) focuses on the degree society reinforces, or does not reinforce, the traditional masculine work role model of male achievement, control, and power. A Low Masculinity ranking indicates that the country has a low level of differentiation and discrimination between genders. In these cultures, females are treated equally to males in all aspects of society.

Uncertainty Avoidance Index (UAI) focuses on the level of tolerance for uncertainty and ambiguity within society, i.e. unstructured situations. A High Uncertainty Avoidance ranking indicates that the country has a low tolerance for uncertainty and ambiguity. This creates a rule-oriented society that institutes laws, rules, regulations, and controls in order to reduce the amount of uncertainty. A Low Uncertainty Avoidance ranking indicates that the country has less concern about ambiguity and uncertainty and has more tolerance for a variety of opinions.

Long-term Orientation (LTO) focuses on the degree society embraces, or does not embrace, long-term devotion to traditional, forward-thinking values. A high Long-term Orientation ranking indicates that the country prescribes to the values of long-term commitments and respect for tradition. A low Long-term Orientation ranking indicates the country does not reinforce the concept of long-term, traditional orientation. In this culture, change can occur more rapidly as long-term traditions and commitments do not become impediments to change.

Geert Hofstede's work and beliefs correlate almost perfectly with one of the central aims of localisation: adapt (foreign original) digital content to reflect not the culture of the source culture, but that of the target culture. In his view "Culture is more often a source of conflict than of synergy. Cultural differences are a nuisance at best and often a disaster."(Hofstede 2008). Thus, not only can sales be increased but nuisance at best and often a disaster can be avoided if material is localised appropriately.

The one problem digital publishers and localisers alike are faced with is how to translate Hofstede's findings into guidelines for the localisation of digital content. How can individualism, masculinity or uncertainty avoidance levels be adapted when localising, for example, a web site developed originally in the USA for the Chinese market? How are or how should web sites be designed on the basis of a country-dependent deep cultural value system?

These and similar questions have been examined by authors in the US-based publication Multilingual (www.multilingual.com), by tutors and presenters at recent Unicode conferences and in a study published in the Journal of Web Engineering (De Troyer et al. 2006). They looked at web sites published by multinational corporations and attempted to map Hofstede's categories to these sites. They essentially adapted Hofstede's findings as a blueprint for guidelines on cultural adaptation. (For details of recent Unicode conferences and tutorials on culture in the context of internationalisation and localisation, visit www.unicode.org.)

Here are some examples of their findings:

Power distance

The following elements on a web site were seen as an indicator for a high power distance, i.e. a large inequality between people in a country's society:

- Focus on hierarchy
- Focus on leaders
- Focus on tradition and/or religion

An indicator for a low power distance, in turn, was detected if the following elements could be found on a web site:

- Focus on equality between leaders and population at large
- Mutual respect between inferiors and superiors
- Focus on personal development

Along these lines, pictures of individuals, flags and heraldic signs were all interpreted as indicators of a high power distance type country.

Individualism
The presence of the following elements on a web site was seen as an indicator for a high individualism ranking:

- Focus on freedom
- Focus on personal development and self-realisation
- Focus on individual interests rather than those of the collective

An indicator for a low individualism ranking, in turn, was detected if the following elements could be found on a web site:

- Focus on consensus
- Focus on tradition and/or religion
- Focus on collective interests rather than those of the individual

Accordingly, the presence of many pictures of individuals on a site were interpreted as an indicator of an individualist culture, the presence of many group pictures as an indicator of a collectivist culture.

Masculinity
The presence of the following elements on a web site was seen as an indicator for a high masculinity ranking:

- Gender (men / women) are addressed separately
- Focus on ambition, competition, material success
- Women shown as tender/modest/caring, men as hard/ambitious/assertive

An indicator for a low masculinity ranking, in turn, was detected if the following elements could be found on a web site:

- Gender is addressed indiscriminately
- Focus on equality, solidarity, quality of life
- Women can be hard/ambitious/assertive, men can be tender/modest/caring

Accordingly, the presence of many pictures of nurturing, home-staying, cooking, cleaning, caring women on a site and the presence of many pictures of decisive, leading, competitive, fighting men were interpreted as an indicator of a masculine culture.

Uncertainty avoidance
The presence of the following elements on a web site was seen as an indicator for a high uncertainty avoidance ranking:

- Focus on formality and rigid rules
- Focus on precision and punctuality
- Focus on tradition and religion

An indicator for a low uncertainty avoidance ranking, in turn, was detected if the following elements could be found on a web site:

- Focus on flexible rules and tolerance for informality
- Focus on tolerance of ambiguity or vagueness
- Focus on evolution and change

Accordingly, the presence of a clear and simple navigation system on a web site was interpreted as an indicator of a low uncertainty avoidance culture.

Long-term orientation
This category has not been considered by most researchers, probably because it represents a concept which they found difficult to translate into elements present on a typical web site.

European Union-funded projects
There have also been European Union-funded projects trying to find the answer to the question of how to adapt web sites to different countries (Vickers 2005). The projects evaluated existing sites and came up with templates for different countries, apparently modelled along the lines of their cultural value system and preferences.

- Germany: well structured, laid out in tables
- Scandinavian countries: nature loving, lots of trees and lakes
- Mediterranean countries: very lively, lots of strong colours

A trendy Web site in France will have a black background, while bright colors and a geometrical layout give a site a German feel. Dutch surfers are keen on video downloads, and Scandinavians seem to have a

Three Coca-Cola web sites: South Africa, China and India (from left to right). Designers of global web sites try to reflect the cultural preferences of their local target audience.

soft spot for images of nature - wrote Ben Vickers (2005) in the European edition of the Wall Street Journal reporting on the EU-funded Multilingual Digital Culture (MUDICU) project, coordinated by Helene Abrand, an internet consultant working for Real Media France, the French subsidiary of US-based Real Media Inc.

The result of this kind of adaptation effort is that a programme or a web page sends out all the right signals to the user using a chameleon-like strategy. But, in many cases, users know and, most of all, feel that something is not quite right. Because no matter how much you change the colour of a web site, the hand signals, the symbols and the sounds used, the content will always remain the same.

The cultural dilemma

The approaches advocated by Hofstede and his followers, as well as by the researchers working in the EU-funded project quoted above, lead to a cultural dilemma.

When you travel to Spain, do you really want to find out from a web-based, localised US travel guide where to eat in Santiago, Madrid or Seville? Or when you travel to the Middle East, read up on the history of the region on a localised US web page? - Sadly, this is what you will most likely be offered when you search the web for this kind of information.

Search for background information on any region of the world and the likelihood is that you will be presented with information not coming from the region itself but from US-based, localised websites which are registered with the major search engines and which are listed at the top of your list of search results.

Like travel writer Macon Leary in Anne Tyler's book The Accidental Tourist, who hates both travel and anything out of the ordinary, many eContent publishers dislike diversity, difference and divergence from standards - for the simple reason that it makes their life more difficult and, very importantly, their projects more expensive. They create perfect 'accidental' web sites, which are acceptable to every global citizen's taste, beliefs and customs. There are no surprises, there is no deviation from the norm - there is an almost clinical feel of global political over-correctness to them.

Travel writer Macon Leary needed Muriel, a deliciously peculiar dog-obedience trainer, to end his insular world and thrust him headlong into a remarkable engagement with life. Local content producers and local cultures need the technical experts to bring their content to the world so that the world can enjoy the different perspectives and approaches offered by them. What is needed is more local content and better access to this content - not more localisation of content originating from one single culture.

It remains to be proven how successful and appropriate current mainstream localisation approaches to cultural adaptation are. There are, however, even at this early stage of investigation, strong indicators suggesting that an adaptation strategy that effectively hides the origin of digital content could be at best subversive, misleading the consumers of this material and making them believe that what they are looking at was produced by someone with a similar value system to their own, thus at least potentially reducing their level of alertness and critical reflection on the content. At worst, at least from the perspective of the digital publishers, this strategy could have a devastating effect on the saleability and commercial success of the product or service in question by removing what could have been its most attractive selling point: strangeness.

Strangeness: how it works

There are many examples and lessons to be learned from international marketing strategies of how strangeness works for sales - lessons, which have surprisingly not even been considered yet by the localisation industry. (It is interesting to briefly refer here to the notions of "domesticating" and "foreignising" which have been widely discussed in translation theory for example by Venuti (1995: 19-20) who traces the roots of these strategies back to Schleiermacher's 1813 treatise on translation, Über die verschiedenen Methoden des Übersetzens. Foreignising strategies are adopted by translators who want to bring the foreign culture to the fore for a reader. House (2006) also gives a concise overview of this approach.)

One example of how strangeness as a marketing strategy works is by association. Certain products are associated with certain countries and cultures: for example, perfume, fashion and romance are associated with France; technology and engineering are associated with Germany; Italians are perceived as sophisticated coffee drinkers, which is why today a dictionary is required if one wants to buy a coffee, or rather a Macchio, Café con Latte, or an Espresso. Strangeness, far from being a cause of disruption and chaos as suggested by Hofstede, can be a source of attraction and differentiation. It works using existing stereotypes, or creating new ones. It works, not by adapting digital content to the culture and the language of the target country but by doing exactly the opposite. It works using reverse localisation.

Reverse localisation

We define reverse localisation as keeping or intentionally introducing linguistic or cultural strangeness into digital content for a particular target locale with the aim of intentionally differentiating a digital product or service from the dominating culture in that locale. The effect of reverse localisation is manyfold: the product or service in question is certainly set apart from potential competitors; the values and connotations associated with it play on a spirit of adventure, sometimes they just plainly take advantage of existing stereotypes, and almost always cause a sense of curiosity and heightened sense of attention.

Examples of reverse localisation can currently be found primarily in advertising, less in other types of digital content publishing, such as in audio, video or web site publishing. We suspect that the reason for this might be the strong business focus in the marketing and advertising industries, the concentration of a high level of creativity and the funds made available to them to realise their ambitions.

Following are some examples of reverse localisation. (Sound and video files of these are available but could not be included in the paper.)

Audio

Volkwagen

This advertisement was broadcasted in the U.K. and Ireland to promote cars manufactured by the German company Volkswagen. It played on the image of German engineering excellence and featured a manager and a designer trying to come up with yet another technical improvement for a Volkswagen - an impossibility as it turns out. The interesting aspect of this ad in our context is the distinguished German accent of the actors and the odd German word thrown in, in particular the almost universally recognised German word nein.

D.I.D.

D.I.D., the Irish Do-it-Yourself store, commissioned a series of advertisements also playing on the universally recognised German word nein and, in addition, on its phonetic closeness to the English word nine. Again, these ads feature an actor with a heavy German accent. While the first ad in the series introduced the main theme and setting, the second ad expanded on it.

Video

Toyota

Building on the engineering theme associated with the advertisement campaigns of many German car manufacturers, and specifically on the tag line used by Audi in their English promotional shoots Vorsprung durch Technik, this advertisement is set in a board room where the Chief Executive or Chairman shows a video of a perfectly engineered, seemingly flawless car on a test drive whose only problem is that it is not built by us. Throughout the video, the actor speaks German which is translated for the English speaking target audience using subtitles.

Stella Artois

The Belgium beer manufacturer commissioned a whole series of highly successful and extremely expensive, multi-million euro advertisements by prominent directors involving well-known professional actors. All of them play on Stella Artois' tagline reassuringly expensive; all of them are presented as short films (rather than advertisements) and

are set in strange settings; none of them are in English, the latest one has no spoken dialogue whatsoever. Examples are Last Orders (entirely in French), The Pilot (featuring dialogues in German, French and English), and Ice Skating Priests (without any dialogue at all).

Finches

Possibly the most striking campaign is that of the wholly Irish-owned soft drink manufacturer Finches, whose main products are (orange) soft drinks. This Irish company advertises its products with a campaign that is entirely presented in French, including the French tagline pour l'amour de l'orange. Two examples are The Poet and Orange Betty.

What are the lessons the localisation industry could learn from the advertisement industry in its efforts to linguistically and culturally adapt digital products? This is the question we will attempt to answer in the final section.

Lessons

The principal lesson to be learned by localisers is, arguably, that adaptation does not necessarily and exclusively mean the need to implement a lowest common denominator to comply with as many cultural requirements as possible, to remove all potentially controversial cultural aspects of a product or service, and to introduce elements reflecting the traditional cultural value systems of particular countries identified as target markets. To the contrary: if the aim of localisation is, similar to that of the marketing and advertising industries, to make a product or service as attractive as possible, then playing on the strangeness factor could prove to be at least as successful as trying to follow the traditional line of avoiding any potential source of conflict by avoiding appearances of cultural differences.

One of the more recent examples of how not to localise is probably that of the Dubai-based network cable company MBC who bought the US-produced comic series The Simpsons from their US-owners for broadcast in Arabic countries.

According to a report by Brian Whitaker (2005) in the Guardian Newspaper, the famously dysfunctional family from small-town America suddenly have all learned Arabic and started talking like Egyptians. The Simpsons have changed their name to Shamsoon. Bart, the skateboarding, gum-chewing delinquent has become Badr. Homer, his slobbish dad, has become Omar and has given up Duff beer and pork sausages, at least for the duration of Ramadan. Immediately, questions were asked about the appeal a sanitised version of the intentionally over-the-top appalling character of Homer (aka Omar) could possibly have, even to an Arab audience? Some viewers, indeed, are sceptical about the series, according to Whitaker. "I watched a promo segment and it was just painful", a blogger known as The Angry Arab wrote. "They were so unfunny and so annoying, those Arab actors ... the guy who played Homer was one of the most unfunny people I ever watched. Just drop the project."

Would it not have been infinitely more entertaining to the Arab audience (and, consequently, infinitely more profitable to the company) to see the original, exaggerated version of a "typical", defunct, ridiculous and chaotic US-type family? Would a humorous approach to US-"culture" not have had the potential to foster an understanding between the two cultures? As some commentators pointed out, it is a well known fact that people find it hard to hate what or who they are laughing about.

Our recommendation to localisers is to abandon the well-trodden path of localisation as we know it, with all its emphasis on the avoidance of even the smallest sign of cultural diversity and cultural differences. Instead, get the imagination working, get your audience involved and take advantage of their curiosity for the unknown, interest for the different, desire for exclusivity, spirit of adventure, and ambition to discover. Encourage your clients to get to know local customs, learn about other languages, taste the delights of foreign cuisine, wear strange clothes, learn how to play the instruments of other cultures; most of all encourage them to bring all their cultural baggage and enjoy the clash of cultures wherever they go and whatever they do.

This strategy will not only make digital products and services much more interesting and colourful, and therefore most likely increase sales revenues, it will also reflect in a much more appropriate way the world that we live in and thus help to prepare people to deal with it. Today, we are living in a world that is no longer defined by artificial country borderlines and cultures belonging to antiquated nation-states. After all, successful Chinese business people surely have much more in common with their European counterparts than with their cousins in the largely underdeveloped Chinese countryside. There are a

large number of similar cross-nation state categories, all of them ignored by Hofstede's framework, among them age (do the teenagers of the world not have more in common with each other than with their parents?), income (does income not largely determine access to information technology and therefore access to knowledge and information?), and education (is the ability to understand complex arguments and continually learn not more important than geographical location?).

Anyone still living under the illusion of the existence of "one state, one language and one culture" just has to take a trip on any of the public transport systems in any major European city, visit any of its restaurants, study at its university, visit its theatres, or go shopping in one of its food stores to realise that today we are living in a globalised, multi-cultural and multilingual society.

It seems that this time around, it is the localisation industry that has to be adapted, not the digital content it is dealing with.

References

Catechism of the Catholic Church (2008). Part two, The Celebration of the Christian Mystery, Section One, the Sacramental Economy, Chapter Two, The Sacramental Celebration of the Paschal Mystery, Article 2, Liturgical Diversity and the Unity of the Mystery, Liturgical Traditions and the Catholicity of the Church. http://www.vatican.va/archive/catechism/p2s1c2a2.htm (last accessed 01.07.2008).

Communication to author (1992). Concerns express by French publishing house Hachette while preparing a translation of a guide book to Ibiza, communicated to the author of the guide book, Reinhard Schäler, by its original German publisher.

De Troyer, Olga et al. (2006). On Cultural Differences in Local Web Interfaces. Journal of Web Engineering. Vol 5. No. 3. pp. 246-264. Rinton Press.

Del Galdo, E. M. and Nielsen, J. (Eds.) (1996). International User Interfaces. John Wiley & Sons, New York.

Diller, Philip (2008). Globalization, Localization, Internationalization and Translation. http://philip.pristine.net/glit/ (last accessed: 01.07.2008).

Hofstede, G. (1977). Cultures and Organizations: Software of the Mind. McGraw-Hill, New York.

Hofstede, G. and Hofstede G.-J. (2005). Cultures and Organizations: Software of the Mind, 2nd Edition. McGraw-Hill, New York.

Hofstede, G. (2008). "Culture is more often a source of conflict than of synergy. Cultural differences are a nuisance at best and often a disaster." Prof. Geert Hofstede, Emeritus Professor, Maastricht University. http://www.geert-hofstede.com/ (last accessed: 01.07.2008).

House, J. (2006). Text and Context in Translation. LRC Summer School. University of Limerick. http://www.localisation.ie/resources/courses/summerschools/2006/text_context.ppt (last accessed: 01.07.2008)

McKenna, M. (2005). Cultural User Interface Design. Twenty-Eighth Internationalisation and Unicode Conference. Orlando, Florida, USA (17-19 September 2005).

Schäler, R. (2005a). The cultural Dimension in Software Localisation, in: Localisation Reader 2003-2004. http://www.localisation.ie/publications/reader/2003/index.htm (last accessed 26.10.2005).

Schäler, R. (2005b): Cultural Adaptation is the Holy Grail of Localisation. Twenty-Eighth Internationalisation and Unicode Conference. Orlando, Florida, USA (17-19 September 2005).

Schleiermacher, F. (1813/1992). 'On the different methods of translating', in R. Schulte and J. Biguenet (eds.) (1992), pp. 36-54.
Sheridan, E.F. (2001). Cross-cultural Web Site Design: Considerations for developing and strategies for validating locale appropriate on-line content. MultiLingual Computing & Technology. October/November 2001. Number 43. Volume 12. Issue 7.

Singh, N. and Baack, D. (2004). Web Site Adaptation: A Cross-Cultural Comparison of U.S. and Mexican Web Sites. JCMC 9 (4) July 2004. http://jcmc.indiana.edu/vol9/issue4/singh_baack.html (last accessed: 01 July 2008).

Venuti, L. (1995). The Translator's Invisibility: A History of Translation. London and New York, Routledge.

Vickers, Ben (2005). Internet Becomes Battleground. For Europe to Defend Culture. Wall Street Journal Europe (26 March 2001). http://www.mudicu.org/common/wsj.html (last accessed: 01.07.2008).

Whitaker, Brian: How Homer became Omar. The Guardian (17 October 2005). http://www.guardian.co.uk/international/story/0,3604,1593794,00.html (last accessed: 01.07.2008).

This article is the revised version of a paper first published in the proceedings of Translating and the Computer 27, ASLIB, London, November 2005 (http://www.aslib.com/conferences/proceedings.html).

The author would like to thank Dr Sharon O'Brien and Carla diFranco for their review of this article and the extremely valuable suggestions they made to improve it. All statements made (and errors left) in this article are, of course, the sole responsibility of the author.

Guidelines for Authors

Localisation Focus
The International Journal of Localisation
Deadline for submissions for VOL 7 Issue 1 is 15 September 2008

Localisation Focus -The International Journal of Localisation provides a forum for localisation professionals and researchers to discuss and present their localisation-related work, covering all aspects of this multi-disciplinary field, including software engineering and HCI, tools and technology development, cultural aspects, translation studies, human language technologies (including machine and machine assisted translation), project management, workflow and process automation, education and training, and details of new developments in the localisation industry.

Proposed contributions are peer-reviewed thereby ensuring a high standard of published material.

If you wish to submit an article to Localisation Focus-The international Journal of Localisation, please adhere to these guidelines:

- Citations and references should conform to the University of Limerick guide to the Harvard Referencing Style
- Articles should have a meaningful title
- Articles should have an abstract. The abstract should be a minimum of 120 words and be autonomous and self-explanatory, not requiring reference to the paper itself
- Articles should include keywords listed after the abstract
- Articles should be written in U.K. English. If English is not your native language, it is advisable to have your text checked by a native English speaker before submitting it
- Articles should be submitted in .doc or .rtf format, .pdf format is not acceptable
- Article text requires minimal formatting as all content will be formatted later using DTP software
- Headings should be clearly indicated and numbered as follows: 1. Heading 1 text, 2. Heading 2 text etc.
- Subheadings should be numbered using the decimal system (no more than three levels) as follows:

  Heading
  1.1 Subheading (first level)
  1.1.1 Subheading (second level)
  1.1.1.1 Subheading (third level)
- Images/graphics should be submitted in separate files (at least 300dpi) and not embedded in the text document
- All images/graphics (including tables) should be annotated with a fully descriptive caption
- Captions should be numbered in the sequence they are intended to appear in the article e.g. Figure 1, Figure 2, etc. or Table 1, Table 2, etc.

More detailed guidelines are available on request by emailing LRC@ul.ie or visiting www.localisation.ie

Localisation Focus
The International Journal of Localisation
VOL. 6 Issue 1 (2007)

CONTENTS