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

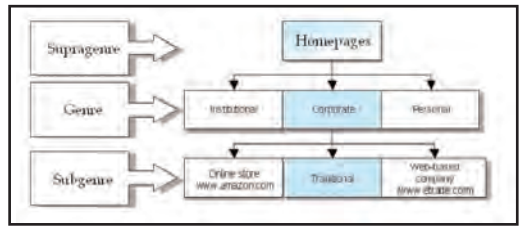


FIG.1. HOMEPAGE GENRES

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

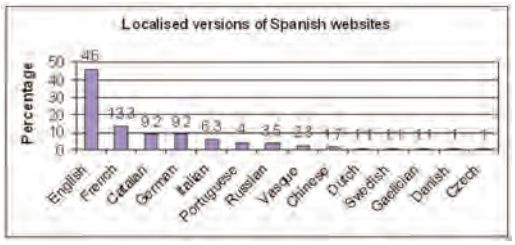
Global corpus statistics: 172 corporate Spanish (es-ES) websites		
Type	**Total**	**Average**
Web pages	27,852	161.93 Pages/site
Words in main body text	5,757,289	205.99 Words/page
Total words	9,929,302	356.50 Words/page
Words in <Meta>	934,361	33.54 Words
Words in <alt>	317,695	11.40 Words
Words in <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.

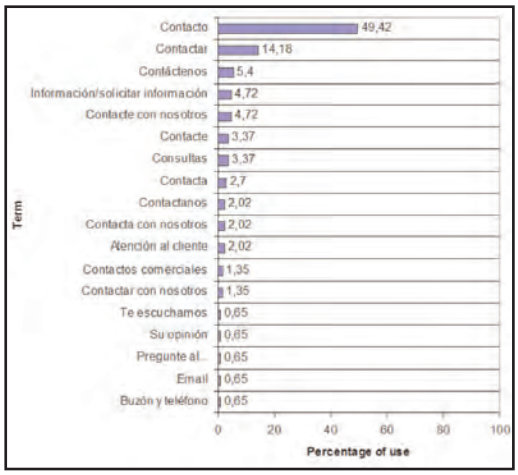
The Internet opened a new platform for information distribution that lead to a new model of communication, the so-called Interactive Mass Communication Model (Janoshka 2003). In this new model the flow of information does not only flow from the sender to the receiver; the communicative acts can be established 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

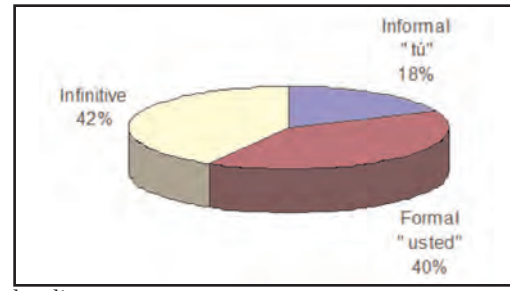


corpus are:

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.