Meta
Journal des traducteurs
Translators' Journal

META

Applying Corpus Data to Define Needs in Web Localization Training

Miguel Ángel Jiménez-Crespo et Maribel Tercedor

Volume 56, numéro 4, décembre 2011

URI : https://id.erudit.org/iderudit/1011264ar
DOI : https://doi.org/10.7202/1011264ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Jiménez-Crespo, M. á. & Tercedor, M. (2011). Applying Corpus Data to Define Needs in Web Localization Training. *Meta*, *56*(4), 998–1021. https://doi.org/10.7202/1011264ar

Résumé de l'article

La localisation prend de plus en plus d'importance dans les programmes universitaires d'enseignement de la traduction. Cependant, peu de recherches empiriques se penchent sur des questions telles que la relation entre localisation et traduction, les champs de compétence spécifiques à la localisation ou sur la manière de rendre compte des différences culturelles en rapport avec le genre des productions numériques, les types textuels et les conventions en usage. Se basant sur les recherches antérieures menées sur les compétences traductionnelles, le présent article vise à proposer les fondements de l'étude des compétences en localisation. La recherche a mis en oeuvre une étude empirique contrastive, fondée sur des corpus, des traductions d'étudiants, ainsi que des données provenant d'un corpus comparable, constitué de textes originaux en espagnol et de textes localisés en espagnol. Notre objectif est d'identifier les différences dans la production des textes numériques localisés par des étudiants et des professionnels, d'une part, et des textes originaux, de l'autre. Cette analyse contrastive vise à mieux comprendre comment les compétences en localisation sont en lien avec le concept inclusif de compétences traductionnelles, afin de relever les aspects sur lesquels devrait porter la formation en localisation au niveau universitaire.

Applying Corpus Data to Define Needs in Web Localization Training

MIGUEL ÁNGEL JIMÉNEZ-CRESPO
Rutgers University, New Brunswick, United States
miguelji@rci.rutgers.edu

MARIBEL TERCEDOR
University of Granada, Granada, Spain
itercedo@ugr.es

RÉSUMÉ

La localisation prend de plus en plus d'importance dans les programmes universitaires d'enseignement de la traduction. Cependant, peu de recherches empiriques se penchent sur des questions telles que la relation entre localisation et traduction, les champs de compétence spécifiques à la localisation ou sur la manière de rendre compte des différences culturelles en rapport avec le genre des productions numériques, les types textuels et les conventions en usage. Se basant sur les recherches antérieures menées sur les compétences traductionnelles, le présent article vise à proposer les fondements de l'étude des compétences en localisation. La recherche a mis en œuvre une étude empirique contrastive, fondée sur des corpus, des traductions d'étudiants, ainsi que des données provenant d'un corpus comparable, constitué de textes originaux en espagnol et de textes localisés en espagnol. Notre objectif est d'identifier les différences dans la production des textes numériques localisés par des étudiants et des professionnels, d'une part, et des textes originaux, de l'autre. Cette analyse contrastive vise à mieux comprendre comment les compétences en localisation sont en lien avec le concept inclusif de compétences traductionnelles, afin de relever les aspects sur lesquels devrait porter la formation en localisation au niveau universitaire.

ABSTRACT

Localization is increasingly making its way into translation training programs at university level. However, there is still a scarce amount of empirical research addressing issues such as defining localization in relation to translation, what localization competence entails or how to best incorporate intercultural differences between digital genres, text types and conventions, among other aspects. In this paper, we propose a foundation for the study of localization competence based upon previous research on translation competence. This project was developed following an empirical corpus-based contrastive study of student translations (*learner corpus*), combined with data from a comparable corpus made up of an original Spanish corpus and a Spanish localized corpus. The objective of the study is to identify differences in production between digital texts localized by students and professionals on the one hand, and original texts on the other. This contrastive study allows us to gain insight into how localization competence interrelates with the superordinate concept of translation competence, thus shedding light on which aspects need to be addressed during localization training in university translation programs.

MOTS-CLÉS/KEYWORDS

formation en localisation Web, compétence en localisation, conception de cours, évaluation, corpus d'apprenants, corpus comparable
web localization training, localization competence, course design, evaluation, learner corpus, comparable corpus

1. Introduction

Despite the rapid growth of localization training at both graduate and undergraduate levels, there is still a lack of empirical studies that would shed some light on the definition of the object of study, localization and localization competence, as well as on the identification of the specific issues that should be incorporated into these courses. This lack of delimitation with respect to the object of study and its relationship to translation competence has meant that localization training tends to concentrate on the acquisition of instrumental competences (PACTE 2005). These mainly deal with the proficient use of technology tools, thus reinforcing the industry's claim that localization is primarily a complex technological process (Dunne 2006; Jiménez-Crespo 2011). Nevertheless, the design and implementation of localization teaching materials also need to address specific cognitive, communicative and textual aspects that characterize the translation of digital genres (Jiménez-Crespo 2008b). Such aspects include, among others, potential processing problems, interference at the production stage resulting in lexical calquing or syntactic interference, intercultural differences in genre or text-specific conventions or the impact of tools on the translation process. The goal of this study is to contribute to this research area with a combination of corpus-based top-down and bottom-up approaches. As a starting point, a foundation for localization competence based on current empirical research on translation competence (PACTE 2005) is presented. This proposal was developed based on the results of the contrastive analysis of a learner corpus of localized web forms and a comparable corpus of original and localized similar texts.

The current gap in focus between professional practice and academic training perspectives needs to be addressed in the conceptualization of localization and the subsequent planning of localization training (Dunne 2006; Pym 2006). In order to bridge this gap, this study focuses on the acquisition of localization competence by novice translators, and it acknowledges the fact that localization competence can also be acquired in the opposite direction: when a localization/internationalization engineer or developer is trained to become an *interlinguistic localizer*. Therefore, the first step in this study is to establish a theoretical framework for the cited empirical corpus-based study. We will then report on the methodology and results of the threefold analysis of learner, professionally localized and original web texts.

2. Localization Competence as a Specialized Subtype of Translation Competence

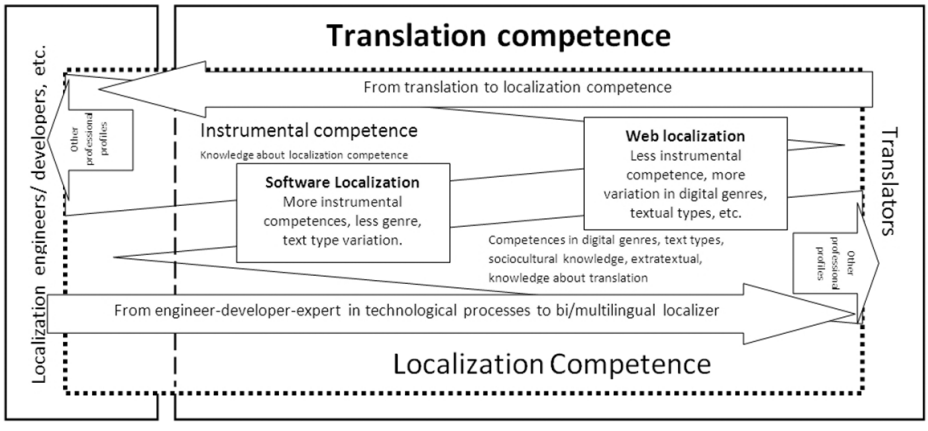
Defining and delimiting translation competence has been the object of an increasing number of theoretical and empirical studies (e.g., Bell 1991; Kiraly 1995; PACTE 2005). The relevance of this research is paramount for translation training, as these empirical studies and models help trace the progression from bilinguals to novice and then professional translators, thus providing a framework for translation training programs. The field of localization is in dire need of the development of a model of localization competence, as this can consequently serve as a foundation for training efforts in this area. So far, this notion has only been mentioned in two previous studies according to an in-depth literature review (Pym 2006; Wright 2004). In

establishing curricula, this notion can prove to be more productive than trying to define *localization* or the role of the *localizer* as such, as an industry-centered approach usually pursues. In this approach, localization is merely defined in opposition to a notion of *translation* based on linguistic equivalence models reminiscent of the early stages in the development of Translation Studies. As a result, localization is normally defined by highlighting certain aspects that are also inherent to any translation process, such as intercultural adaptation, or even by reducing the latter to the simple process that involves just *language* (Brooks 2000), *text* (Sprung 2000) or *words* (Lommel 2007; Esselink 2006). Localization is thus portrayed as a more complex process that deals with intercultural, communicative and technological aspects. This study argues that instead of grounding localization training efforts on this fuzzy notion of localization opposed to general translation, an approach based on localization competence as a specialized subset of translation competence can provide a more reliable and solid foundation.

Figure 1 shows a proposed framework upon which to develop localization competence models. This notion is understood as a specialized subset of translation competence that, nevertheless, includes a set of skills related to both instrumental subcompetences (localization tools, technological processes, etc.), and other subcompetences related to knowledge of pragmatic, socio-linguistic, textual and lexical-grammatical knowledge associated with digital genres, text types, conventions, etc. Following the PACTE (2005) translation competence model, the graphic separates the bilingual, extratextual and translation knowledge competences from instrumental competence. Instrumental competence alone can account for the industry training approach, more focused on the acquisition of technology-related competences. The graphic reflects the two current pathways into localization, from translator or translation novice to multilingual localizer or from localization/internationalization engineer or developer to localizer. In the latter, trainees might be extremely proficient in their instrumental competences, but they might still need to acquire the remaining subcompetences related to general translation competence, such as contrastive knowledge of the language pair, knowledge of general principles of translation as a process and as a profession, etc. In the opposite direction, translation trainees who have already acquired the basics of translation competence need to concentrate not only on advanced instrumental competences, but also on specific issues related to digital genres, their macrostructural and microstructural levels, formats, degree to which a product is to be localized, etc. Following this framework, the study intends to investigate which specific textual and communicative aspects in this area need to be addressed during translation training aimed at developing localization competence.

FIGURE 1

Towards a Definition of Localization Competence



Following Esselink (2006), the continuum in the acquisition of localization competence varies depending on the emphasis and progression of the training, prototypically software, web and videogame localization. Software localization usually requires more advanced instrumental competences, but the potential variation in digital genres and text types can be quite limited, thus allowing trainers to concentrate mostly on technological aspects without losing sight of other translation aspects related to the software product as a unitary digital genre. On the other side of the spectrum, web localization requires less advanced instrumental subcompetences, usually limited to the use of tag editors or tag protecting CAT tools as well as working knowledge of HTML and script languages, but the potential variety of textual and digital genres and hence, textual and linguistic difficulties, is much wider. This is due to the fact that even when most digital genres, such as corporate pages, networking sites, etc. are highly conventionalized (Kennedy and Shepherd 2005, Jiménez-Crespo 2009a), hypertexts are by nature open structures and any other text or genre can be uploaded and incorporated into a site. In this sense, Storrer (2002) draws a very useful distinction, separating hypertexts, open hyperlinked genres, from e-texts, i.e., any text that has been produced for any medium or purpose and is simply posted onto a website.

Finally, this framework accounts for the fuzzy area between a localization expert with a translation background and the multilingual developer who can produce *natural translations* (Harris 1977) but who, nevertheless, intends to become a localizer. This is represented by the gap between translation competence and the localization engineer profession that, however, overlaps in some areas with localization competence. In this regard, any of the two potential profiles, translators and developer-engineers can always concentrate on expanding and enhancing their acquisition of the specific competences they are lacking, such as more instrumental or bilingual pragmatic, textual, sociocultural competences respectively. As in any other professional field, any expert who overlaps job profiles can, with specific advanced training in his/her weaker competences, potentially become an expert in both areas. This is indicated in the graphic by the arrows that read *other professional profiles*.

Once the proposed framework for the development of localization competence has been described, the following sections focus on the objectives, methodology and results of the corpus-based study.

3. Objectives

In a previous study, we reported on the acquisition of instrumental competences in web localization training using the same dataset as in this paper (Jiménez-Crespo and Tercedor 2009). It was found that the areas in which translation students needed to concentrate were: (1) the back-end or *hidden* areas in any webpage, such as the heading or scripts, (2) localizing text embedded in images or flash animations, (3) dealing with tagged files in order not to corrupt the files, (4) leaving segments untranslated or (5) deleting, renaming or moving folders that are critical to hypertext function. In this study, the focus is mostly on the analysis of textual, pragmatic discursive and communicative aspects that need to be tackled and that often are assumed to be acquired by translation trainees prior to engaging in localization training. These are mostly related to bilingual, extratextual and translation knowledge competences in the PACTE (2005) model.

In order to deal with the acquisition of the aspects that need to be specifically addressed during localization training, this investigation draws data from learner, professionally localized and original web corpora (see section 4). In general, the role of learner corpora in translation training has been extensively researched (e.g., Zanettin 1998, 2001; Bowker 2002; Zanettin, Bernardini *et al.* 2003). Learner corpora can be of use to translation trainers to improve or adapt an ongoing course (Bowker 2002: 19), but most importantly, they represent a highly valuable tool for curriculum design. López and Tercedor (2008) explored the use of corpora for the design of teaching materials and knowledge acquisition in scientific and technical translation. This study addresses the issue of conceptual and textual design of localization courses at the university level. Thus, we focus on the analysis of combined data from three types of compiled texts so as to shed some light on whether students' behavior replicates that of professionals as well as to infer some potential processing problems at the conceptual level. Our research questions are: can the combination of different corpora assist in assessing the needs for a comprehensive approach to localization training? What type of issues related to bilingual and extratextual subcompetences need to be addressed during localization training in addition to what is assumed to be already possessed by novice translators?

4. Methods

The methodology used in this study combines a bottom-up approach with a top-down focus. To begin with, a learner corpus was compiled consisting of 76 Spanish localizations by fourth year students working from English into Spanish in a localization and audiovisual translation course as part of the BA in Translation and Interpreting at the University of Granada, Spain. The text localized by students that was selected for this study is an English original HTML document containing contact information as well as a contact form with drop-down lists and controls. During the course of the study, the students worked on the global site and its different sections, and this specific

web subgenre was chosen among the several other assignments that were part of the course. The only modification to the original text was the pre-translation of the main navigation menu so as to observe whether students would be aware of hypertextual local and global coherence ties during their translation process (Storrer 2002). Students had to turn in their localization project within a week and submitted the entire assignment in electronic format. They used the technology tools of their choice, but worked with no previous translation memory. They received course credit for their work.

The comprehensive descriptive study of the corporate website genre in Jiménez-Crespo's doctoral dissertation (2008a) and publications (2009a, 2009b, 2008b, 2010, 2011) provided the quantitative and qualitative comparative data for the top-down approach. These corpus-based contrastive studies offered a compendium of the main textual, discursive and terminological differences between original and localized corporate websites. Among the many *moves* (Swales 1990; Askehave and Nielsen 2005) or *communicative sections* (Gamero Pérez 2001) in a corporate website, these previous studies provided a detailed contrastive analysis of *contact us* pages with online contact forms, the document translated by students. A comparable corpus of these pages was extracted from the Spanish Web Comparable Corpus (Jiménez-Crespo 2008a) with the objective of contrasting student production with a representative collection of these two distinct textual populations. They are representative of the target genre and therefore can be considered adequate for evaluation purposes (Bowker 2001: 352). Additionally, lemmatized wordlists from all corpora were produced. The following table describes the composition of the three corpora:

TABLE 1

Statistics of the Comparable and Learner Corpus of *Contact us* Pages

		Comparable Corpus	
	Learner Corpus	Original Corpus	Localized corpus
Number of contact pages	76	114	95
Number of forms	76	43	21
Words total	52,529	12,090	25,908

Following a cognitive and textual approach, we will address the most significant issues revealed by the threefold analysis of students' renderings in the learner corpus, as contrasted with the data in the original and the localized subcorpora comprising the comparable corpus. These issues are:

- textual conventions and superstructural coherence;
- how to address conceptual differences at the interlinguistic level;
- lexical priming and interference at the word level;
- space constraints in localization;
- improving the quality of the original text.

The results sections will be followed by a discussion and the potential application to localization training.

5. Textual Conventions, Superstructural Coherence and Discursive Moves

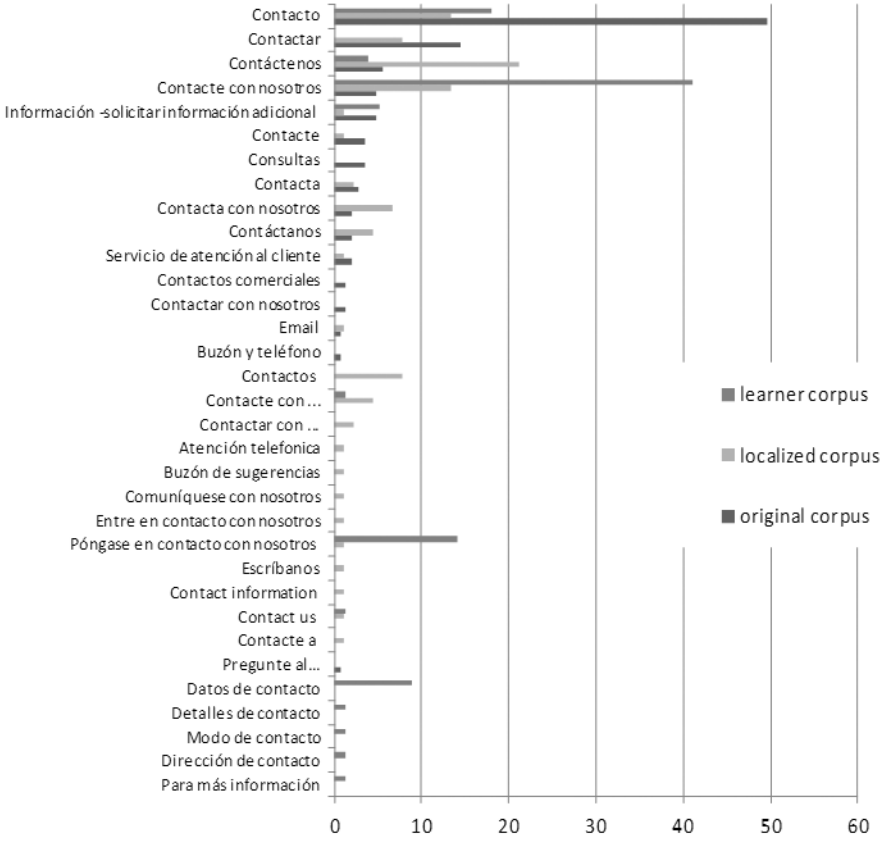
In this section, we analyze interlinguistic differences at the super, macro and micro-structural levels. The analysis begins with the most frequent communicative section in any website, the *move* that provides the contact information.

5.1. Contact us

In the source document, a reference to this move can be found both in the navigation menu and as the heading on the 'Contact us' page. We assume that each lexical unit in a navigation menu or sitemap denotes a superstructural concept in the global hypertext (Bouffard and Caignon 2006). Descriptive studies of corporate websites in English have identified that *Contact us*, the lexical unit that appears in our source text, can be considered the conventional lexicalization of this hyperstructural concept (Nielsen and Tahir 2002). It has also been found that this convention tends to be reproduced in localized sites through calques of this source unit (Jiménez-Crespo 2009a; Bouffard and Caignon 2006). Figure 2 shows the analysis of usage frequencies for lexical units that correspond to this metatextual concept in the three corpora.

FIGURE 2

Contrastive Results for the Translation of Terms Associated with the *Contact us* Move



The significant denominative variation found in students' translations shows that the textual structure of hypertexts is to a great degree ignored: even though the lexical unit contacto represents a metatextual concept in the internal structure of hypertexts, only 17.94% of students identified the relationship between both segments in the source text. In this case, the range of denominative variation shows similarities with the localized corpus, although the frequencies are more concentrated than in the latter corpus. However, as can be observed, the frequencies in each corpus tend to distinctively concentrate around a specific lexical unit, such as contacto (50%) in the original corpus, the direct calque contacte con nosotros (41%) in the learner corpus, or contáctenos (21%) in the localized corpus. Additionally, both translational corpora show greater dispersion of frequencies, a reflection of greater denominative variation as a result of creativity.

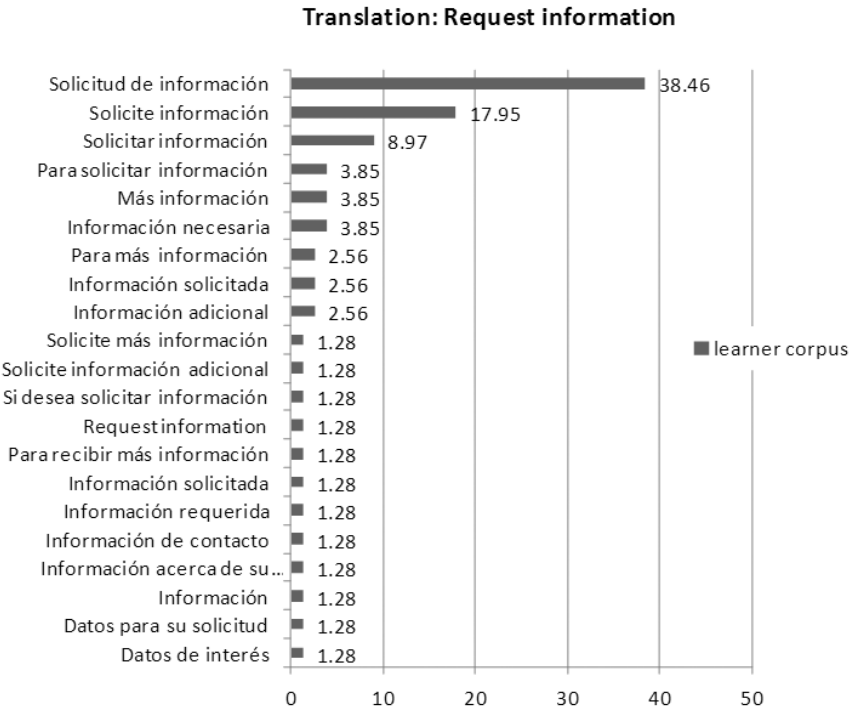
First of all, these results suggest that students need to familiarize themselves with basic textual aspects of digital genres, such as different levels of coherence in hypertexts, both at the global and local level (Storrer 2002), or in other words, with regards to the global site or the specific page. In this regard, the analysis of prototypical textual structures of digital genres can assist in providing the required superstructural awareness. Moreover, this type of pre-translation task could concentrate on the analysis of conventional features of these genres, as terminology use in students' renderings clearly diverges from the conventional terms used in original texts. This necessity is stressed by the fact that students produce an array of linguistically correct expressions such as póngase en contacto con nosotros (14.1%) and datos de contacto (8.97%), which nonetheless contradict the prescription of being concise in order to facilitate screen reading (Nielsen and Loranger 2006).

It is also interesting that 14.09% of student renderings are lexical units neither found in the original nor in the localized corpora, thus indicating that most students did not identify this segment as a genre-specific conventional feature and proposed creative solutions. The frequency of lexical units that do not appear in the original corpus is similar to that found in the localized corpus (18.88%), thus pointing to similarities between learner and localized corpora. These translations could be considered as features of translated language, that is, adequate and valid translated forms which, however, do not appear in spontaneously produced exemplars of the same genre in the target language.

5.2. Request Information

The *Request Information* heading occupied a key position within the page and, since it functions as a title, the more adequate translation would reflect Spanish conventions for titles, where a noun phrase is more conventional than the use of a verb in infinitive or imperative mood.

FIGURE 3
Student Translations for the *Request information* Heading

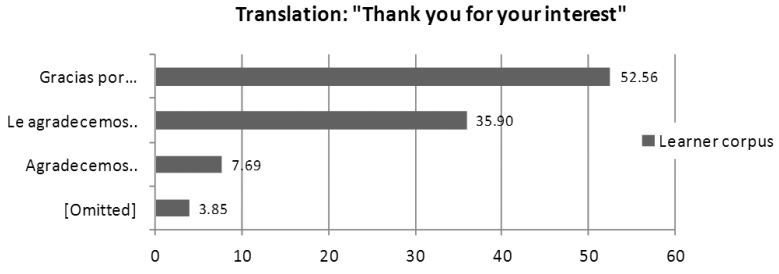


The graph shows that, even though there is a clear preference for the translation of the imperative mood into a noun phrase, the use of the infinitive and imperative is also significant, with 35.89% of renderings including verbal forms. It is also of interest that a single source segment led to a great variety of target segments, twenty-one in all.

5.3. Thank you for your interest

The translation of this segment allowed us to observe whether students were aware of intercultural differences revealed by the rhetorical structure of this type of document. In the original Spanish corpus, only one text (0.87%) included a move thanking the users for their interest in requesting information. When this segment occurs in original Spanish contact pages, it appears at the end of the paragraph and not at the beginning as in the source text. Thus, segments containing lexical structures thanking the users for their interest could be considered to be the result of a discursive strategy that is unconventional in the target culture.

FIGURE 4
Student's translations of the segment *Thank you for your interest*



Despite this near absence in the Spanish original corpus, students adopted a literal approach to its translation, with only 3.85% of the students omitting this segment. This latter strategy would provide a more adequate option for the production of a pragmatic translation suited to the target sociocultural conventions. The other possible solution provided by students, using a construction with the Spanish verb agradecer (43.59%), can be found in 1.72% of original texts. This fact suggests that localization trainees in our study directly cloned the textual structure of the source text, even though in doing so, they might incorporate non-conventional discursive strategies in the target language. In this case, localization training also has to account for the impact of translation memory tools during the translation process favoring a simple sentence-by-sentence replacement, disregarding discursive and textual aspects (Jiménez-Crespo 2009b; Shreve 2006).

5.4. Are you an existing customer?

This segment was quite problematic as shown by student translations. The main difficulty arises from the two parallel communicative processes that are established in any website: interactivity between the client and the website and interaction between the client and the company (Janoschka 2003). The communicative context of all forms can be characterized by a dialogic exercise in which a sender and the receiver exchange turns in asking and answering (Gülich 1981: 329). However, the sender of any interactive segment can be the company or the website, and consequently, a valid translation for this segment requires the identification of the participants in this communicative setting (Jiménez-Crespo 2010). Out of all students, 10.53% consider that context involves the client-website interaction and wrongly translate it as ¿Es un cliente registrado? (*Are you registered as a client?*), ¿Está ya registrado? (*Are you already registered?*), ¿Está registrado? (*Are you registered?*). This indicates that some students might be unaware of the specific characteristics of the communicative situation in which these websites are contextualized. The implications of this behavior point out that students need to develop awareness about the parallel communicative processes that take place during the interaction with any website.

Some students identify the participants in this communicative exchange, but still present problematic renderings of this segment such as ¿Es un cliente conocido? (*Are you a known client?*), ¿Es un cliente habitual? (*Are you a regular client?*), ¿Es un antiguo cliente? (*Are you an old client?*). As in most other segments analyzed, there are

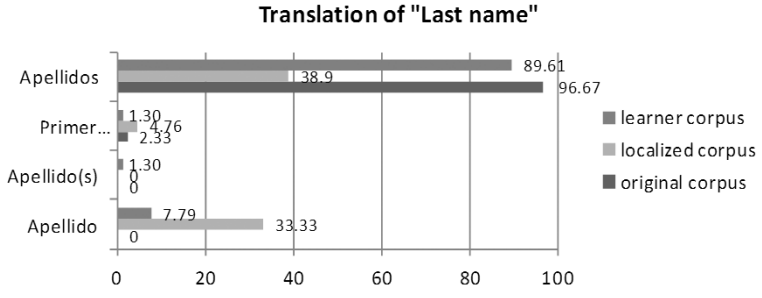
some creative solutions such as ¿Forma ya parte de nuestro grupo de clientes? or ¿Figura entre nuestro grupo de clientes? (*Are you already part of our group of clients?*), ¿Es cliente nuestro? (*Are you a client of ours?*).

6. Addressing Conceptual Differences at the Interlinguistic Level

One of the difficulties while translating web genres is how to address format problems when there is a conceptual gap in one of the languages. This is the case of the translation of the lexical unit *Last name* as a field in the form, as Spanish sociocultural norms require the use of two last names. The following table shows the contrastive analysis of lexical units used for this field.

FIGURE 5

Comparative Analysis of the Translation of the Segment *Last name*



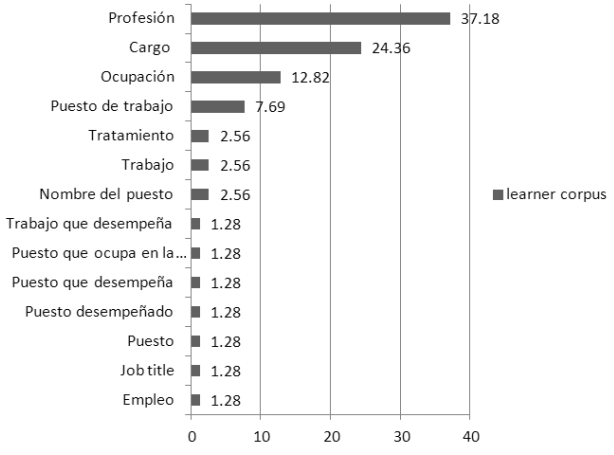
Interestingly, the social convention of using two last names in the Spanish speaking countries is overlooked as shown by the localized corpus (33%), and to a lesser degree, the learner corpus (7.79%).

The analysis of students' renderings shows that they provide a wide range of possible solutions, such as (a) using only one last name, (b) both last names, (c) indicating the field requires only the *first last name*, or (d) using parenthesis to show both possibilities, as in apellido(s). The last two options represent creative solutions to the translation problem, as renderings intend to reconcile both source and target sociocultural norms. When contrasted with the original corpus, it should be noted that the combination of a field with both *Name and Last name* is preferred over a distinct field for each component. This example illustrates how technology often hinders the production of the most suitable solution in the superstructural level, since the translator is constrained by the preexisting database structure, and the technology tools employed favor the mere cloning of the source text structure.

6.1. Job Title

It was expected that the translation of the lexical unit in this field would be somewhat problematic due to the near absence of this segment in the original corpus. This section appears only in 11.63% of original websites and in 38% of localized ones. In both corpora the most frequent Spanish lexical unit is cargo, while students renderings are as follows:

FIGURE 6
Students' Translations of the Segment *Job title*



As shown in the figure, only 24% used the correct term, cargo, whereas the most frequent solution is the inadequate term profesión (37.18%). Again, options such as puesto que ocupa en la empresa, and puesto que desempeña, though correct from a linguistic point of view, are too lengthy for the conventions of this web genre. Given that student translations of this segment resulted in seven different nouns, it can also be deduced that a non-conventional item may lead to higher variation in the range of target terms used. The data point to the appropriateness of using descriptive genre studies of this particular kind of subgenres in order to familiarize students with textual conventions in different languages.

7. Lexical Priming and Interference at Word Level

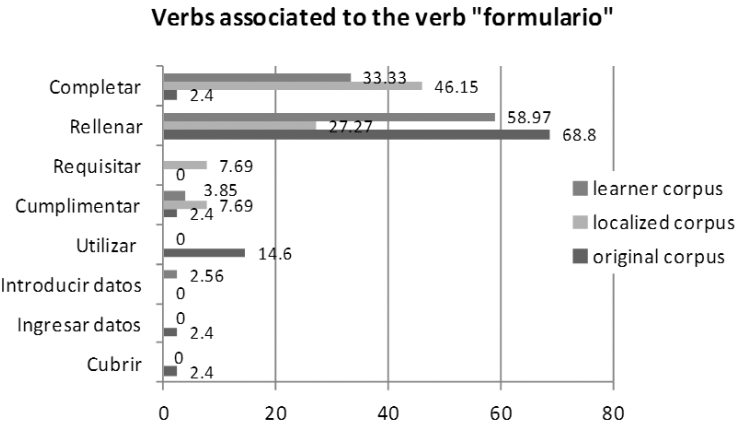
Lexical priming, i.e., lexical facilitation understood as a psychological phenomenon (cf. among others, Feldman, Soltani *et al.* 2004; Hoey 2006), is a phenomenon analyzed at the interlinguistic level in this study. The hypothesis proposed is that data from the localized corpus and from students show a preference towards producing graphically similar words in the translated texts, a phenomenon that is observed when comparing specific renderings in both localized corpora with those in the original corpus. This points out that the phenomenon is inherent in translation-oriented processing tasks relating to source texts and becomes exacerbated by onscreen translation and the fact that the translator usually works with one segment at a time.

For instance, the choice of the term compañía (28.58%) in the localized corpus is a clearly facilitated word because of its graphic similarity with the original *company*. This is however an absent option in the original Spanish corpus. In the latter the conventional term associated with this move is empresa, even though these two terms are considered synonyms. This behavior shows a striking similarity in student solution, as compañía appeared in 24.35% of translations while empresa appeared in 74.75%.

This lexical priming effect can also be observed in the translation of verbs indicating interaction between the web sender and the user. For the conventional collocation *Complete the form*, the following renderings were found:

FIGURE 7

Contrastive Analysis of Verbal Forms in Spanish Associated to the Collocation *Complete the form*



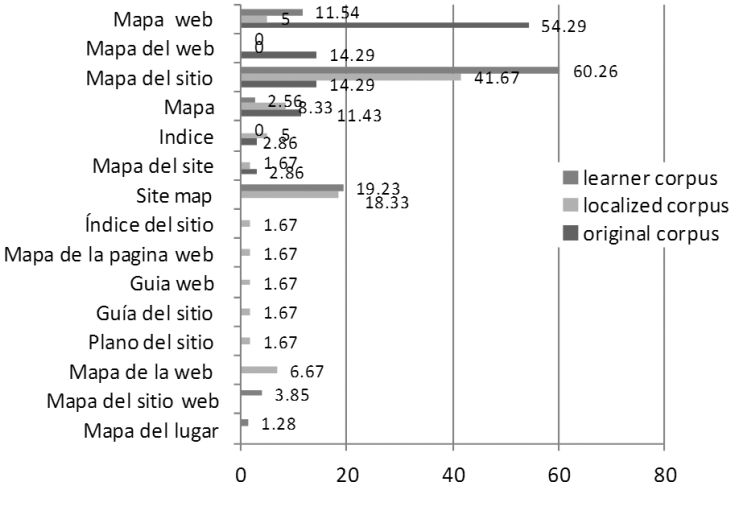
Completar, the closest Spanish rendering for the English verb *complete,* is not significantly represented in the original corpus (2.84%), but is very common in the learner (33.33%) and localized corpora (46.15%). The graph also shows the presence of lexical units in the original and localized corpora which are not present in the learner corpus: requisitar, cubrir.

7.1. Site Map

This segment shows a similar tendency to the previously analyzed move. The majority of students use the direct calque mapa del sitio (60%), a very rare terminological option in the original corpus, which shows a preference for mapa web (55%). As seen by its correspondence with the localized corpus, the preferred lexical unit by students could be considered a direct calque of *Site map.*

FIGURE 8

Contrastive Analysis of Denominative Variation for the Term *Site map*



7.2. Industry Sector

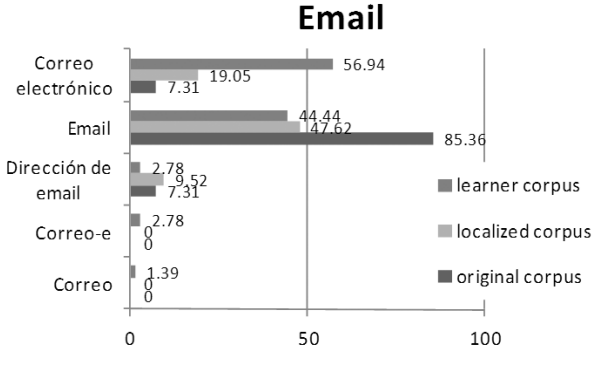
Student renderings for this lexical unit also show the extent to which the strategies used corresponded to direct calques of the source text, even when *industry* in the original text context corresponds in Spanish to empresarial. The noun industria or the adjective industrial was inadequately included in 57.69% of translations, thus indicating that, in case of difficulty, students tended to propose close calques of the source term.

7.3. Email

The opposite tendency to lexical priming has also been observed, as shown when the graphically closest form to the original is avoided, as seen in the following cases: (a) the translation of the form field *Email* and (b) the term *support*. In the first case, the use of the borrowing *email* is the preferred term in the original Spanish corpus (85.36%), while both student and localized forms show a clear preference for the translation correo electrónico. Student choices therefore show an even more pronounced tendency to eliminate the English form *email*, the form favored in the original corpus. This tendency might be a result of overcorrection in student renderings.

FIGURE 9

Contrastive Analysis of Denominative Variation for the Term *Email*



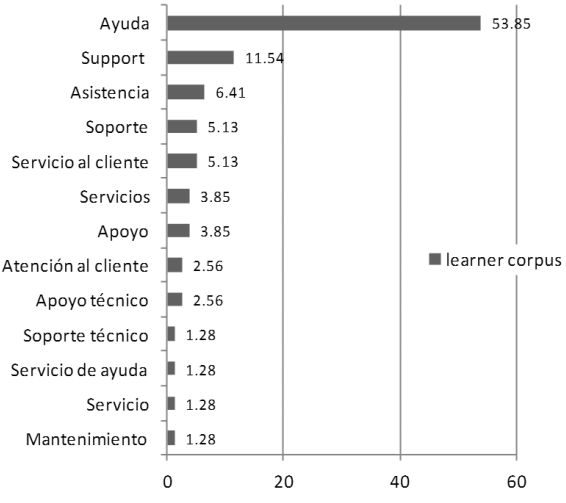
The preference towards the use of the accepted borrowing *email* in the original corpus might also be a reflection of the tendency towards linguistic economy, a conventional linguistic feature of most web genres.

7.4. Support

The translation of the English term *support* for soporte in Spanish, instead of asistencia o ayuda can be considered a recurring anglicism in localized texts, appearing in the twenty-first position in the overall lemmatized list for the localized corpus. This term appeared in the foot navigation menu and its translation was of interest in order to observe the extent to which students would incorporate this recurring anglicism. The results show that students mostly chose the noun ayuda (53%), while a mere 6.41% decided to use the otherwise frequent anglicism soporte in their translations.

FIGURE 10

Student Translations for the Term *Support*



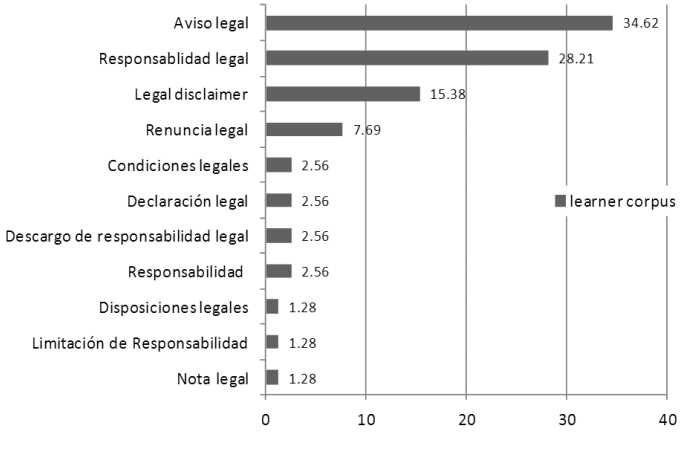
This shows that, as in the cases of *Last name*, *Email* or *Complete the form*, students produced more adequate renderings than the average professionally localized site. Nevertheless, the difficulty in this segment can be observed by the fact that 26.96% of instances in the student renderings are units found neither in the original nor in the localized corpus. As observed with the segments *Job Title* or *Reset*, it could be the case that difficulty in the translation results in a higher number of possible alternatives produced by students, even when there might be a conventional item in the target language to formulate the same communicative goal.

7.5. Legal Disclaimer

Approaching the translation of this heading is not an easy task, since conventions for such a move vary greatly from locale to locale and data from the original and localized corpora show important differences.

FIGURE 11

Student Translations for the Segment *Legal Disclaimer*



Once again, the greater tendency towards terminological variation in student translations can be associated with a higher degree of difficulty in the resolution of this problem. Only 34.62% of students identified the most conventional form in Spanish corporate websites, aviso legal (Jiménez-Crespo 2008a), while the remaining solutions represent non-conventional lexical units in Spanish navigation menus. Students refer to this type of conventional legal genre in corporate websites using seven different nouns: aviso, renuncia, descargo, declaración, condiciones, nota and disposiciones. In this case, we would like to suggest the need to incorporate the use of carefully selected virtual or disposable corpora (Varantola 2003) during localization training. If students had had access to a corpus of original websites, they could have quickly identified the preferred collocations of the word legal, providing quantitative and qualitative data to discard non conventional items such as renuncia legal or declaración legal, non-existent in spontaneously produced texts of this particular digital genre. As an example, the following figure presents an extract from the concordance analysis of the word legal in the original contact page subcorpus.

FIGURE 12

Extract from the Concordance Analysis for the Word *legal* in the Corpus of Original Spanish Websites

N	Concordance
1.	- All Rights Reserved | Página Principal | Aviso **Legal** | El hecho de cumplimentar este
2.	El presente aviso legal (en adelante, el "Aviso **Legal**") regula el uso del servicio del portal de
3.	Todos los derechos reservados *AVISO **LEGAL*** Carburos se reserva el derecho de
4.	de los términos que conforman este Aviso **Lega**l, así como cualquier cuestión relacionada con
5.	era de las obligaciones derivadas del presente Aviso **Legal,** de las correspondientes condiciones
6.	de todas las condiciones incluidas en este Aviso **Legal**, debiendo leer el Usuario atentamente el
7.	l Portal de conformidad con la Ley y el presente Aviso **Legal,** debiendo responder de los daños y
8.	el Portal de conformidad con la ley, el presente Aviso **Legal,** las Condiciones Particulares de ciertos
9.	a los términos que se estipulan en este Aviso **Legal**. Por ello, si
10.	sustituyen, completan y/o modifican el presente Aviso **Legal**. Por lo tanto, con anterioridad a la 64
11.	sustituyen, completan y/o modifican el presente Aviso **Legal**. Por lo tanto, con anterioridad a la
12.	sustituyen, completan y/o modifican este Aviso **Legal.** Por tanto, con anterioridad al acceso y/o
13.	Sociedad de la Información. Aviso e Información **Legal** NETQUEST es una marca registrad
14.	Consulting (Worldsites network) Información **Legal** - Protección de Datos "= ""4.0"" and
15.	atículos | Buscar | Contacto Contacto | Ayuda | Nota **legal** *Está en*:
16.	de Madrid. CERRAR Contacto | Ayuda | Nota **legal** *Está en*:

These concordances show that the only two possible alternatives appearing in original texts are aviso legal, and less frequently nota legal.

8. Space-constrained Segments in Localization

One of the idiosyncrasies of software and web localization alike is the need to fit translation into a limited space for both programming purposes and to facilitate reading in space-constrained screens. In this section we analyze data from two typical sections containing such idiosyncratic feature: (i) translation of controls and (ii) embedded text in graphics.

8.1. Translation of Controls

Translation of controls poses a difficulty, mainly due to the lack of contextual knowledge to identify the real function of the control, the space limit, and the differences in register at the interlinguistic level.

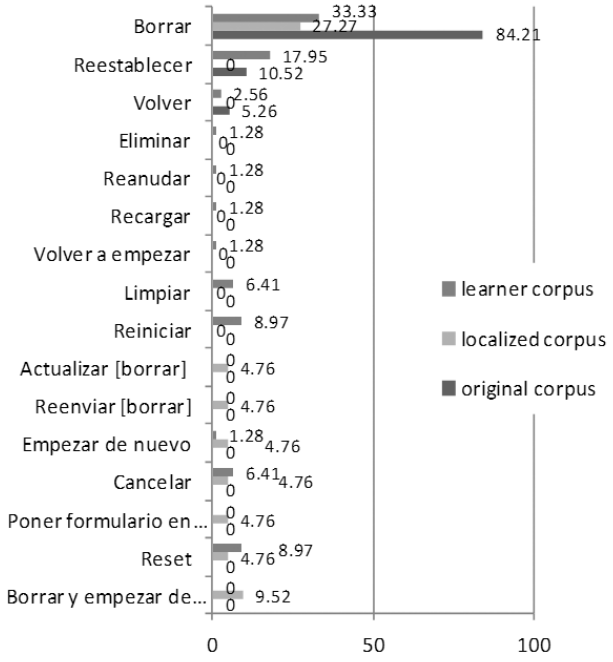
8.1.1. Submit

The button *Submit* was translated as Enviar (83.3%), Submit (9%), Aceptar (5.1%), Continuar (1.3%) and Suscribirse (1.3%). Thus, students' solutions seem to clearly concentrate on the most conventional form in original Spanish online forms that expresses this communicative purpose, enviar (93%).

8.1.2. Reset

However, the opposite can be said of the button *Reset*. It entailed a higher level of difficulty as demonstrated by the greater number of possible renderings. In all, students produced the following translations:

FIGURE 13
Contrastive Analysis of the Segment *Reset*



Conventionally, original Spanish forms show a clear preference for borrar (84%), while the frequency of this term is significantly lower in both the learner (33%) and the localized (27%) corpora. These latter corpora show a similar distribution of frequencies among a number of alternative lexical units, an exclusive phenomenon in localized texts that could again be identified within the appearance of non conventional items.

Additionally, a morphological priming effect can be observed in this case as 43.3% of student renderings of *Reset* included a target lexical unit with the prefix *re-*, such as restrablecer, reiniciar or recargar.

8.2. Embedded Text in Graphics

Even when students were instructed to localize the whole HTML document into Spanish, only 54.23% of students adapted the embedded text appearing in an image.

FIGURE 14
Localization of Text Embedded in Graphics with the Different Options in Student Renderings

Among those images that were localized by students, 25.92% of students expanded the original four word segment (27 characters) in English to five words (40 characters). Only one student reduced the number of words to suscripción, producing a more pragmatically adequate solution regarding web usability guidelines (Nielsen and Loranger 2006) in this case.

We believe that a reflection upon the instrumental competences (PACTE 2005) in order to be able to work with issues such as adaptation of text in graphics is needed as well as the study of means for facilitating such tasks at the design level.

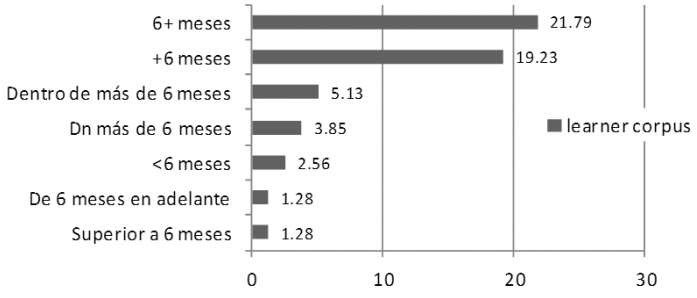
9. Improving the Quality of the Source Text in the Translation

The lack of typographic consistency is a recurring issue in most web texts (Jiménez-Crespo 2009b, 2008a). The typography of the source text provided was not consistent concerning the use of lower or upper case for the segments in the dropdown menus. There were seven drop-down menus in the source text, where five segments were capitalized (*Select*) and two segments were not (*select*). Since students used the tools of their choice, we predicted that this lack of consistency would be maintained when the 100% segment match would be suggested. However, 65.38% of students (51) were consistent in the use of lower case, thus improving the quality of the source text, while the remaining 34.62% replicated the inconsistency in their translations. This result stresses the need to pay attention to consistency and appropriateness in the translation, regardless of the inconsistencies in the source text.

9.1. Typographic Anglicisms

The question *When do you plan to make a decision*? contained several options in a drop-down list menu, such as *Immediately*, *1-3 months*, *3-6 months* and *6+ months*. The translation options of this latter item are shown below:

FIGURE 15
Analysis of Typographic Loans in Students' Translations of *6+ months*



The graphic anglicisms (Martínez de Sousa 2003) *+6 months* and *6+ months* were used by 19.23% and 21.79% of students respectively, while 2.56% of students used the more than (>) sign in order to convey this meaning. The use of the typographic anglicism 6+ meses was especially significant given that this follows the English syntactic construction, *six or more months*, while the Spanish syntax would require placing the plus sign in front of the number, +6 meses, and the most frequent solution was the use of the preposition más instead of the plus sign, más de 6 meses (43.49%).

9.2. Technical Conventions: Numbers

Knowledge of technical conventions plays an essential role in localization and technical translation. In the English source text used for the learner corpus, the letter *k* is used as an abbreviation to indicate thousand units, such as in *500k*. 15.38% of students transferred this convention to their target text and wrongly included the letter *k* in their Spanish translations. Others (5.26%) interpreted the amount as pounds, and some (14.47%) converted the amounts into euros. In some cases the conversion between dollars and pounds and euros was erroneous. These renderings stress the need to incorporate aspects of treatment of technical conventions during localization training, as conventions for the translation of currencies may differ from the conventions in other contexts. Additionally, the confusion between dollars, pounds and euros clearly indicates the importance of providing students with detailed *translation briefs* (Nord 1997) or cues to identify contextual issues prior to engaging in actual localization practices.

9.3. Cognitive Shifts

There is a tendency towards metonymically calling the *site* (sitio) página (*page*) in Spanish as reflected by student renderings, in line with what happens in oral use. In the translation of the segment *visit other [company's] sites*, 46.26% of students used the term página for *site*. This finding stresses the need to teach the idiosyncrasies of the globality of a website, the hypertextual features and its implications in the teaching of term coherence. Students used an array of inadequate options to denote the concept *site* such as, página (46.26%), webs (5.9%), portal (2.9%), dominio (1.49%) and versión (1.49%), even when the source text clearly indicated the word *site*.

9.4. Adaptation to a Global Audience

A key issue regarding the adaptation to particular users is the identification of the range of potential users of a website. In principle, any website can be accessed by anyone from anywhere, which means that the access point to a website, and hence issues such as phone country codes, should not be taken for granted. Spanish is a world language, and users will seek out the Spanish version of the website from many different places in the world. The internationalized source text included a phone number that contained the country code for Spain, and how students would handle this issue was identified as a variable related to knowledge about adaptations for global audiences. 13% of students' translations excluded the country code for Spain (+34), even when some final users of the localized text would have to use this code in order to call.

10. Discussion

Through the combination of learner, localized and original corpora, this study set out to study the various textual, pragmatic and discursive issues that pose difficulties to translation students. Variation in student terminological options as contrasted to the original and localized corpora has been studied in light of their difficulty in identifying conventional forms and as a result of creativity.

It has been shown that that there is a general tendency towards lexical priming in the learner corpus, as students consistently chose certain lexical units that resemble graphically the source form, such as compañía for *company*, even when original texts in Spanish overwhelmingly prefer the synonym empresa. This is also the case with certain collocations such as *complete the form* that some students and professional translators rendered as complete el formulario. Nevertheless, this tendency does not always run parallel to that of professionally localized texts as, in certain cases, such as the calque of *support*, students produce more frequently correct renderings such as ayuda or atención al cliente.

Another aspect of interest is that the analysis of prototypical textual structures of digital genres must be addressed, as it has been shown that using web pages as textual units in localization training could be inadequate (Jiménez-Crespo 2008b). This focus does not provide students with the overall superstructural and macrostructural knowledge associated with the different levels of coherence in hypertexts, from the global site to the macrostructural elements within a page. This implication results from the fact that even when students worked throughout the semester with the entire site, in 17.94% of the students' localization there was a lack of intratextual coherence between the navigation menu and the title of the page, *Contact us*. Additionally, the use of entire websites as the minimal textual unit in localization training (Jiménez-Crespo 2008b) was connected to the translation of the segment *visit other [company] sites* that students interpreted as páginas, dominios o versiones, illustrating that the nature of the entire website as a textual unit was not identified. This has clear implications in terms of hyperstructural coherence in websites (Storrer 2002).

As a teaching tool, we suggest the use of virtual or disposable corpora during localization training, in part to discourage students from exclusively using the Web as a corpus in order to validate their solutions (Jiménez-Crespo and Tercedor 2010).

In general, the prevailing anglocentric focus of the Web can be detrimental to the goal of finding adequate solutions to localization problems. Currently, most inadequate translations, such as direct calques, unconventional terminology, translationese, etc., appear with high frequencies on the Web as witnessed by the texts represented in our localized corpus. Students frequently accept high frequencies in terminology or phraseological searches on the Web as a straightforward validation to their proposed solutions. In order to control the influence of translated texts on the web, we have suggested the use of carefully constructed corpora made up of spontaneously produced texts in the target language in localization training. The translation of *legal disclaimer* provided a clear example of how this could assist in providing more natural translations: using legal in a word search in the original corpus provided the two most adequate solutions, aviso legal and nota legal, as well as information about the higher frequency of use of the former. The use of corpora could also be extremely helpful in order to adjust the discursive structure of the target text, as revealed in the case of *thank you for your interest*. In this case, a simple search for gracias *thanks* or the lemma agradec* *to thank* on the original corpus could help students identify that this discursive strategy is unconventional in original Spanish texts.

Finally, even though all students have completed courses in technical translation, there was a tendency towards calquing not only lexical and syntactic structures of the source text, but also technical and typographic conventions such as the use of *K* to indicate thousands or the structure *6+ months*. It is therefore essential not to lose sight of technical translation skills, such as using technical writing and Web style manuals. The need to use these manuals is evidenced by the students' tendency to produce longer translated segments when conciseness and brevity are key to effective Web style (Nielsen and Loranger 2006).

11. Conclusions

Localization training is increasingly being incorporated into most university training programs and the importance of the Web as a source of translation materials cannot be underestimated. The youth of this new situation has meant that trainers have usually organized their courses in a relative void due to the lack of empirical studies and solid theoretical frameworks. As functional aspects of localized texts can be more objectively assessed and identified than linguistic or pragmatic issues (Dunne 2006), most training programs have concentrated on the acquisition of the instrumental competences required for this translation type, such as knowledge of CAT and localization tools, document formats, encoding issues, etc. This study has shown that the combination of different corpora can offer a more fine-grained analysis of the aspects that shape the specific linguistic and extralinguistic translation proper subcompetences needed for a comprehensive approach to localization training. The results obtained have shown that textual, discursive and communicative aspects of digital texts require systematic training as clear differences have been identified between original texts and students' translations, as well as patterns of problems and solutions that do not always mirror those of professionally localized texts.

As a concluding remark, we would like to mention that the role of Web texts in society will only continue to grow. As a result, localization training as part of trans-

lation training needs to fulfill the demand for highly qualified translators who become expert not only in the use of technology tools, but also in the intercultural and conceptual differences at the linguistic, textual and discursive levels. We hope that this study will set the foundation for more empirical research into localization training, as the training to produce functional localized texts requires a holistic view of the localization process within translation and the under-researched notion of localization competence.

ACKNOWLEDGMENTS

This research has been partly funded by grant FFI2011-23120 from the Spanish Ministry of Science and Innovation.

REFERENCES

Askehave, Inger and Nielsen, Anne E. (2005): Digital genres: a challenge to traditional genre theory. *Information Technology and People*. 18(2):120-141.

Bell, Robert T. (1991): *Translation and Translating: Theory and practice*. London: Longman.

Bouffard, Paula and Caignon, Philippe (2006): Localisation et variation linguistique. Vers une géolinguistique de l'espace virtuel francophone. *Meta*. 51(4):806-823.

Bowker, Lynne (2001): Towards a Methodology for a Corpus-Based Approach to Translation Evaluation. *Meta*. 46(2):345-364.

Bowker, Lynne (2002): Computer-Aided Translation Technology: A Practical Introduction. Ottawa: University of Ottawa Press.

Brooks, David (2000): What Price Globalization? Managing Costs at Microsoft. *In*: Robert. C. Sprung, ed. *Translating into Success. Cutting-edge Strategies for Going Multilingual in a Global Age*. Amsterdam/Philadelphia: John Benjamins, 42-59.

Dunne, Keiran (2006): Putting the cart behind the horse: rethinking localization quality management. *In*: Keiran Dunne, ed. *Perspectives on Localization*. Amsterdam: John Benjamins, 95-117.

Esselink, Bert (2006): *The evolution of localization*. *In*: Anthony Pym, Alexander Perekstenko and Bram Starink, eds. *Translation Technology and its Teaching*. Tarragona: Intercultural Studies Group, 21-30.

Feldman, Laurie B., Soltani, Emilie G., Pastizzo, Matthew J., *et al.* (2004): What do graded effects of semantic transparency reveal about morphological processing? *Brain and Language*. 90(1-3):17-30.

Gamero Pérez, Silvia (2001): *La traducción de textos técnicos*. Barcelona: Ariel.

Gülich, Elisabeth (1981): Formulare als Dialoge. *In*: Ingulf Radtke, ed. *Deutsche Akademie Für Sprache und Dichtung, Der öffentliche Sprachgegrauch (Band II): Die Sprache des Rechts und der Verwaltung*. Stuttgart: Klett-Cotta, 322-356.

Harris, Brian (1977): The importance of natural translation. *Working Papers on Bilingualism*. 12(9):6-114.

Hoey, Michael (2006): *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Janoschka, Anna (2003): *Web Advertising*. Amsterdam/Philadelphia: John Benjamins.

Jiménez-Crespo, Miguel A. (2008a): *El proceso de localización web: estudio contrastivo de un corpus comparable del género sitio web corporativo*. Doctoral dissertation, unpublished. Granada: University of Granada, Spain. Visited on 10 May 2010, <http://hera.ugr.es/tesisugr/17515324.pdf>.

Jiménez-Crespo, Miguel A. (2008b): Web texts in translator training. *Current Trends in Translator Teaching and Training*. 2:30-68.

Jiménez-Crespo, Miguel A. (2009a): Conventions in localization: a corpus study of original vs. translated web texts. *Jostrans: The Journal of Specialized Translation*. 12:79-102. Visited on 10 May 2010, <http://www.jostrans.org/issue12/art_jimenez.php>.

Jiménez-Crespo, Miguel A. (2009b): The effect of Translation Memory tools in translated web texts: evidence from a comparative product-based study. *Linguistica Antverpiensia*. 8:213-232.

Jiménez-Crespo, Miguel A. (2010): The intersection of localization and translation: a corpus study of Spanish original and localized web forms. *Translation and Interpreting Studies*. 5(2):186-207.

Jiménez-Crespo, Miguel A. (2011): To adapt or not to adapt in web localization: A contrastive genre-based study of original and localised legal sections in corporate websites. *Journal of Specialised Translation*. 15:1-27. Visited on 4 March 2012, <www.jostrans.org/issue15/art_jimenez.pdf>

Jiménez-Crespo, Miguel A. and Tercedor, Maribel (2010): Theoretical and methodological issues in web corpus design and analysis. *International Journal of Translation*. 22(2):37-57.

Jiménez-Crespo, Miguel A. and Tercedor, Maribel (2009): Evaluation in Localization Training: Assessing the acquisition of instrumental and transfer competence through a corpus-based study. Conference paper presented at Monterrey Forum on evaluation 2009, Monterey Institute of International Studies (Monterey, April 8-9, 2009).

Kennedy, Alistair and Shepherd, Michael (2005): Automatic Identification of Home Pages on the Web. *Proceeding from the XXXVIII Annual Hawaii International Conference on System Sciences*. (Maui, Hawaii) Los Alamitos, CA: IEEE-Computer Society.

Kiraly, Don (1995): *Pathways to Translation: Pedagogy and Process*. Kent: Kent State University Press.

Lommel, Arle, ed. (2004): *Localization Industry Primer.* 2nd ed. Geneva: The Localization Industry Standards Association (LISA).

Lommel, Arle, ed. (2007): *Localization Industry Primer.* 3rd ed. Geneva: The Localization Industry Standard Association.

López, Clara Inés and Tercedor, Maribel (2008): Corpora and students' autonomy in scientific and technical translation training. *Jostrans: The Journal of Specialized translation*, 9:2-19. Visited on 10 May 2010, <http://www.jostrans.org/issue09/art_lopez_tercedor.php>.

Martínez de Sousa, José (2003): Los anglicismos ortotipográficos en la traducción. *Panacea: boletín de medicina y traducción*. 11(4):1-6.

Nielsen, Jacob and Loranger, Hoa (2006): *Prioritizing Web Usability*. Indianapolis: News Riders.

Nielsen, Jacob and Tahir, Marie (2002): *Homepage usability: 50 Websites deconstructed*. Indianapolis: News Riders.

Nord, Christiane (1997): *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome.

PACTE (2005): Investigating Translation Competence: Conceptual and Methodological Issues. *Meta*. 50(2):609-619.

Pym, Anthony (2006): *Localization, Training, and the Threat of Fragmentation*. Visited on 10 May 2010, <http://www.tinet.org/~apym/on-line/translation.html>.

Shreve, Gregory M. (2006): Corpus Enhancement and Localization. *In*: Keiran Dunne, ed. *Perspectives on Localization*. Amsterdam/Philadelphia: John Benjamins, 309-331.

Sprung, Robert C. (2000): Introduction. *In*: Robert C. Sprung, ed. *Translating into Success: Cutting-edge Strategies for Going Multilingual in a Global Age*. Amsterdam/Philadelphia: John Benjamins, ix-xxii.

Storrer, Angelika (2002): Coherence in text and Hypertext. *Document Design*. 3(2):157-168.

Swales, John (1990): *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Varantola, Krista (2003): Translators and Disposable Corpora. *In*: Federico Zannetin, Silvia Bernardini and Dominic Stewart, eds. *Corpora in Translator Education*. Manchester: St. Jerome, 55-70.

WRIGHT, Sue Ellen (2004): Localization Competence for Translation and Project Management. *In*: Eberhard FLEISCHMANN, Peter. A. SCHMITT and Gerd WOTJAK, eds. *Translationskompetenz*, Tübingen: Stauffenburg, 581-595.

ZANETTIN, Federico (1998): Bilingual Comparable Corpora and the Training of Translators. *Meta*. 43(4):616-630.

ZANETTIN, Federico (2001): Swimming in words: corpora, language learning and translation. *In*: Guy ASTON, ed. *Learning with Corpora*. Houston: Athelstan, 177-197.

ZANETTIN, Federico, BERNARDINI, Silvia and STEWART, Dominic, eds. (2003): *Corpora in Translator Education*. Manchester: St. Jerome Publishing.