The effect of Translation Memory tools in translated Web texts: Evidence from a comparative product-based study

Miguel A. Jiménez-Crespo
Rutgers University, The State University of New Jersey

Translation Memory tools have been widely promoted in terms of increased productivity, quality and consistency, while translation scholars have argued that in some cases they might produce the opposite effect. This paper investigates these two related claims through a corpus-based contrastive analysis of 40,000 original and localized Web pages in Spanish. Given that all Web texts are localized using TM tools, the claim of increased quality and consistency is analyzed in contrast with Web texts spontaneously produced in Spanish. The results of the contrastive analysis indicate that localized texts tend to replicate source text structures and show higher numbers of inconsistencies at the lexical, syntactic and typographic levels than non-translated Web sites. These findings are associated with lower levels of quality in localized texts as compared to non-translated or spontaneously produced texts.

1. Introduction: Translation Memory tools in a digital age

In recent years the use of translation tools has become an imperative for translation professionals (Alcina, 2008; Bowker, 2002). While most tools are promoted in terms of their productivity, consistency and quality, less attention has been paid to the constraints they impose on the cognitive and textual process that the translator carries out (Reinke, 2004; Neubert & Shreve, 1992; Lörcher, 1991). In fact, if translation is understood as "a communicative event which is shaped by its own goals, pressures and context of production" (Baker, 1996, p. 175), which in turn will produce text with observable linguistic features different from text originally produced in any language (Baker, 1995; Baker, 1996; Kenny, 2001; Laviosa, 2002), it is logical to claim that the use of technology tools will leave a trace in translated texts that can be quantitatively observed using a corpus-based methodology.

In the context of corpus-based translation research, the seminal work of Baker (1999, p. 285; 1995) has placed the emphasis on social, cultural, cognitive and ideological constraints. Nevertheless, it has been argued that the introduction of Translation Memory tools (TM) has changed the nature of the task they intend to facilitate (Bowker, 2002), and consequently, the impact of technological constraints should also be taken into consideration. These technological constraints have been associated in previous studies to the reduction of the translation task to a mere "sentence replacement" activ-

ity (Bédard, 2001, p. 29), in which there is a partial decontextualization of the constituent segments that make up the holistic text (Bowker, 2006). Other scholars have argued that TM use leads to a translation process that tends to operate at a microtextual level (Shreve, 2006; Macklovitch & Russel, 2000), with clear implications in terms of style, coherence, cohesion or structural configuration (Heyn, 1998). These claims by translation scholars are somewhat related to the industry's assertion that TM tools lead to higher levels of quality, which according to Reinke (2004) is simply associated with the notion of increased *consistency*. According to companies marketing TM tools, given that pre-established terminology and repeated segments will be equally translated, the target text will display a more coherent and cohesive nature, and therefore higher levels of *quality*.

Nevertheless, several scholars have argued that the above-mentioned technological constraints might produce less coherent and cohesive texts. The rationale behind this claim is that translators might incorporate sentences from texts belonging to diverse domains or genres (Shreve, 2006), produced by several translators with unique styles (Bowker, 2006, p. 181) or that translators might avoid certain cohesive devices, such as anaphoric and cataphoric references, in order to promote future reuse (Heyn, 1998, p. 135). Additionally, it is fair to argue that the industry's rationale for promoting these tools cannot be fully supported in new translation modalities, such as software or Web site localization, given that these translations are always performed with TM tools. In these modalities, there would be no *tertium comparationis* against which to contrast their improved consistency or quality.

Consequently, it seems pertinent to propose that the quality benefits of TM tools could be evaluated through a contrastive analysis with those texts spontaneously produced in any language and therefore not subject to the technological constraints of a technology-driven translation process. Such is the approach taken by this paper: the corpus of texts that will be analyzed includes original Spanish Web sites alongside localized Web sites that have been subject to the pressures of translation as an industrial activity (Sager, 1989, p. 91). A second possible approach would be comparing a corpus of Web sites that have been localized using TM tools with a corpus of texts localized without them. Nevertheless, this would be virtually impossible, mostly due to the difficulty in identifying a large representative body of professionally localized Web sites following this latter approach.

Despite recent interest in the empirical investigation into TM use (Wallis, 2008; Gow, 2003; Reinke, 2004; Sommers, 1999; etc.), the above-mentioned claims on potential effects in the process and products have not been fully empirically researched. Therefore, the goals of this paper are twofold: on the one hand, it intends to research from an empirical perspective whether TM sentence-based processing has an impact in the final textual, pragmatic and discursive configuration of the target Web sites; on the other hand, it will investigate whether localized texts are more consistent than those spontaneously produced in any given language or sociocultural

context. The first goal will be accomplished through a contrastive superstructural analysis of a comparative corpus of 40,000 original and localized Web pages. Following a genre-based model (Göpferich, 1995; Gamero, 2001; Askehave & Nielsen, 2005), a contrastive superstructural and macrostructural analysis of the corpus will be performed in order to observe whether localized Web sites maintain the textual structure of source Web sites. The second objective will be accomplished through a contrastive analysis of *inconsistencies* in localized Web sites identified through a previous study (Jiménez-Crespo, 2008a). These inconsistencies are lexical (analyzing intratextual denomintative variation, such as translated and untranslated borrowings), syntactic (addressing the user using formal-informal forms in the same text) or typographic (inconsistent capitalization of titles and neologisms).

2. Translation memory and the claim of improved quality and consistency

Translation memory tools have been used for over two decades. The number of publications that describe their possible uses in professional practice is steadily growing (L´Homme, 1999; Esselink, 2000; Austermülh, 2001; Bowker, 2002; Bowker, 2005; Corpas & Varela, 2003; Reinke, 2004; Freigang, 2005; Días Fouçes & García González, 2008), with several researchers focusing on TM evaluation and selection depending on the working environment (Höge, 2002; Zerfaß, 2002a; Zerfaß, 2002b; Rico, 2000; Webb, 1998) or the impact of TM use in translator training (Alcina, 2008; Kenny, 1999). Additionally, empirical studies on different aspects of TM use are steadily appearing (i.e. Wallis, 2008; O'Brien, 2007). Generally, most research into translation memory has adopted a process-oriented view, both from the tools' or users' perspective. Nevertheless, even when it has been previously suggested that TM tools bring about increased quality and consistency (Ahrenberg & Merkel, 1999), there is a scarcity of product-based empirical studies that compare texts translated using TM tools with those produced without them in order to validate some of these underlying assumptions.

2.1. TM tools benefits and the notion of *quality*

The use of TM has been generally associated to benefits in terms of quality, consistency, speed, improvements in the quality of the translator's experience or terminology management (O'Brien, 1998, p. 119; Webb, 1998, p. 20; Bowker, 2002, p. 117; Reinke, 2004; etc.). In particular, the evaluation of TM systems from an academic perspective has not concentrated on the claim of improved quality,[1] as this notion is controversial *per se* in Translation Studies literature (Wright, 2006; Bass, 2006). According to standards such as the ISO 9000, quality is defined as the ability to comply with a set

of parameters predefined by the customer (Bass, 2006). These definitions have been criticized as it is theoretically and methodologically impossible to predefine the notion of *quality* in all translated texts: for this reason, common definitions of quality usually focus on the procedural aspects of processes as opposed to establishing what could be considered a *quality* translated text. Basically, such definitions govern procedures for achieving quality rather than providing normative statements about what constitutes quality (Martínez Melis & Hurtado, 2001, p. 274). The only standard that describes the properties of a quality translated text is the German norm DIN 2345. This norm defines quality as the property of a translated text that is complete, terminologically coherent, uses correct grammar, appropriate style and adheres to an agreed-upon style guide. Thus, it explicitly links the notion of quality to that of *coherence* and *consistency* normally used in TM marketing efforts.

In the case of Web and software localization, the notion of quality has also been associated by the Localization Industry Standards Association with a translated text that "looks like it has been developed in-country" (Lommel, 2004, p. 11), and therefore, it is fair to argue that a localized Web site would need comply to the set of expectations shared by the target discourse community, such as the conventional genre superstructure[2] and macrostructure (Nord, 1997). However, and as mentioned previously, the textual segmentation process primes the translator to consciously or subconsciously maintain the overall textual structure of the source document, ignoring that textual structure is culture-specific (Neubert & Shreve, 1992), and exemplars of a similar genre in two different cultures might show structural differences (Shreve, 2006; Gamero, 2002). In this sense, this paper agrees with Bowker (2002, p. 117) in that "The rigidity of maintaining the same order and number of sentences in the target text as are found in the source texts may affect the naturalness and quality of the translation." Therefore, it is assumed that a translated text that maintains the source text structure might produce a negative impact in the text receiver's appreciation of quality. The fact that translated texts maintain the source text organization has been previously put forward as a general problem in the evaluation of translated texts (Larose, 1998), and it has also been associated with the perception of lower quality by the receivers of target texts (Nobs, 2005). Thus, the first working hypothesis in this study is that:

> Hypothesis 1: The use of TM tools in the localization of Web sites will increase the tendency in translation to produce texts whose superstructure and macrostructure are somewhat different from that of spontaneously produced texts in the target language due to their sentence-based processing.

This hypothesis would be also based on previous scholars' assertions that the use of TM tools might produce the opposite effect of that touted; they

might somewhat hinder the production of a *quality* target text (Wallis, 2008; Bowker, 2002; Bowker, 2006; Bédard, 2000; Heyn, 1998).

The second claim that this paper intends to investigate is that of increased consistency, a concept related to the linguistic notion of lexical and syntactic coherence (de Beaugrande & Dressler, 1981). Given that *coherence* is considered to be the most important standard of textuality in hypertexts (Fritz, 1999; Storrer, 2002), the following section reviews this notion and the importance of maintaining consistency in hypertext localization.

2.2. Texts, hypertexts and the claim of consistency

The notion of a unitary text in which translation activity is based is central in Translation Studies. From a text-linguistic approach, it has been defended that texts are the central defining issue in translation, and texts and their situation define the translation process (Neubert & Shreve, 1992). Nevertheless, the extensive use of TM tools has lead to a gradual disappearance of the single text as the operative unit in which the translation task is based. Several scholars have noted that the use of translation memory tools or global management systems (GMS) are forcing translators to work with disaggregated textual units that are not necessarily "the totality of communicative signals used in a communicative interaction" (Nord, 1991, p. 14),[3] or complete coherent and cohesive texts (Bowker, 2006). Instead, translators gradually process subtextual units that are part of a complete text that is sometimes unavailable (Mossop, 2005). From the perspective of the translation process, this has clear implications in terms of cohesion, coherence and contextual cues during source text comprehension and the subsequent textual production stage.

The decontextualization of textual segments is especially significant during the localization process (Dunne, 2006; Esselink, 2000), a translation modality that mostly deals with hypertexts. One of the main characteristics of hypertexts is the need to divide the global text into interrelated nodes or lexias that can be read on a computer screen (Landow, 1997). Due to this disaggregation of hypertexts in smaller subtextual units (the Web pages themselves or their different components such as banner ads), together with the fact that any hypertext can be accessed directly through any node and read in whichever sequence, it has been defended that coherence is their most important standard of textuality (Fritz, 1998; Storrer, 2002). This suggests that maintaining terminological coherence and consistency is crucial in order to produce high quality localizations.

In establishing coherence in hypertexts, it should be mentioned that each hypertextual page includes both *content text,* the new content that is included in each page and makes it an information and retrieval unit (Nielsen & Loranger, 2006), and *interface text*, the textual segments whose function is to structure the global unitary hypertext (Price & Price, 2002). Interface text can be identified as navigation menus, breadcrumb navigation menus, webmaps, and news columns, etc. It promotes global hypertextual

coherence through lexical repetition found in all these textual segments[4] (Jiménez-Crespo & Tercedor, in press). Given that their main function is to structure the entire hypertext as a single textual unit,[5] TM tools would in principle assist the translators in maintaining the same translation for each of the terms associated to the hypertextual superstructure, such as the conventional lexical units: *contact us*, *about us*, *privacy policy*, etc.

Nevertheless, a previous study by the author found a high number of inconsistencies in navigation terminology in localized Web sites (Jiménez-Crespo, 2008a). These were mostly found when a source lexical unit can potentially be translated in several ways in the target language, such as "about us", which can be translated using two synonymous prepositions in Spanish: *acerca de nosotros* and *sobre nosotros*. These inconsistencies were also found whenever any segment of the overall hypertext included a reference to a specific page in the global hypertext, such as "Please refer to our *privacy policy* for more information [...]". In these cases, the problem resides in the fact that the sentence-based operation of TM tools does not fully allow for sub-sentence matches to be presented to the translator (Macklovitch & Russel, 2000; Gow, 2003). Thus, even when the translation of the lexical unit *privacy policy* might be stored in the TM database as a segment;[6] a sentence that contains a reference to this page might not trigger the previously stored translation. Thus, as pointed out previously by Macklovitch & Russel (2000), many repetitions might be subsentential and, therefore, difficult to locate while localizing Web sites.

Another case of recurrent inconsistencies attributed to difficulties in retrieving subsentence matches is the translation of borrowings and calques, such as *email*, *link* or *online*. In Spanish, the use of these loanwords from English is highly extensive (Cabanillas *et al.*, 2007), but nonetheless, the translator has to constantly decide whether to use the loanword or to insert the variety of possible Spanish neologisms, such as *correo electrónico* or *dirección electrónica*, *enlace* - *hipervínculo* - *vínculo* or *en línea* respectively. It would be expected that a Web site translated with TM and terminology management tools might consistently use the same choice, and that any inconsistency could be attributed to translators' behavior during the translation task. Thus, intratextual denominative variation in the case of borrowings and calques can constitute a valid variable in order to research whether TM tools provide higher consistency at the subsentential level than spontaneously produced texts.

Finally, another case of inconsistencies while translating into Romance languages entails differences in register as reflected in the use of formal and informal pronouns and verbal forms. In localization, the overwhelming majority of translations take place from English into other languages (Lommel, 2004), and in the case of Spanish, translators facing a direct appeal to the user have to constantly decide whether to use *tú* or *usted* forms (Jiménez-Crespo, 2008a). In these cases, given that this problem is only related to pronominal and verbal choices, potential matches in the TM database would be at the sub-sentence level, and therefore, the use of TM

tools might not be useful in maintaining a consistent tone. This would lead to a syntactically and stylistically inconsistent target text.

After this brief description of potential inconsistencies, the second working hypothesis is that:

> Hypothesis 2: Due to the current inability of TM tools to effectively provide sub-segment matches and maintain consistency at certain levels, localized texts will display higher percentages of lexical, syntactic and typographic inconsistencies than texts spontaneously produced in a given language.

2.3. Web site localization and "pre-translation" TM mode

Before continuing with the description of the empirical study, it should be mentioned that globalized Web sites are normally updated using global management systems or GMS (LISA, 2007), a process that in TM terms has been identified as *pre-translation* (Wallis, 2008) or *batch mode* (Bowker, 2002, p. 112). In this case, whenever a Web site is updated, the GMS compares the entire text to the database of previous translations and extracts only those segments that do not have an exact match. This process might further accentuate the lack of consistency given that the target Web site is the product of an increasing number of translators with differentiated styles, preferences, etc. Additionally, it should be mentioned that in a previous empirical study the use of pre-translation has been preliminarily shown to produce lower levels of quality than normal interactive translations (Wallis, 2008).

3. Empirical study

The methodology used to test both hypotheses is based on the Spanish Comparable Web Corpus[7] made up of 267 original and localized corporate Web sites (Jiménez-Crespo, 2008a). This Web genre was selected as it has been previously identified as the most conventional digital genre (Kennedy & Shepherd, 2005). The Web corpus was compiled in the context of a wider research project that deals with the effects of the technological context of production of localized Web texts (Jiménez-Crespo, 2008a), and it consists of two sections: a corpus of original Spanish corporate Web sites (172 sites) and another corpus of all Web sites localized into Castilian Spanish[8] from the largest 650 US companies according to the Forbes list (95 sites). The corpus was downloaded synchronically during one single day in 2006. All texts were systematically selected from two directories, the Spanish Google Business directory and the Forbes list, so as to guarantee that the corpus would be representative of the textual population targeted. A detailed description of the corpus compilation process and composition have been

given elsewhere (Jiménez-Crespo 2008a; 2008b; 2009), and therefore, only the most important characteristics will be highlighted in the following table.

Table 1: Spanish Web Comparable Corpus description

	Original Section		*Localized Section*	
	Total	*Average*	*Total*	*Average*
Web sites	178		95	
Web pages	19,102	111.5 per site	21,322	224.3 per site
Words in page body	4,945,103	258.87 page	8,871,512	416.07 page
Words total	8,659,856	453.34 page	12,562,894	589.50 page

In order to test the first hypothesis, a textual genre model was adopted in a modified form (Gamero, 2001; Askehave & Nielsen, 2005). Each thematic unit in a Web site represented in the navigation menu or sitemap, such as *contact us* or *about us*, is identified as a unique *move*[9] in the overall structure of the hypertext (Askehave & Nielsen, 2005; Jiménez-Crespo, 2008c). Moreover, each move is subdivided into *steps*, such as the conventional *history*, *location* or *mission* pages inside the section that describes the company in corporate Web sites. Each localized Web site will be analyzed and all entries in navigation menus and webmaps will be assigned to a move or step in order to quantify the frequency of use. This will provide a detailed statistical analysis of the frequency of use of all moves and steps. This methodology was previously applied to the corpus of original Spanish Web sites, providing a descriptive quantitative and qualitative foundation for this contrastive study (Jiménez-Crespo, 2008b; 2008c). By applying this same analysis to the localized section of the corpus, it will be possible to contrast the structure of localized texts using segment-based TM tools to that of original texts produced without them.

As for the second hypothesis, the intratextual analysis of inconsistencies requires a smaller sample of texts for a more controlled analysis. This led to the creation of a smaller comparable subcorpus made up of ten original and ten localized Web sites that were randomly selected and extracted. Each Web site will also be converted to .txt format and analyzed with the lexical analysis software Wordsmith Tools.

Once this smaller sample subcorpus is compiled and processed, each Web site will be subject to the following intratextual analysis: (1) consistency analysis of the all concepts associated with the hypertextual superstructure as represented in navigation menus or sitemaps; (2) analysis of intratextual denominative variation for borrowings and calques; (3) consistency analysis of the use of upper case letters in navigation menus and neologisms; and finally, (4) a consistency analysis of the use of formal vs. in-

formal verbal and pronominal forms. The results from the original and localized texts will be compared and contrasted.

Table 2: Description of comparable subcorpus extracted from Spanish Web Comparable Corpus

	Original Section		*Localized Section*	
	Total	*Average*	*Total*	*Average*
Total Web sites	10		10	
Web pages	1984	198.4 per site	3141	314.1 per site
Words in body of pages	680,031	342.75 page	1,278,225	406.94 per page

4. Results

The results will be presented following the two distinctive stages in this study that correspond to each formulated hypothesis. The contrastive analysis of the textual superstructure will be presented first, followed by the intratextual consistency analysis designed to test the second hypothesis.

4.1. Contrastive analysis of the hypertextual structure

First of all, the contrastive quantitative analysis of the superstructure of original and localized Web sites shows that both textual profiles share the same number of possible moves or thematic units. In fact, all moves identified in the previous descriptive study on original Spanish Web sites (Jiménez-Crespo, 2008b; 2008c) appear in both corpora. This indicates that, to some extent, the internationalization of this Web genre has led to a similar number of possible moves and steps in original Spanish sites and those localized into this same language. However, the most significant finding relates to substantial differences in the frequency of appearance for several moves, such as *privacy policy* or *terms of use*. Given that, in principle, all texts are directed towards the same target audience and sociocultural context, this study assumes that any differences between both textual profiles can be attributed directly to the replication of the source text structure.

The following bar chart presents the contrastive analysis of the frequency of appearance for each move and step, and it clearly illustrates the superstructural differences between both textual profiles. It is organized according to the difference in the frequency between original and localized Web sites: the darker segment of each column represents the average frequency for moves or steps in original Web sites (FrO), the frequency of use in localized sites for the same move is represented by the total figure in each column (FrL), while the lighter segment represents the variable that reflects the difference in frequency (DF) between both textual profiles.

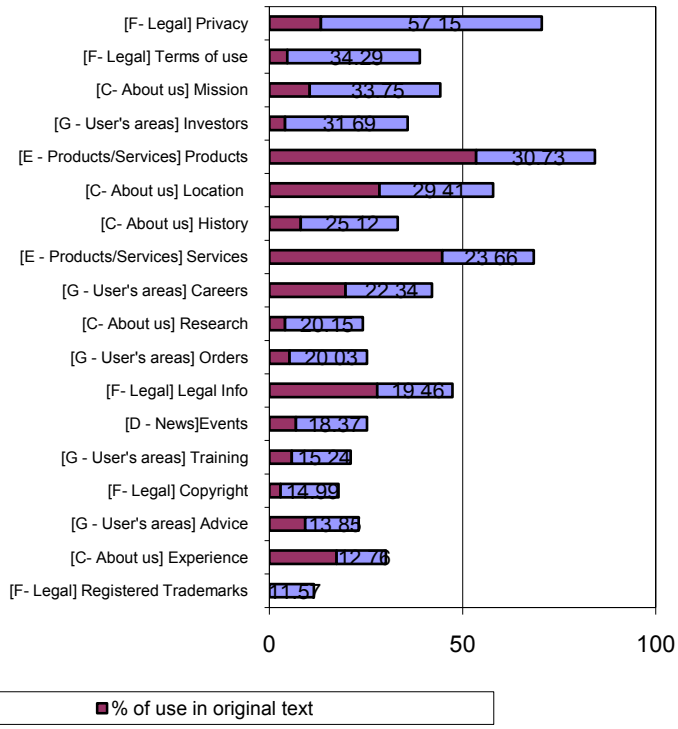

Figure 1: Superstructural differences between original and localized Web sites

The superstructural differences between both textual profiles are mostly concentrated in two moves: *legal* (F) and *about us* (C). In the latter move, it is of interest the higher frequency of the step that contains the values or mission of the company (DF=33.75%). This could be indicative of a conventionalized feature in US corporate Web sites reflecting the need to appeal to tradition and values in the US market. Nevertheless, and as shown by these results, this type of information is not conventionally offered in original Spanish sites (FrO=10.46%).

The most significant differences are concentrated in all moves or thematic units related to legal content, such as *privacy policy* (DF=57.15%), *terms of use* (DF=35.29%), *legal information* (DF=19.46%), *copyright* (DF=14.99%) and *registered trademarks* (DF=11.59%). It is fair to assume that the value of the variable DF reflects differences in the prototypical superstructure in this genre between the source and target sociocultural context, in particular, differences due to their legal systems. This finding is consistent with the results from an earlier study on corporate Web sites concluding that the most consistent difference between US corporate sites and other national sites was that in the former privacy Web pages were more frequent (Robbin & Stylianou, 2003). In fact, online privacy protec-

tion in the United States is self-regulated by companies under the guidance of the Federal Trade Commission, while this is regulated in Spain by the Spanish Data Protection Act of 1999. This means that US Web sites are required to explicitly formulate a full privacy policy, while Spanish sites simply indicate that their online privacy practices are in compliance with the above mentioned Spanish law. This can explain the high frequency (FrL=79%) of localized North American corporate Web sites including an independent privacy policy page while very few Spanish sites do (FrO=10.46%). These results prove that once US sites are localized, the structure of the source text is somewhat replicated in the target text. This indicates that the use of TM tools that operate at the page and sentence levels might promote or contribute to the *cloning* of the superstructure of source texts. This is consistent with what Larose (1998) refers to as *cloned texts*, that is, translated texts whose superstructure is fully maintained in the target text regardless of intercultural macrostructural differences for the same textual genre. Thus, the hypertextual page by page structure is somewhat maintained during the localization of Web sites, regardless of the conventional mental model of the genre structure shared by the target discourse community as represented by the Web sites produced by members of that community.

An additional analysis was performed in order to observe whether the macrostructure of pages containing legal information is also maintained in the translation process. Thus, it was observed that the average number of words in the pages with legal content was 2415.69 in localized sites, while the same average for original Spanish sites was 1074.94. This significant difference in the average number of words in legal pages (+224.72%) also points out the fact that the same sentence structure of the source texts could have been maintained. In this respect, it should be mentioned that the Web localization process is even more constrained that the translation of non-digital texts, as it requires the tag protection functionality offered by most TM tools. This additional issue could also discourage translators from seeking or implementing changes to the textual structures (as tags and/or programming code would also require restructuring).

4.2. Lexical, syntactic and typographic consistency

As mentioned previously, the consistency analysis at the lexical, syntactic and typographic levels was performed in the smaller comparable subcorpus consisting of ten original Web sites and ten localized Web sites. Following the progression noted in the methodology section, the description of the results will start with the contrastive study of lexical and terminological consistency.

4.2.1. Lexical consistency

The first analysis in this category involves the analysis of intratextual consistency for lexical units that denote a superstructural category in Web sites. This is represented in navigation menus, webmaps and page titles both at the top of the browser and at the top of the content itself. As an example, the analysis showed that in a single localized Web site the concept that denotes the move *contact us* is translated using four different lexical units; *contáctenos*, *contacte con nosotros*, *póngase en contacto con nosotros* and *contacto con nosotros*. Another example would be the translation of the concept *privacy policy* that is referred to in the same translated Web site as *política de privacidad*, *declaración de privacidad* and *normativa de privacidad*.

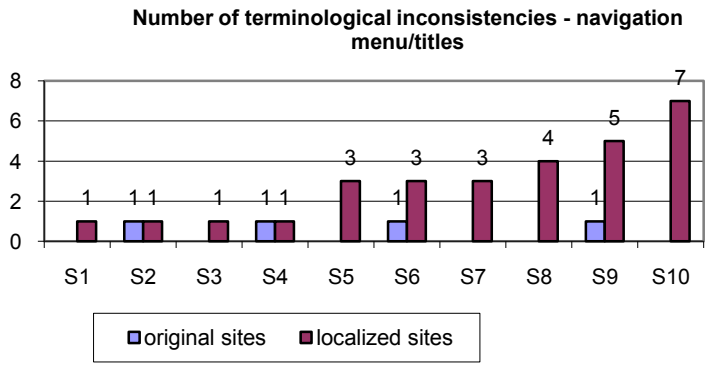

Figure 2: Contrastive analysis of inconsistent terminology for the same hypertextual concept in the same text

As shown in Figure 2, the contrastive analysis revealed that translated Web sites show a greater percentage of terminological inconsistencies in concepts related to the hypertextual structure of the site, mostly in cases in which the source lexical unit has several valid translations in the target language. In the localized section of the corpus, the average number of inconsistencies in superstructural terminology is 2.9 per Web site, with 100% of the Web sites containing this type of inconsistency. In original Web sites, the same analysis yields 0.4% of inconsistencies per site, with 40% of Web sites including this type of inconsistency.

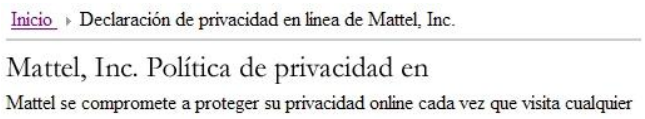

Figure 3: Example of intratextual denominative variation in the translation of "privacy policy" and "online". (Mattel Web site)

Thus, the results show that localized texts are on average less terminologically consistent than original Web sites, even when TM tools would in principle assist in providing consistent translation for these segments. This finding also points out that concepts related to the superstructure of Web sites might not be routinely standardized in terminology databases prior to the actual translation.

The second type of lexical inconsistency analyzed entails the presence of denominative variation in the loanwords and calques *link*, *online* and *email*. The following list illustrates the range of denominative variation found in the subcorpus for each concept. In this list, the use of quotations in order to indicate that the word is a neologism was identified as potential variation, together with the possibility of capitalizing the loanword or calque:

Email [9]: *correo electrónico, correo, dirección electrónica, dirección de correo, dirección de email, email, e-mail, E-mail, mail.*
Link [5]: *enlace, hiperenlace, vínculo, hipervínculo, link.*
Online [6]: *en línea, "en línea", online, On-line, on-line, "on-line".*

The point of interest for this paper is to observe which Web sites use two or more variants for each concept, and more importantly, whether localized sites are more inconsistent than original sites in this respect.

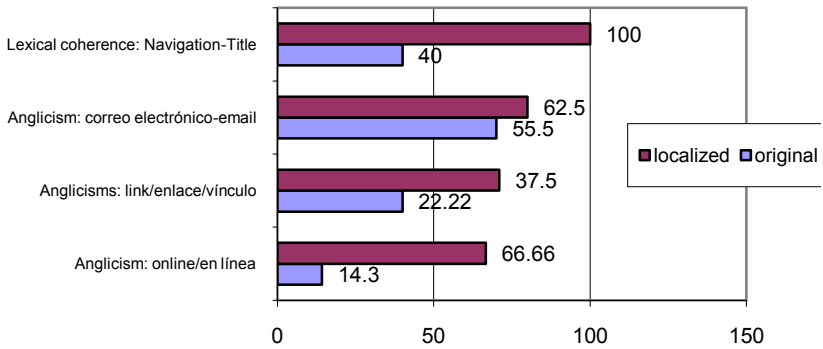

Figure 4: Analysis of lexical inconsistencies in Web sites

Figure 4 shows the contrastive analysis on lexical inconsistencies in which the bar represents the percentage of Web sites that include two or more variants for the same concept. It can be observed that localized sites consistently show higher levels of denominative variation in the same Web site when referring to the same concept. The tendency to use two or more variants for the same term is very similar in the case of the term *email*, with 80% of localized sites and 70% of original sites using the loanword. In the case of *online*, original sites are much more consistent in their use while localized sites show a high percentage of inconsistencies.

4.2.2. Typographic inconsistencies

The third analysis comprises the contrastive study of typographic inconsistencies, another aspect that TM tools cannot fully assist in controlling. In Spanish, capitalization in titles and listed items can be considered a typographic borrowing from English (Martínez de Sousa 2000). The analysis shows that 60% of localized sites use inconsistent capitalization in titles and lexical units in navigation menus, while only 10% of original sites show this type of inconsistency. Similar results are found in the case of inconsistent capitalization of the neologisms *web* and *internet*; localized sites also show higher percentages of sites that interchangeably use these terms both in upper and lower case.

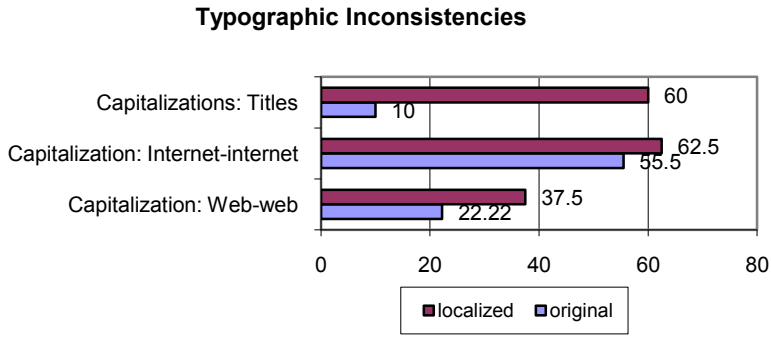

Figure 5: Contrastive analysis of typographic inconsistencies

Once again, it can be observed that localized sites show higher inconsistency levels than original Web sites.

4.2.3. Syntactic inconsistencies: politeness

The last analysis performed deals with the syntactic level. An intratextual analysis of each Web site was carried out searching for formal and informal second person pronominal forms such as *tú* / *usted* and *te / le /se*, possessive adjectives, *tu / su*, and verbal forms such as *haz click* or *haga click* 'click here'. Surprisingly, the percentage of Web sites that address the user both in the *tú* and *usted* form is higher in original sites (70%) than in localized sites (60%).

This finding would in principle contradict the second hypothesis in that original sites would present higher levels of inconsistency in the use of formal and informal markers. However, this can be explained in terms of the different audiences that a Web site might target, such as corporate or personal clients, and the subsequent variation on the power distance depending on the content of different sections of the Web site. In fact, it has been previously observed that Spanish corporate Web sites have a tendency to address customers formally using *usted*, but this tendency changes in the move *job–career,* as the targeted audience would potentially be part of the organization that released the text, and therefore, the power distance relationship would vary (Jiménez-Crespo, 2008a). Nevertheless, it should be noted that this analysis showed that original Web sites would switch between formal and informal tone depending on differences in the targeted audience of sections of the Web site (customer/future employee, regular customer/companies), while inconsistent localized texts would address the same user in both ways regardless of the user's status.

Following these three analyses at the lexical, typographic and syntactic levels, it can be clearly observed that localized sites are on average more inconsistent than original sites, and therefore, these results would in principle confirm the second hypothesis of this study: localized sites show

higher levels of lexical, syntactic and typographic inconsistencies than texts originally produced in the target language. The results therefore demonstrate that, despite industry's claims, TM tools cannot fully control the different dimensions of consistency, a key quality issue in software and Web development (Nielsen, 2001).

5. Conclusions

During the last twenty years, TM tools have been widely promoted in terms of quality and consistency, while translation scholars have argued that technological constraints on the translation task might produce the opposite effect. The goal of this paper was to empirically investigate two related claims: whether sentence-based processing might promote or lead to the replication of source text structures and whether TM tools can guarantee the production of consistent texts. The empirical study was founded on the premises that technological tools would leave a trace that would be observable using a corpus-based methodology (Baker, 1995). Among several possible approaches, the study chose a comparable corpus methodology in which texts translated using TM tools were contrasted with texts originally produced in the target language. Two working hypothesis were formulated in this study. In the first case, the first hypothesis has been validated: original and translated texts from the same genre show significant differences in their prototypical superstructure. This has been explained in terms of the replication of the source text structure during a translation process subject to specific constraints (Baker, 1999, p. 285; Baker, 1995), some of which are related to the impact of TM use on the translation task. This effect has been observed not only in the higher frequency of certain thematic units that respond to source sociocultural norms and conventions, such as differences in the legal systems, but also in the higher number of average words in the same thematic unit, which could be due to a replication of the source sentence structure.

The second hypothesis related to lower consistency levels at the subsentence level was also validated: TM translated texts consistently showed lower levels of lexical and typographic consistency as compared to texts spontaneously produced in the target language. Again, it should also be noted that this effect cannot be fully attributed to TM use, but rather to combination with other factors, such as the presence of multiple translators, not following a pre-established style guide, or an inefficient editing process. Nevertheless, it has been clearly observed that current professional TM use cannot guarantee similar levels of consistency to that of original texts not subject to a technology-driven translation process. Nevertheless, in the case of the variable chosen to validate the syntactic consistency hypothesis – the inconsistent use of politeness markers – original Web sites were on average less consistent than translated ones. This was explained in terms of the specific communicative situation that defines corporate Web sites, as different

sections of the Web site might be addressed at different audiences. A closer analysis found that translated texts were inconsistent when addressing the same user, such as a personal client, as opposed to original Web sites, which varied their politeness levels according to the type of user (such as personal clients or corporate clients).

As a final remark, it should be mentioned that the methodology used in this paper assumes that TM tools have changed the nature of the task that it intends to facilitate (Bowker, 2002). The differences observed between texts translated using TM tools and original texts could also be identified as a general tendency in translated digital texts or as a potential case of a new translation universal (Baker, 1995; Mauranen & Kujamäki, 2004) that would require further study. Thus, the revolutionary impact of TM tools on the translation practice might challenge some basic assumptions in Translation Studies, such as the individualistic character of translation or that translation necessarily entails an operation involving complete and unitary texts. This empirical study has shown that further research into the effects of TM in translation processes and the translation products themselves is needed. It is our hope that additional empirical investigations in this under-researched area promote the development of TM tools that could potentially account for domain and genre-specific intercultural variation or improvements in the retrieval of subsentential matches.

Bibliography

Ahrenberg, L. & Merkel, M. (1996). On translation corpora and translation support tools: A project report. In K. Aijmer, B Altenberg & M. Johansson (Eds.). *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies* (pp. 185-200). Lund: Lund University Press.

Alcina, A. (2008). Translation technologies: Scopes, tools and resources. *Target*, *20*(1), 79-102.

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

Austermühl, F. (2001). *Electronic tools for translators*. Manchester: St. Jerome Publishing.

Baker, M. (1999). The role of corpora in investigating the linguistic behaviour of professional translator. *International Journal of Corpus Linguistics*, *4*(2), 281-298.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager* (pp. 175-186). Amsterdam/Philadelphia, PA: John Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target 7*(2), 223-243.

Bass, S. (2006). Quality in the real world. In K. Dunne (Ed.), *Perspectives on localization* (pp. 69-84). Amsterdam/Philadelphia, PA: John Benjamins.

Bédard, C. (2000). Mémoire de traduction cherche traducteur de phrases. *Traduire*, *186*, 41-49.

Biau, G., Ramón, J., & Pym, A. (2006). Technology and translation (a pedagogical overview). In A. Pym, A. Perekstenko, & B. Starink (Eds.), *Translation technology and its teaching* (pp. 5-20). Tarragona: Intercultural Studies Group.

Bowker, L. (2006). Translation memory and "text.". In L. Bowker (Ed.), *Lexicography, terminology and translation* (pp. 175-187). Ottawa: University of Ottawa Press.

Bowker, L. (2005). Productivity vs quality? A pilot study on the impact of translation memory systems. *Localisation Focus*, *4*(1), 13–20.

Bowker, L. (2002). *Computer-aided translation technology: A practical introduction*. Ottawa: University of Ottawa Press.

Cabanillas, I., Tejedor, C., Díez, M., & Redondo, E. (2007). English loanwords in Spanish computer language. *English for Specific Purposes*, *26*(1), 52-78.

Corpas Pastor, G., & Varela Salinas, M. (Eds.). (2003). *Entornos informáticos de la traducción professional: Las memorias de traducción*. Granada: Editorial Atrio.

De Beaugrande, R-A & Dressler, W.U. (1981). *Introduction to text linguistics.* London/New York, NY: Longman.

Días Fouçes, O. & García González, M. (2008). *Traducir (con) sofware libre*. Granada: Comares.

Dunne, K. (2006). Putting the cart behind the horse: Rethinking localization quality management. In K. Dunne (Ed.), *Perspectives on Localization* (pp. 95-117). Amsterdam/Philadelphia, PA: John Benjamins.

Esselink, B. (2001). *A Practical guide to localization*. Amsterdam/Philadelphia, PA: John Benjamins.

Freigang, K. (2005). Sistemas de memorias de traducción. In D. Reineke (Ed.),*Traducción y localización. Mercado, gestión, tecnologías* (pp. 95-122). Las Palmas de Gran Canaria: Anroart Ediciones.

Fritz, G. (1998). Coherence in hypertext. In W. Bublitz, U. Lenk & E. Ventola (Eds.), *Coherence in spoken and written discourse: How to create it and how to describe it* (pp. 221-234). Amsterdam/Philadelphia, PA: John Benjamins.

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

Göpferich, S. (1995). *Textsorten in naturwissenschaften und technik. Pragmatische typologie-kontrastierung-translation*. Tubinga: Gunter Narr.

Gow, F. (2003). *Metrics for evaluating translation memory software*. MA thesis, School of Translation and Interpretation, University of Ottawa, Ottawa, ON.

Heyn, M. (1998). Translation memories: Insights and prospects". In L. Bowker, M. Cronin, D. Kenny & J. Pearson (Eds.), *Unity in diversity? Current trends in translation studies* (123-136). Manchester: St. Jerome Publishing.

Höge, M. (2002). *Towards a framework for the evaluation of translators' aids systems.* PhD thesis, Department of Translation Studies, Helsinki University, Helsinki.

Jiménez-Creso, M.A. (2009). Conventions in localisation: A corpus study of original vs. translated web texts. *Jostrans: The Journal of Specialized Translation*, *12*, 79-102. Retrieved August17, 2009, from http://www.jostrans.org/issue12/art_jimenez.php

Jiménez-Crespo, M.A. (2008a). *El proceso de localización web: estudio constrastivo de un corpus comparable de género sitio web corporativo*. PhD thesis, Departamento de Traducción e Interpretación, Universidad de Granada, Granada. Retrieved August 17, 2009, from http://hera.ugr.es/tesisugr/17515324.pdf

Jiménez-Crespo, M.A. (2008b). Caracterización del género 'sitio web corporativo' español: Análisis descriptivo con fines traductológicos. In M. Fernández Sánchez & R. Muñoz Martin (Eds.), *Aproximaciones cognitivas al estudio de la traducción e interpretación* (pp. 259-300). Granada: Comares.

Jiménez-Crespo, M.A. (2008c). Web genres in localization: A Spanish corpus study. *Localization Focus – The International Journal of Localization*, *6*(1), 4-14.

Jiménez-Crespo, M.A. & Maribel Tercedor, M. (in press). Theoretical and methodological issues in web corpus design and analysis. *International Journal of Translation.*

Kenny, D. (2001). *Lexis and creativity in translation. A corpus-based study*. Manchester: St. Jerome.

Kenny, D. (1999). CAT tools in an academic environment: What are they good for? *Target*, *11*(1), 65-82.

Larose, R. (1998). Méthodologie de l'évaluation des traductions. *Meta*, *43*(2), 163-186.

Laviosa, S.(2002). *Corpus-based translation studies*. Amsterdam: Rodopi.

L'Homme, M.(1999). *Initiation à la traductique*. Brossard, QC: Linguatech éditeur.

Lommel, A. (Ed.) (2004). *Localiztion Industry Primer, 2nd Edition*. Geneva: The Localization Industry Standards Association (LISA).

Lörscher, W. (1991). *Translation performance, translation process, and translation strategies A psycholinguistic investigation*. Tübingen: Gunter Narr.

Macklovitch, E. & Russell, G. (2000). What's been forgotten in translation memory. In J. White (Ed.), *Envisioning machine translation in the information future* (pp. 137-146). AMTA 2000: Proceedings of the 4th Conference of the Association for Machine Translation in the Americas; Cuernavaca, Mexico, October 10-14, 2000. Berling: Springer.

Martínez de Sousa, J. (2000). *Manual de estilo de la lengua española*. Gijón: Trea.

Martínez Melis, N. & Hurtado Albir, A.(2001). Assesment in translation studies: Research needs. *Meta 47*(2), 272-287.

Mauranen, A. & Kujamäki, P. (Eds.). (2004). *Translation universals: Do they exist?* Amsterdam/Philadelphia, PA: John Benjamins.
Neubert, A. & Shreve, G. (1992). *Translation as text*. Kent, OH: Kent State University Press.
Nielsen, J. & Loranger, H. (2006). *Prioritizing web usability*. Indianapolis, IN: News Riders.
Nielsen, J. (2002). *Coordinating user interfaces for consistency*. San Francisco, CA: Morgan Kaufmann.
Nobs, M. (2006). *La traducción de folletos turísticos: ¿Qué calidad demandan los turistas?*. Granada: Comares.
Nord, C. (1991). *Text analysis in translation*. Amsterdam:Atlanta, GA: Rodopi.
O'Brien, S. (2007). Eye-tracking and translation memory matches. *Perspectives: Studies Translatology*, *14* (3), 185-205.
O'Brien, S. (1998). Practical experience of computer-aided translation tools in the software localization industry. In L. Bowker, M. Cronin, D. Kenny & J. Pearson (Eds.), *Unity in diversity? Current trends in translation* studies (pp. 115-122). Manchester: St. Jerome Publishing.
Price, J. & Price, L. (2002). *Hot text. Web writing that works*. Berkeley, CA: News Riders.
Reinke, U. (2004). *Translation memories: Systeme – konzepte – linguistische*. Frankfurt am Main: Peter Lang.
Rico, C. (2000). Evaluation metrics for translation memories. *Language International*, *12*(6), 36-37.
Robbins, S. & Stylianou, A. (2003). Global corporate web sites: An empirical investigation of content and design. *Information & Management*, *40*, 205-212.
Sager, J.(1989). Quality and standards: The evaluation of translations. In C. Picken (Ed.), *The translator's handbook* (pp. 91-102). London: ASLIB.
Shreve, G. (2006). Corpus enhancement and localization. In K. Dunne (Ed.),*Perspectives on localization* (pp. 309-331). Amsterdam/Philadelphia, PA: John Benjamins.
Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, *14*(2), 113-157.
Storrer, A. (2002). Coherence in text and hypertext. *Document Design*, *3*(2), 157-168.
Swales, J. (1990). *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.
Wallis, J. (2008). Interactive translation vs. pre-translation in TMs: A pilot study. *Meta*, *53*(3), 623-629.
Webb, L. (1998). *Advantages and disadvantages of translation memory: A cost/benefit analysis*. MA thesis, Graduate Division, Monterey Institute of International Studies, Monterey, CA.
Wright, S. (2006). Language industry standards. In K. Dunne (Ed.), *Perspectives on localization* (241-278). Amsterdam/Philadelphia, PA: John Benjamins.
Zerfaß, A. (2002). Comparing basic features of TM tools. *Multilingual Computing and Technology*, *13*(7), 11-14.

---

[1] With the exception of Wallis (2008) that compared the quality of translated texts using interactive translation vs. pre-translation in TM.

[2] In this study, the superstructure of a textual genre is defined as the prototypical pattern that comprises a number of thematic or communicative textual sections whose hierarchical order is fixed to a certain degree (Göpferich, 1995, p. 127; Hurtado Albir, 2001, p. 495).

[3] Including graphics, typography, layout, animation sequences or functionality associated to each textual segment.

[4] Storrer (2002) identifies the function of lexical units in navigation menus as *global* and *local* coherence cues that assist users in navigating the hypertext by providing a the necessary coherence in order to identify a unitary text as such.

[5] Only in the case of hypertexts understood as a thematic, functional and textual unit (Storrer, 2002). E-texts, that is, printed texts simply uploaded to the WWW or linked on a Web site and hyperwebs, such as portals, do not share this characteristic (Jiménez & Tercedor, 2008).

[6] The lexical units in navigation menus or Web page titles cannot be strictly be defined as sentences (Bowker, 2002), even when TM systems consider them as a segment and stores their translation accordingly.

[7] In this study, a comparable corpus is understood as a representative collection of texts spontanously produced in one language alongside similar texts translated into that language (Baker, 1995).

[8] Only the *locale* Spanish-Spain or es-ES was selected in order to exclude the effect of dialectal variation in all Spanish varieties or cultural differences among the different areas in which Spanish is spoken.

[9] For our purposes, a *move* is defined as a "unit of discourse structure which presents a uniform orientation, has specific structural characteristics and has a clearly defined function" (Swales, 1990, p. 140).