ELUA
ISSN 2171-6692
Núm. 38, 2022, págs. 117-136
https://doi.org/10.14198/ELUA.21900

Fine-tuning machine translation quality-rating scales for new digital genres: The case of user-generated content

Adaptación de las escalas de calidad de la traducción automática a los nuevos géneros digitales: el caso del contenido generado por el usuario

Miguel A. Candel-Mora
Universitat Politècnica de València, España
mcandel@upv.es
https://orcid.org/0000-0001-8754-6046

Abstract
With the active participation of users in product review platforms, online consumer-generated content, and, more specifically, user-generated reviews, have become a clear reference in purchasing decision-making processes, which sometimes exceed the impact of advertising campaigns. A common feature of most tourism review platforms is the use of machine translation (MT) systems to immediately make reviews available to users in various languages. However, the quality of the MT output of these reviews varies greatly, primarily due to the subjective and unstructured nature of this digital genre. Different studies confirm that there are no universal quality rating scales. The assessment of MT output quality usually depends on factors such as the purpose of the text or the value given to the immediacy of the translation. New neural MT systems have been a revolution in the quality increase of the translated output;

Resumen
Con la participación activa de los usuarios en las plataformas de reseñas de productos, los contenidos online generados por los consumidores, y más concretamente, las opiniones de los usuarios se han convertido en una clara referencia en los procesos de decisión de compra, que en ocasiones superan el impacto de las campañas publicitarias. Una característica común de la mayoría de las plataformas de reseñas turísticas es el uso de sistemas de traducción automática para poner inmediatamente las reseñas a disposición de los usuarios en diferentes idiomas. Sin embargo, la calidad de la traducción automática de estas reseñas varía en gran medida debido a la subjetividad y a la naturaleza no estructurada de este género digital. Diferentes estudios confirman que no existen escalas universales de valoración de la calidad y que la evaluación de la calidad del resultado de la MT suele depender de factores como la finalidad del texto o el

To cite this article: Candel Mora, M. A. (2022). Fine-tuning machine translation quality-rating scales for new digital genres: The case of user-generated content. *ELUA*, (38), 117-136. https://doi.org/10.14198/ELUA.21900

Recibido: 07/02/2022 Aceptado: 03/05/2022

however, new lines of research are opening up to verify whether the quality of this new paradigm of MT can be assessed with the existing scales, mainly from previous rule-based systems and statistical translation, or whether it is necessary to develop new quality metrics specifically for these new intelligent systems. On the other hand, one of the questions that remain to be resolved in this new context of neural MT is whether the use of large amounts of textual data in the training of these systems is as effective as the use of less data but of higher quality and better-adjusted to the specialty and type of text for which it is used. Based on the hypothesis that each genre requires specific quality rating scales, this work identifies the error patterns and textual characteristics of online user reviews from a corpus-based approach analysis that will contribute to adapting quality rating scales to this specific digital genre.

KEYWORDS: machine translation; post-editing; quality assessment; user-generated content; online reviews

valor que se da a la inmediatez de la traducción. Los nuevos sistemas de traducción automática neuronal han supuesto una revolución en el incremento de la calidad del texto traducido, sin embargo, se abren nuevas líneas de investigación para verificar si la calidad de este nuevo paradigma de traducción automática se puede valorar con las escalas existentes, procedentes en su mayoría de los anteriores sistemas basados en reglas y traducción estadística, o si es necesario desarrollar nuevas métricas de calidad acordes con estos nuevos sistemas inteligentes. Por otro lado, una de las cuestiones que quedan por resolver en este nuevo contexto de traducción automática neuronal es si la utilización de grandes cantidades de datos textuales en el entrenamiento de estos sistemas es igual de eficaz que un uso de menos cantidad de datos pero de mayor calidad y más ajustados a la especialidad y el tipo de texto en el que se utiliza. Partiendo de la hipótesis de que cada género requiere escalas de valoración de la calidad específicas, este trabajo identifica patrones de error y características textuales de las reseñas de usuarios en línea a partir de un análisis basado en un corpus que contribuirá a adaptar las escalas de valoración de la calidad a este género digital específico.

PALABRAS CLAVE: traducción automática; posedición; evaluación de la calidad; contenido generado por el usuario; reseñas en línea

1. INTRODUCTION

With Web 2.0 and the active participation of users, online consumer-generated reviews have become a widespread reference in purchasing decision-making processes, which on occasions, exceed the impact, reliability and authenticity of advertising campaigns. According to Schemmann (2011: 1), "…seven in every ten Internet users worldwide trust consumer opinions and peer recommendations posted online".

Following Ricci and Wietsma (2006: 297), "product reviews can be described as a subjective piece of non-structured text describing the user's product knowledge, experiences and opinions, together with a final product rating". However, Vásquez (2012: 107) acknowledges the constraints of carrying out research on user reviews solely based on the language used since other nonlinguistic cues also play an essential role.

User reviews have already been studied to a large extent from the point of view of marketing, business, tourism and information technology (Schemmann 2011; Popović 2021) in areas such as their influence on decision-making (Ricci and Wietsma, 2006) or the

characteristics of the textual genre (Vásquez 2014). User reviews may appear in different formats and structures (Vásquez 2014): as evaluations of a product, in the form of dialogue in a forum, or, as in the case of the reviews selected for this study, as unstructured free text for the evaluation of a tourism product, and more specifically in this work, hotel reviews. Other genre-specific features include intertextuality, reference to previous comments, the personal profile of the reviewer and paralinguistic elements, mainly "orthographic strategies designed to compensate for the impersonality of written discourse" (Pollach 2006: 8), such as capitalization, spelling, and punctuation. Among other aspects that Pollach (2006) notes are emoticons, the use of capital letters, and overuse of exclamation marks and acronyms. However, Pollach also holds that the use of non-verbal cues was not that common in the corpus she analyzed, perhaps because reviewers mostly take their tasks seriously and use neutral, non-emotive language.

A common feature of most hotel review platforms is the use of machine translation systems to immediately make that review available to as many users as possible in different languages, which has also been studied in recent years by different authors (Castilho *et al.* 2018; Gerlach *et al*. 2013; Gorög, 2014; Jiang, Way and Haque 2012; Lommel 2018). These studies have focused mainly on improving the translation engines and language resources used in order to optimize the MT output (Aranberri 2014; Koby *et al.* 2014; Specia, Raj and Turchi, 2010; Temnikova, 2010), but so far it has not been studied in depth from the point of view of identifying error patterns in specific types of texts to fine-tune MT output quality-rating scales.

From the point of view of MT output quality assessment and keeping in mind the purpose of the target text, post-editing (PE) or work to improve MT output consists in repairing and accommodating the MT output, to varying degrees - to the appropriate target language linguistic conventions (O'Brien 2005), or the expectations of users (Allen 2003; Castilho *et al.* 2018: 11). The discussion seems to be focused on the different definitions of translation quality assessment (TQA); as Castilho *et al*. (2018) put it, "the meaning of quality can vary considerably for different individuals, groups, and contexts" (Castilho *et al.* 2018: 11).

Several studies confirm that there are no universal MT quality-assessment scales (Allen 2003; Lommel 2018; Popović 2018; TAUS 2010), and according to Castilho et al. (2018: 30), there is "a serious lack of standardization in TQA for both HT and MT". In sum, each type of text requires specific quality rating scales, and the most important factor in translation is, therefore, purpose; quality is related to the user's judgement (Castilho *et al.* 2018: 14). Thus, the objective of this work is to identify the standard textual conventions of Consumer-Generated Reviews (CGR) and determine the level of quality of their MT output, and then propose a specific classification of error patterns common to this genre. One potential use of this classification would be educational purposes, where it could serve as guidance and practical training material for future post-editors. According to different authors (Gaspari, Almaghout, and Doherty 2015; Kenny and Doherty 2014; Mellinger 2017), such training is imperative to be current with industry requirements. In particular, Mellinger (2017: 281) highlights that "the incorporation of machine translation (MT) into translation curricula is a growing trend, as demonstrated by several recent scholarly works on MT pedagogy".

As Mellinger (2017: 285) also points out, "it would be prudent to expose students to machine-translated texts in domain-specific translation courses so that the evaluation stages can be addressed appropriately". The high availability of CGR on different products, set-

tings and fields of specializations such as consumer electronics, tourism products, or film reviews poses an excellent opportunity to reflect on translation quality assessment and the specificity of error classifications.

2. METHODOLOGY

In order to achieve the objective of this study, two corpora were designed: CorpusCGR-EN, which consists of 100 consumer-generated hotel reviews originally written in English, and posted online on TripAdvisor. The second corpus, CorpusCGR-MT-ES, is formed by the Spanish MT output of the reviews in CorpusCGR-EN, that is, the 100 consumer-generated hotel reviews mentioned before exactly as they appeared in the platform after their MT processing by Google Translate, which was the default MT engine provided by the platform at the time, September 2020.

The two corpora, CGR-EN and CGR-MT-ES, with a total size of 14.511 and 14.790 words, respectively (see Table 1), were then segmented into sentences and aligned to facilitate processing and manual assessment tasks, totaling 812 segments.

Finally, the quality assessment of the Spanish MT output was determined following the TAUS Dynamic Quality Framework (DQF) (TAUS, 2016; Van der Meer et al., 2017) and its typology of errors: Accuracy, Fluency, Terminology, Style, Design, Locale conventions and Verity, which concentrate on more specific items such as syntax errors, word agreement, word order, use of articles, or mistranslations due to ambiguity or omission, among others.

2.1. Corpus analysis

Both corpora were processed using Wordsmith Tools 7 in order to identify characteristic elements in these types of texts, such as underlying patterns, sentence length, wordlists and keyword lists.

text file	CorpusCGR-EN.txt	CorpusCGR-MT-ES.txt
tokens (running words) in text	14,511	14,790
types (distinct words)	1,949	2,200
type/token ratio (TTR)	13.61	15.06
standardized TTR	41.21	42.91
standardized TTR std.dev.	53.51	52.16

Table 1. Corpus statistics.

TAUS DQF (2016) distinguishes two levels of quality determined by two main criteria: the quality of the MT raw output and the expected end quality of the content. These levels are "good enough" quality, and quality "similar or equal to human translation". TAUS's "good enough" level is defined as comprehensible and accurate but not very convincing concerning style. Based on these levels proposed by TAUS, for this work, we propose a third level labelled "Unacceptable", in which the message is not accurate due to incorrect grammar or lexical usage, or unusual syntax or mistranslation, among other issues. In order

to differentiate our quality classification from TAUS's proposal, for this work TAUS's "good enough" level was labelled "Acceptable" (accurate but not fully convincing or with minor errors), and TAUS's "similar or equal to human translation" was labelled "Correct" (without any errors). Table 2 shows the distribution of the initial MT output quality estimation of the segments in the corpus with the proposal of labels to classify them throughout this work.

Correct	Acceptable	Unacceptable	Total segments
305	324	183	812
38%	40%	22%	

Table 2. Distribution of preliminary MT output quality estimation.

Secondly, the categories labelled as Unacceptable and Acceptable underwent a thorough manual revision and analysis by expert linguists to identify specific recurrent error patterns, which will be described in detail in the analysis and discussion section below and that will serve as the basis for the proposed items to adapt quality assessment scales in line with the objective of this study.

3. MT QUALITY ASSESSMENT SCALES

The concept of quality in translation has always been a topic of interest for research in translation, both human and machine translation, for the industry and the academia alike (Castilho et al., 2018). In the case of translation quality assessment of MT, several variables determine the approach to its assessment. It seems almost impracticable to find common ground that serves as a starting point for proposing universal quality evaluation criteria. According to Valli (2015: 128), among the reasons for this situation are the "lack of transparent evaluation criteria, the difficulty of finding the right metrics, the lack of standardization and the need for different quality levels".

Other variables include the increase in the quality of MT output from recent MT engines, especially since the emergence of neural machine translation systems, their widespread use in professional settings (Torres-Hostench, Presas, and Cid-Leal 2016), the restricted number of methodologies and criteria on how to train post-editors or perform post-editing tasks (Kenny and Doherty, 2014). Finally, some other constraints encountered during the course of this research include the lack of consensus on and diverse definitions of translation quality assessment (Castilho *et al.* 2018) or different definitions of error categories (Popović 2018), which are not always accessible for research purposes for confidentiality reasons, which hinders the possibility of conducting a more general overview of existing post-editing guidelines.

Some authors note that quality is conditioned by the purpose of the MT product (Allen 2003; O'Brien, 2011; TAUS, 2010), i.e., whether the translation is intended to be published and disseminated or if the translation is only aimed at guiding the reader as to the text's overall meaning. All this leads to reflecting on the changing nature of MT output quality assessment and post-editing, and the obstacles to proposing a universal tool that is applicable in any context. As Allen (2003) points out, aspects such as the specifications of the client, the volume of documentation expected to be processed, or expectations with regard to the level of quality,

among others, might influence the exhaustiveness of a post-editing project since "differing percentages of MT accuracy have even been found when applied to different subdomains and different document types within the same technical domain" (Allen 2003: 303).

Among the most common categories of errors are terminology errors, lexical ambiguity, syntax, omission, word agreement and punctuation errors. What is more, different types of metrics also assign a different weight to each error type (Guzmán 2007; Mitchell *et al.* 2014; SAE International 2001; TAUS 2010), in addition to more general criteria such as readability and acceptability of MT output. However, metrics especially consider whether the objectives of the text type are met (Stymne and Ahrenberg 2012; TAUS 2016; Van der Meer *et al*. 2017).

For this work, the TAUS Dynamic Quality Framework (TAUS 2016; Van der Meer *et al.* 2017) was chosen because it is a widespread and consolidated quality metrics and its thoroughness and flexibility facilitate achieving the aim of this study in that it offers a global vision of the translation process.

In this context of CGR, it is especially significant Allen's (2003: 300) consideration of the use of MT in the Web 2.0 and user participation context, which has led to a "change in expectations with regard to the type and quality of translated material". In addition to translation as a high-quality text product for important documents on user safety or commercial information, for example, there is an increased demand for gisting translation where users simply need to understand the main idea of the text in their own language.

Thus, having concluded that there is no universal quality assessment scale and since MT quality-assessment scales cannot be used directly on all types of texts, TAUS DQF seemed to be the most flexible and appropriate instrument to carry out manual translation quality assessment and the identification of the most common error patterns in the corpus of consumer-generated reviews because it takes into account the changing landscape taking into account different content types. However, limitations of this manual methodology, such as time-consuming processes or the subjectivity of assessors, should be pointed out here. In addition, based on Valli's findings (2015: 132) and due to the distinctive features of the genre being examined here and the user's specific needs, the error typology can be more or less granular. Finally, the pass/fail threshold is flexible and depends on content type.

3.1. TAUS Dynamic Quality Framework (DQF)

TAUS DQF proposes seven types of errors: Accuracy, Fluency, Terminology, Style, Design, Locale conventions, and Verity. Since TAUS DQF allows a more or less complete post-editing task depending on the context for which the content is translated, this study has focused on the first four error types: accuracy, fluency, terminology and style, given the characteristics of the reviews and the fact that the objective is to propose specific error categories for this digital genre and help translators in identifying them and becoming familiar with them in post-editing. To this end, TAUS DQF is also enriching since it offers a global vision of the most common types of errors and is a reference in the language industry.

In brief, the description of the categories (TAUS 2016:11) and the specific errors they incorporate are as follows: Accuracy is defined as the situations when "the target text does not accurately reflect the source text, allowing for any differences authorized by specifications". We find specific errors within this type of error: addition, omission, mistranslation, over-

translation, under-translation, and untranslated text. Fluency considers formal or content aspects, not necessarily related to translation, but rather to the use of punctuation, spelling and syntax, as well as the register and coherence of the text. After these two factors are assessed, the MT output of Terminology is evaluated, the main error identified here being the inconsistent use of terminology. Design refers to problems relating to design aspects (vs linguistic aspects) of the content, while Style rates the error as awkward style, in the case of not fulfilling the genre-specific characteristics.

As noted above, the Locale conventions category seems to be more oriented to the translation of more specialized texts and considers error types such as date format, currency format, or measures. Finally, Verity and the Culture-specific reference category assess whether the reference will be understandable to the intended audience.

For the assessment of consumer reviews processed with MT, it is necessary to keep in mind from the beginning that the users of this translation might be less demanding in terms of quality, therefore, the type of error Locale conventions has been disregarded as it targets much more specialized texts. From the category Verity, we have selected some illustrative examples from the *Culture-specific references* section to highlight the most severe errors since, in the translation of tourist texts, the cultural load associated would significantly exceed the scope of this study.

4. RESULTS AND DISCUSSION

Conventional translation quality scales include error annotation and calculation of the proportion of errors with the total amount of words in the translated text. However, in the case of consumer reviews, with an average of 128,05 words per review (see Table 3), the error proportion would be higher, and low-quality translation would be more noticeable.

	CorpusCGR EN	CorpusCGR MT-ES
Average number of words	128.05 words	137.02 words

Table 3. Average number of words per review.

With the design of specific guidelines to assess the MT quality of CGR in mind, the most relevant findings are presented below. Firstly, Figure 1 shows an overview of the most frequent types of error according to the TAUS DQF scale, which will help post-editors anticipate and plan their PE strategy. Next, a detailed analysis is presented for each of the classifications selected: Accuracy, Fluency, Terminology and Style, concluding with a singular contribution which consists in the manual identification of errors in the original text not included in TAUS DQF, which, in turn, produce errors in the MT output which are not common in specialized MT environments.

4.1. Most common error categories (based on TAUS DQF)

Figure 1 presents the overall percentages of the most common error categories, which will provide the basis for an approach to the planning of post-editing tasks in consumer reviews. After that, the main recurrent error patterns of this genre will be detailed and illustrated with examples.

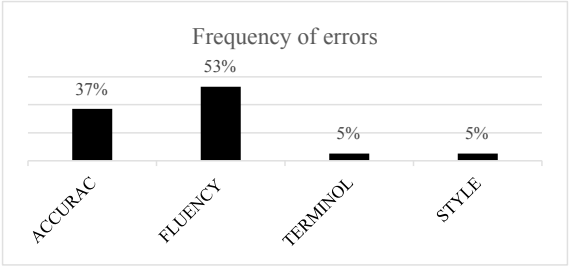

Figure 1. Frequency of error types.

As can be seen, the most recurrent type of error in the Spanish MT output corresponds to the Fluency category, with 509 errors, 53% of the total, followed by errors in Accuracy (354 errors), Terminology (50) and Style (49). Given the amount of text of the sample analyzed, 100 reviews comprising 14,511 words, this finding reveals which error typology would require more effort to be input in a future post-editing project and the skills required from the translator willing to specialize in this type of texts.

4.2. Detailed analysis of recurrent error patterns in CGR

Each of the sections of the following detailed analysis starts with a global view of the distribution of errors from the TAUS DQF, and then, each type of specific error is analyzed and illustrated with examples taken from the corpus. Although errors have been classified following the TAUS typology, on many occasions, the same segment contains different types of errors and has been computed in other sections.

4.2.1. Accuracy

As can be seen below (Figure 2), within this type of error, the first place is occupied by mistranslations - when the translated content does not correspond exactly to the original - with a total of 280, which represents 79% of the total, followed by 47 omission errors (13%), and 13 under-translation errors (4%). Untranslated, text addition and over-translation errors are practically negligible, with 6, 6, and 2 instances, respectively.

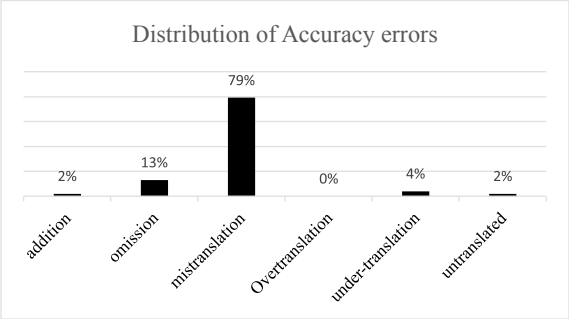

Figure 2. Distribution of Accuracy errors.

4.2.1.1. Addition

Regarding addition errors, only six instances were identified, and as shown below (Examples 1 to 3), they do not affect the content but rather contribute to the naturalness of the reviews.

	Original English	Spanish MT output
Ex. 1.	No complaints all happy and fine.	No tengo ninguna queja todos contentos y **muy** bien.
Ex. 2.	The hotel is conveniently located within walking distance of Buckingham Palace, Big Ben, Parliament and many other **sites**.	El hotel está convenientemente ubicado a poca distancia a pie del Palacio de Buckingham, el Big Ben, el Parlamento y muchos otros **lugares de interés**.
Ex. 3.	It is located in Belgravia just steps from Buckingham Palace.	Está situado en la **elegante zona de** Belgravia a sólo unos pasos del Palacio de Buckingham.

The addition of content in the target text, not present in the source, is not, therefore, common, and in this digital genre does not seem to have the severity it would in other genres such as legal or technical texts.

4.2.1.2. Omission

Overall, 47 errors were identified in which content is missing from the translation but is present in the source text. However, as can be seen in the examples below, the textual elements omitted are, in most cases, modifiers which do not affect the overall content of the translation. However, in the case of tourist product reviews, these modifiers have a substantial role in the user's assessment. Thus, for instance, in example 4, the omission of "as well" in a very favorable opinion of one of the hotel services ("Breakfast was superb") removes the referent and the consequent favorable opinion of the other services valued within the same review.

	Original English	Spanish MT output
Ex. 4.	Breakfast was superb **as well**.	El desayuno era excelente.
Ex. 5.	could not see any **issues** or problems.	no pudimos ver ningún problema.
Ex. 6.	We recently spent 3 nights at Hotel 41, before a long haul flight **home**.	Recientemente pasamos 3 noches en el hotel 41, antes de un largo vuelo.
Ex. 7.	I cannot praise this hotel and the brilliant staff **enough**.	No puedo alabar este hotel y el personal estupendo.

Some authors trace the origin of "online consumer reviews" or "electronic word of mouth" precisely to the traditional word of mouth, previous to computer-mediated communication, when users communicated orally their experience of a tourism product (Ricci and Wietsma 2006: 297). Holgado and Redio (2013: 94) and Yus (2011: 19), when referring specifically to written language in an electronic medium, refer to "oralized written text" due to its hybrid nature and the use of oral and written features in the same medium, and highlight the use of capitals and repetitions as strategies for the compensation of the loss of nonverbal

features. In this study, the approach to capitalization and repetitions within the *omission* category was also identified as a fundamental feature of this digital genre. As shown below, for a representative number of cases from the corpus, both capitalization and repetition have been neutralized in the Spanish MT output, which in fact seems to have an effect on acceptability, which according to Castilho *et al.* (2018: 20) "In the context of TQA, it refers to the degree to which the target or output text meets the needs and expectations of its reader(s) or user(s)".

	Original English	Spanish MT output
Ex. 8.	Hotel 41 played a **HUGE** part.	Hotel 41 jugó un papel muy importante.
Ex. 9.	I really can't fault this Hotel, it's perfect in every way and delivers on **ALL** levels and exceeds expectations regularly.	Realmente no puedo quejarme de este hotel, es perfecto en todos los sentidos y entrega en todos los niveles y sobrepasa las expectativas con frecuencia.
Ex. 10.	Everything is **so So SO** amazing.	Todo es tan increíble.
Ex. 11.	we did the 'historic London' bus tour which was **HORRIBLE**.	hicimos la excursión en autobús de 'Londres histórico que era horrible.

4.2.1.3. Mistranslation

As mentioned above, this category concentrates the highest number of errors (280). One of the main difficulties during this research work was their inclusion in the error classification, since sometimes when the target content does not accurately represent the source content, it could be due to ambiguity or polysemy (as in examples 12, 13, 14, 15, 16 and 17), word order, or a combination of several errors (see examples 18 and 19).

	Original English	Spanish MT output
Ex. 12.	Weather was perfect and London was absolutely a **gas**.	El tiempo era perfecto y Londres era absolutamente un **gas**
Ex. 13.	**Home** baked biscuits.	**Inicio** galletas horneados
Ex. 14.	We WILL be back for Christmas and have started **saving** already.	Volveremos para Navidad y ya hemos comenzado a **salvar**.
Ex. 15.	We dined **too**.	Cenamos **demasiado**.
Ex. 16.	The rooms are cosy but space in London is a **premium**.	Las habitaciones son acogedoras pero el espacio en Londres es una **prima**.
Ex. 17.	I had selected a few **plays** (from internet) and they booked them for me.	Había seleccionado algunos **juega** (desde Internet) y que había reservado para mí.
Ex. 18.	Warm, welcoming personnel.	Dando la bienvenida al personal cálidos.
Ex. 19.	They provided excellent restaurant recommendations and tips on sightseeing.	Nos dieron excelentes recomendaciones y consejos sobre restaurantes en turismo.

4.2.1.4. Over-translation

The errors in this category are practically negligible compared to the rest of the error types. As shown in the example below (Example 20), the target text is more specific than the source text, but it does not have a significant effect on the quality of the text in Spanish.

	Original English	Spanish MT output
Ex. 20.	The breakfast **included** in the room was HUGE - we actually couldn't eat it all.	El desayuno incluido en el **precio de la habitación** era enorme, de hecho no podía comer todo.

4.2.1.5. Under-translation

In this category, a total of 13 instances were identified, some of which were also labelled as omission, although they also correspond to TAUS's definition of under-translation (2016): "The target text is less specific than the source text". The cases in which the exclamations and capital letters are omitted, thus removing the emphasis added by the author of the original, have also been included within this category because, in fact they have an effect on the acceptability of the target text and entail a loss of specificity.

	Original English	Spanish MT output
Ex. 21.	We recently spent 3 nights at Hotel 41, before a long haul flight **home**.	Recientemente pasamos 3 noches en el hotel 41, antes de un largo vuelo.
Ex. 22.	but **all** the other thing were great!	pero la otra cosa era genial!
Ex. 23.	**Very** poor.	Pobre
Ex. 24.	Enjoy**!!!!!!!**	Disfrute
Ex. 25.	Cheap and **NOT WORTH IT.**	Barato y no vale la pena.

4.2.1.6. Untranslated

In this category, three scenarios were identified: The first one, the most common, consists in the non-translation of certain elements that contain proper names or proprietary services of a tourist establishment which, when capitalized, could be interpreted as a proper name. In the second category, parts of the text remain untranslated, as in examples 26 and 27. The last scenario refers to the vocabulary of tourism since certain terms in English sometimes have a greater degree of acceptance when used as a loan word or Anglicism than they do in Spanish (Examples 28, 29).

	Original English	Spanish MT output
Ex. 26.	The Executive lounge is a brilliant concept and works so well as the **'hub'** of this hotel.	El Executive Lounge es un brillante concepto y funciona tan bien como la '**hub**' de este hotel.
Ex. 27.	The **Executive Lounge** is a lovely place to sit and it's very hard not to be tempted by the many complimentary treats that await the weal willed (such as myself).	El **Executive Lounge** es un lugar encantador para sentarse y resulta muy difícil no ser tentado por las muchas delicias gratuitas que le esperan el bienestar querida (como yo).
Ex. 28.	Service, champagne at **check-in**, the pantry to plunder - all perfect.	El servicio, champán en el **check-in**, la despensa para que saquearan - todo perfecto.

	Original English	Spanish MT output
Ex. 29.	On check in you are led to a beautiful Hogwarts like library and ask which on of four **champaignes** you would like with canapes.	En el registro llevado a una bonita Hogwarts como biblioteca y pide que en cuatro **champaignes** desea con canapés.

4.2.2. Fluency

This category takes into consideration punctuation, spelling and grammatical errors. It also includes a section called internal inconsistency, but this has not been included in the analysis given the shortness of the reviews studied. However, to study the behavior of the machine translation engine, a minor revision is presented as an illustration of examples of internal inconsistency within the whole corpus to assess the different choices for the same translation issue.

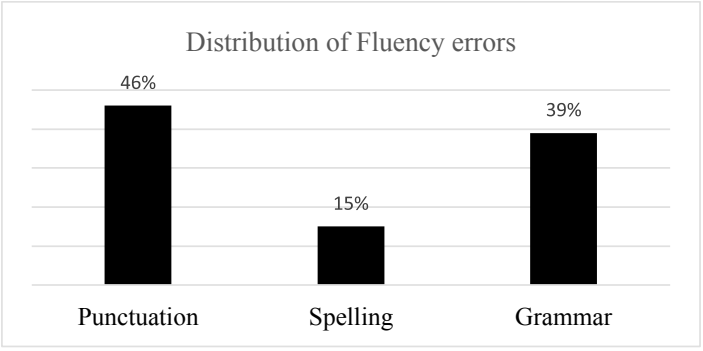

Figure 3. Distribution of Fluency errors.

It should be noted that there are no instances of spelling errors in the Spanish MT output in this category. However, it has also been considered the spelling errors in the original, which in many cases has had an effect on the translated text and which, therefore, should be part of the post-editing strategy for this type of text.

4.2.2.1. Punctuation

Exclamation marks are a widespread feature of this genre, used to compensate for the impersonality of reviews and express emotions. There are several scenarios: exclamation marks in the Spanish MT output are left as in English (with a single exclamation at the end); exclamation marks are adapted to the norms of Spanish and, therefore, correct (with exclamation at the beginning and at the end); Several exclamation marks are used - which is ultimately incorrect in both languages, but characteristic of this digital genre.

Only in 10 examples of the 363 identified has MT correctly processed the use of exclamation marks in Spanish (see examples 33 and 34). In most cases, the use of exclamations marks reproduces the original text in English.

	Original English	Spanish MT output
Ex. 30.	but all the other thing were great!	pero la otra cosa era genial!
Ex. 31.	I'm going back!	Voy a volver!
Ex. 32.	We loved it!	¡Nos encantó!
Ex. 33.	I will definitely be back!	¡Sin duda volveré!
Ex. 34.	WOW!	¡GUAU!
Ex. 35.	Again, thanks to all of you!!!!	Una vez más, gracias a todos!!!!
Ex. 36.	Thanks to the management and the staff for an unforgettable 2nd stay. !!!!	Gracias a la dirección y el personal para una segunda estancia inolvidable .!!!!

4.2.2.2. Grammar

Among the grammar and syntax errors of the text, the most common errors identified are word agreement (99 instances), word order (57), use of articles (54), collocations (49), pronouns (46), and verb tenses (37).

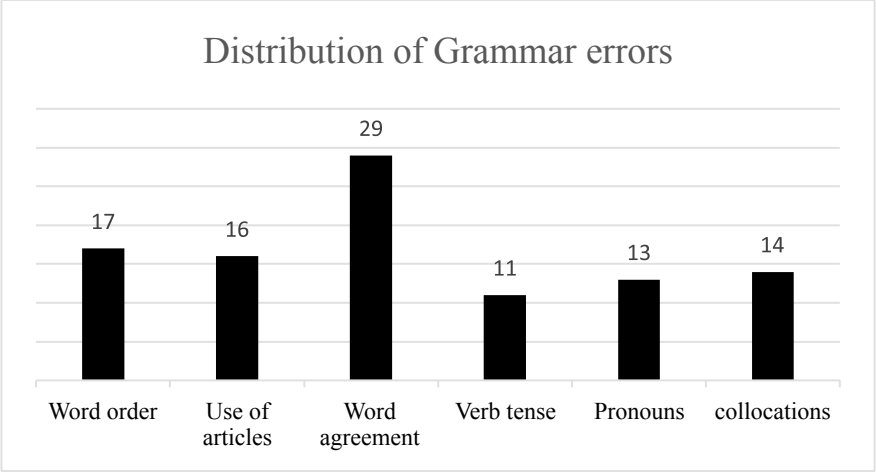

Figure 4. Distribution of grammar errors.

Editing translations made by a human translator is a common practice in professional translation processes. Although this type of error is much less frequent in human translation than in machine translation, it is one of the most laborious tasks and requires more thoroughness. In sum, grammar errors are usually less frequent in human translation but more easily identifiable.

As in the previous cases, the accumulation of errors in the same segment has hindered the inclusion of errors in a single specific category. However, for illustration purposes, some examples of the most frequent types of grammatical errors found in this work are presented below:

	Original English	Spanish MT output	Type of error
Ex. 37.	a very personal, small hotel.	una muy personal, un pequeño hotel.	Word order
Ex. 38.	My husband travels to England every so often for business.	Mi marido viaja a Inglaterra cada tan a menudo por negocios.	Word order
Ex. 39.	Hotel 41 has very good service.	Hotel 41 tiene un muy buen servicio.	Use of articles
Ex. 40.	It's London centre after all.	Es centro de Londres después de todo.	Use of articles
Ex. 41.	There's no getting round it, a stay here is not cheap.	No hay que ir, una estancia aquí no es barato.	Word agreement
Ex. 42.	We felt comfortable, relaxed but with luxury too.	Nos sentimos cómodos, relajado pero con el lujo.	word agreement
Ex. 43.	We were never left waiting and the staff really couldn't do enough for us.	Nunca nos estará esperando y el personal no podía hacer lo suficiente para nosotros.	Verb tenses
Ex. 44.	Will definitely book to stay here again	Sin duda reservar aquí de nuevo	Verb tenses
Ex. 45.	They were genuinely **pleased to** be part of our special occasion.	Estaban verdaderamente **satisfechos a** ser parte de nuestra ocasión especial.	Collocations
Ex. 46.	Simply put, the best hotel I've ever had the pleasure of visiting.	Sencillamente, el mejor hotel en el que he tenido el placer de visitar.	Collocations

4.2.2.3. Grammatical register

The only instances of incorrect grammatical register consist of using informal pronouns and verb forms when their formal counterparts are required (see examples 47 to 49) or the combination of a formal and informal register within the same segment as in example 50.

	Original English	Spanish MT output
Ex. 47.	The rooms although small have all the luxury that you need, with luxury items better than a 5* hotel.	Las habitaciones, aunque pequeñas tienen todo el lujo que usted necesita, con artículos de lujo mejor que un hotel de 5 *.
Ex. 48.	You are always greeted with an appetizer.	Usted siempre son recibidos con un aperitivo.
Ex. 49.	Treats are also served in your room during the late afternoon.	También se sirven Treats en su habitación durante la tarde.
Ex. 50.	Don't bother going elsewhere , it's almost like having your own private stylish London residence in town.	No te molestes ir a otro lugar, es casi como tener su propio Londres elegante residencia privada en la ciudad.

4.2.2.4. Inconsistency

This item has not been considered since the length of the reviews is excessively short. However, for illustrative purposes, several inconsistencies have been identified in the corpus, such as the different translations of snacks as "tentempiés, bocadillos or aperitivos"; check-in as "registro, check-in", Champagne as "champaña, champaigne", Buckingham Palace as "Palacio de Buckinghamm, Buckingham Palace".

	Original English	Spanish MT output
Ex. 51.	some delicious **snacks** are available for all hotel guests.	algunos deliciosos **tentempiés** están disponibles para todos los huéspedes del hotel.
Ex. 52.	The small lounge area is so lovely and the complimentary **snacks** are gorgeous.	La pequeña zona del salón es tan encantadora y los **bocadillos** de cortesía son preciosas.
Ex. 53.	The atmosphere was relaxed and we had another glass and some **snacks** before being taken to our room.	El ambiente era relajado y tuvimos otra copa y algunos **aperitivos** antes de ser llevado a nuestra habitación.
Ex. 54.	Every aspect of our stay was managed with care and attention, from check-in to **check-out.**	Cada aspecto de nuestra estancia fue manejado con cuidado y atención, desde la llegada hasta la **salida**.
Ex. 55.	Service, champagne at **check-in,** the pantry to plunder - all perfect.	El servicio, champán en el **check-in**, la despensa para que saquearan - todo perfecto.

4.2.3. Terminology

This is not a common type of error, as only 50 were found (5% of the total). In this section, the analysis paid special attention to the vocabulary related to the hotel industry and the usual lexicon of this sector. The main types of error found were due to polysemy of some words, such as "bar" (establishment / counter / candy), "play" (sport / theatre), "ticket" (train / theatre), "glass" (receptacle / material), or common English verbs that have two forms in Spanish: "to be" (*ser* or *estar*), "to have" (*tener* o *tomar*), "to miss" (*perder* o *echar de menos*) (see Examples 56 to 59).

On occasions, the term in English remains untranslated, however, the language of tourism seems to be quite receptive to importing Anglicisms and loan words. Thus, the resulting Spanish MT output does not seem completely unacceptable, for example, in the translation of "amenities", "check-in", "check-out", "lobby", "lounge", and "staff" (see example 60).

	Original English	Spanish MT output
Ex. 56.	The hotel also booked theatre **tickets** for me.	El hotel también reservamos **billetes** de teatro para mí.
Ex. 57.	After a long day of travel to London it was delightful to be offered a **glass** of chilled champagne while our bags were taken to our room.	Después de un largo día de viaje a Londres era una delicia ofrecerá un **vaso** de champaña helada mientras nos llevaron las maletas a la habitación.

	Original English	Spanish MT output
Ex. 58.	There were no **wardrobe** or **cupboard** doors nor any drawers as there was simply not enough room to open them.	No había **armario o armario** puertas ni ningún cajones, ya que simplemente no había suficiente espacio para abrirlos.
Ex. 59.	**Check in**, **lobby** and dining are all combined in one area centrally located among the rooms.	El **registro**, el **vestíbulo** y los restaurantes están combinados en una zona con una ubicación céntrica entre las habitaciones.
Ex. 60.	You **check in** on the fifth floor as these are the only two floors in the hotel and you are greeted with a glass of champagne.	El **check-in** en el quinto piso, ya que son las únicas dos plantas del hotel y te dan la bienvenida con una copa de champán.

4.2.4. Style

The majority of errors in this section affect the naturalness of the texts and, therefore, the purpose of the genre: reliability and credibility. In sum, the style errors identified are mainly due to awkward choice of words or the addition of articles and pronouns which show a lack of naturalness, although the segments are apparently grammatically acceptable.

	Original source text	Spanish MT output
Ex. 61.	Quite unique and refreshing.	Muy *único* y refrescante.
Ex. 62.	We love it.	*¡A* nosotros nos apasiona!
Ex. 63.	I felt at home, and welcome, even though I was alone.	Me sentí como en casa, y bienvenido, aunque yo estaba solo.
Ex. 64.	This is the secret to this hotel: detail.	Este es el secreto para este hotel: los detalles.

4.2.5. Verity: culture specific references

The only culture-specific references consist in the translation of currency, measures and on occasions, food, which can be considered as acceptable within the context of the hotel industry because they add an exotic flavor to the travel experience. However, in professional translation and localization processes, the lack of adaptation to the locale would be labelled as incorrect. In addition, the target reader of the review platform is not clearly identified with the Latin America or Spain versions of Spanish, which would be the other source of errors. Overall, verity errors do not need a detailed revision as they contribute to the characteristic style of the language of tourism.

	Original source text	Spanish MT output
Ex. 65.	Worth every cent - and the executive lounge is a great place to sit and have a drink after a day out in the city.	Vale la pena cada centavo - y el salón **Executive** es un gran lugar para sentarse y disfrutar de una bebida después de un día en la ciudad.
Ex. 66.	Thoroughly recommend this place and worth every penny	Recomendaría este lugar y vale la pena cada **centavo**

4.2.6. Errors in original text

This category is not included in the TAUS DQF scale since its use is mainly intended for specialized professional translation processes where it is not common to find errors in the original text. However, one of the characteristics of consumer-generated reviews lies in the spontaneity and the variety of different user profiles.

In the corpus analyzed, a total of 315 segments contained errors in the original which consequently affected the quality of the Spanish MT output. Considering the pedagogical implications of the post-editing tasks suggested in this work, it is necessary that the future post-editor takes into consideration the potential effect of an incorrect text in the source language and its subsequent improvement in the post-editing process. Here we also find different scenarios: most of the time the MT output directly reproduces the misspelt word (Examples 67 to 73). In contrast, in others, it corrects the error in the original (Examples 74 and 75).

	Original English	**Spanish MT output**
Ex. 67.	All the staff **wen** up and above to get you what ever you needed.	Todo el personal **Wen** y por encima de todo lo que necesitamos.
Ex. 68.	the staff **where** exceptionally professional.	el personal **donde** excepcionalmente profesional.
Ex. 69.	Is one of those hotels that you don't mind what you'll pay at the end **if** the day.	Es uno de esos hoteles que no te importa lo que tendrás que pagar al final **si** el día.
Ex. 70.	**Breaks aft** did not try and other facilities most of the time in city and entertaing guests.	**Escapadas aft** no probamos y otras instalaciones más del tiempo en la ciudad y tumbonas.
Ex. 71.	Would definitely stay here **ahain**!	Sin duda me alojaría aquí **ahain**!
Ex. 72.	All the **staff** I cam across were very helpful and friendly.	Todo el personal me **leva** a través era muy servicial y amable.
Ex. 73.	**Prob** nicest and friendliest staff I've come across.	**Prob** nicest and friendliest staff I've come across.
Ex. 74.	We had a very **quite** room facing the courtyar	Teníamos una habitación muy **tranquila** con vistas al patio.
Ex. 75.	We give this **ony** 4 stars because of the size of the room, and the lack of a lobby at street level,	Debemos dar este **solamente** 4 estrellas debido al tamaño de la habitación, y la falta de un vestíbulo en el nivel de la calle,

5. CONCLUSIONS

The extensive production of content generated by user reviews posted on platforms of all kinds (tourism, restaurants, consumer electronics) and the use of machine translation to make this content available in different languages have paved the way for new research opportunities. This work has explored the exploitation of a comparable corpus of consumer-generated reviews written originally in English and processed into Spanish in order to identify the most important textual conventions and error patterns of this new digital genre. In the first place, post-editors should be aware that there is no universal translation quality

scale or set of guidelines that apply to all scenarios. The decision as to whether a more or less detailed post-editing effort is appropriate depends on the use and purpose of the translated document. Thus, the characteristics of the textual genre should be considered and the quality-rating scale should be adapted accordingly. Therefore, a comprehensive analysis of the textual genre must be taken into consideration when training future post-editors, along with training in common post-editing techniques and guidelines. For example, in the case of reviews of a tourist product, apart from language, other genre-specific features include naturalness, reliability and credibility.

The nature of user reviews conditions, to a large extent, the need to propose specific quality-rating scales. Most translation quality scales include error annotation and the calculation of the proportion of errors as a function of the total amount of words in the translated text; however, in the case of consumer reviews, which consist of free text of reduced dimensions, the error proportion would be higher, and low-quality translation would be more visible. The results of this work will help to fine-tune MT quality assessment scales according to other parameters, such as the length of the text, for example.

The training of future post-editors would benefit from the identification and extraction of MT-related data obtained from a comparable corpus of the same text type and domain. Among the MT-related data extracted from comparable corpora are collections of translation equivalent fragments of text, such as terminological expressions, frequent multi-word expressions, or usual content words and collocations.

One of the most valuable resources for the development and improvement of machine translation systems is parallel texts; however, in the case of consumer-generated content, it is rather complicated to find a parallel corpus since most of the consumer-generated content is translated by MT systems without any further post-editing. Thus, further exploitation might include using that translation knowledge and adding bilingual dictionaries of previously extracted translation pairs to the parallel corpus of the MT system to improve its overall performance, a well-known practice in domain adaptation and training MT engines.

Among the future research lines identified during this work is the application of this methodology of genre-specific features in user-generated content and contrast it with other types of texts, such as social media content, reviews and testimonials, blog posts, video content, or Q&A forums, which are also susceptible to being processed by machine translation systems, or with other combinations of languages, especially low resource languages.

REFERENCES

Allen, J. (2003). Post-editing. In H. Somers (Ed.), *Computers and Translation. A translator's guide* (pp. 297-317). John Benjamins.

Aranberri, N. (2014). Posedición, productividad y calidad. *Tradumàtica: Tecnologies de la Traducció, 12*, 471-477. https://doi.org/10.5565/rev/tradumatica.62

Castilho, S., S. Doherty, F. Gaspari & J. Moorkens. (2018). Approaches to Human and Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari and S. Doherty (Eds.). *Translation Quality Assessment: From Principles to Practice* (pp. 9-38). Springer.

Gaspari, F., H. Almaghout & S. Doherty. (2015). A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives,* 23(3), 333-358. https://doi.org/10.1080/0907676X.2014.979842

Gerlach, J., V. Porro Rodriguez, P. Bouillon & S. Lehmann. (2013). Combining pre-editing and post-editing to improve SMT of user-generated content. In S. O'Brien, M. Simard and L. Specia (Eds.). *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice,* (pp. 45-53). Nice (France) 2 Sept., 2013. https://archive-ouverte.unige.ch/unige:30952

Guzmán, R. (2007). Manual MT Post-Editing: If it's not Broken, don't Fix it. *Translation Journal,* 11(4). http://translationjournal.net/journal/42mt.htm

Holgado Lage, A. & A. Recio Diego. (2013). La oralización de textos digitales: usos no normativos en conversaciones instantáneas por escrito. *Caracteres: estudios culturales y críticos de la esfera digital*, 2(2), 92-108. http://hdl.handle.net/10366/124906

Jiang, J., A. Way & R. Hague. (2012). Translating user-generated content in the social networking space. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas*, (pp: 1-9). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.299.5601&rep=rep1&type=pdf

Kenny, D. & S. Doherty. (2014). Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer,* 8(2), 276-294. https://doi.org/10.1080/1750399X.2014.936112

Koby, G., P. Fields, D. Hague, A. Lommel & A. Melby. (2014). Defining Translation Quality. *TradumàTica: Tecnologies de la Traducció,* 12, 413-420. https://doi.org/10.5565/rev/tradumatica.76

Lommel, A. (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice,* (pp: 109-127). Springer.

Mellinger, C. (2017). Translators and machine translation: knowledge and skills gaps in translator pedagogy. *The Interpreter and Translator Trainer,* 11(4), 280-293. https://doi:10.1080/1750399X.2017.1359760

Mitchell, L., S. O'Brien & J. Roturier. (2014). Quality evaluation in community post-editing. *Machine Translation,* 28, 237-262. https://doi.org/10.1007/s10590-014-9160-1

O'Brien, S. (2011). Towards Predicting Post-Editing Productivity. *Machine Translation,* 25, 197-215. https://doi:10.1007/s10590-011-9096-7

Pollach, I. (2006). Electronic word of mouth: A genre analysis of product reviews on consumer opinion web sites. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences-Volume 03. IEEE Computer Society,* (pp: 51-53). IEEE. https://doi.org/10.1007/s10590-014-9160-1

Popović, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari, & F. Doherty (Eds.) *Translation Quality Assessment: From Principles to Practice,* (pp. 129-15). Springer.

Ricci, F. & R. Wietsma. (2006). Product reviews in travel decision-making. In M. Hitz, M. Sigala and J. Murphy (Eds.), *Information and communication technologies in tourism* (pp. 296–307). Springer.

SAE International. (2001). *SAE J2450: Translation Quality Metrics*. Society of Automotive Engineers.

Schemmann, B. (2011). A Classification of Presentation Forms of Travel and Tourism-Related Online Consumer Reviews. *e-Review of Tourism Research,* 2. http://ertr.tamu.edu/enter-2011-short-papers

Specia, L., D. Raj & M. Turchi. (2010). Machine Translation Evaluation Versus Quality Estimation. *Machine Translation,* 24, 39–50. https://doi.org/10.1007/s10590-010-9077-2

Stymne, S. & L. Ahrenberg. (2012). On the practice of error analysis for machine translation evaluation. In *Proceedings of the LREC 2012 Conference. Istanbul: European Language Resources Association* (pp. 1785-1790). http://www.lrec-conf.org/proceedings/lrec2012/pdf/717_Paper.pdf

TAUS. (2010). *Error Typology Guidelines.* https://www.taus.net/academy/best-practices/evaluate-best-practices/error-typology-guidelines

TAUS. (2016). *Quality Dashboard White Paper.* https://www.taus.net/think-tank/reports/evaluate-reports/taus-quality-dashboard-white-paper

Temnikova, I. (2010). Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. In *Proceedings of the LREC 2010 Conference*, (pp. 17-23). Valletta: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2010/pdf/437_Paper.pdf

Torres-Hostench, O., M. Presas & P. Cid-Leal. (2016). *El uso de la traducción automática y la posedición en las empresas de servicios lingüísticos españolas: Informe de investigación ProjecTA* [The Use of Machine Translation and Post-editing among Language Service Providers in Spain]. UAB. http://ddd.uab.cat/pub/estudis/2016/148361/usotraaut_2016.pdf

Valli, P. (2015). The TAUS Quality Dashboard. In *Proceedings of the 37th Conference Translating and the Computer,* 127–136. London, UK, November 26-27.

Van der Meer, J., Görög, A., Dzeguze, D., & Koot, D. (2017). *TAUS Quality Dashboard White Paper*. https://www.taus.net/insights/reports/taus-quality-dashboard-white-paper

Vásquez, C. (2012). Narrativity and involvement in online consumer reviews. The case of Tripadvisor. *Narrative Enquire,* 22 (1), 105-121. https://doi.org/10.1075/ni.22.1.07vas

Vásquez, C. (2014). *Online consumer reviews*. Bloomsbury.

Vilar, D., J. Xu, L. F. d'Haro & H. Ney. (2006). Error analysis of statistical machine translation output. In *Proceedings of the LREC 2006 Conference* (pp. 697-702). Genoa: European Language Resources Association, http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.

Yus, F. (2011). *Cyberpragmatics*. John Benjamins.