Qf Lingüístics

ojs.uv.es/index.php/qfilologia/index

Rebut: 11.06.2022. **Acceptat**: 08.09.2022

Per a citar aquest article: Candel-Mora, Miguel Ángel. 2022. "Big data to assess genre-specific features of the machine translation output of online travel reviews in Spanish". *Quaderns de Filologia: Estudis Lingüístics* XXVII: 49-69.

doi: 10.7203/QF.27.24667

Big data to assess genre-specific features of the machine translation output of online travel reviews in Spanish

Big data para la evaluación de características propias del género de reseñas de turismo en la traducción automática al español

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

Resumen: El análisis de datos masivos (*big data*) como el contenido generado por el usuario, y concretamente las opiniones *online* de los consumidores, ha atraído una atención considerable en los últimos años debido a sus numerosas oportunidades de investigación y aplicaciones comerciales en casi todos los campos del conocimiento. El origen de este género digital en la tradición oral pone de relieve los rasgos espontáneos de la lengua hablada que se reflejan en el texto escrito y que tiene características propias según la cultura y la lengua del usuario. Mediante la comparación de un corpus de 2.000 reseñas, este trabajo propone la identificación y el análisis de las características singulares de este nuevo género digital para determinar el comportamiento de la traducción automática de las opiniones de los usuarios al español. Así, el objetivo de este trabajo es el estudio de las reseñas turísticas traducidas al español e identificar cómo aborda la TA las principales características que confieren a este género naturalidad y credibilidad.

Palabras clave: traducción automática; redes sociales; contenido generado por el usuario; datos masivos; reseñas turísticas.

Abstract: Big data analysis such as user-generated content and specifically online consumer reviews has attracted considerable attention in recent years due to its numerous research opportunities and commercial applications in almost all fields of knowledge. The origin of this digital genre in the oral tradition highlights the spontaneous features of the spoken language that are reflected in the written text which has its own characteristics depending on the user's culture and language. By comparing a corpus of 2,000 reviews, this paper proposes the identification and analysis of the unique characteristics of this new digital genre to determine the behavior of machine translation of users' reviews into Spanish. Thus, the aim of this work is to study the tourism reviews translated into Spanish and identify how MT handles the main characteristics that confer naturalness and credibility to this genre.

Keywords: machine translation; social media; consumer-generated content; big data; tourism reviews.

1. Introduction

Big data analysis has attracted considerable attention in recent years due to its numerous research opportunities and business applications for almost every field of knowledge. Among the advantages of processing big data, Chen et al. (2014: 1) point out that "big data also brings new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and incurs new challenges". The literature on big data has identified different fields of study such as structured data analysis, text data analysis, website data analysis, multimedia data analysis, network data analysis, and mobile data analysis (Chen et al., 2014: 61), all of them with an inherent potential to be analysed, processed, and interpreted depending on the tools and technology used, as well as its purpose.

According to Chen et al. (2014: 1), big data refers to "datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time".

For Minelli et al. (2013: 1), the production of big data is the logical consequence of four major global trends: Moore's Law (technology always gets cheaper), mobile computing (widespread use of smart phones or mobile devices), social networking (Facebook, Twitter, Instagram, Pinterest, etc.), and cloud computing.

Big data is usually described in terms of the high volume of data to process with current tools, the speed at which it is produced to be stored and indexed properly, and its varied nature to fit into a rigid schema (Suciu, 2013: 7). Similarly, the Web 2.0 and the active participation of users has favoured one of the largest sources of big data susceptible of research from the point of view of natural language processing, computational linguistics, and machine translation (MT): user-generated reviews (UGR) (Candel-Mora, 2015).

For this research work, text data analysis appears as a challenging approach to explore features of UGR that could not be addressed otherwise. In terms of text mining methodologies, common sources are email communication, business documents, web pages, and social media (Chen et al., 2014: 61), this latter includes online consumer reviews in the form of tourist reviews, the object of this study, which generates approximately 988 million interactions annually only in one platform: TripAdvisor (Tripadvisor, 2021).

According to data from the Spanish National Commission on Markets and Competition (CNMC, 2019), tourism occupies four of the ten areas of activity with the highest percentage of e-commerce transactions in Spain, with approximately 25.9 % of the total turnover. In this line, the role of travel review platforms is crucial in the users' decision-making process, as Schemmann notes (2011: 1), "seven in every ten Internet users worldwide trust consumer opinions and peer recommendations posted online".

In order to reach a global audience, travel review platforms use machine translation engines, which poses an interesting research possibility to study the resulting MT output of unstructured free text in the context of UGR (Castilho, Doherty & Gaspari, 2018; Gerlach et al., 2013; Jiang, Way & Haque, 2012; Lommel, 2018). Most studies on consumer-generated content and machine translation have focused primarily on improving translation engines and language resources used to optimize the MT output (Aranberri, 2014; Koby et al., 2014; Specia, Raj & Turchi, 2010; Temnikova, 2010), but so far it has not been studied in depth from the point of view of the analysis of the MT output to verify whether the translation of new digital genres complies with the end-user expectations in terms of its core features of acceptability.

As Allen points out (2003, 300) the use of MT in the Web 2.0 and user participation has led to a "change in expectations with regard to the type and quality of translated material" and the increased demand for gisting translation: users simply need to understand the main idea of the text in their own language.

Considerable research attention has been devoted to online consumer reviews and its digital genre characteristics (Schemmann, 2011; Pollach, 2006; Vásquez, 2014); however, the effects of machine translation on the reliability and transfer of UGR main features from its origins as word-of-mouth personal communications in the MT output seem to be unexplored.

Therefore, the objective of this work is to analyse tourist reviews MT-processed into Spanish and identify whether the main characteristics that confer this genre naturalness and credibility are transferred adequately to the end-user in Spanish.

From a corpus of 1,000 hotel reviews originally written in English and its comparison with its MT output as it appears in the review platform, this work attempts to categorise the unique features of this new digital genre that the Spanish audience finally receives through machine translation.

In addition, to obtain evidence of the native Spanish speaker's production of reviews, a reference corpus of 1,000 reviews written originally in Spanish was compiled. The analysis of this corpus will reveal whether the digital genre in Spanish follows similar patterns to English reviews and machine translation fully transmits the intention of the reviews.

The first part of this work will explore the origins of UGR as word- of-mouth and its main features to conclude that, since reviews are not only transmitted through language, some of the characteristics of this digital genre such as the reviewer's profile, the intertextuality or reference to other reviews, and paralinguistic elements that contribute to the reliability and credibility of user reviews, for example, cannot be transmitted solely through machine translation.

First, we proceeded to review the literature on user reviews to identify their characteristics, patterns, linguistic resources, and pragmatic purpose, and from there establish the method of analysis of the distinctive features of this genre. The second part presents the analysis and discussion of the study of the corpus of online consumer reviews in order to identify the characteristics that confer naturalness and credibility to this user-generated content reflected in the written text of the reviews in Spanish.

2. Origins and characteristics of online consumer reviews

Several authors trace the origin of online consumer reviews in the tradition of communicating orally consumers' experiences, as evidenced by the variety of designations found in the literature to refer to the evaluation by users posted on travel review sites on their experience: "electronic word-of-mouth" or "eWOW" (Pollach, 2006), "online consumer reviews" (Vásquez, 2012) "user-generated product reviews", "product reviews" or "user opinions" (Ricci & Wietsma, 2006).

UGR has been researched extensively from perspectives such as the potential roles of product reviews in the decision-making process (Ricci & Wietsma, 2006); the role of reviewers (Vásquez, 2014); the characterization of online reviews (Schemmann, 2011); or the improvement of review platforms (Pollach, 2006).

Therefore, all this seems to indicate the emergence of a new digital genre, which thanks to online review platforms and the considerable number of reviews posted online deserves a more detailed analysis from the language perspective. Especially because this genre, traditionally transmitted orally and without a specific structure, did not exist in written format before.

Ricci & Wietsma (2006: 297) define user reviews as “a subjective piece of non-structured text describing the user’s product knowledge, experiences and opinions, together with a final product rating”. With regards to its pragmatic purpose, according to Pollach (2006: 3) the objective of reviews “…is to inform potential buyers of the strengths and weaknesses of consumer products”.

There is a wide variety of formats in which user reviews are presented (Vásquez, 2014): as evaluations of a product, as dialogue in a forum, or, as in the case of the reviews object for this study, as unstructured free text for the evaluation of a tourism product. Other genre-specific features include intertextuality –or reference to previous comments, the personal profile of the reviewer and paralinguistic elements, mainly “orthographic strategies designed to compensate 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 punctuation marks and acronyms.

Although at first glance, reviews may appear as free text without a defined structure, the literature on consumer-generated reviews (Vásquez, 2014) identifies some common characteristics and patterns: they are written in chronological sequence (8 phases); reviews usually include references to other opinions; paralinguistic elements such as punctuation, capitalization, spelling, emoticons and abbreviations are quite frequent; reviews use indirect style; humour, details and personal experience tend to be included in the reviews; and finally, due to its oral origins, users develop strategies designed to compensate for the impersonality of written discourse.

Therefore, for Vásquez (2014) the main structural features of the review as a genre are the summary, the background (the reason for travelling, the people sharing the trip…), the explicit evaluation, the interactions with the hotel staff, the resolution (check-out/price) and the personal advice, suggestions or warnings. However, this author adds that for the review to be reliable and have credibility, some characteristics associated with the user’s participation in the discourse must be taken into account: the indirect style, the introduction of narration, and the use of deictics, which are ultimately responsible for the establishment of links between participants. Finally, Vásquez (2012: 107) recognizes the limitations of conducting this type of research based solely on language, as there are other non-linguistic cues that also play an important role.

Among the linguistic resources used to evaluate the experience, Vásquez (2014: 22) distinguishes three levels: the lexical level, with the use of evaluative adjectives and adverbs of epistemic-evidential evaluation; the discourse level, with the use of colloquial language and interrogative forms; and rhetorical strategies, including the experience of others or the expression of the reason for their evaluation.

Holgado & Recio (2013: 94) and Yus (2011: 19) describe the written language in an electronic medium as "oralized written text" due to its hybrid nature and the use of oral and written features in the same medium. Holgado & Recio (2013) focus their work on the deviations of standard Spanish at the phonetic and syntactic level from a corpus of conversations on Facebook, Skype, and WhatsApp. These authors identify what they call strategies for the compensation of the loss of nonverbal features. Likewise, Yus (2011: 175) recognizes a deviation from the neutral text that compensates for the feeling of shared conventions or oralization, of an increase in sociability, or in humorous effects.

Holgado & Recio (2013: 93) conclude that the characteristics of the oral language in a written format can be studied from three levels: lexical, grammar and discourse. At the lexical level, there is a tendency to use a low lexical density and general vocabulary. At the syntactic level, there is a tendency to ellipsis, the use of short sentences, or the use of active verb forms. Finally, from the discourse perspective, there is a constant reformulation of statements and repetitions and a high proportion of markers of interpersonal dynamics and hedging. Nevertheless, what seems common to all the characterizations of reviews is the aim to transmit naturalness and reliability to potential users.

As mentioned earlier, most travel review platforms use machine translation engines to deliver reviews to as many users as possible and, most of the times, without human supervision. Therefore, the object of the research question raised in this work focuses on the potential effects of reviews posted in English and its Spanish machine-translation output on the conventions of the Spanish-speaking tourist review community. More concretely, this work addresses how machine translation transmits the acceptability, reliability and credibility features established as the main characteristic of the UGR and its influence in the style of the digital genre in Spanish.

Scholars like Pollach (2006) suggest corpus linguistics techniques and textual analysis to identify the main rules and conventions established by the genre community. Pollach focuses her work on the analysis of structure, content, audience appeals, sentence style, and word choice. Similarly, Vásquez (2014) proposes the study of frequency lists to identify grammatical and lexical elements associated with evaluation, the study of the lexical combinations that appear most frequently, the use of slang and jargon, interjections, rhetorical questions, and reference to other users to establish credibility.

In sum, the authors consulted (Ricci & Wietsma, 2006; Vásquez, 2014; Schemmann, 2011; Pollach, 2006) coincide in the identification of common characteristics of this genre in English: there is a chronological sequence of events; from the language point of view, the use of reported speech, story prefaces and deictic shifts is preferred; reviews contain frequently humour details and personal experience; non-linguistic cues such as punctuation and use of specific orthotypographic and paralinguistic elements; there is a high degree of intertextuality and reference to previous comments; and, finally, reviews make use of strategies designed to compensate the impersonality of the written discourse. However, these authors also highlight that the objective of reviews is to achieve reliability and credibility closely related to the origins of UGR as word-of-mouth, and thus influence other users.

3. Methodology

Upon identifying the main features of orality from the literature on online mediated communication, together with the main features of consumer-generated reviews, a corpus of one thousand user reviews originally written in English and published during the same period of time (and their machine translation output produced by Google translate within the review platform itself) was compiled from TripAdvisor, one of the leading online travel review platforms operating in 49 markets and in 28 languages. TripAdvisor stores more than 988 million reviews on more than 8 million properties and businesses, in more than 300 thousand destinations (Tripadvisor, 2021). The criteria to include the reviews in the corpus was based on their authenticity and their representativeness (based on the authors consulted, the amount of reviews used for similar studies ranges from 250 reviews (Pollach, 2006) to 1,000 reviews (Vásquez, 2014). In addition, with the aim of extracting common features by hand, the number of reviews needed to be limited and approachable. Then, the reviews were aligned at the sentence level with their corresponding Spanish MT output to facilitate the study and the extraction of examples to illustrate the analysis and discussion section: the aligned corpus contains a total of 7,803 sentences.

Syntactic analysis	• extension of reviews • use of ellipsis • incomplete sentences
Lexical analysis	• verbs denoting orality • evaluation adverbs • filler words and expressions • use of colloquial language and idiomatic expressions
Discourse analysis (focus on features of online discourse)	• use of abbreviations, acronyms, and clippings • repetitions • discourse markers and intensifiers • rhetorical questions • paralinguistic items (exclamation marks, emoticons, parenthesis, typographic innovations, quotation marks) • genre-specific features (evaluation, advice, intertextuality)

Table 1. Framework for analysis of corpus (own source)

Based on previous studies on the linguistic characteristics of written language in an electronic medium and online reviews (Vásquez, 2014; Pollach, 2006; Holgado & Recio, 2013; and Yus, 2011), we have developed our own analysis framework and structured the analysis for this study around three levels: syntactic, lexical, and discursive. Different analyses have been carried out within each level to validate the degree of machine translation performance and offer a global view in the conclusion sections (see table 1).

In order to contrast the general linguistic features of UGR in Spanish with the findings in the Spanish-MT corpus, another corpus of 1000 reviews originally written in Spanish was compiled. Finally, corpus analysis techniques were used with AntWordProfiler 1.5.1 (Anthony, 2021) for the extraction of frequency lists and computations of the most frequent lexical combinations, and the rest of the language resources were extracted manually by linguists. For the calculation of some lexical items, it was necessary to compile *ad-hoc* word lists of verbs denoting orality and evaluation adverbs.

The next section on analysis and discussion presents the main results of how the oral features of reviews are handled by machine translation and how its distinguishing characteristics reach the Spanish travel review community.

4. Analysis and discussion

The approach to the study of the digital genre of users' opinions adopted in this work focuses mainly on the linguistic resources used to evaluate their experience, the aspects that highlight the user's credibility and identity, intertextuality and involvement, and personal experience (Vásquez, 2014; Pollach, 2006; Holgado & Recio, 2013; and Yus, 2011). Since the aim of this paper is to see how machine translation handles the expression of the pragmatic function of reviews, we will focus exclusively on the linguistic resources used by reviewers and how they are processed by machine translation systems.

With all the above, some of the analyses proposed by the authors studied have been selected (Pollack, 2006; Vásquez, 2014) and carried out on the corpus of user reviews. The analysis has been structured in three sections, syntactic, lexical, and discursive resources, with the main orality features of the genre of online reviews, object of this study, in mind.

The syntactic analysis takes into consideration aspects such as the extension of reviews, the use of ellipsis, use of simple phrases or incomplete sentences. Within the lexical analysis, emphasis is made on lexical density, filler words and expressions, verbs denoting orality, use of abbreviations, acronyms and clippings, use of colloquial language and general vocabulary and idiomatic expressions. The third analysis considers repetitions, discourse markers and intensifiers along with artifacts such as typographic innovation and genre-specific features such as naturally occurring stylistic preferences to express evaluations, advice and intertextuality or reference to other reviews.

4.1 *Syntactic analysis*

Corpus linguistics research methods applied to large amounts of text provide first-hand insights into the differences of this digital genre depending on whether they are originally written in English or Spanish. A preliminary word count of the corpus reveals the first difference between reviews written originally in English or in Spanish: the extension of the texts. In general, with the same number of reviews studied (1,000) the corpus in English contains 130,439 words, while the Spanish corpus has 80,939, that is, almost 50,000 words less (see table 2).

Although the type/token ratio or the STTR are not accurate measures for lexical density, the results in table 2 reveal a higher degree of lexical variety in the corpus of reviews originally written in Spanish, and curiously, in the MT corpus. UGRs are not written by specialists and thus vocabulary is recurrent. The short extension of reviews justifies using STTR, though the cause of the low lexical variety may reside in the imitation of previous reviews, which is a well-known feature of this genre.

Text file	*CorpusUGR_EN*	*CorpusUGR_MT-ES*	*CorpusUGR_ES*
Tokens	130,439	139,160	80,939
Types	6,796	8,462	7,145
Type/token ratio	5.26	6.14	8.91
Standardised TTR	41.21	42.17	42.25
Standardised TTR std.dev.	57.97	57.24	55.74

Table 2. Corpus data

In terms of the extension of the individual reviews (table 3), the average number of words per review is significantly higher in English, with 128.5 words per review, while in Spanish is almost 50 words shorter: 78.5 words per review. However, the most revealing fact is that 50 % of the total number of reviews studied in Spanish range from 8 to 50 words, while in English, the range is between 20 and 90 words. Some authors attribute this difference to the familiarity and tradition of Anglo-Saxon cultures with giving personal opinions and evaluations in public (Vásquez, 2014).

	CorpusUGR EN	*CorpusUGR MT-ES*	*CorpusUGR ES*
Average number of words	128.05 words	137.02 words	78.85 words
Shortest review	26 words	30 words	9 words
Longest review	1196 words	1284 words	980 words

Table 3. Sentence length

In terms of the use of simple sentences, both languages make use of this artefact with the same proportion. In fact, there is strong parallelism in the length of the sentences in Spanish and in English, both with many simple sentences.

However, the use of ellipsis and incomplete sentences, although not very frequent (129 cases in total from the 7803 sentences studied) is much more noticeable in English, and non-existent in Spanish. Some examples to illustrate the use of ellipsis in English are: *Now for the bad points ....; The room ...... Oh dear light fitting hanging, toilet handle broken*. Interestingly, ellipsis only appears in negative reviews.

4.2 *Lexical analysis*

At first sight, the study of frequency lists with the most recurring words used in the corpus of reviews reveals a considerable difference between the aspects of the tourist experience evaluated by users who write in English and those who write in Spanish. Words like *hotel, room, staff,* and *location* occupy the first positions of both frequency lists, with more or less the same number of occurrences in both languages. However, a more detailed study of the frequency list reveals that Spanish speakers rank first aspects such as *precio* (price), *calidad* (quality), *metro* (underground), *limpieza* (cleaning), *vistas* (views), *terraza* (balcony) and *decoración* (decoration), which are not so frequent in the English corpus. In contrast, English-speaking reviewers give more relevance to aspects such as *bed, shower, food, tea, bar,* and *coffee,* all of which have a much lower frequency in Spanish.

Reminiscent of the oral tradition, reviews make extensive use of verbs denoting orality such as *say, tell, comment, speak, ask, answer, reply, explain,* or *suggest,* with 2 % of the total verbs. However, the most common verbs in the corpus, as expected, are descriptive verbs, and verbs related to the hotel experience. Thus, verbs such as *be, have, stay, like,* or *walk* occupy the first positions with a total usage with respect to the total number of verbs of 27 %.

On the other hand, evaluative adverbs to express certainty, attitude or judgement also throw light on the characteristic style of reviews. The 10 most frequent evaluative adverbs found in the corpus *really, definitely, actually, probably, simply, clearly, unfortunately, kindly, fairly,* and *obviously* reveal a trend towards the intention of reviewers to emphasize their opinion and transmit certainty and veracity with their review.

So far, the Spanish MT output transmits accurately verbs and adverbs. However, in the case of filler words and expressions (see table 4), mainly used in spoken conversations to signal pauses while speaking, which are culturally bound, the results show a higher degree of mistranslations.

Closely related to filler words, discourse markers and intensifiers contribute to the oral mode of written communication in digital genres. Reviews are posted online and read asynchronously by other users, which makes unnecessary the use of conversation markers to check for the answer of the receiver, despite their high frequency in the corpus.

	Original English	*Spanish MT output*
Ex. 1	**Ok**, so the worst part of our one night stay	**Ok**, por lo que la peor parte de nuestra estancia de una noche
Ex. 2	The place was disgustingly dirty, **I mean** filthy from head to toe.	El lugar estaba asquerosamente sucio, **me refiero** sucio de pies a cabeza.
Ex. 3	This is a great location **so** walking was was a good option to several areas.	Esta es una gran ubicación para caminar **se** era una buena opción para varias áreas.
Ex. 4	**Oh** and the train passes next to the rooms, with very poor window quality.	**Ah**, y el tren pasa junto a las habitaciones, con muy mala calidad de la ventana.

Table 4. Filler words and expressions

As expected, the use of colloquial language and discourse markers such as *by the way*, *anyway*, *well* or *in sum* appear quite frequently in the corpus, confirming thus the use of orality features in reviews to confer authenticity to the opinions expressed (table 5).

	Original English	*Spanish MT output*
Ex. 5	They said the housekeeping folks couldn't get it and it came from the front desk. **Huh?**	Dijeron que la gente de limpieza no pudieron conseguirlo y que provenían de la recepción. ¿Eh?
Ex. 6	**Oh by the way**, no breakfast.	**Ah, por cierto**, no hay desayuno.
Ex. 7	**In sum,** we had an enjoyable time staying at this hotel.	**En suma,** tuvimos un tiempo agradable estancia en el hotel.
Ex. 8	the receptionist was **really** helpful	el recepcionista era **muy** servicial

Table 5. Discourse markers and intensifiers

To conclude this lexical analysis, one of the linguistic aspects that poses a challenge for MT is the use of idiomatic expressions. As shown in table 6, examples 9 to 11 have been transferred as a word-for-word translation, thus the original meaning is mistranslated and the reviewer's intention is lost. The idiomatic expression in example 9 refers to an extremely good feeling (*feel like a million dollars*), while the MT output simply reproduces the million dollars metaphor, which is not used in Spanish. Similarly, *Nothing short of perfect* (example 10) is used in English to emphasize a situation or quality, while the MT output in Spanish is an incoherent sequence of words, which may even confuse the reader and express a negative opinion on the hotel. Nevertheless, as shown in examples 12 and 13, in other cases, the MT output correctly deciphers the idiomatic expression.

	Original English	*Spanish MT output*
Ex. 9	This place makes you feel a million dollars	Este lugar te hace sentir un millón de dólares
Ex. 10	Nothing short of perfect.	Nada menos que perfecto.
Ex. 11	You want for nothing.	Usted quiere para nada.
Ex. 12	The hotel itself is beautifully appointed and ideally located, close to the **Tube**	El hotel en sí está muy bien equipado y muy bien situado, cerca del **metro**,
Ex. 13	My wife was particularly annoyed by the lack of a **cuppa** and it really wouldn't cost much to put this right.	Mi esposa estaba particularmente molesto por la falta de una **taza de té** y realmente no le costaría mucho poner este derecho.

Table 6. Idiomatic expressions

General vocabulary and vague expressions represented by the use of *stuff, things, kind of, sort of*, or the suffix *-ish* are also characteristic of the spoken language, and many times are also culturally bound. The solution adopted by the MT engine is not appropriate in almost all cases, as the Spanish reader would not perceive the natural and fluent transition expected in an authentic review (see table 7).

	Original English	*Spanish MT output*
Ex. 14	So, with all this positive **stuff**	con todas estas **cosas** positivas
Ex. 15	We didn't drink/eat all the complimentary **stuff** in the mini bar	No beber / comer todas **las cosas de** cortesía en el minibar
Ex. 16	It was great to have daily tea **things**, including biscuits.	Fue genial tener **cosas** de té diarias, incluyendo galletas.
Ex. 17	Quiet **ish** at night!	Ish Tranquilo por la noche!
Ex. 18	It was **cheapish** £100 a night for 2 people."	Era muy **caprichoso** £ 100 por noche para 2 personas."
Ex. 19	Everyday we had **some sort of** delicacy from the hotel chef to enjoy at evenings.	Todos los días teníamos **una especie de** delicadeza del chef del hotel para disfrutar en las noches.

Table 7. General vocabulary and vague expressions

4.3 *Discourse analysis*

According to Vásquez (2014: 25), from the point of view of discourse analysis, the study of evaluation and reviews is very complex since orthographic, lexical, syntactic, and discursive resources are involved, in addition to the differences in meaning that these resources acquire according to the context in which they are found: "No word or set of words is inherently positive or negative. Instead, it is the particular linguistic and social context that determines whether a word, set of words or expressions are to be interpreted in positive or negative terms" (2014: 26).

One of the common linguistic resources in online discourse is the use of abbreviations, acronyms, and clippings, not only for economic reasons but also to reflect the current and modern use of language and as a marker of community membership (Yus, 2011: 31). As shown in table 8, the Spanish MT output only reproduces established abbreviations (see examples 26 to 28) while the rest are left in English, and therefore the target user misses both the meaning and the intention of the reviewers.

	Original English	*Spanish MT output*
Ex. 20	**Thx** to all staff at the Montague	**Thx** a todo el personal en el Montague
Ex. 21	**Prob** nicest and friendliest staff	**Prob** mejor personal y amable
Ex. 22	the room had be tidied, our **pjs** folded	habitación había ser arreglado, nuestros **pjs** cruzadas
Ex. 23	**BTW** looking at some of the pictures as bad as they are they look better than the room I stayed in.	**BTW** mirando algunas de las fotos tan malas como son se ven mejor que la habitación que me alojé.
Ex. 24	How **FAB** is that??!!!!!	Cómo **FAB** es eso ??
Ex. 25	70.00 for this **inc** breakfast was just fine	70.00 por este desayuno **inc** estaba bien
Ex. 26	and ineffective **aircon**	el **aire acondicionado** ineficaces
Ex. 27	Breakfast **ok** for 15 pounds.	El desayuno está **bien** por 15 libras.
Ex. 28	Atmosphere is **fab** in the gorgeous quaint setting.	El ambiente es **fabuloso** en el magnífico entorno pintoresco.

Table 8. Use of abbreviations

With regards to repetitions as intensifiers, English reviews use this artifact quite frequently, up to 150 examples were found in the corpus (table 9).

The MT solutions are varied: remove the repetition (example 29); transfer word-for-word the repetition (examples 30 to 31); mistranslate the repetition and produce an incorrect sentence in Spanish (example 33). In addition, the removal of capitalization in the Spanish MT output should be considered as a translation error of omission or mistranslation as the pragmatic use of capitals is well-established in digital genres as a marker of orality or to add emphasis (Yus, 2011:118), which is removed in the Spanish translation.

	Original English	*Spanish MT output*
Ex. 29	do not EVER, EVER book this hotel.	nunca reservar este hotel.
Ex. 30	AVOID AVOID AVOID.	Evitar evitar evitar.
Ex. 31	Breakfast was ok too, plenty of eggs, cereal, cheese, yogurts, muffins, croissants etc etc.	El desayuno también estuvo bien, un montón de huevos, cereales, queso, yogures, magdalenas, croissants, etc, etc.
Ex. 32	The hotel has a very English feel and charming in many many ways - stop in for tea or a meal	El hotel tiene un ambiente muy Inglés y encantador en muchas muchas maneras - parar en para el té o una comida
Ex. 33	Thank you so so much to the management and staff at the Milestone	Gracias por lo tanto a la dirección y el personal en el Milestone

Table 9. Use of repetitions

Among the paralinguistic features or aspects of communication that do not involve words, the most common strategies found are the use of emoticons, exclamation marks, parenthesis, or capitalization of words. The number of emoticons or punctuation emphasis found in the corpus of Spanish reviews was very low, with the exception of the use of several exclamation marks common in digital genres: *Recomendable :); Felicitaciones a los propietarios y personal !!!*. However, the analysis of English reviews reveals very high use of typographic innovations, and in many cases, a combination of several innovations within the same sentence (see table 10).

Reminiscent of the oral origins of reviews, there are several instances of emphasis artifacts common in a spoken language like the use of parenthesis within the review to spark the reactions of other users such as in examples 41 and 42.

The other use of parenthesis corresponds to the addition of details or explanations to contextualize the review: *Our room (a standard) is precious* or

Prices for food and beverage are "hotel" (ie., expensive). Similar use is found in the Spanish reviews: *muy bien comunicado, cerca del metro, de tiendas, de restaurantes, y de teatros (ideal para ir a ver el Rey León); El desayuno (aunque no estaba incluido en el precio de la habitación) es totalmente recomendable incluirlo en ella*. Overall, the use of parenthesis is very similar in English, with 367 occurrences, and in Spanish (295) revealing a consolidated strategy of this digital genre in both languages.

	Original English	*Spanish MT output*
Ex. 34	We will be back!!!	¡¡¡Volveremos!!!
Ex. 35	Nothing is too much trouble, and the staff make your stay extra special - thank you :-)	Nada es demasiado problema, y el personal hace que su estancia sea especial - gracias :-)
Ex. 36	WOW!	¡GUAUU!
Ex. 37	A totally English experience!!!!!!	Una experiencia totalmente Inglés !!!!!!
Ex. 38	THANK YOU 41	GRACIAS 41"
Ex. 39	Nice place! :)	¡Buen lugar! :)
Ex. 40	make sure Fran gets a raise please! :)))	asegúrese de Fran consigue un aumento de sueldo por favor! :)))
Ex. 41	The view through our Room (96) window was to lots of Toilets, basins and storage area (FUNNY isn't it)	La vista a través de nuestra habitación (96) era la ventana a un montón de inodoros, lavabos y área de almacenamiento (divertido ¿no es)
Ex. 42	While our room was very small (by American standards), it was lovely.	Aunque nuestra habitación era muy pequeña (para los estándares americanos), que era una maravilla.
Ex. 43	This ““hotel”“ is a DISGRACE.	Este ““hotel”“ es una vergüenza.
Ex. 44	where you apparently get a ““quality hotel”“ for a bargain price.	donde al parecer obtiene un ““hotel de calidad”“ para un precio de ganga.

Table 10. Paralinguistic artifacts

Finally, special attention should be given to genre-specific features such as naturally occurring stylistic preferences to express evaluations, advice and intertextuality or reference to other reviews, which are essential for the naturalness, reliability, and credibility of consumer reviews (see table 11).

These stylistic preferences reveal the most distinguishable aspect of this culture-specific approach to reviews. The sequence of Spanish reviews, and overall syntactic and stylistic artefacts, or even the use of loan words from

English such as *check-in, amenities*, or *conveniente* to express well located or adequate, indicates a strong influence of English reviews in Spanish users. Thus, the expression of genre-specific features demonstrates a completely different style in Spanish.

	Original English	*Spanish MT output*	*Genre-specific features*
Ex. 45	Loved this place	Me encantó este lugar	Evaluation
Ex. 46	Other than that, a fabulous stay.	Aparte de eso, una estancia fabulosa.	Evaluation
Ex. 47	Definitely a 5 star experience.	Sin duda una experiencia de 5 estrellas.	Evaluation
Ex. 48	I would recommend this hotel highly.	Yo recomendaría altamente este hotel.	Advice
Ex. 49	I highly recommend staying at The Montague on The Gardens.	Recomiendo encarecidamente alojarse en el Soho Hotel.	Advice
Ex. 50	I will definitely revisit!!!	Definitivamente voy a volver !!!	Advice
Ex. 51	Overall, recommended for short stay for a couple only.	En general, recomendado para una estancia corta de sólo un par.	Advice
Ex. 52	as the previous reviewer said	como se dice en la crítica anterior	Reference to other reviews
Ex. 53	Other reviews have just about said it all.	Otras críticas han dicho casi todo.	Reference to other reviews
Ex. 54	I read and reread lots of previous reviews	He leído y releído muchas críticas anteriores	Reference to other reviews
Ex. 55	As others have said,	Como han dicho otros,	Reference to other reviews

Table 11. Genre specific features

The most remarkable aspect of this analysis is the recurrent use in Spanish of a punctuation scale to express evaluations, which does not appear so frequently in English. English reviews conclude with a more personal overall impression such as examples 45 to 47. However, Spanish reviews conclude with an evaluation based on a mark such as: *No suelo dar 10 pero no puedo poner una sola pega al servicio ofrecido., "la habitación otro 10, todo de maxima nota; Solo puedo darle la mejor puntuación tanto a su personal como al hotel; "Mi puntuación, sin dudarlo, un 10!!!.*

To conclude this section on discourse resources, it should be noted that although the performance of MT engines on repetitions, discourse markers and paralinguistic items is quite accurate and the quality of the MT output is not affected to a large extent, it seems that the conventions of online discourse in English have an effect on the conventions in Spanish and readers might not perceive a lack of linguistic quality in the MT processed reviews. However, with regards to idiomatic expressions or paralinguistic artifacts, there is no homogeneous behavior of MT engines to handle, and the resulting MT output appears with a variable quality. This would demonstrate that for the 2.0 community authenticity is more highly valued than quality and that usefulness prevails over grammatical correction.

5. Conclusions

The key role of travel reviews in commercial decision making, the profusion of evaluation platforms, and the active participation of users make user-generated content a novel and challenging approach to identifying the characteristics of new digital genres. Likewise, the analysis of large amounts of data with different tools makes it possible to study patterns that would otherwise remain unnoticed.

The unquestionable influence of English in scientific, academic, and commercial environments is also evident in the new digital genres. However, in the case of travel reviews, this study shows that despite the great influence of English and machine translation in the delivery of reviews on online platforms, travel reviews in Spanish preserve different distinguishing characteristics, both from the approach to the evaluation experience and the use of language-specific resources.

Due to the extensive production of content generated by user reviews, along with the increase of review platforms of all kinds: tourism, restaurants, consumer electronics, and the fact that there is no general-purpose machine translation system, to facilitate that the MT output matches the style expected in the genre of user reviews, the exploitation of a comparable corpus will facilitate training these engines with the most relevant textual conventions.

Although there is no universal translation quality scale, or set of guidelines that apply to all scenarios, the characteristics of the textual genre should be taken into account. For example, in the case of reviews of a tourist product review, apart from language, other genre-specific features include naturalness, reliability and credibility.

In addition, most 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, consisting of free text of reduced dimensions, the error proportion would be higher and low-quality translation would be more visible. The results will help to fine-tune MT quality assessment scales according to other parameters such as the extension of the text, for example.

The main characteristics that confer naturalness and authority to this user-generated content are precisely the spontaneous features of the spoken language that are reflected in the written text of the review, which are not necessarily transmitted completely when processed by machine translation. Reviews are not only transmitted through language, some of the characteristics of this digital genre such as the reviewer's profile, intertextuality or reference to other reviews, or paralinguistic elements contribute to the reliability and credibility of user reviews.

The scarce use of paralinguistic features in Spanish reviews indicates the care of reviewers to avoid appearing unprofessional. No emoticons or punctuation emphasis were found in the Spanish corpus, with the exception of the use of several exclamation marks common in digital genres. However, as a reminiscence of its oral origins, there are several cases of common emphasis artifacts in spoken language.

The use of simple and short sentences with general vocabulary allows a quite reliable machine translation processing of most of the oral features of travel reviews, however, it appears that the 2.0 community values authenticity more than quality, and utility prevails over linguistic correctness. The efficiency of machine translation engines shows varying degrees of quality depending on the resource processed: while the lexical and grammatical level is translated with ease, in the case of abbreviations, colloquial language and ellipsis is not completely accurate.

As in the development of applications to detect emotions in user reviews, it would be thought-provoking to design a detector of possible machine translation errors that will identify the reviews that due to the use of certain linguistic resources may be susceptible of error, such as a detector of potential situations in which the context acquires more importance than the element itself.

Unquestionably, the quality of MT output currently, although not perfect, is quite far from previous machine translation stereotypes, since sometimes the quality reaches relatively high levels, and in the case of the MT of tourism reviews, reaching the adequacy and linguistic correction of the texts produced by humans.

Research and design of bilingual spoken corpora or bilingual corpora with oral features need to be encouraged as machine translation research is also aimed towards real-time translation of spoken interactions such as video conferencing or automatic subtitling.

References

Allen, Jeff. 2003. Post-editing. In Somers, Harold (ed.) *Computers and Translation. A translator's guide*. Amsterdam: John Benjamins, 297-317.

Anthony, Laurence. 2021. AntWordProfiler (Version 1.5.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Aranberri, Nora. 2014. Posedición, productividad y calidad. *TradumàTica: Tecnologies de la Traducció* 12: 471-477. doi: 10.5565/rev/tradumatica.62

Candel-Mora, Miguel A. 2015. Evaluation of English to Spanish MT Output of Tourism 2.0 Consumer-Generated Reviews with Post-Editing Purposes. In Esteves, Joao; Macan, Juliet; Mitkov, Ruslan & Stefanov, Olaf (eds.) *Proceedings of Translating and the Computer 37*. London: Editions Tradulex, 37-47.

Castilho, Sheila; Doherty, Stephen, & Gaspari, Federico. 2018. Approaches to Human and Machine Translation Quality Assessment. In Moorkens, Joss; Castilho, Sheila; Gaspari, Federico, & Doherty, Stephen (eds.) *Translation Quality Assessment: From Principles to Practice*. New York: Springer, 9-38.

Chen, Min; Mao, Shiwen; Zhang, Yin, & Leung, Victor. 2014. *Big Data: Related Technologies, Challenges and Future Prospects*. New York: Springer.

Comisión Nacional de los Mercados y la Competencia. 2019. *Datos estadísticos del comercio electrónico en España, 2019*. http://data.cnmc.es/datagraph [Accessed 15-10-2021]

Gerlach, Johanna; Porro Rodríguez, Victoria; Bouillon, Pierrette, & Lehmann, Sabine. 2013. Combining pre-editing and post-editing to improve SMT of user-generated content. In O'Brien, Sharon; Simard, Michel & Specia Lucia (eds.) *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*. Nice: AMTA, 45-53.

Holgado, Anais & Recio Diego, Álvaro. 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.

Jiang, Jie; Way, Andy, & Haque, Rejwanul. 2012. Translating user-generated content in the social networking space. In Foster, George (ed.) *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas*. San Diego: AMTA, 1-9.

Koby, Geoffrey; Fields, Paul; Hague, Daryl; Lommel, Arle, & Melby, Alan. 2014. Defining Translation Quality. *TradumàTica: Tecnologies de la Traducció* 12: 413-420.

Lommel, Arle. 2018. Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In Moorkens, Joss; Castilho, Sheila; Gaspari, Federico, & Doherty, Stephen (eds.) *Translation Quality Assessment: From Principles to Practice*. New York: Springer, 109-127.

Minelli, Michael; Chambers, Michele, & Dhiraj, Aambiga. 2013. *Big data, big analytics: emerging business intelligence and analytic trends for today's businesses*. New Jersey: John Wiley & Sons.

Pollach, Irene. 2006. Electronic word of mouth: A genre analysis of product reviews on consumer opinion web sites. In Mohsin, Mansoor; Cavin, David; Sasson, Yoav; Prakash, Ravi & Schiper, André (eds.) *Proceedings of the 39th Hawaii International Conference on System Sciences*. IEEE Computer Society, 51-61.

Ricci, Francesco & Wietsma, René. 2006. Product reviews in travel decision-making. In Hitz, Martin; Murphy, Jamie, & Sigala, Marianna (eds.) *Information and communication technologies in tourism*. Wien: Springer, 296-307.

Schemmann, Brita. 2011. A Classification of Presentation Forms of Travel and Tourism-Related Online Consumer Reviews. *e-Review of Tourism Research* 2: 7.2.

Specia, Lucia; Raj, Dhwaj, & Marco Turchi. 2010. Machine Translation Evaluation Versus Quality Estimation. *Machine Translation* 24: 39-50. doi:10.1007/s10590-010-9077-2.

Temnikova, Iirina. 2010. Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. In Calzolari, Nicoletta (ed.) *Proceedings of the LREC 2010 Conference*. Valletta: European Language Resources Association, 17-23.

Tripadvisor. 2021. Tripadvisor's Year in Review: All the good in 2021 Dec 14, 2021. https://ir.tripadvisor.com/news-releases/news-release-details/tripadvisors-year-review-all-good-2021

Vásquez, Camilla. 2012. Narrativity and involvement in online consumer reviews. The case of Tripadvisor. *Narrative Enquire* 22(1): 105-121.

Vásquez, Camilla. 2014. *Online consumer reviews*. London: Bloomsbury.

Yus, Francisco. 2011. *Cyberpragmatics*. Amsterdam: John Benjamins.